Order-Theoretic Combination Techniques and the Electronic Schrödinger Equation

Dissertation

zur

Erlangung des Doktorgrades (Dr. rer. nat)

der

Mathematisch-Naturwissenschaftlichen Fakultät

der

Rheinischen Friedrich-Wilhelms-Universität Bonn

vorgelegt von

James Nicholas John Barker

aus

Adelaide, Australien

Bonn, 2023

Angefertigt mit Genehmigung der Mathematisch-Naturwissenschaftlichen Fakultät
der Rheinischen Friedrich-Wilhelms-Universität Bonn.

1. Gutachter: Prof. Dr. Michael Griebel

2. Gutachter: Prof. Dr. Jochen Garcke

Tag der Promotion: 02.02.2024

Erscheinungsjahr: 2024

Abstract

Most standard constructions of the *combination technique* [M. Griebel et al., *Iterative Methods in Linear Algebra*, Elsevier, North Holland, p. 263] manipulate families of functions organised by downward-closed subsets of $\mathbb{N}^d$. We introduce instead an alternative formulation, with functions indexed from a more general kind of partially ordered set (poset). The combinatorial and order-theoretic machinery of *Möbius inversion* helps us to construct combination sums of functions organised by *order ideals* of a *poset grid*. An adaptive algorithm is given for the quasi-optimal assembly of such an order ideal. This *order-theoretic combination technique* (OTCT) formalism is applied in the quantum-chemical setting of the high-dimensional *electronic Schrödinger equation*. Here, the OTCT allows us to connect, understand, and improve on a number of existing approaches.

We consider first a selection of existing extrapolative *composite methods*. Extending on the idea of the CQML approach [P. Zaspel et al., *J. Chem. Theory Comput.*, 15(3), 2018], an application of basically just the standard version of the combination technique leads to a *generalised composite method* (GCM). This approach is systematically improvable and appears comparable, if not yet truly competitive with standard composite methods from the perspectives of both accuracy and of cost.

We turn then to *energy-based fragmentation methods*, which are often founded upon a truncated *many-body expansion* (MBE). It is well-known that Möbius inversion can provide non-recursive expressions for the individual MBE terms, and so the OTCT delivers by construction a framework for the adaptive truncation of MBE-like formulae. The same also functions for a class of related graph-based decompositions described in the existing literature. We term these in our context as SUPANOVA (*SUbgraph Poset ANOVA*) decompositions, and motivate them as extensions to the BOSSANOVA decomposition [M. Griebel et al., *Extraction of Quantifiable Information from Complex Systems*, Springer, Cham, 2014, p. 211; F. Heber, Dissertation, Rheinische Friedrich-Wilhelms-Universität Bonn, 2014]. We identify a subtle technical issue that afflicts BOSSANOVA in certain cases, and apply instead an adaptive SUPANOVA decomposition defined for *convex subgraphs*.

Finally, we combine the GCM and SUPANOVA ideas to obtain a poset grid that recovers many existing *multilevel fragmentation methods*. We extend the ML-BOSSANOVA method [S. R. Chinnamsetty et al., *Multiscale Model. Simul.*, 16(2), 2018], exploring now also a hierarchy of *ab initio* theories. Although an initial assessment is inconclusive, this ML-SUPANOVA formulation appears well-founded and paves the way to a number of interesting possible applications in the future.

For and in memory of my grandfather

Gordon Prujean Williams
1921–2016

Acknowledgements

I begin by thanking Prof. Dr. Michael Griebel and Dr. Jan Hamaekers for offering me the opportunity to pursue my doctoral research in Germany, and for their considerable support and encouragement throughout.

I am grateful for past positions at the Institute for Numerical Simulation (funded in parts by the DFG, via the HCM and the SFB1060) and with Fraunhofer SCAI. A long roster of colleagues made these pleasant and stimulating working environments; since I do not have space to list everyone by name, I resort here to a heartfelt expression of collective appreciation. In particular, however, I thank Dr. Astrid Maaß for many patient explanations of matters chemical. I also extend thanks to Dr. Bastian Bohn, Prof. Dr. Jochen Garcke, Dr. Christopher Kacwin and Johannes Rentrop for discussions regarding the combination technique.

Many thanks are also due to a number of academics outside Bonn who provided helpful discussions on various topics both directly and indirectly related to the work covered here. In particular and in alphabetical order, I thank here Dr. Daniel Claudino, Prof. Dr. Virginie Ehrlacher, Prof. Dr. David Feller, Prof. Dr. Luca Ferrari, Prof. Dr. Michael Herbst, Prof. Dr. Antoine Levitt, Dr. Michael Obermaier, and Prof. Dr. Alistair Rendell.

I am grateful to have had access to the high-performance computing resources of Fraunhofer SCAI. I thank past and present members of SCAI-IT, particularly Andre Gemünd and Ralph Thesen, for their professional support and flexibility. I also thank the developers of the quantum chemistry software packages used in the preparation of this thesis, both in general and for some helpful technical assistance.

I am indebted to Valerie Barker, Dr. Heinz-Jürgen Flad, Dr. Jan Hamaekers, Dr. Astrid Maaß, Tobias Olbrich and Johannes Rentrop, who were all generous enough to proof-read and comment on earlier manuscripts of this document, either in part or in full. Any errors in this work are certainly mine alone.

Without intention to compare to the experiences of others, the most difficult aspect of the pandemic for me personally was a long and unplanned separation from my family. I thank them for a lifetime of love and support, and am very much looking forward to both seeing them again and meeting them for the first time.

And finally, my deepest gratitude goes to F. H., who already knows why.

Contents

1. Introduction

Many topics of interest in chemistry, such as the structure of chemical compounds and their behaviour as they react with each other, reduce to the study of changes in the total energies of molecular systems [Tul00; LeB05; HOJ13; Jen17]. The field of *quantum chemistry* investigates these topics from the perspective of the Schrödinger equation: a linear partial differential eigenproblem [SO89; Can+03]

$$H\Psi = E\Psi, \tag{1.1}$$

where H is a Hamiltonian operator representing the pairwise interactions between the M nuclei and N electrons which together make up the system, and each eigenfunction Ψ is a wavefunction representing a valid quantum state of the system, with a total energy given by its eigenvalue E.

Application of the Born-Oppenheimer approximation [BO27; Tul00] to (1.1) produces the *electronic Schrödinger equation*, sometimes called the *electronic problem* [SO89]. Solutions to this latter provide, or rather, would provide the total electronic energies of molecular systems under fixed nuclear conformations. However, analytic solutions cannot generally be had [GH07; Yse10]. Since the primary impediment to the numerical solution of the electronic Schrödinger equation is its very high dimensionality [LeB05], so the "curse of dimensionality" [BG04] predetermines standard numerical techniques as unworkable in most cases [GH07]. Instead, quantum chemists use a variety of alternative approximation schemes that are heavily and explicitly tailored to the electronic problem [SO89; HOJ13; Jen17]. Those schemes which are based on particular assumptions about the structure of the electronic wavefunction are referred to as *ab initio* methods [LeB05].

The standard approach of *ab initio* quantum chemistry is to discretise the electronic wavefunction in terms of a set of *Slater determinants*, each an antisymmetrised product of members of a family of lower-dimensional functions [SO89; Can+03; EA07; HOJ13]. These functions are themselves discretised as finite linear combinations of members of a shared *basis set* of functions. Many templates exist for constructing such basis sets [Huz85; Hil12], but the size of such a set as used in practice is always proportional to the number of atoms in the system. For a particular basis set, a Galerkin approximation of the electronic wavefunction in terms of all so-derivable Slater determinants is best possible, but the number of such determinants, and thus the computational cost of obtaining that approximation, scales roughly exponentially in the number of basis functions [Sch09]. Several families of *ab initio* algorithms allow the consideration of wavefunctions discretised in terms of only particular subsets of all possible Slater determinants [SO89; HOJ13;

Jen17]. As the granularity of the discretisation increases and the basis set becomes in some sense complete, and as fewer of the available Slater determinants are neglected, the resulting approximations trend reliably towards the exact solution [Can+03; Sch09].

The fundamental *ab initio* Hartree-Fock approximation [Roo51; Hal51; SO89; EA07] presents a cost of solution that scales formally as $\mathcal{O}(K^4)$ in the size of the discretising basis set. Further approximations to the accompanying *correlation energy* error term are usually based in techniques either drawn from Møller-Plesset perturbation theory [MP34; SO89] or founded in the coupled cluster approximation [CS00; Sch09]. The cost scaling expressions of these techniques begin at $\mathcal{O}(K^5)$, and rise dramatically for more accurate approximations [SO89; HOJ13]. For instance, the cost of calculating the well-regarded CCSD(T) approximation to the correlation energy [PB82; Rag+89], which is regularly referred to as setting the "gold standard" for accuracy [ŘH13], scales basically as $\mathcal{O}(K^7)$ [BM07; HOJ13].

The energetic values produced by a CCSD(T) calculation may still fail to agree with experimentally-observed results to within the exacting tolerance requirements of the quantum chemist [Kar16; KSM17]. To reliably obtain this level of accuracy, even more comprehensive *ab initio* methods must be used, and the scaling terms of these involve even higher-order polynomials. Extrapolative *composite methods* offer a more computationally plausible possibility [Pop+83; Pop+89; Taj+04; CRR07b; Kar+06; DeY+09; Kar16]. These schemes ultimately come down to a certain assumption of additivity of error, see, e.g., [PFD12; RS15; Jen17; CGH18]. Here, sets of approximate solutions are calculated, each stressing accuracy in a particular way, then deconstructed and carefully rebuilt into one. But while composite methods have found favour in various settings, their use always involves calculations of the correlation energy, so the implied scaling as mentioned above still extinguishes their utility for large systems [RS15].

The success and applicability of *reduced scaling* protocols like *energy-based fragmentation methods* [Gor+11; CB15; RS15; Her19] is related to the famous principle of the *nearsightedness of electronic matter* as recognised by Kohn [Koh96; PK05]. Surely the most well-known such method involves the *many-body expansion* (MBE) of some full-system energetic property as modelled by a potential function $V : (\mathbb{R}^3 \times \mathbb{N})^N \to \mathbb{R}$ [HMS70; SDS09; RH12; CGH18; Her19],

$$V(X_1, \ldots, X_M) = \sum_{A=1}^{M} \tilde{V}^{(1)}(X_A) + \sum_{A<B}^{M} \tilde{V}^{(2)}(X_A, X_B) + \cdots + \tilde{V}^{(M)}(X_1, \ldots, X_M). \quad (1.2)$$

Here, each $X_A = (R_A \in \mathbb{R}^3, Z_A \in \mathbb{N})$ gives the spatial coordinates and charge of the nucleus of the Ath-indexed atom, and each $\tilde{V}^{(k)} : (\mathbb{R}^3 \times \mathbb{Z})^k \to \mathbb{R}$ is a *k-body potential*. A truncation of (1.2) after all terms $\tilde{V}^{(k)}$ with $k \leq n$ for some n is called an *n-body expansion*. If $|\tilde{V}^{(k)}|$ decays pointwise sufficiently quickly with increasing k — an assumption which seems at least empirically plausible — then an n-body expansion for some sufficiently low n can provide a reduced scaling approximation to V that may be accurate enough

for practical purposes [CGH18; Her19].

Of course, the number of k-body terms in (1.2) is generally superlinear in M for $k > 1$. Many implementations of n-body expansions and closely related fragmentation methods also take advantage of another expected decay in the magnitudes $|\tilde{V}^{(k)}(X_1, \ldots, X_k)|$ as the nuclei $\{X_A\}_{A=1}^k$ grow more and more distant from each other [OCB14; OB16; LH17]; most commonly, "distant" is understood here in purely Euclidean terms, but more abstract concepts of graph-based connectivity can also be used [DC05; WHM10; GHH14; Heb14]. Both forms of decay can be considered a manifestation of Kohn's nearsightedness principle [OB16; CGH18].

Our thesis is that many composite methods and energy-based fragmentation methods can be usefully viewed and understood as *combination techniques*. A well-known and comprehensively studied mechanism for the efficient approximation of high-dimensional functions, the combination technique [GSZ92] originally emerged from the theory of sparse grid approximation [BG04]. The approximation of functions using sparse grids can lead to a significant discount in the computational cost of a solution taken to a certain accuracy, when that cost is measured relative to conventional techniques. However, true sparse-grid approximation techniques are intrusive [TW18], in the sense that their application generally requires the reconfiguration and thus reimplementation of existing algorithms [BG04; OB21]. It was observed by Griebel et al. [GSZ92] that it is possible to instead combine certain sets of results obtained using more conventional numerical formulations in a non-intrusive fashion according to particular weighted summation formulae [BG04; Gar12b]. The resulting family of approximations display an asymptotic complexity behaviour that is comparable, up to a factor logarithmic in the dimensionality of the problem under consideration, to that which could be achieved by equivalent families of sparse grids.

The functions involved in a standard combination technique sum can be indexed by and organised according to particular partially ordered subsets of $\mathbb{N}^d$ [GG03; Heg03; HGC07; Har16a; Won16]. We present here a generalisation of the standard combination technique that operates, not necessarily in terms of the standard index space $\mathbb{N}^d$, but instead in terms of a general partially ordered set Π. Only slight restrictions are placed on the structure of this set. The construction of index sets drawn from Π, as well as expressions for combination sums of functions indexed by the members of those sets, is formulated with reference to machinery drawn from the toolboxes of order theory and combinatorics, and particularly that of *Möbius inversion* [Sta12]. Just as in the standard combination technique case and related applications [Gri98; GG03; Heg03; CGH18; TW18], combination-sum index sets may be grown adaptively by consideration of the benefits and costs associated with individual terms.

Similarities between the standard combination technique and both composite methods and energy-based fragment methods have been previously noted and exploited [CGH18; Zas+18]. It is also historically well-known, and not coincidental, that the very tools which we use to construct our *order-theoretic combination technique* can also be applied to the

derivation of the MBE [DFS04], as well as other, related decompositions [Dom74], such as the *chemical graph-theoretic cluster expansion* (CGTCE) of Klein [Kle86], and also the *lattice fundamental-measure theory* (LFMT) of Lafuente and Cuesta [LC05]. We will build upon these observations and use our generalisation of the combination technique to variously construct and evaluate direct implementations, analogues, and extensions of the energy-based fragmentation methods and composite methods mentioned above, and also to investigate and formalise certain similarities between particular instances of them.

1.1. Thesis structure

The main body of this document is structured as follows:

- We begin in Chapter 2 by summarising some key definitions and fundamental techniques from molecular quantum chemistry. We assemble for later use an *abstract cost model* by which the relative computational costs of certain approximation methods applied to a particular problem can be compared.
- In Chapter 3, we revise and extend the theory of the *combination technique* for the numerical approximation of high-dimensional functions. The focus of this chapter is the extension of the standard combination technique from the conventional d-dimensional index space $\mathbb{N}^d$ to the more abstract setting of a direct product of almost arbitrary partially ordered sets. We develop an adaptive algorithm that progressively assembles *order ideals* of these *poset grids* and thus index sets for quasi-optimal combination sums, again providing an extension of similar algorithms already known in the standard setting and in related work.
- Chapter 4 briefly reviews several *composite methods* which are widely used in practical quantum chemistry. Extending on the idea of the CQML approach of Zaspel et al. [Zas+18], we construct a *generalised composite method* (GCM): a straightforward adaptive mechanism that aims to obtain high-quality results similar to those delivered by existing conventional composite methods, but in a systematically-improvable manner.
- In Chapter 5, we investigate the MBE in detail, as well as a variety of other *energy-based fragmentation methods*. We focus particularly on different mathematical viewpoints on these methods. The order-theoretic combination technique provides an *adaptive many-body expansion*; we investigate its ability to accurately approximate the total energies of clusters of water molecules.
- Chapter 6 considers a class of decompositions which we call in our context *SUPANOVA decompositions*. This class is essentially a special case of Klein's very general CGTCE [Kle86], but we develop it with respect to the BOSSANOVA approach

of Heber and co-workers [Heb14; GHH14], revisited in the order-theoretic context. Discovering certain technical issues with the original BOSSANOVA construction, we consider a corrective extension, which we ground in the theory of abstract convex structure. Case studies examine the application of this *convex SUPANOVA* decomposition to two medium-sized molecules, and also to two proteins.

- Finally, we turn in Chapter 7 to a synthesis of the preceding chapters: we consider a three-dimensional poset grid that combines composite method-style grids with general SUPANOVA decompositions, and corrects and extends in turn on the ML-BOSSANOVA method [CGH18]. Preliminary testing of this multilevel *ML-SUPANOVA* setup is performed on two systems which are small enough to allow the calculation of reference total energies.
- The thesis concludes in Chapter 8 with a brief summary of our findings, and a discussion of various questions which remain open and may suggest potential future work.

The following supplementary materials are included as appendices:

- Appendix A contains technical details relating to calculations reported in the main text.
- Appendix B provides brief algorithmic descriptions for certain pieces of functionality required for a practical implementation of the adaptive algorithm.

1.2. Outline of contributions

The primary contributions of this thesis are as follows:

- The formal and technical development of the order-theoretic combination technique, and of an accompanying adaptive algorithm for the construction of quasi-optimal combination sums and order-ideal index sets; see Chapter 3.
- The detailed investigation of the GCM, SUPANOVA and ML-SUPANOVA classes of decomposition, and also an adaptive formulation of the standard MBE, from the particular perspective of the order-theoretic combination technique; see Chapters 4 through 7.
- The identification of a subtle technical issue with the BOSSANOVA and ML-BOSSANOVA methods. This involves a slight generalisation on an observation previously made by Lafuente and Cuesta [LC05] in the context of their LFMT, which turns out to be generally important in the setting of the order-theoretic combination technique; see in particular Sections 5.2.3, 6.3, and 6.5. Although the

core idea of the correction, specifically the idea of a decomposition constructed with reference to convex subgraphs, can be found suggested in [Kle86], we are not aware of an explicit development or application of this idea in the specific context of energy-based fragmentation methods. As we will point out, however, some recent works, e.g., [RHI18; RI18; KI19; RI20; Zha+21], can be recognised as using a particular special case of this setup; see Sections 6.5, 7.1, and 7.2.

Considering the standard combination technique from the perspective of lattices and order theory is not in and of itself novel; see, e.g., [Heg03; HGC07; Har16a; Won16], as well as discussion in Section 3.2. The adaptive algorithm we construct is a direct extension of existing approaches, particularly inspired by one given in the context of the ML-BOSSANOVA method [CGH18]. Moreover, we acknowledge explicitly that the basic observation that Möbius inversion can be used to construct ANOVA-like decompositions like ML-BOSSANOVA was first brought to our attention by Griebel [Gri19]. However, the order-theoretic combination technique that we describe here provides, to our knowledge, the first fully general coalescence of these ideas to appear in the literature.

In the same vein, we are careful to emphasise that the various GCM, SUPANOVA, and ML-SUPANOVA decompositions that we consider in this thesis are deliberately similar to, and in some cases either formally identical to or directly recoverable from, other decompositions and expansions that are described in the existing literature. We will make these connections very clear in the pages to come, but for now, we acknowledge particularly strong connections to previous work in [Heb14; GHH14; CGH18; Zas+18]. Also, several such existing decompositions are themselves explicitly constructed using Möbius inversion; see, e.g., [Dom74; Ess+77; Kle86; DFS04; LC05]. The abstract formal and algorithmic structure of the order-theoretic combination technique thus reproduces these as immediate special cases. This allows us to directly link such explicitly Möbius inversion-based decompositions with many others in the more recent literature, and also to generally and rigorously introduce concepts of quasi-optimal adaptivity to these decompositions.

2. Molecular quantum chemistry

We begin by reviewing some fundamental concepts in molecular quantum chemistry. We proceed at a high level: our goal here is only to fix our problem, locate our tools, and to provide sufficient justification for the *abstract cost model* we construct at the end of the chapter. We refer here to the standard textbooks by Szabo and Ostlund [SO89], Helgaker et al. [HOJ13], and Jensen [Jen17]. We also make general reference thoughout to the more mathematically-oriented reviews by Cancès et al. [Can+03], see in conjunction [LeB05; LL05; Yse10], and by Echenique and Alonso [EA07], and further to work on the projected coupled cluster method by Schneider [Sch09]. We have previously reviewed some of the topics covered in Sections 2.1, 2.2, and 2.4 in [Bar09], there more from the perspective of implementation. We will choose, adjust, and convert our notation by preference and for consistency, usually without explicit comment. The use of Dirac's bra-ket notation is deliberately avoided.

2.1. The electronic Schrödinger equation and the Born-Oppenheimer potential

The *electronic Schrödinger equation* [SO89; Can+03; EA07],

$$H\Psi(x_1, \ldots, x_N) = E\Psi(x_1, \ldots, x_N), \tag{2.1}$$

sometimes called the *electronic problem*, emerges from the application of the *Born-Oppenheimer approximation* [BO27; Tul00] to the full time-independent Schrödinger eigenproblem. Here, H is the electronic Hamiltonian operator, and $\Psi : (\mathbb{R}^3 \times \{\pm 1/2\})^N \to \mathbb{C}$ is the spin-inclusive *electronic wavefunction*, hereafter just *wavefunction*, of a molecular system containing M nuclei and N electrons. Each $x_i = (r_i, \sigma_i)$ is a (composite) variable for the ith electron; each $r_i \in \mathbb{R}^3$ is a spatial location, and each $\sigma_i \in \{\pm 1/2\}$ a spin state. Both H and Ψ carry an implicit dependence on the family of nuclear parameters $\{X_A = (R_A \in \mathbb{R}^3, Z_A \in \mathbb{N})\}_{A=1}^{M}$, where each $R_A \in \mathbb{R}^3$ is the location of the Ath nucleus in the system, and each $Z_A \in \mathbb{N}$ is its charge; the nuclei are said to be "clamped" [EA07, p. 3060] in place. The wavefunction is mandated to be antisymmetric in the electronic variables, and should in principle be normalised,

$$\langle \Psi, \Psi \rangle_{L^2} = 1, \tag{2.2}$$

where integration over a spin variable is understood to mean summation over spin states. In what follows, inner products will always be taken in L^2, so we will omit the subscript.

The eigenvalue $E = \langle \Psi, H\Psi \rangle$ corresponding to an eigenfunction Ψ of the Hamiltonian is the *electronic energy* of the electronic state represented by Ψ. Both electronic energies and the Hamiltonian itself are conventionally given in atomic units. For the former, the relevant unit is the Hartree (E_h). The latter takes the form

$$H = -\sum_{i=1}^{N} \frac{1}{2} \nabla_{r_i}^2 - \sum_{A=1}^{M} \sum_{i=1}^{N} \frac{Z_A}{\|r_i - R_A\|} + \sum_{i=1}^{N} \sum_{j>i}^{N} \frac{1}{\|r_i - r_j\|}. \tag{2.3}$$

As per Schneider [Sch09] and also with reference to [Can+03; Ham09], the appropriate solution space for the weak form of (2.1) is

$$\mathcal{V} := H^1((\mathbb{R}^3 \times \{\pm 1/2\})^N, \mathbb{C}) \cap \bigwedge_{i=1}^{N} L^2(\mathbb{R}^3 \times (\pm 1/2), \mathbb{C}), \tag{2.4}$$

where $H^1((\mathbb{R}^3 \times \{\pm 1/2\})^N, \mathbb{C})$ denotes the Sobolev space of those square-integrable trial wavefunctions Ψ with also square-integrable first derivatives with respect to the spatial locations of the electronic variables $\{x_i\}_{i=1}^N$, and the antisymmetrised product space $\bigwedge_{i=1}^N L^2(\mathbb{R}^3 \times (\pm 1/2))$ locates functions satisfying the requirement of antisymmetry in those variables. The *ground-state electronic energy* of the molecular system is then defined as the minimising eigenvalue

$$E_0 = \inf_{\substack{\Psi \in \mathcal{V} \\ \langle \Psi, \Psi \rangle = 1}} \langle \Psi, H\Psi \rangle, \tag{2.5}$$

again as in [Sch09], up to the requirement of normalisation, or almost equivalently [Can+03]. Under certain conditions, the infimum is well-defined; see, e.g., [Fri03] and references within. An energy-minimising wavefunction Ψ_0 which corresponds to the ground-state total energy is called a *ground-state wavefunction*.

Following, e.g., [CGH18], we define the *ground-state Born-Oppenheimer potential*, $V^\mathrm{BO} : (\mathbb{R}^3 \times \mathbb{Z})^M \to \mathbb{R}$, as

$$V^\mathrm{BO}(X_1, \ldots, X_M) := \sum_{1 \leq A < B \leq M} \frac{Z_A Z_B}{\|R_A - R_B\|} + \inf_{\substack{\Psi \in \mathcal{V} \\ \langle \Psi, \Psi \rangle = 1}} \big\langle \Psi, H\big[\{X_A\}_{A=1}^M\big]\Psi \big\rangle, \tag{2.6}$$

where we make explicit the dependence of the electronic Hamiltonian on $\{X_A\}_{A=1}^M$; something notationally similar is done in [RH13]. The initial summation term includes the *nuclear repulsion energy*, to make V^BO a better approximation of the true total energy of the molecular system; see and cf., e.g., [SO89; Can+03]. A number of practical applications require also or instead the gradient $\nabla_{\{R_A\}_{A=1}^M} V^\mathrm{BO}$; see, e.g., [Sch82; Can+03;

Sch03; EA07; CGH18]. For simplicity, however, we shall not deal explicitly with the evaluation of nuclear gradients in this thesis.

A ground-state solution to the electronic Schrödinger equation can be had analytically in the case of an isolated hydrogen-like atom [SO89; Yse10, Chap. 8], but this is not possible for larger systems [GH07]. Here, the high dimensionality of $\mathcal{V}$ frustrates the application of standard numerical techniques to (2.1); such schemes quickly run afoul of the "curse of dimensionality" [GH07, p. 216] as the number of electrons N increases.

Quantum chemistry prefers instead to approach the electronic problem via techniques that ultimately derive from certain approximating assumptions made about the structure of the solution. We will now outline four such *ab initio* methods, sometimes also called *wavefunction methods* [LeB05]. The latter name serves to distinguish them from a separate class of formal approaches, which explicitly treat not the wavefunction Ψ of a system, but instead the *electron density* $\rho(r) : \mathbb{R}^3 \to \mathbb{R}$. The most important density-based method is *density functional theory* (DFT). However, since we shall use and refer to it only in passing, we omit any explicit description of DFT, and refer the unfamiliar reader instead to the textbook of Parr and Yang [PY94], or [Can+03] for a more mathematical approach.

2.2. Standard *ab initio* methods

Most of the following material is basic background in quantum chemistry; we give only as quick a summary as possible. For deeper detail, we defer again to our general references, that is, [SO89; Can+03; EA07; Sch09; HOJ13; Jen17]. The lecture notes of Toulouse [Tou17] and of Benedikter and Sok [BS17] were also helpful general resources for this section.

Before we begin, we recall some particularly basic definitions and terminology from, e.g., [SO89; Can+03; EA07]; historically, see [Roo51]. Given an orthonormal collection of functions $\{\chi_i \in H^1(\mathbb{R}^3 \times \{\pm 1/2\}, \mathbb{C})\}_{i=1}^N$, called in context *(molecular) spin orbitals*, their *Slater determinant* [Sla29] is defined as

$$\Psi_{\mathrm{SD}}(x_1, \dots, x_N) = \frac{1}{\sqrt{N!}} \begin{vmatrix} \chi_1(x_1) & \chi_2(x_1) & \cdots & \chi_N(x_1) \\ \chi_1(x_2) & \chi_2(x_2) & \cdots & \chi_N(x_2) \\ \vdots & \vdots & \ddots & \vdots \\ \chi_1(x_N) & \chi_2(x_N) & \cdots & \chi_N(x_N) \end{vmatrix}, \tag{2.7}$$

which is antisymmetric in $\{x_i\}_{i=1}^N$ by construction and normalised consistent with (2.2). Each spin orbital can be decomposed as [EA07, (35)]

$$\chi_i(x) = \chi_i(r, \sigma) = \psi_i^\alpha(r)\alpha(\sigma) + \psi_i^\beta(r)\beta(\sigma); \tag{2.8}$$

Here, the *spin functions* $\alpha(\sigma) \neq \beta(\sigma)$ are valued in $\{0, 1\}$, and are multiplied by functions $\psi_i^\alpha, \psi_i^\beta \in H^1(\mathbb{R}^3, \mathbb{C})$ called *(molecular) spatial orbitals*.

2.2.1. The Hartree-Fock method (HF)

We refer here particularly to the conventional treatments in [SO89; EA07]. For alternative formulations in terms of the electron density, see [PY94; Can+03].

The *Hartree-Fock method* [Roo51; Hal51] obtains an approximate solution to (2.1) by constraining the wavefunction Ψ to membership of the set of Slater determinants [SO89; Can+03; EA07]. The *Hartree-Fock ground-state energy* is defined minimally over members of this set,

$$E_0^{\mathrm{HF}} = \inf_{\Psi_{\mathrm{SD}}} \{\langle \Psi_{\mathrm{SD}}, H\Psi_{\mathrm{SD}} \rangle\}. \tag{2.9}$$

Assuming for the moment its existence and uniqueness, the minimising Slater determinant is called the *Hartree-Fock ground-state wavefunction*, written Ψ_0^{HF}.

A constrained optimisation of $\langle \Psi_{\mathrm{SD}}, H\Psi_{\mathrm{SD}} \rangle$ in terms of the set of spin orbitals $\{\chi_i\}_{i=1}^N$ of a trial Slater determinant Ψ_{SD} produces, after some non-trivial manipulation, a set of nonlinear pseudo-eigenvalue equations, as in [EA07, (64)],

$$\hat{F}_{\mathrm{GHF}}\big[\{\chi_i\}_{i=1}^N\big]\chi_i(x) = \varepsilon_i \chi_i(x), \qquad 1 \leq i \leq N, \tag{2.10}$$

which are called the *canonical (general) Hartree-Fock equations*. Here, $\hat{F}_{\mathrm{GHF}}[\{\chi_i\}_{i=1}^N]$ is the *general Fock operator* parametrised by the set of spin orbitals $\{\chi_i\}_{i=1}^N$; we leave a full definition for [EA07]. For details and discussion on the existence and uniqueness of solutions to (2.9) and (2.10), see, e.g., [Can+03; Fri03; LL05; EA07; BS17] and references therein. In brief, however, a result of Lieb and Simon [LS74; LS77] guarantees an energy-minimising solution to (2.9) and transitively (2.10), at least when the molecule under study is not negatively charged, and that the spin orbitals of that solution are also the solutions to (2.10) with the lowest-lying eigenvalues.

In any case, approximate solutions to the Hartree-Fock equations are usually obtained by a Galerkin-style discretisation of $H^1(\mathbb{R}^3, \mathbb{C})$ formulated with reference to a single linearly independent family of K basis functions, $\{\phi_\mu \in H^1(\mathbb{R}^3, \mathbb{C})\}_{\mu=1}^K$ [SO89; Can+03; EA07]. The basis functions are called in context *atomic orbitals*. Each spatial orbital in a solution to (2.10) is approximated as

$$\psi_i \approx \sum_{\mu=1}^{K} c_{\mu i} \phi_\mu, \tag{2.11}$$

that is, as a *linear combination of atomic orbitals* (LCAO) [Roo51; Hal51]. We will come back to the precise choice of the atomic orbitals in Section 2.4.2 below.

In the *restricted Hartree-Fock* (RHF) formalism, which is applicable only for a *closed-shell* system where N is even, insertion of (2.11) into a slightly different form of (2.10) leads to a set of K equations in the LCAO coefficients $\{c_{\mu i}\}_{\mu,i=1}^K$ and the eigenvalues ε_i [SO89; EA07]. When these *Roothaan-Hall equations* [Roo51; Hal51] are considered in matrix form, they represent a single nonlinear generalised eigenproblem, which can be

approximately solved via iterative *self-consistent field* (SCF) methods. Numerical well-behaviour of the naïve SCF method as given in, e.g., [SO89] is not guaranteed [Can+03], but see, for instance, [CL00; Lev12].

Sets of equations analogous to those of Roothaan and Hall can be constructed and solved similarly in other, more general formalisms [PN54; EA07; JHS11]. The side-lengths of the involved matrices are in all cases small constant multiples of the size of the basis K, so the solution cost of the generalised eigenproblem encountered at each iteration of an SCF procedure scales as $\mathcal{O}(K^3)$. However, the actual formation of these matrices requires the evaluation of certain integrals taken over basis functions, particularly the *two-electron integrals* [SO89] or *electron-repulsion integrals* (ERIs) [Gil94]. Using standard notation, these are

$$(\mu\nu \,|\, \lambda\sigma) := \int_{\mathbb{R}^3} \int_{\mathbb{R}^3} \phi_\mu^*(r_1)\phi_\nu(r_1) \frac{1}{\|r_1 - r_2\|} \phi_\lambda^*(r_2)\phi_\sigma(r_2) \,\mathrm{d}r_1 \,\mathrm{d}r_2 \tag{2.12}$$

for $1 \leq \mu, \nu, \lambda, \sigma \leq K$. We will return briefly to their evaluation in Section 2.4.3 below, but note for now that the technical cost scaling of the Hartree-Fock method thus goes as $\mathcal{O}(K^4)$, and this expense is usually incurred repeatedly in practice, once at each SCF iteration [AFK82].

The full set of coefficients $\{c_{\mu i}\}_{\mu,i=1}^K$ produced by a converged SCF approach specifies some number $K \geq N$ of orthonormal spin orbitals [EA07].[1] Of these, the *occupied* orbitals with the lowest eigenvalues ε_i in (2.10) are taken as providing the best available approximation to Ψ_0^{HF}. The *virtual* spin orbitals left over are utilised by the various *post Hartree-Fock* [Can+03] techniques which we now consider.

2.2.2. Full configuration interaction (FCI)

The error in the ground-state energy that can be ascribed to the Hartree-Fock approximation,

$$E_0^{\mathrm{corr}} := E_0 - E_0^{\mathrm{HF}}, \tag{2.13}$$

is conventionally referred to as the *correlation energy* [Löw55; SO89; Can+03]. A better approximation to the true wavefunction, and thus the true ground-state energy including the correlation energy, can be obtained via the method of *full configuration interaction* (FCI), which involves a Galerkin discretisation of the full solution space $\mathcal{V}$. We outline this method following primarily Schneider [Sch09], again with some general reference to [SO89].

Presupposing an orthonormal set of spin orbitals $\{\chi_i\}_{i=1}^K \subset H^1(\mathbb{R}^3 \times \{\pm 1/2\}, \mathbb{C})$ for some $K \geq N$, Schneider constructs an accompanying subspace $\mathcal{V}_K \subset \mathcal{V}$ spanned by all Slater determinants composed of N distinct spin orbitals χ_i [Sch09]. The *FCI ground-state*

[1] This is an abuse of notation. The precise number of spin orbitals delivered depends on the HF formalism used, but like the side-lengths of the involved matrices, it is always a small constant multiple of K.

energy is then just the best approximation to the true ground-state energy available within $\mathcal{V}_K$ [Sch09, (7)],

$$E_0^{\mathrm{FCI}} := \min_{\substack{\Psi' \in \mathcal{V}_K \\ \langle \Psi', \Psi' \rangle = 1}} \{ \langle \Psi', H\Psi' \rangle \}, \tag{2.14}$$

and an *FCI ground-state wavefunction* is some $\Psi_0^{\mathrm{FCI}} \in \mathcal{V}_K$ that provides the minimum in (2.14), that is, a Ritz-Galerkin approximation of Ψ_0 in $\mathcal{V}_K$ under the constraint (2.2). Under some conditions, the FCI ground-state energies and wavefunctions provided by similarly-constructed members of a family $\{\mathcal{V}_K\}_{K \geq N}$ converge to E_0 and Ψ_0 quasi-optimally as $K \to \infty$ [Sch09, Thm. 3.1].

An explicit construction of the full set $\mathcal{V}_K$ is performed in terms of a *reference wavefunction*, itself a particular Slater determinant built from N of the spin orbitals and thus a member of $\mathcal{V}_K$ [Sch09; RS13]. The remaining Slater determinants are conceptually organised in terms of their derivation from the reference determinant via the substitution of one or more of its involved spin orbitals with spin orbitals drawn from the complement of the reference set. The standard choice of reference determinant is the ground-state Hartree-Fock wavefunction Ψ_0^{HF}, and in practice, an LCAO-discretised approximation to it. The full set of K spin orbitals are here then the N_{occ} occupied orbitals used to form Ψ_0^{HF}, which are conventionally indexed as $1 \leq i, j, k, \ldots \leq N_{\mathrm{occ}}$, as in, e.g., [CS00], and also the N_{virt} virtual orbitals produced via solution of the Roothaan-Hall equations or equivalent, indexed $N_{\mathrm{occ}} + 1 \leq a, b, c, \ldots \leq K$. In this context, the *singly-excited determinants* are those produced by the exchange of an occupied orbital indexed i with a virtual orbital indexed a,

$$\Psi_i^a := a_a^\dagger a_i \Psi_0^{\mathrm{HF}}, \tag{2.15}$$

written using the standard creation and annihilation operators of *second quantisation*; see, e.g., [SO89; CS00; HOJ13] for a full development of this formalism. Similarly, the *doubly-excited determinants* take the form

$$\Psi_{ij}^{ab} := a_a^\dagger a_b^\dagger a_j a_i \Psi_0^{\mathrm{HF}}, \tag{2.16}$$

and so on.

Using notation consistent with [SO89, (4.2a)], this leads to an explicit form for a trial FCI wavefunction $\Psi^{\mathrm{FCI}} \in \mathcal{V}_K$ as

$$\Psi^{\mathrm{FCI}} = c_0 \Psi_0^{\mathrm{HF}} + \sum_{i,a} c_i^a \Psi_i^a + \sum_{\substack{i<j \\ a<b}} c_{ij}^{ab} \Psi_{ij}^{ab} + \sum_{\substack{i<j<k \\ a<b<c}} c_{ijk}^{abc} \Psi_{ijk}^{abc} + \cdots, \tag{2.17}$$

in terms of coefficients which are written using the same indexing style as the excited determinants. For practical details on how the coefficients of Ψ_0^{FCI} can be obtained, see, e.g., [Dav75; SO89]. It suffices to say, however, that the cost scaling of an FCI calculation is prohibitive in K, i.e., as larger and larger collections of atomic orbitals are used to

discretise $H^1(\mathbb{R}^3, \mathbb{C})$, since the dimensionality of the problem is $\binom{K}{N}$ and thus roughly proportional to $\mathcal{O}(K^N)$ for large K [Sch09].

The elision of all terms in (2.17) beyond single and double excitations leads to a more tractable approximation technique referred to as CISD, for *configuration interaction, singles and doubles*; beyond triple excitations, CISDT, and so on; see, e.g., any of [SO89; Can+03; HOJ13; Jen17; Tou17]. These are, however, problematic in their application, since it is well-known that they lose the two very desirable properties of *size-consistency* and *size-extensivity*; see for details also [Bar81; CS00].

2.2.3. Møller-Plesset perturbation theory (MPn)

An alternative approach to the calculation of an approximate solution to the Schrödinger equation that includes an approximation to the correlation energy is provided by *Møller-Plesset perturbation theory* [MP34]. In the interest of brevity, we do not give any development of either Møller-Plesset theory or the underlying *Rayleigh-Schrödinger perturbation theory*, and refer the reader instead to [SO89, Secs. 6.1 and 6.5; Can+03, Sec. 35; Jen17, Sec. 4.8]; the concise summary in [Tou17] is also informative. We simply state that Møller-Plesset perturbation theory constructs a particular expansion of each eigenvalue E_i in (2.1) as a formal power series

$$E_i = \sum_{j=0}^{\infty} \lambda^j E_i^{(j)}, \tag{2.18}$$

in terms of a parameter $\lambda \in \mathbb{C}$ [SO89; Can+03; Jen17]. Assuming that this series converges, it provides an exact solution for the target eigenproblem for $\lambda = 1$.

Truncated sums of the terms $E_i^{(j)}$ can be used as approximations to the target eigenvalue E_i. In particular, the first-order Møller-Plesset ground-state energy is $E_0^{\mathrm{MP1}} = E_0^{(0)} + E_0^{(1)} = E_0^{\mathrm{HF}}$ [SO89; Can+03; Jen17]. The second-order Møller-Plesset term can be (re)written as

$$E_0^{(2)} = \frac{1}{2} \sum_{i=1}^{N} \sum_{j=1}^{N} \sum_{a=N+1}^{K} \sum_{b=N+1}^{K} \frac{[ia \,|\, jb][ai \,|\, bj] - [ia \,|\, jb][aj \,|\, bi]}{\varepsilon_i + \varepsilon_j - \varepsilon_a - \varepsilon_b}, \tag{2.19}$$

by slight manipulation of [SO89, (6.73)], where we denote as usual the *two-electron integrals* in terms of the spin orbitals as, with reference to [SO89, Tab. 2.2],

$$[ij \,|\, kl] := \sum_{\sigma_1, \sigma_2 \in \{\pm 1/2\}} \iint \chi_i^*(x_1) \chi_j(x_1) \frac{1}{\|r_1 - r_2\|} \chi_k^*(x_2) \chi_l(x_2) \,\mathrm{d}r_1 \,\mathrm{d}r_2. \tag{2.20}$$

Higher-order Møller-Plesset terms can in principle also be explicitly calculated, but the convergence behaviour of the series of MPn corrections can be problematic; see, e.g., [Ols+96; Lei+00; Cre11; HOJ13].

Note that the number of terms in (2.19) scales as $\mathcal{O}(K^4)$ in the size of the total number of occupied and virtual orbitals K, which is proportional to the number of atomic orbital basis functions. As we shall discuss in Section 2.5 below, the practical evaluation of $E_0^{(2)}$ and thus the MP2 total energy $E_0^{\mathrm{MP2}} = E_0^{\mathrm{HF}} + E_0^{(2)}$ in the LCAO setting requires an additional $\mathcal{O}(K^5)$ computational step; see, e.g., [SO89].

2.2.4. Coupled cluster approximations (CC)

We turn now to approximations of the true electronic energy and wavefunction via *coupled cluster* approaches. We follow here again most strongly [Sch09], but use some second-quantisation notation more like that in, e.g., [CS00; Jen17]. For related rigorous mathematical treatments, see [RS13; Roh13]; for a detailed development with a practical focus, see [CS00].

An FCI trial wavefunction built from a reference Hartree-Fock wavefunction Ψ_0^{HF} and a set of $K = N_{\mathrm{occ}} + N_{\mathrm{virt}}$ spin orbitals as in (2.17), which is subject not to (2.2) but instead to the *intermediate normalisation* constraint that $c_0 = 1$, can alternatively be written in terms of a *cluster operator*,

$$T = \sum_{k=1}^{\min(N_{\mathrm{occ}},\, N_{\mathrm{virt}})} T_k, \tag{2.21}$$

where each subsidiary operator T_k embodies k-fold excitations [CS00; Jen17],[2]

$$T_1 = \sum_{i,a} t_i^a a_a^\dagger a_i, \tag{2.22}$$

$$T_2 = \sum_{\substack{i<j \\ a<b}} t_{ij}^{ab} a_a^\dagger a_b^\dagger a_j a_i, \tag{2.23}$$

and so on. Rather than an FCI trial wavefunction $\Psi^{\mathrm{FCI}} = (1+T)\Psi_0^{\mathrm{HF}} \in \mathcal{V}_K$, the coupled cluster approximation considers one of the form [CS00; Sch09]

$$\Psi^{\mathrm{CC}} = e^T \Psi_0^{\mathrm{HF}} = \left(\sum_{k=0}^{\infty} \frac{T^k}{k!}\right) \Psi_0^{\mathrm{HF}} = \left(1 + T + \frac{1}{2}T^2 + \frac{1}{6}T^3 + \cdots\right) \Psi_0^{\mathrm{HF}}, \tag{2.24}$$

which is ultimately in terms of the complete set of values $\{t_i^a, \ldots, t_{ij}^{ab}, \ldots\} = \{t_\mu\}_{\mu \in \mathcal{I}}$, for some index set $\mathcal{I}$ that enumerates the various possible excitations, as in [Sch09]. The values t_μ are in context referred to as *amplitudes*. The rewriting is legitimate, in the sense that any function $\Psi_K \in \mathcal{V}_K$ can be written in terms of a cluster operator

[2] Noting that the presentation in [CS00] uses unrestricted indices and a compensatory prefactor for each T_k.

$\Psi_K = \Psi_K^{\mathrm{CC}} = e^T \Psi_0^{\mathrm{HF}}$ defined in terms of some particular set of amplitudes [Sch09, Thm. 4.3].

Since the trial wavefunctions Ψ^{CC} inhabit a nonlinear manifold within $\mathcal{V}$ rather than a linear subspace $\mathcal{V}_K$ [Sch09], a Ritz-Galerkin minimisation cannot be used to locate a ground-state couple cluster wavefunction.[3] Instead, a collection of *amplitude equations* are obtained by noting that, if $\{t_\mu\}_{\mu\in\mathcal{I}}$ is a set of amplitudes defined by $\Psi_0^{\mathrm{CC}} = e^T \Psi_0^{\mathrm{HF}} = \Psi_0^{\mathrm{FCI}}$, then [Sch09, Prop. 4.7]

$$\langle \Psi_\mu,\ e^{-T} H e^T \Psi_0^{\mathrm{HF}} \rangle = 0, \quad \forall \mu \in \mathcal{I}, \tag{2.25}$$

where each Ψ_μ is an appropriately excited determinant.

Much as in the CI case, the excitations included in the cluster operator T can be restricted, leading to a smaller set of amplitude equations. The problem obtained by a truncation of T as T_1+T_2, thus considering single and double excitations only, is referred to as *coupled cluster singles and doubles* (CCSD) [PB82]. Truncation as $T_1+T_2+T_3$, so with the explicit inclusion of triple excitations, is called CCSDT [NB87], etc. Here, however, the properties of size-consistency [CS00] and size-extensivity [BM07] can basically be retained.

This *projected coupled cluster* method (or rather, family of methods) can be viewed as a Galerkin approximation in a suitably-constructed space of amplitude vectors; see once more [Sch09] for such a construction, and for related existence and optimality results. In practice, approximate solutions to the (truncated) amplitude equations are calculated using iterative quasi-Newton schemes [Sch09; HOJ13]. In general, the computational cost required to evaluate the amplitude equations for a cluster operator truncated after nth-order excitations scales as $\mathcal{O}(N_{\mathrm{occ}}^n N_{\mathrm{virt}}^{n+2})$ per iteration [BM07], although there is significant scope for adjusting the involved prefactors and non-dominant terms; see, e.g., [KR97; CS00; KS01; Hir03; EH11].

Deep connections exist between coupled cluster theory and perturbation theory, and a variety of *perturbative correction* terms have been formulated which aim to improve the accuracy of truncated coupled cluster energies [BM07; CS00; HOJ13]. Of these, the most widely-used is the CCSD(T) correction of Raghavachari et al. [Rag+89], which compensates for a subset of the triple-excitation contributions omitted by a CCSD truncation relative to a CCSDT truncation. We will not describe the details of this correction here, but remark that, given a set of CCSD amplitudes, the cost of evaluating the CCSD(T) correction $\Delta E^{\mathrm{CCSD(T)}}$ scales as $\mathcal{O}(N_{\mathrm{occ}}^3 N_{\mathrm{virt}}^4)$ [BM07].

The principle of the CCSD(T) correction has been extended, first to a CCSDT(Q) correction [Bom+05], and then to a general series of $(n+1)$th-order perturbative corrections which can be applied to arbitrary nth-order coupled cluster truncations [KG05; KG08]. These are denoted CCSDT(Q), CCSDTQ(P), etc., and can be calculated at a cost that

[3] See, however, discussion of variational coupled cluster formulations in [CS00].

scales as $\mathcal{O}(N_{\text{occ}}^{n+1} N_{\text{virt}}^{n+2})$ [KG05].[4]

The various coupled-cluster approximations obtained by truncating the cluster operator T to a particular order, with or without the application of a perturbative correction to an obtained result, suggest a natural hierarchy of *ab initio* computational methods with regularly increasing cost and hopefully increasing accuracy; see, e.g., [CS00; Zas+18]. For example, similarly to as in [BM07], but with the addition of an entry for Hartree-Fock calculations and including cost scaling terms in the total number of atomic orbitals K:

$$\begin{aligned} &\text{HF } [\mathcal{O}(K^4)] \to \text{MP2 } [\mathcal{O}(K^5)] \to \text{CCSD } [\mathcal{O}(K^6)] \to \text{CCSD(T) } [\mathcal{O}(K^7)] \\ &\quad \to \text{CCSDT } [\mathcal{O}(K^8)] \to \text{CCSDT(Q) } [\mathcal{O}(K^9)] \to \cdots \to \text{FCI } [\mathcal{O}(K^N)]. \end{aligned} \tag{2.26}$$

2.3. Derived energetic properties

Although the ground-state energy is a fundamental observable in the quantum-mechanical sense [Yse10], practical applications of quantum chemistry focus more on *derived energetic properties* that are obtained secondarily from either calculated energies or the characteristics of the approximate wavefunction; see, e.g., [HOJ13, Chap. 15]. A key example of such a property is the *total atomisation energy* (TAE, or just the *atomisation energy*) of a molecule, which is "the energy required to dissociate the molecule into separate atoms in their electronic ground states" [Mar21, p. 6]. Formally, we define the (non-relativistic) TAE of a molecular system with M atoms as

$$E_{\text{atom}} := \left(\sum_{i=1}^{M} E_0^{(i)} \right) - E_0, \tag{2.27}$$

where $E_0^{(i)}$ is the ground-state electronic and thus total energy of the ith atom of the system; see, e.g., [HOJ13, Sec. 15.7.1; Mar21], up to choice of notation. Throughout this thesis, a post-inclusion of the nuclear repulsion energy in E_0, as in (2.6), will be implicit.

Atomisation energies are usually given in the chemical literature in units of $\text{kcal}\,\text{mol}^{-1}$ or $\text{kJ}\,\text{mol}^{-1}$; see, e.g., [FPH11; Kar16]. As mentioned above, total energies are measured and calculated in terms of Hartree per molecule (E_h). Energies in Hartree per molecule can be converted into values per mole through multiplication by Avogadro's constant [NIS19], that is, $1\,E_\text{h}\,\text{mol}^{-1} = 6.022\,140\,76 \times 10^{23}\,E_\text{h}$. Using the further relationships that $1\,E_\text{h} = 4.359\,744\,722\,207\,1(85) \times 10^{-18}\,\text{J}$ [NIS19] and that $1\,\text{cal} = 4.184\,\text{J}$ [TT08], we obtain to sufficiently high precision for our purposes here the conversion factors $1\,\text{kJ}\,\text{mol}^{-1} \approx 0.000\,380\,88\,E_\text{h}$ and $1\,\text{kcal}\,\text{mol}^{-1} \approx 0.001\,593\,60\,E_\text{h}$. A calculated energetic value which agrees with a reference value (either experimental or theoretical) to within a tolerance of $1\,\text{kcal}\,\text{mol}^{-1}$ is said to have obtained *chemical accuracy* [Mar96; FPH11; Lao+16]. We shall return to the topic of obtaining *ab initio* results to chemical accuracy in Chapter 4.

[4] Technically, the scaling as given in [KG05] is $\mathcal{O}(N_{\text{occ}}^{n+1} N_{\text{virt}}^{n+2} + N_{\text{virt}}^{n+1} N_{\text{occ}}^{n+2})$, but we assume here that N_{occ} is sufficiently smaller than N_{virt} that the first term dominates.

Approximately-calculated atomisation energies, and reaction energies more generally, can be more accurate than the individual total energies used in their derivation [Bak+00; HOJ13, Chap. 15]. This effect is usually attributed to a cancelling of the errors afflicting those total energies; in the terminology of [Bak+00], these include the *basis-set error* due to an incomplete discretisation of $H^1(\mathbb{R}^3, \mathbb{C})$, or the *intrinsic error* associated with, e.g., a truncation of the cluster operator in (2.24). The mathematical details of such error cancellations are not well-understood [Can+03, p. 31]. In any case, the beneficial effect of error cancellations is less strongly observed in the calculation of atomisation energies than in that of other types of reaction energies [Taj+04; KSM17], and so atomisation energies provide a useful test for the quality of high-accuracy quantum-chemical methods; see, e.g., [Fel13].

2.4. Computational details

We now briefly outline some important computational details that must be considered in the practical application of the *ab initio* methods outlined above. Again, this is basic material in applied quantum chemistry, and we limit ourselves to only the amount of detail needed to establish a clear context for later discussion.

2.4.1. The frozen-core approximation

All of the various post-HF methods outlined above either are, or in the case of MP2, can be [SO89] phrased in terms of sets of excited determinants. Contributions to the correlation energy E_0^{corr} associated with the excitation of those occupied orbitals (in the mathematical, Hartree-Fock sense) that can be identified with the core, i.e., non-valence orbitals (in the more standard chemical sense) tend to depend only weakly on the geometric conformation of a molecular system [HOJ13, Sec. 8.3.1; ŘH13]. The restriction of the excited determinants used in a post-HF method to only those corresponding to excited valence orbitals is called the *frozen-core* (FC) approximation [Jen17, Sec. 4.1]. Calculations performed in the absence of this approximation are in context referred to as *all-electron* (AE) calculations; see usage in, e.g., [HOJ13; Tho+21].

When calculating derived properties such as atomisation energies, the error due to the FC approximation may therefore be mostly nullified by a cancellation effect as mentioned above [HOJ13; Jen17]. The benefit is a potentially decisive reduction in computational cost; see, e.g., [Hal+03]. In the MP2 case, for example, that the FC approximation amounts to a restriction of the summation indices i which run over occupied orbitals in (2.19) can be seen from inspection of (6.71) in [SO89]; the evaluation cost of the resulting frozen-core summation is thus reduced by a prefactor quadratic in the number of occupied orbitals under consideration.

In standard practical application of the frozen-core approximation, some number N_{frozen} of occupied orbitals to be excluded from excitation may be chosen, usually equal to the

total number of expected core orbitals in the system under study; see for instance the documentation for the MP2 capabilities of the NWChem software package [Apr+20; NWCDoc]. It is often assumed that the core orbitals can be identified with the N_{frozen} canonical Hartree-Fock spin orbitals with the lowest eigenvalues ε_i in (2.10), and so just these orbitals are then removed from consideration. This assumption is, however, not always legitimate [Pet98].

2.4.2. Atomic orbitals and basis sets

A key consideration in the implementation of the *ab initio* methods described above is the finite collection of basis functions $\{\phi_\mu\}_{\mu=1}^K \subset H^1(\mathbb{R}^3, \mathbb{C})$ used to discretise the spatial orbitals according to (2.11). We give here a brief summary of three classes of basis functions and their use in the construction of such *basis sets*, following from the outset [HOJ13, Chaps. 6 and 8] and also making general reference to [Can+03, Secs. 23 and 24; Gil94; EA07].

We seek, in effect, a suitable complete basis for $H^1(\mathbb{R}^3, \mathbb{C})$ that can be finitely sampled in a way that allows systematic improvement of the solution space with a correspondingly regular increase in computational cost [Can+03; HOJ13]. For the moment, we consider only basis sets for monoatomic problems. An historically important class of basis functions in this setting are the *Slater-type orbitals* (STOs) [Sla30; HOJ13, Sec. 6.5.5]. Each STO is defined in spherical coordinates as $\Psi^{\text{STO}}_{lmn}(r, \theta, \phi; \zeta) = R^{\text{STO}}_n(r; \zeta) Y^m_l(\theta, \phi)$, the product of a radial part and an angular part. The latter is a spherical harmonic; the former is given by [HOJ13, (6.5.26)]

$$R^{\text{STO}}_n(r; \zeta) = \frac{(2\zeta)^{\frac{3}{2}}}{\sqrt{(2n+1)!}} (2\zeta r)^{n-1} \exp(-\zeta r) \tag{2.28}$$

parametrised by some $\zeta > 0$. Both radial and angular parts are also parametrised by m, n, $l \in \mathbb{Z}$, where $n \geq 1$, $0 \leq l < n$, and $-l \leq m \leq l$. The STOs are complete in $L^2(\mathbb{R}^3, \mathbb{C})$ for any particular choice of ζ.

A similar radial form, but with an exponential in terms of r^2 [HOJ13, (6.6.7)],

$$R^{\text{SH-GTO}}_{nl}(r; \alpha) = \frac{2(2\alpha)^{\frac{3}{4}}}{\pi^{\frac{1}{4}}} \sqrt{\frac{2^{2n-l-2}}{(4n-2l-3)!!}} \left(\sqrt{2\alpha} \cdot r\right)^{2n-l-2} \exp(-\alpha r^2), \tag{2.29}$$

produces the *spherical-harmonic Gaussian-type orbitals* (spherical harmonic GTOs) [HOJ13, Sec. 6.6.3], parametrised by $\alpha > 0$ and written $\Psi^{\text{SH-GTO}}_{lmn}(r, \theta, \phi; \alpha) = R^{\text{SH-GTO}}_{nl}(r; \alpha) Y^m_l(\theta, \phi)$. These are also complete for $L^2(\mathbb{R}^3, \mathbb{C})$ for fixed α and varying m, n, l, and are related to another complete family of functions which are defined in Cartesian coordinates rather than spherical, the *Cartesian Gaussian-type orbitals* (Cartesian GTOs) [Boy50; Gil94; HOJ13, Sec. 6.6.7]

$$\Psi^{\text{C-GTO}}_{l_x l_y l_z}(x, y, z; \alpha) = N(\alpha, l_x, l_y, l_z) x^{l_x} y^{l_y} z^{l_z} \exp(-\alpha(x^2 + y^2 + z^2)), \tag{2.30}$$

where $\alpha > 0$ is a parameter as before, $l_x, l_y, l_z \geq 0$, and $N(\alpha, l_x, l_y, l_z)$ is a normalisation constant.

For the STO and GTO families of functions, completeness in $H^1(\mathbb{R}^3, \mathbb{C})$ can also be shown to hold in some cases; see [KB77b; KB77c] and discussion in [Can+03]. There are also other ways to construct derived families of functions that are complete, at least in $L^2(\mathbb{R}^3, \mathbb{C})$; in addition to the above citations, see also [KB77a]. But since in practice a basis set must always be finite, and given that the cost scaling of the various *ab initio* methods is polynomial in the size of the basis set, emphasis is placed instead on constructing specific basis sets that lead to somehow optimal monoatomic solutions for the use of a given finite number of functions [Can+03; HOJ13]. For certain computational reasons to which we will return below, modern basis sets are almost exclusively built from GTOs; see, e.g., [Gil94; Can+03; EA07]. We gloss over an important practical detail here, namely that GTO basis functions are not specified alone, but instead in linear combination; for details and motivation, see, e.g., [HSP69; Raf73; Gil94; HOJ13]. These *contracted* basis functions come in matched sets called *basis shells*; again, for details, see, e.g., [Gil94; HOJ13, Chap. 9].

Turning now to the construction of basis sets for poly- rather than monoatomic systems, we encounter a potentially confusing double usage of the term "basis set" that is standard in the quantum chemical literature. When a particular name is applied, e.g., "the" cc-pVDZ basis set [Dun89], quantum chemists refer in fact to a predefined collection of multiple distinct monoatomic basis sets, each one corresponding to and constructed for a particular atomic species [Can+03; HOJ13]. These monoatomic basis sets can be used to construct a basis set for a calculation over a polyatomic system: for each atom in that system, the functions from the appropriate monoatomic basis set are included in the full-system basis set, with each included function translated so as to be centred at the nuclear coordinates R_A of the relevant atom.

A panoply of such basis sets (i.e., collections of monoatomic basis sets) are available; see, e.g., the reviews of Huzinaga [Huz85] and Hill [Hil12], as well as [HOJ13, Chap. 8; Jen17, Chap. 5]. In this thesis, we will make particular use of the *correlation-consistent* basis sets due to Dunning and co-workers [Dun89; KDH92; WD95]; see Section A.2 for a complete list and full citations. Most important here are the family of cc-pVnZ basis sets [Dun89], where $n \geq 2$; one speaks of each cc-pVnZ as following an *n-tuple zeta* pattern. Again, we defer a detailed description of these, as well as the *augmented* aug-cc-pVnZ [KDH92] and *core-valence* cc-pCVnZ [WD95] families of basis sets, to, e.g., [Hil12; HOJ13; Jen17]. It is, however, particularly important to note that the (aug)-cc-pCVnZ basis sets are suitable for all-electron post-HF calculations, whereas the (aug)-cc-pVnZ basis sets are intended only for frozen-core calculations.

The basis-set errors of calculated energetic quantities obtained with, e.g., the cc-pVnZ sets decrease regularly as $n \to \infty$, and such results are thus commonly said to approach the *complete basis set* (CBS) limit; see, e.g., [FPH11]. The number of contracted basis functions placed on a first-row atom according to cc-pVnZ is given

by $\frac{1}{3}(n+1)(n+\frac{3}{2})(n+2) \sim n^3$ [HOJ13, (8.3.7); Hel+97; CGH18]. Similar formulae can be given for species in other rows, and also for the augmented and/or core-valence correlation-consistent basis sets [HOJ13, Sec. 8.3.4]. As a result, the cc-pVnZ, cc-pCVnZ, aug-cc-pVnZ, and aug-cc-pCVnZ families can be viewed as providing four distinct basis set hierarchies, which either have been constructed or could in principle be constructed according to a well-defined set of rules. These hierarchies are systematically improvable in quality, but such improvement incurs a correspondingly regular increase in cost; see, e.g., an application in [CGH18].

2.4.3. Efficient calculation of two-electron integrals

We refer in this section generally to [Gil94; EA07], and mention again that we are familiar with most of these concepts from our previous investigation in [Bar09]. Assuming the use of the LCAO ansatz (2.11) to discretise the spatial orbitals ψ_i, a practical implementation of any of the methods detailed above requires explicit evaluation of all of the integrals $(\mu\nu \,|\, \lambda\sigma)$ in (2.12) [SO89; Gil94; HOJ13]. The practical value of Cartesian GTOs stems from an ERI calculation procedure discovered by Boys [Boy50]; for various improvements on his basic idea, see, e.g., [MD78; RDK83; OS86], and the review of Gill [Gil94]. For details of computationally efficient implementations of these and related ERI-calculation algorithms, see, e.g., [Bar09; Sun15; PC16]. Since it is possible to rewrite ERIs over spherical-harmonic GTOs in terms of ERIs over Cartesian GTOs [SF95], these approaches can also function in terms of basis sets composed of spherical-harmonic GTOs.

Given a basis set $\{\phi_\mu\}_{\mu=1}^{K}$, there are clearly K^4 formal ERIs $(\mu\nu \,|\, \lambda\sigma)$ defined in terms of those functions. However, permutational symmetries between the indices in (2.12) reduce the number of distinct integrals that must be explicitly calculated [BH71; SO89; Gil94]. In particular, if the basis functions are real-valued, as is usually the case [EA07; HOJ13] and certainly so for the Cartesian GTOs of (2.30), then the number of distinct ERIs reduces to $\frac{1}{8}K(K+1)(K^2+K+2)$; see, e.g., [BH71, (27); Gil94, (6)].

More powerfully, the magnitude of an ERI $(\mu\nu \,|\, \lambda\sigma)$ shows a dependence on the distance between the nuclear centres on which the involved atomic orbitals are placed. As a result, the number of ERIs that are practically non-zero can scale closer to $\mathcal{O}(K^2)$ than to the naïve $\mathcal{O}(K^4)$; see, e.g., discussion and further references in [Gil94; Hea96; EA07]. The *a priori* identification of precisely which ERIs are in this sense regarded as *non-negligible* is performed via *integral prescreening* techniques. For example, a well-known approach due to Häser and Ahlrichs [HA89] that requires only $\mathcal{O}(K^2)$ initial evaluations of (2.12) exploits the fact that the ERI form describes an inner product, leading to a Cauchy-Schwarz inequality [HA89, (11) and (12)],

$$|(\mu\nu \,|\, \lambda\sigma)| \leq \sqrt{|(\mu\nu \,|\, \mu\nu)|} \cdot \sqrt{|(\lambda\sigma \,|\, \lambda\sigma)|}, \tag{2.31}$$

and thence a diagnostic mechanism for at least some of those ERIs $(\mu\nu \,|\, \lambda\sigma)$ which have magnitude below some $\varepsilon_{\mathrm{ERI}} > 0$. Häser and Ahlrichs suggested applying their bound in

terms of basis shells rather than for individual basis functions. Tighter bounds can also be applied in similar fashion [Gil94]. The involved prescreening threshold ε_{ERI} is usually exposed as a tuneable parameter of calculation; see, for instance, the documentation for NWChem [Apr+20; NWCDoc].

2.5. An abstract cost model for *ab initio* calculations

We close the chapter by assembling an *abstract cost model* for total energy calculations performed according to the *ab initio* wavefunction-based methods outlined above. This abstract cost model assigns a dimensionless number to the problem posed by some combination of molecular system, basis set, and *ab initio* method. The ratio between the abstract costs of any two such problems represents an estimate of the relative difference in computational effort between them, in a way that ignores practical details, such as specific computational resources used, as well as the precise implementation of an algorithm in code.

A simple example of such a cost model is given by Chinnamsetty et al. [CGH18], who anticipate the cost of a Hartree-Fock calculation using the cc-pVnZ basis set asymptotically as

$$\mathcal{C}^{\text{HF}}_{\text{cc-pV}n\text{Z}}(X_1, \ldots, X_M) \lesssim M^3 n^9 \tag{2.32}$$

as the number of atoms/nuclei M grows. This follows firstly since the total size of the discretising polyatomic basis set K goes like $K \sim Mn^3$ (see Section 2.4.2 above), and secondly from the $\mathcal{O}(K^3)$ scaling associated with solution of the generalised eigenvalue problem prescribed by the discretised Hartree-Fock equations. Note that (2.32) is predicated on the assumption that the number of non-negligible atomic ERIs that must be calculated scales as $\mathcal{O}(K^3)$ or less.

The cost model we collate here can be applied not only to Hartree-Fock calculations, but also to MP2 calculations and coupled cluster calculations to arbitrary excitation orders. The model is asymptotic, similarly to that of Chinnamsetty et al. [CGH18], and is generally based upon the computational complexities of individual algorithms, with the explicit insertion of some tuneable constant factors. Thus, we are really only just collecting and restating existing formulae.

We place no restrictions on the basis set applied to the system under study; rather, for any input problem, our cost model is defined in the total number of functions K in the resulting full-system basis set. In the practical implementation of this cost model which we have used in this work, we calculate K using functionality provided by the PySCF package [Sun15; Sun+17; Sun+20]. We also require an estimate of the number of non-negligible ERIs $N_{\text{ERI}} \leq K^4$ over those atomic orbitals. An ERI is considered non-negligible if $|(\mu\nu \,|\, \lambda\sigma)| \geq \varepsilon_{\text{ERI}}$, where the threshold is here taken consistently to be $\varepsilon_{\text{ERI}} = 10^{-12}$.

To estimate N_{ERI}, we used a simple Cauchy-Schwarz prescreening approach as per (2.31), applied at the basis shell level and considering the complete permutational symmetry mentioned above. The required ERIs were calculated via the LIBCINT library [Sun15]. The involved ideas are all well-known, for very non-exhaustive example [HA89; Gil94; CS97; Bar09; HOJ13, Sec. 9.12.4; Bar+20], up to the trivial difference that we only count non-negligible ERIs rather than explicitly evaluating and using them. Sketching anyway for completeness, using essentially the notation of [HA89] and up to the consideration of full basis shells instead of lone basis functions, the estimation reduces to first evaluating all $\{Q_{\mu\nu} := \sqrt{|(\mu\nu\,|\,\mu\nu)|}\}_{\mu\leq\nu}$, and then counting for each pair $\mu \leq \nu$ those terms $Q_{\lambda\sigma} \geq \varepsilon_{\mathrm{ERI}}/Q_{\mu\nu}$ for symmetry-appropriate $\lambda \leq \sigma$. An efficient implementation follows readily from first sorting $\{Q_{\mu\nu}\}_{\mu\leq\nu}$ in descending order; this basic idea can be traced back at least to [CS97].

We begin with the optimisation of the ground-state Hartree-Fock wavefunction, which is required both to evaluate a ground-state Hartree-Fock energy, and also as the first step in all of the post-HF methods outlined above and considered in this thesis. We assume the use of a *direct SCF* procedure [AFK82], whereby all required ERIs are reevaluated during the formulation of the Hartree-Fock eigenproblem at each individual iteration. The cost of each SCF iteration thus depends firstly on the number of non-negligible ERIs that must be (re)evaluated, and secondly on the cost of solution of the Roothaan-Hall equations or equivalent. We make here the simplification[5] of assuming that the evaluation of each ERI requires a relatively significant cost that is on average constant, and so introduce an explicit average cost factor f_{ERI} which weights the calculation and inclusion of each ERI. Under the further assumption that an SCF calculation requires on average some $N_{\mathrm{iter}}^{\mathrm{HF}}$ iterations to achieve convergence to some desired level of accuracy, the abstract cost of a Hartree-Fock calculation is modelled as

$$\mathcal{C}^{\mathrm{HF}}(\cdots) = N_{\mathrm{iter}}^{\mathrm{HF}}\big(f_{\mathrm{ERI}}N_{\mathrm{ERI}} + K^3\big). \tag{2.33}$$

Here and for the remainder of this section, we write for simplicity, e.g., $\mathcal{C}^{\mathrm{HF}}(\cdots)$ to indicate that the abstract cost is a function of both system and basis set.

We turn now to the costs associated with post-Hartree-Fock methods for calculating the correlation energy. These are given in terms of the number of correlated orbitals, N_{corr}, that is, the number of occupied spin orbitals N_{occ} that are excited, or not, as per the frozen-core approximation. If the frozen-core approximation is applied, then $N_{\mathrm{corr}} \leq N_{\mathrm{occ}}$; in the case of an all-electron calculation, $N_{\mathrm{corr}} = N_{\mathrm{occ}}$. The total number of frozen orbitals in a polyatomic system is derived as the sum of the monoatomic values given in Section A.3, weighted by the number of times those atoms appear in the full system.

Implementations of all of the various post-HF methods require explicit evaluations of the molecular two-electron integrals $[ij|kl]$ in (2.20). These reduce to the evaluation of

[5] And it is certainly a simplification; cf., e.g., the much more thorough treatments in [Gil94; HOJ13, Chap. 9].

expressions in terms of two-electron integrals over atomic orbitals; in the RHF case, for example, basically $[ij|kl] = \sum_{\mu\nu\lambda\sigma} c^*_{\mu,i} c_{\nu,j} c^*_{\lambda,j} c_{\sigma,l} (\mu\nu|\lambda\sigma)$. Naïve evaluation of all of these quantities would require $\mathcal{O}(K^8)$ operations; however, this can be reduced to $\mathcal{O}(K^5)$ by phrasing the *four-index transformation* as a series of successive tensor contractions, for example [Ben72; WHR96]:

$$[i\nu|\lambda\sigma] = \sum_{\mu} c^*_{\mu,i} (\mu\nu|\lambda\sigma), \qquad [ij|\lambda\sigma] = \sum_{\nu} c_{\nu,j} [i\nu|\lambda\sigma], \tag{2.34}$$

$$[ij|k\sigma] = \sum_{\lambda} c^*_{\lambda,k} [ij|\lambda\sigma], \qquad [ij|kl] = \sum_{\sigma} c_{\sigma,l} [ij|k\sigma]. \tag{2.35}$$

An efficient implementation of these contractions is non-trivial in practice [Raj+17], as is the exploitation of any negligibility of the tensor entries $[(\mu\nu \,|\, \lambda\sigma)]_{\mu\nu\lambda\sigma}$ which may be estimable by the use of integral prescreening techniques [WHR96].

The MP2 contribution (2.19) involves only $\mathcal{O}(N^2_{\text{corr}} N^2_{\text{virt}})$ terms. Thus, it is common to see the cost of an MP2 energy evaluation given as just $\mathcal{O}(K^5)$, as in, e.g., [WHR96; Jen17]. Although technically correct, this is somewhat misleading, since the practical cost of an MP2 evaluation dominates that of the necessarily preceeding Hartree-Fock calculation only slowly as the size of the system and/or basis set is increased; see comments in [SA89; BP02; Jen17].

In fact, the expression (2.19) involves only molecular ERIs like $[ia \,|\, jb]$ (up to the permutational symmetry of [SO89, (2.99b)] that is obtained with real atomic orbitals), where a and b index virtual spin orbitals and i and j index correlated spin orbitals [SA89]. Thus, only a subset of the complete collection of molecular ERIs must be explicitly calculated, as in [WHR96]. If the tensor contraction given above is performed first over the correlated indices i and j, and if the sparsity of the atomic ERIs is considered in the first contraction (on this point, see, e.g., discussion in [SL97]), then a tighter abstract cost for an MP2 total-energy calculation can be given as

$$\begin{aligned} \mathcal{C}^{\text{MP2}}(\cdots) = \mathcal{C}^{\text{HF}}(\cdots) &+ N^2_{\text{corr}} N^2_{\text{virt}} + f_{\text{ERI}} N_{\text{ERI}} \\ &+ N_{\text{corr}} N_{\text{ERI}} + N^2_{\text{corr}} K^3 + N^2_{\text{corr}} N_{\text{virt}} K^2 + N^2_{\text{corr}} N^2_{\text{virt}} K. \end{aligned} \tag{2.36}$$

The first term on the right-hand side is the cost of the initial Hartree-Fock calculation. The second term represents the explicit evaluation of the MP2 energy term. The third term is the cost of calculating the non-negligible atomic-orbital ERIs. The remaining terms represent the cost of each of the four successive contractions required to calculate the molecular ERIs; if $N_{\text{corr}} < N_{\text{virt}}$, as is usually the case, it is easy to see that this represents the optimal contraction order, at least in terms of a raw operation count [Ben72; WHR96; Raj+17]. Variations of this contraction pattern are used by many MP2 implementations, e.g., [HPF88; SA89; WHR96; SL97]. Note that these implementations generally avoid forming complete tensors of atomic and/or molecular

ERIs, in favour of direct or "semi-direct" approaches that balance for intermediate memory/storage costs at the expense of some redundant computation. We also neglect the possibility to consider here permutational symmetries [Ben72].

Finally, we extend our cost model to coupled cluster total energy calculations, both with and without a perturbative correction. The cost of a coupled cluster calculation involving excitations up to level n (where $n = 2$ for CCSD, $n = 3$ for CCSDT, etc.) is modelled simply according to the asymptotic cost scaling given in Section 2.2.4 as

$$\mathcal{C}^{\mathrm{CC}(n)}(\cdots) = f_{\mathrm{ERI}} N_{\mathrm{ERI}} + K(N_{\mathrm{corr}} + N_{\mathrm{virt}})^4 + N_{\mathrm{iter}}^{\mathrm{CC}} N_{\mathrm{corr}}^{n} N_{\mathrm{virt}}^{n+2}. \tag{2.37}$$

The first and second terms on the right-hand side represent calculation of all non-negligible atomic ERIs and their complete transformation to molecular ERIs; for simplicity, we do not model the four-index transform in as much detail as for the case of MP2. $N_{\mathrm{iter}}^{\mathrm{CC}}$ is a constant representing the average number of iterations required to solve the projected coupled cluster amplitude equations. We assume that the impact of the frozen-core approximation can be captured by substituting N_{corr} for N_{occ} in the original asymptotic expression.

The cost of a coupled cluster calculation with an additional non-iterative perturbative correction for excitations at order $n + 1$ (for example, $n = 2$ for CCSD(T)) is then just

$$\mathcal{C}^{\mathrm{CC}(n)(n+1)}(\cdots) = \mathcal{C}^{\mathrm{CC}(n)}(\cdots) + N_{\mathrm{corr}}^{n+1} N_{\mathrm{virt}}^{n+2}. \tag{2.38}$$

The cost of ERI calculation and transformation will likely not be significant for all but the smallest problems, and these terms are explicitly included in (2.37) and (2.38) only for completeness.

For the remainder of this thesis, we apply this abstract cost model with the specific values $N_{\mathrm{iter}}^{\mathrm{HF}} = 15$, $N_{\mathrm{iter}}^{\mathrm{CC}} = 15$, and $f_{\mathrm{ERI}} = 50$. The first two values are chosen to be roughly consistent with the number of iterations that we observed in calculations performed with the three quantum chemistry software packages that were used in the preparation of this thesis, namely MRCC [Kál+20; MRCC], NWChem [Apr+20], and PySCF [Sun15; Sun+17; Sun+20]; see Section A.1. The final value, $f_{\mathrm{ERI}} = 50$, was somewhat arbitrarily chosen as a plausible estimate based upon informal experimentation and the general experience of the author. In any case, the values $N_{\mathrm{iter}}^{\mathrm{HF}}$, $N_{\mathrm{iter}}^{\mathrm{CC}}$, and f_{ERI} are only prefactors and do not change the overall scaling behaviour of the cost model as the number of occupied and virtual orbitals increases.

Naturally, the abstract cost model is by construction only an abstraction, and becomes valid only in the asymptote. A comprehensive assessment of the relationship between abstract costs of calculations and their corresponding true costs as measured in terms of operation count or elapsed "wall time" would require the consideration of many complicated and interacting practical factors, including but not limited to the impact of parallelism, and is beyond the scope of this thesis. However, we mention for completeness that preliminary investigation in this direction suggests, in particular, a non-negligible

disagreement between the rates of growth of the abstract and true costs of some CCSD and CCSD(T) calculations performed with PySCF and NWChem. It seems reasonable to assume that this is simply a preasymptotic expression of computational work ignored by the abstract cost model; indeed, the disagreement does seem to eventually trend towards stability, although only slowly and for quite substantial problems. Nevertheless, a rigorous analysis of this effect is important, and is left for future work.

3. An order-theoretic combination technique

The focus of this thesis is the efficient approximation of energetic properties of molecules by combination of approximate solutions to the Schrödinger equation, obtained using the methods described in the previous chapter. Our primary tool will be the *order-theoretic combination technique* which we will develop in this chapter. As the name suggests, this represents an extension of the "standard" combination technique, a well-known and well-understood multivariate extrapolation method that has particular advantages in some high-dimensional problem settings [GSZ92; BG04; Gar12b; Heg+16; TW18].

We will begin by sketching the standard combination technique to establish context, and then briefly review its history and applications. We will develop a new formulation of the combination technique in a setting of abstract hierarchies of functions, which are organised according to a general class of partial orderings. We will then develop an algorithm for the adaptive assembly of particular subsets of these model hierarchies, which are intended to allow the approximation of a target function in a *quasi-optimal* [NTT15; TW18] manner.

3.1. The standard combination technique

All of the following ideas are well-known in the literature, up to different settings, precise formulations, and notations; we make general reference to, e.g., [GSZ92; Heg03; BG04; Gar12b; Har16a; Heg+16; TW18]. We proceed quickly and informally, and choose the structure of our presentation, and our notation and nomenclature, to lead directly into the latter part of the chapter. We deliberately de-emphasise the usually stressed connections between the combination technique and the theory of sparse grids.

Let V be a vector space of d-dimensional functions, and pick some interesting $f \in V$. As in, e.g., [Heg03; Har16a; Heg+16; TW18] up to exact notation, we introduce multiindices written like $\mathbf{m}, \mathbf{n} \in \mathbb{N}^d$, equipped with a componentwise partial ordering such that $\mathbf{m} \leq \mathbf{n}$ whenever $m_i \leq n_i$ for all $1 \leq i \leq d$. We assume that there exists a particular family of functions $\{f_{\mathbf{m}} \in V\}_{\mathbf{m} \in \mathbb{N}^d}$, which we picture as approximations to f that become somehow more accurate as the components m_i of their multiindices $\mathbf{m}$ increase. We will call each such $f_{\mathbf{m}}$ a *model function* of f, and f itself the *target function*.

Consistent with the idea of, e.g., [Heg+16, Prop. 1], we define, recursively, the *hier-*

archical surplus of each $f_{\mathbf{m}}$ to be

$$\tilde{f}_{\mathbf{m}} := f_{\mathbf{m}} - \sum_{\mathbf{m}'<\mathbf{m}} \tilde{f}_{\mathbf{m}'}; \tag{3.1}$$

cf. also [PZ99; BG04]. For each $k \in \mathbb{N}$, let S_k be the truncation of the formal sum $\sum_{\mathbf{m}\in\mathbb{N}^d} \tilde{f}_{\mathbf{m}}$ after only those surplus terms $\tilde{f}_{\mathbf{m}}$ for which all multiindex components $m_i \leq k$, that is,

$$S_k = \sum_{\substack{\mathbf{m}\in\mathbb{N}^d \\ \|\mathbf{m}\|_\infty \leq k}} \tilde{f}_{\mathbf{m}}. \tag{3.2}$$

Trivially and by construction, then, $S_k = f_{(k,k,\ldots,k)}$, and we can view the terms of the sequence $(S_k = f_{(k,k,\ldots,k)})_{k=0}^{\infty}$ as approximations of f that are systematically improvable by increasing k, of course up to the lack of precision in our informal setup. This also holds if we take instead an alternative, non-recursive definition of the hierarchical surplus, cf., e.g., [PZ99; Hul14, (5.60); NTT15, (3); Heg+16, (5); Won16, Def. 1.2.25],

$$\tilde{f}_{\mathbf{m}} = \sum_{\substack{\mathbf{n}\in\{0,1\}^d \\ \mathbf{n}\leq\mathbf{m}}} (-1)^{\|\mathbf{n}\|_1} f_{\mathbf{m}-\mathbf{n}}. \tag{3.3}$$

Later in the chapter, we will explicitly rederive the equivalence of (3.1) and (3.3).

The essence of the *combination technique* [GSZ92], see also [BGR94; Bun+94], is the replacement of the terms S_k in such a sequence with differently-truncated partial summations of $\sum_{\mathbf{m}\in\mathbb{N}^d} \tilde{f}_{\mathbf{m}}$. We write, for instance, the canonical choice as

$$S_{I_L} = \sum_{\mathbf{m}\in I_L} \tilde{f}_{\mathbf{m}}, \tag{3.4}$$

differentiated from S_k notationally by the subscript indicating not a natural number but instead a particular *index set*, here $I_L = \{\mathbf{m} \in \mathbb{N}^d \mid \|\mathbf{m}\|_1 \leq L\}$ for some $L \in \mathbb{N}$ [GSZ92; Gar12b]. Under a simple but practically quite strong assumption about the model functions $f_{\mathbf{m}}$, the terms of the sequence $(S_{I_L})_{L=0}^{\infty}$ converge at least pointwise to f at almost the same rate as do those of $(S_k)_{k=0}^{\infty}$, but carry only a much more reasonable cost scaling.

As is standard, we will illustrate this in the two-dimensional case, directly following the original formulation of [GSZ92]; see also [Gar12b]. Let $f : [0,1]^2 \to \mathbb{R}$, and take each $f_{(i,j)}$ to be a discretisation of f onto a rectilinear grid within $[0,1]^2$ with generally anisotropic mesh widths $h_i = 2^{-i}$ and $h_j = 2^{-j}$. Suppose that, pointwise,

$$f - f_{(i,j)} = C_1(h_i)h_i^2 + C_2(h_j)h_j^2 + D(h_i,h_j)h_i^2h_j^2, \tag{3.5}$$

and that there exists some $K > 0$ such that for all h_i and h_j, it holds that $|C_1(h_i)| \leq K$, $|C_2(h_j)| \leq K$, and also $|D(h_i,h_j)| \leq K$. Then it can be shown that $|f - S_{I_L}| =$

$\mathcal{O}(h_L^2 \log(h_L^{-1}))$ [GSZ92; Gar12b]; that is, the combinations S_{I_L} converge pointwise to f at the same rate, up to a logarithmic factor, as do the models $f_{(k,k)}$ and thus the partial sums S_k.

Suppose, though, that the computational cost $\mathcal{C}_{(i,j)}$ of calculating each $f_{(i,j)}$ is $\mathcal{C}_{(i,j)} = \mathcal{O}(h_i^{-1} h_j^{-1})$, so linear in the number of grid points. The summation S_{I_L} involves only those $\mathcal{O}(L^2)$ model functions $f_{(i,j)}$ with $i + j \leq L$, the evaluation of which requires a combined cost of $\sum_{i=0}^{L} \sum_{j=0}^{L-i} \mathcal{C}_{(i,j)} = \mathcal{O}(L \cdot 2^L) = \mathcal{O}(h_L^{-1} \log(h_L^{-1}))$ [Bun+94; GT95], compared to the $\mathcal{O}(h_k^{-2})$ cost incurred to evaluate each $S_k = f_{(k,k)}$. An additional saving, albeit one which does not change the overall cost scaling, is found by noting that S_{I_L} can be written in terms of only $2L + 1 = \mathcal{O}(L)$ model functions [GSZ92; Gar12b],

$$S_{I_L} = \sum_{i=0}^{L} f_{(i,L-i)} - \sum_{i=0}^{L-1} f_{(i,L-i-1)}. \tag{3.6}$$

We will explicitly rederive this identity later in the chapter. Since this is just a particular linear combination of model functions, hence we call S_{I_L} a *combination sum*.

In the generally d-dimensional version of the same setup, writing $[d] = \{1, 2, \ldots, d\}$, suppose that for each $f_{\mathbf{m}}$ there exists an analogous pointwise error expansion

$$f - f_{\mathbf{m}} = \sum_{\mathbf{u} \subseteq [d]} C_{\mathbf{u}}(\ldots, h_{i \in \mathbf{u}}, \ldots) \prod_{i \in \mathbf{u}} h_i^2 \tag{3.7}$$

to that in (3.5), in terms of a collection of 2^d $|\mathbf{u}|$-dimensional functions $\{C_{\mathbf{u}}\}_{\mathbf{u} \subseteq [d]}$ which are all bounded absolutely by some $K > 0$. Then [Rei12, Thm. 5.4], see also [Rüt16], provides that $|f - S_{I_L}| = \mathcal{O}(h_L^2 \log(h_L^{-1})^{d-1})$. This is once more within a logarithmic factor of the $\mathcal{O}(h_k^2)$ pointwise convergence offered by the terms of the sequence $(S_k)_{k=0}^{\infty}$. However, the cost required to evaluate each S_{I_L} scales only as $\mathcal{O}(h_L^{-1}(\log(h_L^{-1}))^{d-1})$ [Gar12b; Rüt16, Sec. 4.2.2], compared to the $\mathcal{O}(h_k^{-d})$ required for each S_k. Generalising (3.6), a well-known expression for S_{I_L} directly as a combination sum of model functions is [Del82; WW95; Rei04; Gar12b; Har16a]

$$S_{I_L} = \sum_{k=0}^{d-1} (-1)^k \binom{d-1}{k} \sum_{\|\mathbf{m}\|_1 = L-k} f_{\mathbf{m}}. \tag{3.8}$$

Again, we will rederive this identity below.

Leaving the explicit grid setting behind and following the more general lead now of [TW18], up to slightly different formulation and also with some reference to [CGH18], suppose that $\mathcal{L} : V \to Y$ is some linear functional from the vector space V into a Banach space Y. Rather than requiring an explicit error decomposition like (3.7) of the model functions $f_{\mathbf{m}}$, suppose instead that the Y-norms of all values $\mathcal{L}[\tilde{f}_{\mathbf{m}}]$ are bounded up to a

universal constant factor by a product of strictly decreasing functions $\{b_i : \mathbb{N} \to \mathbb{R}^+\}_{i=1}^d$, that is,

$$\|\mathcal{L}[\tilde{f}_{\mathbf{m}}]\|_Y \leq K_1 \prod_{i=1}^{d} b_i(m_i) \tag{3.9}$$

for $K_1 > 0$. If it also holds that $\sum_{\mathbf{m}\in\mathbb{N}^d} \prod_{i=1}^d b_i(m_i) < \infty$, then [TW18, Prop. 3.1(i)] provides, firstly, that $\lim_{\min_{i=1}^d m_i \to \infty} \mathcal{L}[f_{\mathbf{m}}]$ exists, and secondly, that $\sum_{\mathbf{m}\in\mathbb{N}^d} \mathcal{L}[\tilde{f}_{\mathbf{m}}]$ converges absolutely (in Y) to this limit. In practice, the model functions $f_{\mathbf{m}}$ are usually chosen such that the limit is known to be some particular property $\mathcal{L}[f]$ of the target function f.

Suppose further that each evaluation $\mathcal{L}[f_{\mathbf{m}}]$ incurs a computational cost $\mathcal{C}_{\mathbf{m}}$ that is similarly bounded by strictly increasing functions $\{w_i : \mathbb{N} \to \mathbb{R}^+\}_{i=1}^d$ [TW18],

$$\mathcal{C}_{\mathbf{m}} \leq K_2 \prod_{i=1}^{d} w_i(m_i) \tag{3.10}$$

for universal $K_2 > 0$. We seek then to find the best possible approximation of $\mathcal{L}[f]$ according to $\|\cdot\|_Y$ that is achievable as a sum of evaluated hierarchical surpluses $\mathcal{L}[\tilde{f}_{\mathbf{m}}]$, with a total evaluation cost that does not exceed some predetermined cost budget $W > 0$.

In the absence of any other information about the models $f_{\mathbf{m}}$ and the target f, we want [GG98; Heg03; Har16a; CGH18; TW18] to find an index set, denoted $I_W \subset \mathbb{N}^d$, that maximises the sum of the *benefits* of the terms in I_W, that is, the sum $K_1 \sum_{\mathbf{m}\in I_W} \prod_{i=1}^d b_i(m_i)$ of the bounds on $\|\mathcal{L}[\tilde{f}_{\mathbf{m}}]\|_Y$ for each $\mathbf{m}$ in the index set [TW18, Sec. 3.1]. This maximisation is made while keeping the *cost* of I_W less than W, as provided by $K_2 \sum_{\mathbf{m}\in I_W} \prod_{i=1}^d w_i(m_i) \leq W$. Some such maximising set is guaranteed to exist, and must be *downward-closed*, in the sense that if $\mathbf{m} \in I_W$, then also $\mathbf{n} \in I_W$ for any $\mathbf{n} < \mathbf{m}$ [TW18, Prop. 3.2].

It is easier to construct instead *quasi-optimal* [NTT15; TW18] index sets by picking all points $\mathbf{m}$ with benefit/cost ratios exceeding some δ_W; this cutoff is then reduced as far as possible while keeping the total cost less than W [TW18]. It follows informally that a sequence of combined quasi-optimal approximations S_{I_W} can be obtained by progressively increasing the cost budget W, and that the evaluated terms $\mathcal{L}[S_{I_W}]$ of this sequence will be at least as cost-effective as those of $(\mathcal{L}[S_k])_{k=1}^\infty$ with equivalent or lesser cost, at least up to the information provided by the bounding functions b_i and w_i. As mentioned in [TW18], this kind of quasi-optimal construction can be generally related to various adaptive algorithms, e.g., those given in [GG98; Heg03; Gar07a; CGH18], which construct index sets via inspection of the norms of the true calculated benefits $\|\mathcal{L}[\tilde{f}_{\mathbf{m}}]\|_Y$ and optionally also their costs.

3.2. History and applications of the combination technique

The *(standard) combination technique* as sketched above was initially shaped by Griebel et al. to engage with PDEs [GSZ92]; see also [BG04; Gar12b; TW18]. Here, while the combination sums S_{I_L} can be in a conditional sense "as good" [GH14, p. 9] as approximate solutions leveraging the well-known *sparse grid* functions [Smo63; Zen91; GSZ92; BGR94; Bun+94; BG04; Gar12b; TW18], their evaluation sidesteps the non-trivial implementational effort necessary to reconfigure algorithms for use with the same [BG04; Gar06; OB21].

Versions of the original PDE-focused formulation of the combination technique have been applied in a variety of settings; see, e.g., [BG04; Heg+16] and references therein. However, the power of the combination technique derives from error decompositions equivalent to (3.7), and these can be difficult to demonstrate in many problems of interest [Heg+16; Lag+20]. Moving beyond the solution of PDEs, certain and similar issues with the standard combination sum [HGC07; Gar12a] often suggest rather *dimensionally-adaptive* combination techniques [Heg03; Gar07a; HGC07]; cf. [Gri98; GG03]. These have been applied in turn in the PDE setting [SG22], and alternative algorithms based on the combination technique that seek full spatial adaptivity rather than just dimensional adaptivity have also been recently reported [OB21; Obe21].

An interesting interpretation of dimensionally-adaptive formulations of the combination technique [Heg03; Gar07a; Gar12b] involves the rewriting of the d-dimensional target function f as an ANOVA decomposition [Kuo+09; Feu10]

$$f = \sum_{\mathbf{u}\subseteq[d]} f_{\mathbf{u}} = f_{\emptyset} + \sum_{i\in[d]} f_{\{i\}} + \sum_{i<j} f_{\{i,j\}} + \cdots + f_{[d]}, \tag{3.11}$$

where each $f_{\mathbf{u}}$ is a $|\mathbf{u}|$-dimensional function defined only on those dimensions indexed by $\mathbf{u} \subset [d]$. If the model functions $f_{\mathbf{m}}$ are chosen in a particular way, then a dimensionally-adaptive combination sum can be viewed as recovering approximations for each $f_{\mathbf{u}}$ individually [Gar07a; Gar12b; Gar12a]; see also, e.g., [Heg+16].

In the quantum chemistry setting, Garcke [Gar98] and Garcke and Griebel [GG00] used the combination technique to approximate solutions to monoatomic problems; see also later work in the context of the *opticom* approach [Gar07b; HGC07]. More recently, Zaspel et al. [Zas+18] developed a multilevel combination-technique strategy for the machine-learning and subsequent prediction of chemical properties. Their scheme is closely related to the *composite methods* which we shall discuss in Chapter 4, and we will provide further details there. Separately, Heber [Heb14] and co-workers [GHH14] produced a technique for approximating ground-state polyatomic total energies that centres upon a modified ANOVA-style decomposition. Their *BOSSANOVA decomposition* is heavily informed by the combination technique; we shall return to it, and its adaptive multilevel ML-BOSSANOVA generalisation [CGH18], at length in Chapters 6 and 7.

An alternative construction of the standard combination technique has been considered by Hegland and co-workers [Heg03; HGC07; HH13]; cf. here the earlier [HP97]. As per [Heg03], their setting is a lattice of tensor-product function spaces $V_{\mathbf{m}} = \bigotimes_{i=1}^{d} V_{i,m_i}$, for $\mathbf{m} \in \mathbb{N}^d$ and $V_{i,1} \subset V_{i,2} \subset \cdots \subset V_{i,m}$. The combination sum is formed in terms of projections $P_{V_{i,j}}$ into the spaces $V_{i,j}$, which provide themselves a lattice of projections $P_{\mathbf{m}} = \bigotimes_{i=1}^{d} P_{V_{i,m_i}}$. Informally, these projections could be used to obtain the model functions $f_{\mathbf{m}}$ considered in our sketch in the previous section; see, e.g., discussion directly prior to Proposition 4.26 in [Har16b].

The lattice-of-projections version of the standard combination technique has been explored in particular detail by Harding [Har16a; Har16b] and Wong [Won16]. In the general context of a fault-tolerant scheme, and building on earlier results of Hegland [Heg03], Harding [Har16b, Sec. 4.2] investigated the structure of combination coefficients like those in front of each $f_{\mathbf{m}}$ in (3.8), but for projections formed in terms of arbitrary downward-closed index sets I; the modification of the combination coefficients under adaptive-style updating of I is also considered. We mention also Harding's treatment and extension of alternative combination techniques for extrapolated solutions in a similar lattice-of-projections setup; see [Har16a, Chap. 4] for further details, as well as several other interesting developments and applications.

Wong [Won16, Chap. 3] constructed a *generalised combination technique*[1] that is defined, rather unusually, for index sets that are not necessarily downward-closed, but are allowed more generally to be meet subsemilattices of $\mathbb{N}^d$, that is, subsets of $\mathbb{N}^d$ that are closed under taking componentwise minima. Wong's characterisation of the combination coefficients is explicitly founded on an inclusion/exclusion-style argument; although a relationship between the combination coefficients and the principle of inclusion/exclusion is well-known, see, e.g., [HGC07; Gar12b], we find Wong's interpretation particularly interesting and will come back it briefly in Chapter 5.

In what follows, we consider a combination technique defined not simply for the lattice $\mathbb{N}^d$, but rather for a much more general class of partially ordered sets. We focus on machinery for the calculation and adaptive updating of the involved combination coefficients based upon the technique of *Möbius inversion* that is fundamental in order theory and combinatorics [Sta12]. Our inspirations here are twofold. Firstly, in structuring a combination technique in terms of partially ordered sets, we are heavily inspired by and start from the same basic idea as [Heg03; HGC07], and there are also strong connections to related work in, e.g., [Har16b; Won16]. In particular, although the development in [HGC07] is focused on a particular kind of lattice which is basically equivalent to $\mathbb{N}^d$, the setup in Section 3.1 of that work allows, in principle, for an arbitrary meet semilattice of spaces, there called an *intersection structure*; also, the possibility to use

[1] We remark in passing that we are aware of at least eight competing usages of the phrase "generalised combination technique" or similar in the literature [GG98, Sec. 5.2; HGC07; Kra07, Sec. 3.3.5; Hol08, Sec. 4.1.3; Har16a, Sec. 1.1.2; Won16, Chap. 3; OB21; SG22, Sec. 4.2].

various lattices other than the standard is at least mentioned in [Heg03]. In general, although we do not explicitly follow the formalisms of [Heg03; HGC07; Har16b; Won16], when our construction is applied to the setting of the standard combination technique, a number of their results relating to the combination coefficients in particular can be easily rederived. This is, however, not our primary purpose.

Secondly, although we build the following in the abstract, we will apply it to computational chemistry. There, we shall be particularly interested in certain ANOVA-like decompositions of high-dimensional functions, including (ML-)BOSSANOVA [GHH14; Heb14; CGH18] but also many others; much more on this in the chapters to follow. The observation that Möbius inversion can be used to construct such decompositions was first brought to our attention by Griebel [Gri19], who also suggested the relevance of Möbius inversion to the standard combination technique in the particular context of ML-BOSSANOVA. Möbius inversion has also been used to explicitly construct other related decompositions, such as those in [Dom74; Ess+77, (7) and (8); Kle86, (7) and (15); DFS04, (3), (4), and (7); LC05, (3.18)]. The basic idea of application of Möbius inversion that leads to decompositions and sums like, e.g., (3.17), (3.19), and (3.20) below is fundamentally the same as used there, up to minor formal details, and so not new. The novelty of our work lies instead in first formally extending this idea to include the standard combination technique, and then bringing across existing and well-understood concepts of quasi-optimal adaptivity from that setting. For now, we present the construction cleanly in full generality, and will comprehensively relate it back to existing literature in the pages to come; see, in particular, Sections 5.2.3 and 6.1 below.

3.3. Poset model hierarchies

We now derive the promised generalisation of the combination technique in terms of families of model functions which are indexed by elements drawn from a very broad class of partially ordered sets, and from "grids" defined using such sets as "axes".

Our primary reference for standard order-theoretic material here and throughout this thesis is the textbook of Stanley [Sta12], particularly Chapter 3 of that source;[2] we make additional general reference to [Rot64; Aig97]. We broadly follow the notational conventions of [Sta12], with some minor adjustments. To make the following reasonably self-contained, and for the convenience of the reader, we shall restate some core definitions and results, particularly at the outset. We assume that the basic notion of a *partially ordered set*, or *poset*, is clear. Given some poset P, we use the notation $s \prec t$ to denote that t is a *cover* of s in P, that is, that $t > s$ and there is no $u \in P$ with $s < u < t$. A poset P is *locally finite* if, for every $s \leq t$ in P, the *interval* defined as $[s,t] := \{u \in P \mid s \leq u \leq t\}$ is finite. Further, the *zero* of a poset, $\hat{0} \in P$, is a unique minimal element of

[2] It is perhaps worth mentioning that Stanley speaks explicitly only of poset theory and lattice theory, rather than order theory, but cf. [Aig97].

P, if such exists.

The first two definitions directly extend ideas seen in the sketch of the standard combination technique given earlier in the chapter. In particular, the second definition can be viewed as a kind of generalisation of that given in [Heg+16, Prop. 1].

Definition 3.3.1 (Poset model hierarchy)**.** Let P be a locally finite poset with a $\hat{0}$, and let $\mathcal{F}_P = \{f_t\}_{t\in P}$ be a family of elements of some vector space V of functions, each of which is indexed by an element of P. We call $\mathcal{F}_P$ a *(poset) model hierarchy* over P, and each f_t a *model* or a *model function*.

Definition 3.3.2 (Hierarchical decomposition)**.** Let $\mathcal{F}_P$ be a poset model hierarchy. The *hierarchical decomposition* of each $f_t \in \mathcal{F}_P$ is given by

$$f_t = \sum_{s\leq t} \tilde{f}_s, \tag{3.12}$$

where each implicitly-defined term $\tilde{f}_s$ is the *hierarchical surplus* or simply *surplus* of the model $f_s \in \mathcal{F}_P$.

We can obtain an explicit recursive definition of the surpluses by rewriting (3.12):

$$\tilde{f}_t = f_t - \sum_{s<t} \tilde{f}_s. \tag{3.13}$$

The sum on the right-hand side of (3.13) is finite, so all functions in the family $\{\tilde{f}_t\}_{t\in P}$ are well-defined.

We introduce now a key standard definition, given in a form that is almost but not quite equivalent to that in [Sta12, Sec. 3.7];[3] see also [Rot64], and cf. the version given in [God18].

Definition 3.3.3 (Möbius function of a poset [Sta12])**.** Let P be a locally finite poset, and K be a field. Define the function $\mu : P \times P \to K$ recursively as follows:

$$\mu(s,t) = \begin{cases} 1 & \text{if } s = t, \\ -\sum_{s\leq u<t} \mu(s,u) & \text{if } s < t, \\ 0 & \text{otherwise, i.e., when } s \not\leq t, \end{cases} \tag{3.14}$$

where 0 and 1 are those of K. Then we call μ the *Möbius function* of P (defined for K).

The specific choice of field K will always be clear in context and we will generally not mention it explicitly. For future reference, suppose that $P \cong Q$ are locally finite posets

[3] Specifically, we make explicit the definition of $\mu(s,t)$ when $s \not\leq t$.

that are *isomorphic*; that is, two posets connected by a bijection $\phi : P \to Q$ that is order-preserving in both directions [Sta12]. We use subscripts to distinguish the Möbius functions of each, i.e., μ_P and μ_Q, both for the same K. Then it is well-known that their Möbius functions are related as $\mu_P(s,t) = \mu_Q(\phi(s), \phi(t))$ for all $s, t \in P$; cf., e.g., [BG75, Cors. 2 and 3].

The following theorem [Hal34; Wei35; Rot64] provides the basis for our generalisation of the combination technique. We recite it verbatim from [Sta12], noting that our slightly extended definition of μ makes here no difference. We recall beforehand that the *principal order ideal* of some $t \in P$ is $\Lambda_t := \{s \in P \mid s \leq t\}$ [Sta12].

Theorem 3.3.4 (Möbius inversion formula; verbatim from [Sta12])**.** *Let P be a poset for which every principal order ideal Λ_t is finite. Let $f, g : P \to K$, where K is a field. Then*

$$g(t) = \sum_{s \leq t} f(s), \tag{3.15}$$

for all $t \in P$, if and only if

$$f(t) = \sum_{s \leq t} g(s)\mu(s,t), \tag{3.16}$$

for all $t \in P$.

Proof. See [Sta12, Prop. 3.7.1]. □

This suggests immediately an alternative expression for the hierarchical surpluses $\tilde{f}_t$. Although the following cannot be had directly from the preceeding theorem as stated, due to the assumption of field-valued functions f, g, it is just an example of Möbius inversion and is in no way novel. Indeed, it follows immediately from an alternative proof of the preceeding theorem that is also given by Stanley [Sta12].

Proposition 3.3.5. *Let $\mathcal{F}_P$ be a model hierarchy as defined above. Then for any $t \in P$,*

$$\tilde{f}_t = \sum_{s \leq t} \mu_P(s,t) f_s. \tag{3.17}$$

Proof. See the alternative proof of [Sta12, Prop. 3.7.1], replacing $g(s)$ and $f(t)$ there with f_s and $\tilde{f}_t$ here, and ensuring that $\delta(s,t)$ is valued in the appropriate field K. □

Next, we define summations of surpluses analogous to the combination sums S_I in (3.4) in Section 3.1 above. We define these summations in terms of particular subsets $I \subseteq P$. We require that, if $t \in I$, then also $s \in I$ for every $s \leq t$ in P. In the language of order theory, such an I is called an *order ideal* of P [Sta12]. Finite order ideals provide generalisations of the downward-closed index sets used in the original adaptive formulations of the combination technique.

Definition 3.3.6. Let $\mathcal{F}_P$ be a model hierarchy, and let I be a finite order ideal of P. Then we define the *I-truncation* of $\mathcal{F}_P$ as

$$S_I := \sum_{t \in I} \tilde{f}_t \tag{3.18}$$

$$= \sum_{t \in I} \sum_{s \leq t} \mu(s,t) f_s. \tag{3.19}$$

We use the name "truncation" to convey the idea that such a summation is a finite approximation of the possibly infinite formal sum

$$S_P = \sum_{t \in P} \tilde{f}_t. \tag{3.20}$$

When we wish to emphasise the poset structure of I, we will refer to it as an "order ideal"; when we wish to emphasise its use as a particular selector of terms $\tilde{f}_t$ for use in a summation, we may refer to it as a "(downward-closed) index set", for consistency with the combination-technique literature.

We will also sometimes refer to truncations S_I as *combination sums* (with respect to either the order ideal or index set I), again when this helps to connect our construction with the standard combination technique. In particular, we view each S_I as a weighted "combination" of the model functions f_s for $s \in I$. In (3.19), each f_s appears exactly once in the summation over $t \in I$ for every $s \leq t$. By collecting terms in s and reordering summations,[4]

$$S_I = \sum_{s \in I} \sum_{\substack{t \geq s \\ t \in I}} \mu(s,t) f_s \tag{3.21}$$

$$= \sum_{s \in I} f_s \sum_{\substack{t \geq s \\ t \in I}} \mu(s,t), \tag{3.22}$$

which motivates the following definition, again by extension of the same for the standard combination technique.

Definition 3.3.7 (Combination coefficient)**.** Let $\mathcal{F}_P$ and I be as in Definition 3.3.6. For any $s \in \mathcal{F}_P$, define

$$D_s^{(I)} := \sum_{\substack{t \geq s \\ t \in I}} \mu(s,t). \tag{3.23}$$

We call $D_s^{(I)}$ the *combination coefficient* of f_s in the I-truncation S_I.

[4] Cf. here again the alternative proof of [Sta12, Prop. 3.7.1].

Note that the latter definition is not restricted to $s \in I$, but is universally zero when this is not the case. In general, it will not be possible to further characterise the values of the combination coefficients without more information about the particular poset P, such as a non-recursive expression for its Möbius function μ_P.[5] Even without such knowledge, however, we can easily obtain a consistency guarantee that generalises one already known for the combination coefficients in the standard combination technique; see, e.g., [Rei04, Lem. 4.3; Har16a, Lem. 6; Won16, Prop. 3.2.20]. This is really just a slight rephrasing of a very basic property of the Möbius function, cf., e.g., [Sta12, Exercise 3.88]. The proof makes trivial use of a different version of Theorem 3.3.4 that is also given in [Sta12]; cf. a similar usage of Theorem 3.3.4 on [Sta12, p. 265] that we will come back to in Section 5.2.3.

Proposition 3.3.8. *Let $\mathcal{F}_P$ and I be as in Definition 3.3.6. Then*

$$\sum_{s \in I} D_s^{(I)} = 1. \tag{3.24}$$

Proof. Define $g(s \in I) = 1$, let $f(s \in I) = D_s^{(I)} = \sum_{t \geq s} \mu(s,t) g(t)$, consider $g(\hat{0})$, and apply the dual form of Möbius inversion as per [Sta12, Prop. 3.7.2].[6] □

On that note, we take a moment to relate the construction back to the standard combination technique as sketched in Section 3.1 above. There, the multiindices $\mathbf{m}$ are drawn from $\mathbb{N}^d$. The functions $f_{\mathbf{m}}$ obtained by progressively increasing these multiindices componentwise can be viewed as more refined approximations in distinct dimensions, often according to some kind of grid. The componentwise partial order on $\mathbb{N}^d = \mathbb{N} \times \cdots \times \mathbb{N}$ is just an example of the standard *product order* that is assigned to the *direct product* $P \times Q$ of two posets P and Q, where $(s,t) \leq_{P \times Q} (s',t')$ whenever $s \leq_P s'$ and $t \leq_Q t'$ [Sta12]. To generalise this idea, we define by analogy a "grid" poset model hierarchy formed in terms of individual "axes".

Definition 3.3.9 (Poset model grid)**.** Let $P_1, P_2, \ldots, P_d$ be locally finite posets, each with a zero $\hat{0}_{P_i}$, and let $\Pi = P_1 \times \cdots \times P_d$ be their direct product. Now let $\mathcal{F}_\Pi = \{f_{\mathbf{p}}\}_{\mathbf{p} \in \Pi}$ be some family of functions. Then we call Π a *d-dimensional (poset) model grid*, and refer to the ith original poset P_i as the ith *(poset) axis* of the model grid. If necessary to remove ambiguity, we may call the family $\mathcal{F}_\Pi$ a *poset grid model hierarchy*.

[5] A slightly different formal application of Möbius inversion as used in [LC05] does, however, lead to a different way of understanding the combination coefficients, which makes their well-known connection with the principle of inclusion/exclusion extremely clear, and also relates them to the construction in [HGC07, Sec. 3.1]. See Section 5.2.3, and in particular, Footnote 4 on page 112.

[6] This could be considered the same basic idea of proof used by Wong in [Won16, Prop. 3.2.20], but applied in a more general setting. See discussion in Section 5.2.3 below.

Let us mention here that this can be more broadly viewed as a generalisation on the tensor-product function space lattices constructed in, e.g., [Heg03; HGC07]. Notationally, we will use boldface characters to indicate entries of a poset model grid, for example, $\mathbf{p} \in \Pi$, as above; to identify the ith component of some $\mathbf{p}$, we will write, e.g., $p_i \in P_i$. The "grid" Π is itself a poset which is locally finite and has a zero $\hat{0}_\Pi = (\hat{0}_{P_1}, \ldots, \hat{0}_{P_d})$, so all of the above definitions and results apply without change to Π and any model hierarchy $\mathcal{F}_\Pi$ defined over it. In particular, we can speak of order ideals, combination coefficients, I-truncations, and combination sums over poset grids just as over poset axes.

This grid definition contains a certain ambiguity. If P and Q are arbitrary posets, then trivially $P \times Q \cong Q \times P$. Consider a poset grid Π constructed as above from some set of poset axes $P_1, \ldots, P_d$. There are $d!$ different ways of arranging the terms in the direct product of the constituent poset axes, and so of building equivalent grids up to isomorphism. Slightly more subtly, if we group the axes into arbitrary subsets and take their direct products first, we can view the result as being a "grid" formed from "subgrids", e.g. $(P_1 \times P_2) \times (P_3 \times P_4)$.

This ambiguity is unproblematic, and is better regarded as flexibility: we are free to choose the product structure of Π in such a way that we can best exploit it. In particular, the calculation of the Möbius function values $\mu_\Pi(\mathbf{p}, \mathbf{q})$ for $\mathbf{p}, \mathbf{q} \in \Pi$ can be cast in terms of the Möbius functions μ_{P_i} of the poset axes. For this, we will need the following standard result, adapted slightly from [Sta12]; see also [Rot64; God18].

Theorem 3.3.10 (Product theorem [Sta12])**.** *If P and Q are locally finite posets, then*

$$\mu_{P\times Q}((s,t),(s',t')) = \mu_P(s,s') \cdot \mu_Q(t,t'). \tag{3.25}$$

Proof. This is essentially [Sta12, Prop. 3.8.2], but with the restriction that $(s,t) \leq (s',t')$ removed, in keeping with our slightly broader definition of the Möbius function. □

We consider now the combination coefficients for two particular posets. The first is, in our terminology, a d-dimensional poset grid $\Pi = P \times \cdots \times P$, where P is an infinite but locally finite chain with a $\hat{0}$; here, a *chain* is just another name for a totally ordered set [Sta12]. Such a P is trivially isomorphic to $\mathbb{N}$, with the necessary bijection $\phi : P \to \mathbb{N}$ provided by the usual *rank function* of P [Sta12, p. 244]; so $\phi(\hat{0}) = 0$, then $\phi(s \succ \hat{0}) = 1$, and so on. Thus $\Pi \cong \mathbb{N}^d$, the setting for the standard combination technique.

The second poset we consider is the powerset $2^{[n]}$ of $[n] = \{1, 2, \ldots, n\}$, with its elements ordered so that $\mathbf{u} \leq \mathbf{v}$ whenever $\mathbf{u} \subseteq \mathbf{v}$ for $\mathbf{u}, \mathbf{v} \subseteq [n]$. This is the well-known (finite) *boolean algebra of rank n*, denoted B_n [Sta12]. Since B_n is finite, it is locally finite, and it has a $\hat{0}$ in the form of the empty set $\emptyset$. The boolean algebra is deeply related to certain approximations of high-dimensional functions; we will come back to this particularly in Chapter 5 below.

Non-recursive expressions for the Möbius functions of these posets are derived in Examples 3.8.3 and 3.8.4 of [Sta12], respectively; we refer to the source for details, but

note that in both cases, the derivations follow directly from the product theorem. We will use the resulting expressions to rederive in turn existing expressions for the combination coefficients of particular order ideals of these two posets. In both examples, we present the rederivations cleanly before explicitly surveying their presence in existing literature, but these are only motivating examples and there is no particular novelty. We will make use of the following supporting identity.

Lemma 3.3.11 ([QG15])**.** *Let $n, k \in \mathbb{N}$ be such that $k \leq n$. Then*

$$\sum_{j=0}^{k} (-1)^j \binom{n}{j} = (-1)^k \binom{n-1}{k}. \tag{3.26}$$

Proof. This is a special case of (1.31) in [QG15]. □

Example 3.3.12 (d-dimensional grid of chain posets $\Pi = P \times \cdots \times P \cong \mathbb{N}^d$)**.** We work directly with $\mathbb{N}^d$ for simplicity, but everything here also translates immediately to Π. We have from [Sta12, Example 3.8.4], up to precise formulation and the use of infinite chains, that

$$\mu_{\mathbb{N}^d}(\mathbf{m}, \mathbf{n}) = \begin{cases} (-1)^{\|\mathbf{n}-\mathbf{m}\|_1} & \text{if } \mathbf{m} \leq \mathbf{n} \text{ and } \mathbf{n} - \mathbf{m} \in \{0,1\}^d, \\ 0 & \text{otherwise,} \end{cases} \tag{3.27}$$

where $\mathbf{n} - \mathbf{m}$ is defined componentwise as standard. Now let $I_L := \{\mathbf{m} \in \mathbb{N}^d \mid \|\mathbf{m}\|_1 \leq L \in \mathbb{N}\}$. Choose an arbitrary $\mathbf{m} \in I_L$, and consider the value of the corresponding combination coefficient according to (3.23),

$$D_{\mathbf{m}}^{(I_L)} = \sum_{\substack{\mathbf{n} \geq \mathbf{m} \\ \mathbf{n} \in I_L}} \mu_{\mathbb{N}^d}(\mathbf{m}, \mathbf{n}). \tag{3.28}$$

Collecting terms by $\|\mathbf{n} - \mathbf{m}\|_1$ leads to

$$D_{\mathbf{m}}^{(I_L)} = \sum_{k=0}^{L-\|\mathbf{m}\|_1} (-1)^k \binom{d}{k}. \tag{3.29}$$

Now using (3.26), we have

$$D_{\mathbf{m}}^{(I_L)} = (-1)^{L-\|\mathbf{m}\|_1} \binom{d-1}{L-\|\mathbf{m}\|_1}. \tag{3.30}$$

This lets us write an explicit form of the combination sum over I_L as

$$S_{I_L} = \sum_{k=0}^{d-1} (-1)^k \binom{d-1}{k} \sum_{\substack{\mathbf{m} \in I_L \\ \|\mathbf{m}\|_1 = L-k}} f_{\mathbf{m}}. \tag{3.31}$$

These last three identities are well-known in the literature surrounding the combination technique and related methods. The earliest formulation of (3.31) of which we are aware is an inductive derivation given by Delvos in the context of Boolean interpolation [Del82, Lem. 2], while forms of (3.29) and (3.30) appear at least in [WW95], and we essentially follow their idea of derivation once (3.28) is established. We observe in passing that the also well-known expression for the standard combination coefficients $D_{\mathbf{m}}^{(I)} = \sum_{\mathbf{z}\in\mathbb{N}^d,\ \mathbf{z}\leq\mathbf{1}} (-1)^{\|\mathbf{z}\|_1} \chi_I(\mathbf{m}+\mathbf{z})$ for an arbitrary downward-closed finite index set $I \subset \mathbb{N}^d$, as given in, e.g., [GG98, Sec. 5.2; Gar12b, (33); Har16a, Prop. 4] and used to begin the very similar proof of [Har16a, Lem. 7], also emerges trivially as a consequence of (3.27) and (3.23). In any case, (3.27) along with Proposition 3.3.5 validates the equivalence of (3.1) and (3.3), and (3.31) confirms (3.6) and (3.8), as promised.

Example 3.3.13 (The boolean algebra B_n)**.** We know from [Sta12, Example 3.8.3], also, e.g., [Rot64; Gre82; Aig97], that the Möbius function of B_n is given by

$$\mu_{B_n}(\mathbf{u},\mathbf{v}) = \begin{cases} (-1)^{|\mathbf{v}-\mathbf{u}|} & \text{if } \mathbf{u} \subseteq \mathbf{v}, \\ 0 & \text{otherwise.} \end{cases} \tag{3.32}$$

That given, consider a model hierarchy $\mathcal{F}_{B_n}$ with models $f_{\mathbf{u}}$, and form the index set $I_L = \{\mathbf{u} \in B_n \mid |\mathbf{u}| \leq L\}$ for some $0 \leq L \leq n$. An explicit expression for the combination coefficient of some $\mathbf{u} \in I_L$ is easily obtained:

$$D_{\mathbf{u}}^{(I_L)} = \sum_{\substack{\mathbf{v}\supseteq\mathbf{u} \\ \mathbf{v}\in I_L}} \mu_{B_n}(\mathbf{u},\mathbf{v}) \tag{3.33}$$

$$= \sum_{k=0}^{L-|\mathbf{u}|} (-1)^k \binom{n-|\mathbf{u}|}{k}, \tag{3.34}$$

where the latter follows from (3.32) by considering each $\mathbf{v}-\mathbf{u}$ by cardinality. Using once more (3.26), we find that

$$D_{\mathbf{u}}^{(I_L)} = (-1)^{L-|\mathbf{u}|} \binom{n-|\mathbf{u}|-1}{L-|\mathbf{u}|}. \tag{3.35}$$

In particular, when $L = n$,

$$D_{\mathbf{u}}^{(I_L)} = \begin{cases} 1 & \text{if } \mathbf{u} = [n], \\ 0 & \text{otherwise,} \end{cases} \tag{3.36}$$

and when $L = 0$, it is easy to see that $I_L = \{\emptyset\}$ and $D_{\emptyset}^{(I_L)} = 1$. Thus, the combination sum over I_L for $L < n$ is given explicitly by

$$S_{I_L} = \sum_{\mathbf{u}\in I_L} \tilde{f}_{\mathbf{u}} = \sum_{k=0}^{L} (-1)^{L-k} \binom{n-k-1}{L-k} \sum_{|\mathbf{u}|=k} f_{\mathbf{u}}. \tag{3.37}$$

Again, these expressions are not novel. Nor are the combinatorial manipulations by which they are derived, certainly not once $\mu_{B_n}(\mathbf{u}, \mathbf{v})$ is explicitly converted into $(-1)^{|\mathbf{v}-\mathbf{u}|}$ in (3.33). Equations (3.34) and (3.35) are exactly [Hul14, (2.82)] up to notation, and are derived using an identical argument. The expression (3.37) can be found at least in [KC06], in the context of a *many-body expansion* (there, more precisely, an *n-mode expansion*) in computational chemistry. Variations on the individual expressions (3.34) and (3.35) are recognisably implicit in the original proof, which also involves (3.26). A more general but not completely explicit version of (3.34) can be seen in [KC16, (15)], which is equivalent to our definition (3.23) of $D_{\mathbf{u}}^{(I)}$ for any order ideal $I \subseteq B_n$. We note with some interest that the derivation of that identity involves an apparently independent reformulation of what is basically the general version of the Möbius function (3.14), as constructed by a counting argument [KC16, (A1)]. The development of [KC16, (15)] thus amounts to an inductive reformulation of Möbius inversion in the specific boolean algebra setting. An expression that is effectively equivalent but not identical to (3.37) was given independently in a closely related context by Richard et al. [RLH14, (2.6)], there also inductively derived and also making use of similar combinatorial arguments and (3.26). Contrast the above also with [LH16, (11) and (13)], noting in particular that [LH16, (13)] corresponds exactly to (3.34) and (3.35). A slight variation on (3.37) is also given in [Kuo+09, (3.2)],[7] in the context of certain ANOVA-like decompositions. We will discuss the settings of these works in detail in Chapter 5 below.

There is an important difference between the combination coefficients here and those shown in the previous example. For the case $L < n$, every coefficient $D_{\mathbf{u}}^{(I_L)}$ is non-zero; contrast this with the vanishing of many of the coefficients for the grid of chains in the previous example. This does not necessarily remain true for an I-truncation over an arbitrary order ideal $I \subseteq B_n$; see, e.g., [KC16].

3.4. Targets, benefits, and costs

We have now put in place a rather complicated formal apparatus for the construction of combination sums in terms of index sets drawn from any instance of a very general class of poset. However, we have not yet introduced anything for the associated model hierarchies to actually model. We remedy this now, drawing inspiration from ideas and constructions in, e.g., [Nob+16; CGH18; TW18].

Definition 3.4.1. Let $\mathcal{F}_P$ be a model hierarchy, and let $f \in V$ be some particular element of the same vector space as the models in $\mathcal{F}_P$. We call f the *target* of $\mathcal{F}_P$. Let further $\mathcal{L} : V \to Y$ be some linear functional into some Banach space Y. We call $\mathcal{L}$ a *property evaluation functional*.

[7] There given without explicit proof; see, however, an extended preprint version of the same paper [Kuo+08], which makes use of very similar combinatorial manipulations to those here.

Since we have as of yet assumed no explicit relationship between f and $\mathcal{F}_P$ beyond the underlying vector space, this definition is purely a linguistic assocation. We shall, in fact, generally want to work in the reverse direction: given a target f, find (or construct) a model hierarchy in such a way that combination sums S_I taken over that hierarchy for a particular sequence of index sets I will provide systematically-improvable approximations for f, in the particular sense that the applications $\mathcal{L}[S_I]$ will provide the same for $\mathcal{L}[f]$.

How this construction is done, and how well it will work, will naturally depend on the forms of both the target f and the model hierarchy $\mathcal{F}_P$. In this thesis, we shall be working in a setting where no error decomposition like (3.7) will be available. Thus, following [TW18], we will consider instead quasi-optimal I-truncations, similar to those discussed in Section 3.1 in the standard case.

In [TW18], Tempone and Wolfers outline a general framework for the analysis of a slightly-extended variant of the standard combination technique. In our terminology, they consider a d-dimensional poset grid composed of infinite chain axes. We now very briefly generalise a small subset of the development in [TW18] to our setting, with intention to motivate the adaptive algorithm which we shall develop in the final part of this chapter; cf. here again also, e.g., [NTT15; Nob+16]. A similar idea in a fixed non-$\mathbb{N}^d$ context was previously considered in [CGH18].

We work in the following with a poset model hierarchy $\mathcal{F}_\Pi$ in terms of a d-dimensional poset grid $\Pi = P_1 \times \cdots \times P_d$, along with a property evaluation functional $\mathcal{L}$. For simplicity, we will assume that each P_i is infinite; this will often not be so, but the necessary adjustments are not difficult. We assume by extension of [TW18] the existence of some family of strictly decreasing functions $\{b_i : P_i \to \mathbb{R}^+\}_{i=1}^d$, in the sense that if $s < t \in P_i$, then $b_i(s) > b_i(t)$. We also assume that there exists some constant $K_1 > 0$ such that

$$\|\mathcal{L}[\tilde{f}_{\mathbf{p}}]\|_Y \leq K_1 \prod_{i=1}^{d} b_i(p_i). \tag{3.38}$$

That is, we assume that each evaluated hierarchical surplus term can be bounded above by a product of the functions b_i, up to a constant factor. Let further $\mathcal{C} : \Pi \to \mathbb{R}^+$ be a cost function, which associates a non-negative number to each $\mathbf{p} \in \Pi$ representing the amount of computational work required to perform the evaluation $\mathcal{L}[f_{\mathbf{p}}]$, and that we have some family $\{w_i : P_i \to \mathbb{R}^+\}_{i=1}^d$ of strictly increasing functions and a constant $K_2 > 0$ such that

$$\mathcal{C}(\mathbf{p}) \leq K_2 \prod_{i=1}^{d} w_i(p_i). \tag{3.39}$$

The elements of each P_i can then be totally ordered, not necessarily uniquely, in decreasing order of $b_i(p_i)$; we denote P_i' to be any such totally reordered set corresponding to P_i. We make now a further assumption, which will be implicit for any poset axis considered for the remainder of this chapter, and thesis more generally: that, for any

$p \in P_i$, the set $\{q \in P_i \mid q \succ p\}$ is finite. Under this assumption,[8] $P_i' \cong \mathbb{N}$, with corresponding bijection $\phi_i : P_i' \to \mathbb{N}$, and so $\Pi' = P_1' \times \cdots \times P_d' \cong \mathbb{N}^d$. Define by its inverse the further bijection $\Phi^{-1}(\mathbf{m} \in \mathbb{N}^d) = (\phi_1^{-1}(m_1), \ldots, \phi_d^{-1}(m_d))$. Let us assume, as in [TW18, Prop. 3.1(i)], that

$$\sum_{\mathbf{m}\in\mathbb{N}^d} \prod_{i=1}^{d} b_i(\phi_i^{-1}(m_i)) < \infty. \tag{3.40}$$

By an equivalent argument as there given, we have that $\lim_{\min_{i=1}^d m_i \to \infty} \mathcal{L}[f_{\Phi^{-1}(\mathbf{m})}]$ exists and is equal to some x_∞, and also that the first sum in

$$\sum_{\mathbf{m}\in\mathbb{N}^d} \mathcal{L}[\tilde{f}_{\Phi^{-1}(\mathbf{m})}] = x_\infty = \sum_{\mathbf{p}\in\Pi} \mathcal{L}[\tilde{f}_{\mathbf{p}}] \tag{3.41}$$

converges absolutely. The latter sum is obtained by reordering the terms of the first.

We will simply make the assumption now that $x_\infty = \mathcal{L}[f]$, that is, the evaluated property of the target function; it is left up to the precise implementation to ensure that this is in fact so. So, by extension of Section 3.1 of [TW18] and for some cost budget $W > 0$, we seek an index set $I \subset \Pi$ that solves the binary knapsack problem of finding, by adaptation of [TW18, (11)],

$$\arg\max_{I\subset\Pi} \quad |I|_b := K_1 \sum_{\mathbf{p}\in\Pi} \prod_{i=1}^{d} b_i(p_i), \tag{3.42}$$

$$\text{such that} \quad |I|_c := K_2 \sum_{\mathbf{p}\in\Pi} \prod_{i=1}^{d} w_i(p_i) \le W, \tag{3.43}$$

and provides a quasi-optimal truncation of (3.41). The following proposition is a tailoring of Proposition 3.2 of [TW18] to our setting, rephrased in our terminology.

Proposition 3.4.2 (Adapted from [TW18])**.** *The following hold:*

(i) The binary knapsack problem (3.42) *and* (3.43) *has a solution* I^* *such that* $|I^*|_b = \max_{I\subset\Pi} |I|_b := B_W^*$.

(ii) Any $I^* \subset \Pi$ *such that* $|I^*|_b = B_W^*$ *is a finite order ideal of* Π.

(iii) Let I_W *be the set*

$$I_W := \left\{ \mathbf{p} \in \Pi \;\middle|\; \prod_{i=1}^{d} \frac{b_i(p_i)}{w_i(p_i)} > \delta_W \right\}, \tag{3.44}$$

[8] To get a sense for why this is helpful, consider the set of non-negative real numbers, partially ordered such that $0 \prec x$ for every non-zero x. This is a locally finite poset with a $\hat{0}$, but clearly cannot be isomorphic to $\mathbb{N}$.

with $\delta_W > 0$ taken smallest possible while maintaining $|I_W|_c \leq W$. Then I_W is a finite order ideal of Π, and

$$|I_W|_b \geq \frac{|I_W|_c}{W} B^*_W. \tag{3.45}$$

Proof. The original proof of [TW18, Prop. 3.2] applies basically unchanged. Note, however, that the necessary bound on the size of $I \subset \Pi$ for any W holds here because of the assumption that each element in Π has only finitely many covers. □

Just as stated in the original [TW18, Prop. 3.2], the set I_W provides a solution to the possibly different binary knapsack problem in terms of the cost budget $|I_W|_c$, but is not guaranteed to be a solution to the problem for the original cost budget W.

3.5. Adaptive construction of poset-grid order-ideal index sets

The mechanism for choosing a quasi-optimal index set I_W outlined in the previous section presupposes an explicit knowledge of families of bounding functions b_i and w_i, but in practice, these may not be precisely known; see, e.g., [CGH18]. However, we assume that it is always possible to know the actual value $\|\mathcal{L}[\tilde{f}_{\mathbf{p}}]\|$ simply by calculating it, at least where this is computationally feasible. Since this necessarily involves calculation of $\|\mathcal{L}[f_{\mathbf{p}}]\|$, we assume that we may also calculate a corresponding abstract cost $\mathcal{C}(\mathbf{p})$ for the same, even if we cannot decompose that cost neatly in terms of the grid axes.

We consider an alternative, adaptive way of choosing an index set I. Our strongest influence here is a similar algorithm described in [CGH18, Alg. 2], and we are also inspired by ideas in [HGC07]. Like most adaptive algorithms that have been developed in combination-technique settings, the basic idea follows the schemes of Griebel and Gerstner [GG03] and of Hegland [Heg03], both of which build in turn on [Gri98]; see also, e.g., [Hol08; Gar07a; Gar12a; Nob+16; SG22]. These algorithms, and ours, are fundamentally greedy in nature: beginning from an empty or somehow minimal index set $I^{(0)}$, indices that maximise some benefit/cost criteria are used to guide the addition of new indices to that set, with appropriate care taken to retain the downward-closure property. The resulting sequence of index sets $I^{(i)}$ for each iteration i is intended to provide a progressively more expensive but hopefully more accurate approximation.

The overall approach of the algorithm at the ith iteration can be summarised as follows. An element $\mathbf{p} \in I^{(i)}$ is considered *active* if there exists some *successor*[9] element $\mathbf{q} \succ \mathbf{p}$ such that $\mathbf{q} \notin I^{(i)}$. Some subset of the active elements of $I^{(i)}$ are selected for *expansion*, based upon the values of the respective benefit/cost ratios $\|\mathcal{L}[\tilde{f}_{\mathbf{p}}]\|/\mathcal{C}(\mathbf{p})$. When an active element is expanded, each successor $\mathbf{q} \succ \mathbf{p}$ which is not yet included in $I^{(i)}$ is tested to

[9] Note that this corresponds with the idea of a *forward neighbour* in, e.g., [GG03]. Similarly, a *predecessor* is basically a *backward neighbour*. Our terminology originally stems from considering the Hasse diagram of a poset P as in [HGC07].

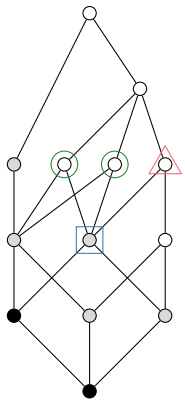

Figure 3.1.: A sample illustration of a possible adaptively-obtained index set I for a small single-axis poset grid $\Pi = P$. The figure shows P as a Hasse diagram [Sta12]: each vertex corresponds to an element of the poset, and an edge between two vertices indicates that the higher-drawn element covers the lower-drawn element. Black-filled vertices (●) indicate non-active members of I, grey-filled vertices (⊙) show active members of I, and unfilled vertices (○) show members of P which have not yet been added to I. One particular active element is highlighted with a blue square (□). Were this member to be expanded, three possible successors would be considered. Two of these are admissible for inclusion in I, and are highlighted with green circles (○). One successor is not immediately admissible, and is highlighted with a red triangle (△). This latter element is inadmissible since one of the elements which it covers is not already in I.

see whether its inclusion in $I^{(i+1)}$ would maintain the invariant that $I^{(i+1)}$ must be an order ideal of Π. This is equivalent to testing whether all of the *predecessors* of that successor, i.e., all elements $\mathbf{r} \prec \mathbf{q}$, are in $I^{(i)}$. If so, then $\mathbf{q}$ is considered *admissible*, and is added to $I^{(i+1)}$. An illustration of a possible adaptively-obtained index set is given in Figure 3.1, visualising the distinction between active and non-active elements of an $I^{(i)}$ for a particular poset.

The detailed description of this algorithm is organised into subsections as follows. In Section 3.5.1, we will reformulate the calculation of combination sums and surplus terms in terms of tensors, which allows an efficient and clean implementation of the algorithm via the use of a sparse tensor data structure. Sections 3.5.2 and 3.5.3 describe the functionality which must be implemented in order to use any arbitrary poset as an axis of a poset grid. Section 3.5.4 outlines three selection strategies that may be used to control the adaptive growth of the index set at each iteration. Section 3.5.5 introduces and motivates an error indicator by which the quality of an adaptively-obtained approximation may be judged. Full pseudocode for the main loop of the algorithm is given and discussed in Section 3.5.6. Finally, a brief assessment of the computational complexity of the algorithm is provided in Section 3.5.7.

3.5.1. Sparse tensor formulation of combination sums

Some formulations of adaptive algorithms (see e.g. [GG03]) form the final combination sum directly from the equivalents of the evaluated surplus terms $\mathcal{L}[\tilde{f}_{\mathbf{p}}]$. Here, as each new element $\mathbf{p} \in \Pi$ is added to the index set, the relevant surplus term is calculated and folded into a running total. That is, if $I^{(i)}_{\text{new}}$ indicates the collection of newly-added elements at the ith iteration of the adaptive algorithm, such that the complete index set after iteration i is formed as $I^{(i)} \leftarrow I^{(i-1)} \cup I^{(i)}_{\text{new}}$, then the combination sum at iteration i is maintained and updated as

$$S_{I^{(i)}} \leftarrow S_{I^{(i-1)}} + \sum_{\mathbf{p} \in I^{(i)}_{\text{new}}} \mathcal{L}[\tilde{f}_{\mathbf{p}}], \tag{3.46}$$

with $S_{I^{(0)}} := \mathcal{L}[\tilde{f}_{\hat{0}}]$. If the algorithm terminates after n iterations, the combination sum $S_{I^{(n)}}$ is therefore immediately available, along with the index set $I^{(n)}$ to which it corresponds.

So calculated, the final combination sum $S_{I^{(n)}}$ involves $|I^{(n)}|$ distinct terms $\mathcal{L}[\tilde{f}_{\mathbf{p}}]$, each itself a sum of potentially many terms $\mu(\mathbf{q}, \mathbf{p})\mathcal{L}[f_{\mathbf{q}}]$. In the standard combination technique case, these latter are non-zero for only a small subset of $I^{(i)}$, cf. (3.3), and there is alternatively a well-known expression for the final $S_{I^{(n)}}$ in terms of models $\mathcal{L}[f_{\mathbf{p}}]$ rather than surpluses; see, e.g., [SG22, (4.2)]. In the general poset-grid setting, however, such an expression for $S_{I^{(n)}}$ will not generally be known, and it may be the case (for example, for some index subsets of $\Pi = B_n$) that each $\tilde{f}_{\mathbf{p}}$ depends explicitly on $f_{\mathbf{q}}$ for each and every $\mathbf{q} \leq_{\Pi} \mathbf{p}$.

The total number of operations involved in the naïve numerical evaluation of $S_{I^{(i)}}$ can therefore be as high as $\sum_{\mathbf{p} \in I^{(i)}} |\Lambda_{\mathbf{p}}|$, which is potentially much greater than in the standard case, and the numerical stability of the sum is correspondingly and negatively affected. Extending slightly on an observation made by Richard et al. [RLH14] in a particular computational-chemistry setting: as well as the usual floating-point sources, an instability can stem here more subtly from any compounding inaccuracies which may be inherent to the underlying model-function evaluations $\mathcal{L}[f_{\mathbf{p}}]$, particularly if those evaluations are provided only to limited precision by iterative-type numerical solvers. We will come back to this topic in Section 5.3.

Rather than apply (3.46), we progressively calculate and update instead the complete set of combination coefficients $\{D^{(I^{(i)})}_{\mathbf{p}}\}_{\mathbf{p} \in I^{(i)}}$. This is achieved by working with representations of the involved quantities in terms of tensors. We use the word "tensor" in the sense of a multidimensional array, and our terminology is chosen accordingly. Following convention set by the NumPy library [Har+20], we give a d-dimensional tensor a *shape*: a sequence $(s_i)_{i=1}^{d}$ of values, each drawn from $\mathbb{N}^+ \cup \{\infty\}$, which describe the lengths of each axis of the tensor. We denote the entries of a tensor T as $T_{\mathbf{m}}$, each indexed by some $\mathbf{m} \in \mathbb{N}^d$ such that each $m_i \leq s_i$. Note that we explicitly allow for tensors with

axes of countably infinite length. Elementwise addition ($+$) and multiplication ($\odot$) are defined in the obvious way. To avoid ambiguity, we explicitly define the tensor product ($\otimes$) as follows: if T and U are k- and l-dimensional tensors respectively, then $T \otimes U$ is $(k+l)$-dimensional and has entries $(T \otimes U)_{(m_1,\cdots,m_k,n_1,\cdots,n_l)} = T_{(m_1,\cdots,m_k)} U_{(n_1,\cdots,n_l)}$.

From this point on, when dealing with a poset grid $\Pi = P_1 \times \cdots \times P_d$, we assume the existence of an *indexing bijection* for each poset axis P_i. This must be a bijection $\phi_{P_i} : P_i \to \mathbb{N}$ (when P is an infinite poset) or $\phi_{P_i} : P \to \{0, \ldots, |P| - 1\}$ (when P is a finite poset with n elements) that maps each element of P to a unique natural index. We then define a composite bijection by $\Phi(\mathbf{p} \in \Pi) = \prod_{i=1}^{d} \phi_{P_i}(p_i)$. By abuse of notation, if we work with a d-dimensional tensor T with shape $(|P_1|, \cdots, |P_d|)$, then when we write $T_{\mathbf{p}}$ for some $\mathbf{p} \in \Pi$, we mean $T_{\Phi(\mathbf{p})}$.

Assuming a fixed choice of poset grid $\Pi = P_1 \times \cdots \times P_d$, we define a family of d-dimensional tensors $\{M^{(\mathbf{p})}\}_{\mathbf{p} \in \Pi}$ as follows. The entries of each $M^{(\mathbf{p})}$ are defined to be such that each element $M^{(\mathbf{p})}_{\mathbf{q}} := \mu_\Pi(\mathbf{q}, \mathbf{p})$. We call each $M^{(\mathbf{p})}$ the *Möbius tensor* of $\mathbf{p}$. Note that although $M^{(\mathbf{p})}$ may have infinitely many entries, depending on the posets P_i, the number of non-zero entries is finite by the definition of the Möbius function, so $M^{(\mathbf{p})}$ can be represented exactly in finite memory; we will discuss such a representation below.

We then define a family of tensors $D^{(I)}$ for every finite order ideal $I \subseteq \Pi$, such that

$$D^{(I)} := \sum_{\mathbf{p} \in I} M^{(\mathbf{p})}. \tag{3.47}$$

Our notational choice is deliberate: so constructed, every entry $D^{(I)}_{\mathbf{p}}$ of each $D^{(I)}$ is just a combination coefficient for $f_{\mathbf{p}}$ in I, cf. (3.21), (3.22), and (3.23). Since I is finite, each $D^{(I)}$ can again be represented exactly in finite memory.

At each iteration i of our adaptive algorithm, some collection of new elements $\mathbf{p}$ are added to the index set $I^{(i-1)}$ to produce the new index set $I^{(i)}$. For each such $\mathbf{p}$, the tensor $M^{(\mathbf{p})}$ is constructed. This can be done by calculating a one-dimensional tensor $m^{(p_i)}$ for each per-axis component p_i of P: a *Möbius vector* defined elementwise such that $m^{(p_i)}_t := \mu_{P_i}(t, p_i)$ for each $t \in P_i$. The entries of the full $M^{(\mathbf{p})}$ are then obtained as

$$M^{(\mathbf{p})} = \bigotimes_{i=1}^{d} m^{(p_i)}, \tag{3.48}$$

as per Theorem 3.3.10.[10] The introduction of the per-axis tensors $M^{(p_i)}$ means that we need only explicitly implement the Möbius functions μ_{P_i} on a per-axis basis.

As remarked above, these tensors will contain a potentially large but always finite number of non-zero terms, even though they may be countably infinite in general. We

[10] We note some inspiration in this construction from an observation on [Sta12, p. 267], where some $\mu_{P \times Q}$ is recognised as a tensor product $\mu_P \otimes \mu_Q$ in a particular algebra of functions. Similarly, Godsil uses a Kronecker product of matrices to prove the product theorem [God18, Lem. 3.1].

rely on a d-dimensional generalisation of a standard sparse matrix data structure; we refer to this as a *sparse tensor*. The basic idea is certainly not new; see, e.g., [BK08]. Algorithmically, we consider a sparse tensor to be "empty" on creation, in the sense that all (potentially infinitely many) entries are assumed to be zero and therefore not explicitly stored. Entries can be assigned to, written $T_{\mathbf{m}} \leftarrow x$ for some appropriate x. In addition to the operations on tensors defined above, we also assume the existence of a reduction operation, which returns the sum of all entries in the sparse tensor:

$$\textsc{Reduce}(T) := \sum_{\mathbf{m}} T_{\mathbf{m}}. \tag{3.49}$$

A Reduce operation need only consider the non-zero entries of the involved sparse tensor T, and since only a finite number of entries of a sparse tensor can be practically set to be non-zero, this operation is always well-defined in an implementation.

Throughout the execution of the adaptive algorithm, three persistent sparse tensor objects are maintained: D, V, and C. Each has shape consistent with the Möbius tensors defined above. The *combination tensor* D stores the combination coefficients at each iteration. The *value tensor*, V, stores the explicitly-evaluated values $\mathcal{L}[f_{\mathbf{p}}]$ for all $\mathbf{p}$ in the index set. The *cost tensor*, C, stores the corresponding values $\mathcal{C}(\mathbf{p})$ of a cost function $\mathcal{C} : \Pi \to \mathbb{R}^+$. If all values indexed by $\Lambda_{\mathbf{p}}$ have already been calculated, then a hierarchical surplus $\mathcal{L}[\tilde{f}_{\mathbf{p}}]$ can be explicitly evaluated as the reduction of the elementwise product between $M^{(\mathbf{p})}$ and V, that is, $\mathcal{L}[\tilde{f}_{\mathbf{p}}] = \textsc{Reduce}(M^{(\mathbf{p})} \odot V)$. The complete combination sum at each iteration i can be likewise evaluated as $S_{I^{(i)}} = \textsc{Reduce}(D \odot V)$. Finally, the cost of the combination sum is obtained as $C_{I^{(i)}} = \textsc{Reduce}(\mathbf{1}[D] \odot C)$, where $\mathbf{1}[D]$ is a sparse tensor with ones everywhere D is non-zero.

3.5.2. Operations on poset axes and poset grids

We assume that we are able to represent each element $t \in P$ of any axis P in the poset grid symbolically, either directly (for posets which admit a numerical representation) or by means of some kind of encoding scheme. With such a representation fixed, we then express the ordering structure of each poset axis in terms of four functions, which provide a consistent mechanism for manipulating and exploring the topology of the poset axis around various elements and for the calculation of Möbius functions. Borrowing from the terminology of computer science, we call this the *poset axis interface*; these functions need to be explicitly constructed for every type of poset axis that may be employed in a problem setting.

The first function we require is $\textsc{AxisIndex}(P; t \in P)$, where P is the poset axis, treated as a parameter in the mathematical sense and fixed for each concrete implementation, and $t \in P$ is a parameter in the computational sense, which varies between calls to the function. This function simply implements an indexing bijection $\phi_P : P \to \mathbb{N}$ or $\phi_P : P \to \{0, \ldots, |P| - 1\}$, as discussed above.

Algorithm 3.1 Poset grid interface functionality.

```
function GridMultiindex(Π = P_1 × ··· × P_d; p = (p_1, ..., p_d) ∈ Π)
    l ← (0, ..., 0) ∈ ℕ^d
    for 1 ≤ i ≤ d do
        l_i ← AxisIndex(P_i; p_i)
    return l

function GridPredecessors(Π = P_1 × ··· × P_d; p = (p_1, ..., p_d) ∈ Π)
    R ← ∅
    for 1 ≤ i ≤ d do
        for all p' ∈ Predecessors(P_i; p_i) do
            R ← R ∪ {(p_1, ..., p_{i-1}, p', p_{i+1}, ..., p_d)}
    return R

function GridSuccessors(Π = P_1 × ··· × P_d; p = (p_1, ..., p_d) ∈ Π)
    R ← ∅
    for 1 ≤ i ≤ d do
        for all p' ∈ Successors(P_i; p_i) do
            R ← R ∪ {(p_1, ..., p_{i-1}, p', p_{i+1}, ..., p_d)}
    return R

function MöbiusTensor(Π = P_1 × ··· × P_d; p = (p_1, ..., p_d) ∈ Π)
    M ← MöbiusVector(P_1, p_1)
    for 2 ≤ i ≤ d do
        M ← M ⊗ MöbiusVector(P_i; p_i)
    return M
```

We require also a set-valued function $\textsc{Predecessors} : P \to 2^P$, such that a call to $\textsc{Predecessors}(P;\, t \in P)$ returns the set of all elements s such that $s \prec t$. Similarly, the function $\textsc{Successors} : P \to 2^P$ must return the set of all elements $s \in P$ such that $t \prec s$. Finally, we need an implementation of MöbiusVector, which must return a one-dimensional sparse tensor representing $m^{(t)}$ as defined above. As observed previously, this sparse tensor need only explicitly contain entries for the finite number of values $\phi_P(s)$ such that $\mu_P(s,t)$ is non-zero.

Given implementations of these four operations for an arbitrary collection of posets $\{P_i\}_{i=1}^d$, we can construct equivalent operations for a complete poset grid $\Pi = P_1 \times \cdots \times P_d$. Each such *poset grid interface* function is analogous to a poset axis interface function. $\textsc{GridMultiindex} : \Pi \to \mathbb{N}^d$ provides the indexing bijection $\phi(\mathbf{p}) = \prod_{i=1}^n \phi_{P_i}(p_i)$. $\textsc{GridPredecessors} : \Pi \to 2^\Pi$ and $\textsc{GridSuccessors} : \Pi \to 2^\Pi$ explore the cover relationship of Π just as Predecessors and Successors explore that of an axis P_i.

MöbiusTensor implements (3.48).

The poset grid interface functions are completely defined by the four poset axis interface functions, so must be implemented only once, regardless of the choices of P_i. For pseudocode demonstrating this, refer to Algorithm 3.1. Note that we implicitly use the obvious fact that, for a direct product $P \times Q$, it holds that $(s,t) \prec (s',t')$ if and only if either $s \prec s'$ or $t \prec t'$, but not both.

3.5.3. Fallback calculation of the Möbius function

The Möbius function μ_P for each of the various poset axes P may be available in an explicitly non-recursive form, as in Examples 3.3.12 and 3.3.13 above. However, we will wish to apply our algorithm to poset grids that contain an axis, or axes, for which no non-recursive expression for μ_P is known. As such, we sketch here how one might implement a "fallback" version of MöbiusVector$(P;t)$, which can be used for an arbitrary poset axis and requires only the functionality specified by the poset axis interface.

In the general case, the number of values $s \in P$ for which $\mu(s,t) \neq 0$ is bounded above by $|\Lambda_t|$, which we denote N for the purposes of this section. Since posets P exist for which $\mu(s,t) \neq 0$ exactly when $s \leq t$ (for example, boolean algebras), and as we shall see, the cost for the insertion of a value into a sparse tensor can be made $\mathcal{O}(1)$ on average, it follows that the cost required for an invocation of MöbiusVector$(P;t)$ in the general case can scale at best as $\mathcal{O}(N)$.

Naïvely calculating each $\mu(s,t)$ by recursive expansion of (3.14) leads however to a cost which may scale exponentially, depending on the structure of P. To avoid this, we can borrow from a slightly alternative characterisation of the Möbius function in purely linear-algebraic terms. We refer to, e.g., [NW78; God18] for more details, but this is basic and well-known. Index each $s_i \in \Lambda_t$ by some $1 \leq i \leq N$, and write Z to be the $N \times N$ matrix with entries $Z_{ij} = \zeta(s_i, s_j)$, where $\zeta(s_i, s_j)$ is the usual *zeta function* [Sta12]; here, if $s_i \leq s_j$, then $\zeta(s_i, s_j) = 1$, and if not, $\zeta(s_i, s_j) = 0$. As noted in [God18], the elements s_i can be indexed such that if $s_i < s_j$, then $i < j$, and so Z is upper-triangular; cf. [NW78, Chap. 25]. Then, if we temporarily repurpose notation and write M to be the equivalently-sized matrix with entries $M_{ij} = \mu(s_i, s_j)$, it holds simply that $M = Z^{-1}$ [NW78; God18]. Calculation of all values $\mu(s_i, s_j)$ thus reduces to a matrix inversion, with cost $\mathcal{O}(N^3)$. However, MöbiusVector$(P;t)$ needs to return only the final column M_N of M, and this can be had by solving $ZM_N = e_N$, where $e_N = \begin{bmatrix}0 & \cdots & 0 & 1\end{bmatrix}^T$ [Cor+22]. Since Z is upper-triangular, this costs only $\mathcal{O}(N^2)$ using back substitution; cf. here Godsil's derivation of the Möbius function in [God18, Lem. 2.2].[11]

[11] We note here also connections to other work related to sparse grids and the combination technique, for example the „Hierarchisierung und Dehierarchisierung" matrices discussed in [Mat14], or the linear systems (6) and (7) in [Heg+16]; cf. here [HGC07, (11)]. Indeed, Harding uses a basically equivalent linear-algebraic construction in [Har16b, Sec. 4.4.1] — but derived from an existing expression for the

Calculating MöbiusVector$(P;t)$ in this way also requires some ancillary work to establish the system for solution. In particular, we must actually obtain all the elements in Λ_t, and calculate the values $\zeta(s_i, s_j)$. Working within the confines of the poset axis interface, the former can be achieved by what amounts to a depth-first traversal of the Hasse diagram [Sta12] of Λ_t, which can be explored via repeated calls to Predecessors$(P;s_i)$, starting from a call to Predecessors$(P;t)$. Under the assumption that the cost of each such call scales linearly in the size of the output,[12] it follows from basic results in computer science that this costs at most $\mathcal{O}(N^2)$; see again [Cor+22]. For the latter, if $s_i \leq s_j$ can be explicitly tested at $\mathcal{O}(1)$ cost, which is the case for all posets we consider in this thesis, the naïve cost of all pairwise comparisons is also $\mathcal{O}(N^2)$. Even if not, that is, if the only known information about the ordering relationship is that given by the poset axis interface, then an algorithm given in [NW78, Chap. 26] can still be used to deliver MöbiusVector$(P;t)$ (actually, the complete matrix Z) at cost $\mathcal{O}(N^3)$.

We mention also that the application of standard techniques in applied computer science can make the cost of calculating MöbiusVector by recursive expansion more tolerable. It is just a restatement of the recursive definition of the Möbius function that

$$\text{MöbiusVector}(P;t) = \mathbf{1}_{\phi(t)} - \sum_{s<t} \text{MöbiusVector}(P;s), \tag{3.50}$$

where $\mathbf{1}_{\phi(t)}$ is an appropriately-sized one-dimensional sparse tensor that contains a single one in the $\phi(t)$th entry, and zeros everywhere else. A straightforward implementation of this expression also based on repeated calls to Predecessors$(P;s_i)$ can be computationally acceptable in repeated use if the results of evaluations of MöbiusVector$(P;s_i)$ are memoised, which incurs, however, an according and possibly non-trivial storage cost. Note here that, in an adaptive calculation, the poset P is explored from the zero up; so, since MöbiusVector$(P;s)$ is then already known for every $s < t$, the cost of evaluation is again at most $\mathcal{O}(N^2)$, under the same assumptions as above.

3.5.4. Adaptive selection strategies

The preceeding subsections have described the manipulation of various quantities related to the adaptive update $I^{(i)} \to I^{(i+1)}$ of the index set at each iteration i of our algorithm. We consider now the details of how new elements are selected for addition to the index set.

We define the subset $I^{(i)}_{\text{act}} \subseteq I^{(i)}$ of *active* elements of $I^{(i)}$ to be all those with at least one covering element which is itself not in $I^{(i)}$, that is,

$$I^{(i)}_{\text{act}} := \Big\{ \mathbf{p} \in I^{(i)} \;\Big|\; \exists \mathbf{q} \in \Pi - I^{(i)} \text{ s.t. } \mathbf{q} \succ \mathbf{p} \Big\}. \tag{3.51}$$

standard combination coefficients.

[12] A strong assumption which may not necessarily be achievable in practice. See examples and comments in Appendix B.

It should be noted that this definition is not necessarily equivalent to that used in previously-described adaptive combination-technique algorithms, such as [GG03; SG22]. The adaptive update process involves the *expansion* of one or more active elements $\mathbf{p} \in I^{(i)}_{\text{act}}$. Here, every successor $\mathbf{q} \succ \mathbf{p}$ is considered in turn; if all predecessors $\mathbf{p}' \prec \mathbf{q}$ are currently in the index set $I^{(i)}$, which we assume to be an order ideal of Π, then clearly also $I^{(i)} \cup \{\mathbf{q}\}$ is an order ideal of Π. We say that such a $\mathbf{q}$ is *admissible*, and select it for inclusion in $I^{(i+1)}$.

Following [GG03], we maintain a *queue* of elements $\mathbf{p} \in I^{(i)}$ which may currently be active; these are ranked in descending order, according to the corresponding benefit/cost ratios $\|\mathcal{L}[\tilde{f}_{\mathbf{p}}]\|/\mathcal{C}(\mathbf{p})$. As each new element $\mathbf{q}$ is added to the index set, its benefit/cost ratio is calculated, and $\mathbf{q}$ is inserted into the queue, ranked accordingly. Selection of elements to expand is achieved by removing maximally-ordered elements one-by-one from the queue, according to one of the three *selection strategies* which we shall describe below.

As each element $\mathbf{p}$ is removed from the queue, its successors are enumerated according to the GridSuccessors function described in Section 3.5.2 above. Each successor $\mathbf{q} \succ \mathbf{p}$ which is not already in the index set is tested for admissibility, by enumeration of its own GridPredecessors in turn. All such admissible successors are compiled into the set of new elements to add to the index set, I_{new}, such that $I^{(i+1)} \leftarrow I^{(i)} \cup I_{\text{new}}$.

If all successors of an active $\mathbf{p}$ are either admissible or already in the index set, then by definition, $\mathbf{p}$ will not be active in $I^{(i+1)}$, and is not considered further. Otherwise, $\mathbf{p}$ is reinserted into the queue. The justification of allowing an element to be thus considered multiple times is that, when an active element is expanded, we would like in principle to add every possible successor of that element to the index set as quickly as possible, including those which are currently inadmissible. If the original element is not requeued, then any such inadmissible element will not be added to the index set until, firstly, all of its own predecessors have been added, and secondly, one of those elements is itself expanded. If the original element is requeued, however, the inadmissible element will then be validly added to the index set at the next iteration after only the first of these two conditions becomes true.

We allow three possible strategies for the selection of the subset of active elements that are eligible for expansion at each step of our algorithm. The first strategy simply expands All of the active elements in the queue, regardless of their benefit/cost ratios. The intention of this strategy is to produce a sequence of approximations equivalent to those that would be obtained with a non-adaptive refinement of a combination sum, as would be obtained by calculating (3.8) for increasing values of L. In this case, the two are indeed equivalent. Care should be taken using this equivalence for intuition, since in the general poset grid case, it is possible that some active elements will be incompletely expanded and hence requeued during an All selection.

The second strategy expands only the Best active element; that is, the active element in the queue with the highest benefit/cost ratio. This strategy represents a faithful

implementation of a greedy solution to the binary knapsack problem; cf. [TW18] and references within. However, as the index set grows larger, the related bookkeeping cost associated with this strategy may become inefficient. As a compromise, we also allow selection of all queued and active elements whose benefit/cost ratios are at or above some THRESHOLD, defined as some factor $0 \leq \alpha \leq 1$ multiplied by the benefit/cost ratio of the BEST active element in the queue; we base this strategy on a similar approach outlined in [CGH18]. Note that here, a selection of $\alpha = 0$ corresponds to the ALL strategy, and $\alpha = 1$ to the BEST strategy, up to the possibility of multiple such best elements. There is a subtlety here. For example, it may be the case that the BEST-ranked active element in the queue may not have any admissible successors. Since this element would be requeued, it would then be selected again in the subsequent iteration, leading to an infinite loop and an unchanging index set. As such, the choice of elements to expand under the BEST and THRESHOLD strategies should more precisely be defined as being relative to the best active element in the queue that has at least one admissible successor that is not already in the index set.

For later use, we introduce a function SELECTELEMENTS, which we assume takes the queue Q, the current index set $I^{(i)}$, and a chosen `selection_strategy`, and returns the set I_{new}. Pseudocode for SELECTELEMENTS is given for the THRESHOLD case in Algorithm 3.2; the cases ALL and BEST are similar but simpler, and are not given explicitly for reasons of space. The main loop of the function removes elements from the queue in descending order, and comparing their benefit/cost ratios with a threshold. The threshold is initially set to a placeholder of $-\infty$; since every benefit/cost ratio is non-negative, every such ratio is above this value. The successors (if any) of each element which are not already in the index set are tested for admissibility, and added to I_{new} whenever so. The first time an element is encountered with at least one admissible successor, the threshold is explicitly updated; this can happen at most once. Once an element is encountered with benefit/cost ratio below the (necessarily updated) threshold, the loop terminates; this element is then requeued, along with any other elements which still have successors that have not been added to the index set, and the function terminates. Note that any non-active element which is removed from the queue is implicitly ignored and removed from any future consideration, since a non-active element can never become active at a future iteration.

3.5.5. Error indicator

Most existing dimension-adaptive combination technique and sparse grid algorithms provide an *error indicator*, see, e.g., [GG03; BG04; Gar07a; Gar12a; Nob+16; CGH18; SG22], and that which we now develop is similar in concept to those.

The adaptive algorithm we give here produces sequences of index sets $(I^{(i)})_{i=1}^{\infty}$, and implicitly, corresponding sequences of tensors $(D^{(I^{(i)})})_{i=1}^{\infty}$ and combination sums $(S_{I^{(i)}})_{i=1}^{\infty}$. Each $I^{(i)}$ is always a finite order ideal of Π. The maximal elements of $I^{(i)}$ form an

Algorithm 3.2 SelectElements for Threshold selection strategy

function SelectElements(Q, $I^{(i)}$, `selection_strategy` = Threshold)
  $\alpha \leftarrow$ a pre-chosen factor between 0 and 1
  $I_{\text{new}} \leftarrow \emptyset$ ▷ *New elements to be added to the index set.*
  $R \leftarrow \emptyset$ ▷ *Existing elements to be requeued.*

  `threshold` $\leftarrow -\infty$ ▷ *Placeholder threshold.*
  while Q is not empty **do**
    Remove element $\mathbf{p}$ with maximal $\|\mathcal{L}[\tilde{f}_{\mathbf{p}}]\|/\mathcal{C}(\mathbf{p})$ from Q.
    if $\|\mathcal{L}[\tilde{f}_{\mathbf{p}}]\|/\mathcal{C}(\mathbf{p}) <$ `threshold` **then**
      ▷ *Found first element below threshold. This element must be requeued.* ◁
      $R \leftarrow R \cup \{\mathbf{p}\}$
      Terminate while loop.
    ▷ *Consider all successors of* $\mathbf{p}$ *not already in the index set.* ◁
    `any_successor_added` $\leftarrow$ `false`
    for all $\mathbf{q} \in$ GridSuccessors($\mathbf{p}$) $- I^{(i)}$ **do**
      if GridPredecessors($\mathbf{q}$) $\subseteq I^{(i)}$ **then**
        ▷ $\mathbf{q}$ *is admissible.* ◁
        $I_{\text{new}} \leftarrow I_{\text{new}} \cup \{\mathbf{q}\}$
        `any_successor_added` $\leftarrow$ `true`
    if GridSuccessors($\mathbf{p}$) $\supset I^{(i)} \cup I_{\text{new}}$ **then**
      ▷ $\mathbf{p}$ *has an inadmissible successor; set aside for requeueing.* ◁
      $R \leftarrow R \cup \{\mathbf{p}\}$
    if `threshold` $= -\infty$ and `any_successor_added` **then**
      ▷ *First and thus best element with at least one admissible successor.* ◁
      `threshold` $\leftarrow \alpha \cdot (\|\mathcal{L}[\tilde{f}_{\mathbf{p}}]\|/\mathcal{C}(\mathbf{p}))$
      Terminate while loop.

  ▷ *Requeue unexpanded or incompletely expanded elements.* ◁
  for all $\mathbf{p} \in R$ **do**
    Reinsert $\mathbf{p}$ into Q, keyed by $\|\mathcal{L}[\tilde{f}_{\mathbf{p}}]\|/\mathcal{C}(\mathbf{p})$.

  return I_{new}

antichain [Sta12] of Π; that is, they are a subset $A^{(i)} = \{\mathbf{a}_1^{(i)}, \mathbf{a}_2^{(i)}, \ldots, \mathbf{a}_n^{(i)}\}$ such that for any $j \neq k$, neither $\mathbf{a}_j^{(i)} \leq \mathbf{a}_k^{(i)}$ nor $\mathbf{a}_k^{(i)} \leq \mathbf{a}_j^{(i)}$. Using the terminology and notation of [Sta12, Sec. 3.1], the antichain $A^{(i)}$ is said to *generate* the order ideal $I^{(i)}$, and we write $I^{(i)} = \langle A^{(i)} \rangle = \langle \mathbf{a}_1^{(i)}, \mathbf{a}_2^{(i)}, \ldots, \mathbf{a}_n^{(i)} \rangle$.[13] This suggests a natural error indicator which is determined purely by the state of the index set $I^{(i)}$ at each iteration, and carries no direct dependency on how precisely the algorithm arrived at that index set:

$$\mathcal{E}_i := \mathcal{L}\left[\sum_{\mathbf{a} \in A^{(i)}} \tilde{f}_{\mathbf{a}}\right] = \sum_{\mathbf{a} \in A^{(i)}} \mathcal{L}[\tilde{f}_{\mathbf{a}}]. \tag{3.52}$$

This definition is motivated by the informal idea that the surpluses for all elements $\mathbf{p} > \mathbf{a}_j^{(i)}$ are usually expected to be "small" and to decay "rapidly", so the sum of all the surpluses $\tilde{f}_{\mathbf{a}^{(i)}}$ can be used to approximately estimate the sum of those elements and thus the error of $S_{I^{(i)}}$; on this latter, see, e.g., [Nob+16; TW18, Sec. 3.1].

Given an arbitrary finite order ideal I, naïvely calculating its generating antichain A requires $\mathcal{O}(|I|^2)$ pairwise evaluations of $\mathbf{s} \leq \mathbf{t}$. We avoid this in our algorithm by maintaining and updating a *generator set* $A^{(i)}$ alongside the index set $I^{(i)}$ at each iteration of the adaptive refinement process. A new element $\mathbf{p}$ can be added in iteration i only if all of its predecessors are contained in $I^{(i-1)}$. As a result, every such element must be a maximal element of $I^{(i)}$, so it can also be immediately added to $A^{(i)}$. Every predecessor of the newly-added $\mathbf{p}$ can then no longer be a maximal element of $I^{(i)}$, and is removed from $A^{(i)}$ if present. This is sufficient to maintain the invariant that $A^{(i)}$ indeed generates $I^{(i)}$ at the completion of the ith iteration.

The explicit calculation of $\mathcal{E}_i$ according to (3.52) is prone to all of the same numerical issues associated with direct evaluation of a combination sum $\mathcal{L}[S_I]$. We avoid this in the same way: we maintain an explicit error-indicator tensor E, analogously to the full combination tensor D. Once the Möbius tensor $M^{(\mathbf{p})}$ of a newly-added element $\mathbf{p}$ has been calculated and summed into the full combination tensor D, it is also summed into the error estimator tensor E. Similarly, when an element is evicted from $A^{(i)}$ for being a predecessor of a newly-added element, its Möbius tensor is either recalculated or retrieved from some form of cache, and subtracted from E. The value of the error indicator $\mathcal{E}_i$ at the end of the iteration can then be obtained by a product-and-reduce operation $\mathcal{E}_i = \textsc{Reduce}(E \odot V)$ against the value tensor V, equivalently to the calculation of the combination sum itself.

[13] Technically, Stanley constructs the concept of a generating antichain only for a finite poset P, by building a bijection between the set of all antichains of P and the set $J(P)$ of all order ideals of P [Sta12, Sec. 3.1]. It is unproblematic in our context to extend the idea to an infinite poset P, and speak only of generating antichains for the members of the set $J_f(P)$ of explicitly finite order ideals of P.

3.5.6. Main loop

Pseudocode for the main loop of our adaptive algorithm is given in Algorithm 3.3. The algorithm begins by initialising empty sets I and A, to represent the current index set and the current generator set respectively. An empty initial queue, Q, is constructed. Four empty sparse tensors D, V, C, and E must also be created; the shapes of these tensors will depend on the poset grid Π, as described in Section 3.5.1.

Each loop iteration begins with the selection of a set I_{new} of new elements to add to the index set. On the first iteration, I_{new} contains only the zero element $\hat{0}_\Pi$ of the poset grid Π (line 8). On all subsequent iterations, the calculation of I_{new} is delegated to the SelectElements function, according to a pre-chosen selection strategy, as described in Section 3.5.4.

Once I_{new} is calculated, each new element $\mathbf{p} \in I_{\text{new}}$ is incorporated into the index set and the approximation in turn. First, $\mathcal{L}[f_{\mathbf{p}}]$ is explicitly calculated, and inserted into the value tensor V at a location according to the indexing bijection Φ (line 13). The (abstract) cost of evaluation $\mathcal{C}(\mathbf{p})$ is also calculated and stored. The Möbius tensor for $\mathbf{p}$ is then evaluated, and accumulated into both the full combination tensor (line 17) and the error indicator tensor (line 18). Finally, the hierarchical surplus $\mathcal{L}[\tilde{f}_{\mathbf{p}}]$ is calculated directly as a reduction of the elementwise product of the Möbius tensor $M^{(\mathbf{p})}$ and the value tensor V (line 20), and an appropriate norm of this quantity is used to calculate a benefit/cost ratio and to enqueue $\mathbf{p}$ for consideration in future iterations.

Once every new element has been treated, the data structures for the index set and the generator set are updated for consistency. As described in Section 3.5.5, the generator set must be further processed; all direct predecessors of any newly-added element $\mathbf{p} \in I_{\text{new}}$ that were in A are removed (lines 25 and 26), and their Möbius tensors are discounted from the error indicator tensor E. This done, the full combination sum S_i, the error indicator $\mathcal{E}_i$, and the complete cost of the algorithm so far C_i can be calculated. The main loop of the algorithm repeats until $\mathcal{E}_i$ and $\mathcal{C}_i$ meet or exceed some particular choice of termination criteria, at which point the final index set is returned.

3.5.7. Data structures and computational complexity

To conclude, we briefly and informally sketch some computational details involved in an implementation of Algorithm 3.3. We give this only in the interest of completeness and to aid an implementation, and do not seek to establish rigorous complexity results for the algorithm in any particular case. For basic computer science concepts, we refer particularly to [Cor+22]. For details of the specific implementation we use throughout the remainder of this work, see Section A.8.

Algorithm 3.3, as well as the subsidiary functions specified by the poset grid interface, involve three distinct kinds of data structures: sets, sparse tensors, and the queue. On the assumption that the involved elements $\mathbf{p} \in \Pi$ admit a suitable hash function, the

Algorithm 3.3 Adaptive construction of an order-ideal index set for a poset grid Π.

$I, A \leftarrow \emptyset$ ▷ *Index set and generator set.*
$Q \leftarrow$ an empty queue ▷ *Expansion queue.*
$D, V, C, E \leftarrow$ empty sparse tensors ▷ *Combination, value, cost, and error tensors.*
$i \leftarrow 0$ ▷ *Iteration counter.*
repeat
  ▷ *Select new elements and include in index set.* ◁
  if $i = 0$ **then**
    $I_{\text{new}} \leftarrow \{\hat{0}_\Pi\}$
  else
    $I_{\text{new}} \leftarrow \textsc{SelectElements}(Q, I^{(i)}, \texttt{selection\_strategy})$
  for all $\mathbf{p} \in I_{\text{new}}$ **do**
    ▷ *Evaluate $\mathcal{L}[f_{\mathbf{p}}]$, and calculate its cost.* ◁
    $V_{\Phi(\mathbf{p})} \leftarrow \mathcal{L}[f_{\mathbf{p}}]$
    $C_{\Phi(\mathbf{p})} \leftarrow \mathcal{C}(\mathbf{p})$
    ▷ *Calculate and accumulate the Möbius tensor of $\mathbf{p}$...* ◁
    $M^{(\mathbf{p})} \leftarrow \textsc{MöbiusTensor}(\mathbf{p})$
    $D \leftarrow D + M^{(\mathbf{p})}$ ▷ *...into the full combination tensor,*
    $E \leftarrow E + M^{(\mathbf{p})}$ ▷ *...and into the error indicator tensor.*
    ▷ *Calculate the surplus for $\mathbf{p}$, and enqueue $\mathbf{p}$ for later expansion.* ◁
    $\mathcal{L}[\tilde{f}_{\mathbf{p}}] \leftarrow \textsc{Reduce}(M^{(\mathbf{p})} \odot V)$
    Insert $\mathbf{p}$ into Q, keyed by $\|\mathcal{L}[\tilde{f}_{\mathbf{p}}]\|/C_{\Phi(\mathbf{p})}$.
  ▷ *Update index set, generator set, and error indicator tensor.* ◁
  $I \leftarrow I \cup I_{\text{new}}$
  $A \leftarrow A \cup I_{\text{new}}$
  $R \leftarrow A \cap \bigcup_{\mathbf{p} \in I_{\text{new}}} \textsc{GridPredecessors}(\mathbf{p})$
  $A \leftarrow A - R$
  $E \leftarrow E - \sum_{\mathbf{p} \in R} \textsc{MöbiusTensor}(\mathbf{p})$
  ▷ *Calculate approximation, error indicator, and cost.* ◁
  $i \leftarrow i + 1$
  $S_i \leftarrow \textsc{Reduce}(D \odot V)$
  $\mathcal{E}_i \leftarrow \textsc{Reduce}(E \odot V)$
  $\mathcal{C}_i \leftarrow \textsc{Reduce}(\mathbf{1}[V] \odot C)$
until $\mathcal{E}_i$ and $\mathcal{C}_i$ meet termination criteria.
return $I^{(i)}$

sets I, A, and R can be readily represented using hash tables [Cor+22, Chap. 11].[14] Informally, this leads to unions, intersections, and set difference operations with cost at worst linear in the size of the largest of the two involved sets; this is achieved in practice by, e.g., the Python `set` built-in datatype [PWTC].

There are several ways to implement the sparse tensor datatype and related operations, depending on the particular sparse storage format used. For a detailed discussion of some possible approaches, see, e.g., [BK08]. However, for our purposes, a completely adequate implementation is obtained by constructing a sparse tensor as an associative array, with each non-zero entry stored by association with its multiindex. This is sometimes referred to as a *dictionary-of-keys* (DOK) format; see, e.g., [ASW07], also the documentation of the `scipy.sparse` package in the SciPy library [Vir+20]. There exists at least one existing Python implementation of arbitrary-dimensional sparse tensors that supports DOK storage [Sparse]. Such a scheme can also be based on hash tables, and so the cost of setting and accessing entries is basically constant; see again [Cor+22]. Assuming that the multiindex/value pairs for the non-zero entries in a sparse tensor can be enumerated at a cost linear in their number, it is trivial to construct naïve implementations of the simple arithmetic operations required here. In particular, the cost of obvious implementations of elementwise addition or multiplication of two sparse tensors T and U scales on average as $\mathcal{O}(|T| + |U|)$, where we abuse notation and mean $|T|$ to be the number of non-zero entries in T, and a Reduce operation over T scales similarly as $\mathcal{O}(|T|)$.

The queue Q can be most easily realised using a MaxHeap data structure [Cor+22, Chap. 6]; see also [GG03]. It is well-known that there exist particular implementations of MaxHeap that guarantee constant-cost insertion of an element, and that extraction of a maximally-keyed element from a heap containing N elements costs $\mathcal{O}(\log N)$; see, e.g., the introduction to [Cor+22, Part V] and references within.

The cost of the SelectElements function is necessarily dependent on both the selection strategy used, and on the structure of the queue. In the worst (or at least, most costly) case, which is equivalent to the All selection strategy, it will involve at least one call to GridSuccessors for each element $\mathbf{p}$ in the queue, followed by calls to GridPredecessors for each such successor. The cost of these latter functions is determined by the structure of the poset axes P_i involved. Although in simple cases, such as the standard combination technique grid $\Pi = \mathbb{N}^d$, these functions may be implemented at constant cost for fixed dimension d, this is not necessarily so for all possible poset axes; see Appendix B.

The cost of each iteration of the loop over new elements (lines 13–21), and thus the inclusion of each new element in the index set, requires a number of operations linear in the number of non-zero values of the Möbius tensor $M^{(\mathbf{p})}$ for each $\mathbf{p} \in I_{\text{new}}$, as well as the explicit evaluation of both $\mathcal{L}[f_{\mathbf{p}}]$ and $\mathcal{C}(\mathbf{p})$ and the Möbius tensor itself. In practice, the cost of these latter operations — which are also dependent on the problem setting and on

[14] If no better hash function is available, one can be constructed using the required indexing bijection ϕ.

the involved poset axes P_i — may be expected to dominate that of the surrounding work, and this is certainly true for the applications we consider for the remainder of this thesis.

4. Composite methods

In the quantum chemistry literature, a particular type of two-dimensional grid-style diagram is commonly used to visualise and organise *ab initio* calculations by their approximation quality, and, implicitly, their computational expense [Hea96; Kar16]; for examples, see [Hea96, Fig. 1; HOJ13, Fig. 15.1]. Such charts, as well as their annotated or adapted extensions, are related to a famous plot by Pople [Pop65, Fig. 1; Kar90], and are sometimes referred to as *(modified) Pople diagrams* [Kar16; Zas+18]; Pople himself discussed just such a diagram in the lecture he gave upon being awarded the Nobel Prize [Pop99, Fig. 1].

On a Pople diagram, calculations are organised along one axis by the precision of the core *ab initio* approximation in use, in steps often basically as in (2.26); along the other, by the granularity of discretisation applied to $H^1(\mathbb{R}^3)$, ranked for instance by the cardinal number n indexing the cc-pVnZ family [Dun89] of basis sets [Kar16]. Some formulations may also include a third axis indexing the quality of approximate solutions of the relativistic Dirac equation, as in the "magic cube" of [Tar+01, Fig. 1]. In any case, it is understood that each step further away from the origin on the underlying grid implies a *level of theory* both offering a higher accuracy and carrying a higher cost [Hea96; Kar16].

Pople diagrams have been used to motivate and explain the structure of several instances of extrapolative *composite methods*; see, e.g., [PTM91, Fig. 1; Pop99, Fig. 2; Kar16, Fig. 1; Jen17, Figs. 5.4 and 5.5; Zas+18, Fig. 1]. Composite methods seek to divide and reunify multiple approximate solutions to the Schrödinger equation in such a way as to recover a single, highly-accurate value [PFD12; RS15; Kar16; Zas+18]. Here, there is a limited assumption of additive separability of the basis-set and intrinsic errors mentioned in Section 2.3; see, e.g., [Pop+83; Pop+89; Pop99; PFD12; RS15; GKC17; Jen17; CGH18]. Plotting the individual subcalculations required by a composite method on a Pople diagram makes visible [Pop99, Fig. 2; Jen17, Figs. 5.4 and 5.5] a sampling from a roughly lower-triangular subset of the points on the diagram grid.

We will begin this chapter by outlining some simple one-dimensional extrapolative techniques for obtaining higher-accuracy quantum chemical results from relatively cheaper instances of the same. We will then review the formulations of a selection of composite methods. After discussing certain similarities between their working equations and the summation patterns of the standard combination technique, as previously noted in [CGH18] and exploited by [Zas+18], we will apply the machinery developed in Chapter 3 to the development of a *generalised composite method* (GCM); this is itself very closely

related to a construction given in [Zas+18]. We will conclude by investigating in detail the performance of the GCM as a systematically-refinable tool for the high-accuracy calculation of energetic properties for three small molecular systems.

Before we proceed, a word on notation. For the remainder of this thesis, we will be interested exclusively in ground-state energetic properties; thus, we elide the zero subscript on, e.g., the ground-state total energy that we used in Chapter 2. We will instead make liberal use of superscripts and subscripts to indicate different energetic quantities calculated at different levels of theory; we are influenced here perhaps most strongly by notation used in works relating to the HEAT methods, e.g., [Taj+04; Bom+06; Har+08]. We will also generally use a prefixed Δ to indicate various correction terms, including correlation energies. Finally, for readability and consistency, we will in most cases report formulae using equivalent but not identical notation to that in cited sources, without further explicit comment.

4.1. Energetic extrapolation

The simplest approach for extracting additional accuracy from calculations drawn from a notional Pople-diagram grid is unidirectional extrapolation. Given some collection of results obtained by ranging along a particular direction on such a grid, such as values obtained with denser and denser basis sets [Mar96; Var07; FPH11; PFD12], or by the use of increasingly involved treatments of electron correlation [Fel93; Goo02; BR04a], and under some assumptions regarding the decay in approximation error, one can seek to extrapolate those results towards an anticipated limit. Unidirectional extrapolation techniques can be applied in each of the two directions on a Pople-diagram grid: both towards the complete basis set (CBS) limit, and towards the best-possible FCI solution. For more detail on the following, as well as empirical investigation of their relative qualities, see, e.g., [FP07; FPD08; HKT08; FPH11; Fel13; Var18], to which we make general reference.

4.1.1. Complete basis set (CBS) extrapolation

We collect here some CBS extrapolation formulae discussed in the comparative study of Feller et al. [FPH11], covering only those formulae which we shall explicitly refer to later in the chapter. The interested reader is referred to that source for further information, and also other reviews mentioned above. Let us note that effectively the same formulae were used in [CGH18] to partially motivate the construction of the ML-BOSSANOVA method, to which we shall return in later chapters.

Use of the well-known and basically empirically-motivated extrapolation formula

$$E_n^{\mathrm{tot}} = E_\infty^{\mathrm{tot}} + Ae^{-\alpha n}. \tag{4.1}$$

traces back to Feller [Fel92; Fel93]. Each E_n^{tot} is here the total (or, alternatively, interaction) energy for a molecular system, as calculated using cc-pVnZ. Fitting the three unknowns E_∞^{tot}, A, and α requires three total-energy values, usually taken as E_n^{tot}, E_{n+1}^{tot}, and E_{n+2}^{tot} for some n [Var18]. Thus, the approximate CBS-limit energy E_∞^{tot} so obtained from (4.1) is referred to as a *three-point extrapolation*. An alternative three-point formula due at least to Peterson et al. [PWD94] extrapolates the total energy as

$$E_n^{\text{tot}} = E_\infty^{\text{tot}} + Ae^{-(n-1)} + Be^{-(n-1)^2}. \tag{4.2}$$

The two formulas given above presuppose exponential decay in the total energy, and thus both in the Hartree-Fock and correlation-energy components, but there is theoretical reason to anticipate that the correlation energy should decay only algebraically [Hel+97; FPH11]. Helgaker et al. [Hel+97] suggested that the correlation energy E^{corr} should therefore be extrapolated in isolation from the Hartree-Fock energy E^{HF}, the latter according to (4.1), and the former with the *two-point* extrapolation formula

$$E_n^{\text{corr}} = E_\infty^{\text{corr}} + An^{-3}. \tag{4.3}$$

The practical evaluation of Feller et al. [FPH11] suggests that, given energies calculated using basis sets up to (aug)-cc-pVnZ, these formulae can produce extrapolated values with accuracy consistent at least with (aug)-cc-pV$(n+1)$Z calculations.

4.1.2. Full configuration-interaction (FCI) extrapolation

Clearly, the form of an extrapolation towards an FCI solution will depend on the type of calculations of correlation energy used in that extrapolation. When these latter are taken as sums of increasingly high-order terms in a Rayleigh-Schrödinger or Møller-Plesset perturbation expansion, the convergence of the underlying power series can be accelerated by the introduction of Padé approximants and similar quantities [BG70; BS77; Wil80; HC96]. For example, Pople et al. [Pop+83] suggest an extrapolative approximation of the correlation energy using second-, third-, and fourth-order Møller-Plesset terms [Pop+83, (4)]:

$$E^{\text{corr}} \approx \frac{E^{(2)} - E^{(3)}}{1 - E^{(4)}/E^{(2)}}, \tag{4.4}$$

which they note amounts to an assumption "that the even and odd terms of the Moller-Plesset series both constitute geometric series with a constant ratio of $E^{(4)}/E^{(2)}$" [Pop+83, p. 312].

If the correlation energy is approximated by a series of truncated configuration interaction calculations (CISD, CISDT, etc.), then it may be possible to apply exponential extrapolative formulae similar to (4.1) above, as also suggested by Feller [Fel93]. Here, the quantities used for fitting are not functions of the basis set cardinal n, but rather of values derived from the individual coefficients in (2.17). We note informally that, since

this approach requires knowledge of the values of the CI coefficients, it is less readily applied in a "black box" manner than are the CBS extrapolations considered in the previous section.

An even more intrusive extrapolation scheme in the context of truncated CI is provided by the *correlation energy extrapolation by intrinsic scaling* (CEEIS) method of Bytautas and Ruedenberg [BR04a]. As CEEIS is relatively complex in the details, we refer the reader to the source for further information. Briefly, however, studies by the original authors suggest that CEEIS approximations to the true FCI energy are in some cases good to at least the mE_h [BR04b; BR05].

The only extrapolation technique of which we are aware that is explicitly defined for series of coupled cluster results is due to Goodson [Goo02]. Here, the FCI energy is decomposed into a series of correction terms, $E^{\mathrm{FCI}} = E^{\mathrm{CCSD(T)}} + \cdots = E^{\mathrm{HF}} + \Delta E^{\mathrm{CCSD}} + \Delta E^{\mathrm{CCSD(T)}} + \cdots$, which are then rearranged into the form of a continued fraction. This continued fraction is itself truncated below all terms which can be explicitly calculated. The continued-fraction extrapolation has been utilised in some applied studies [Pet+06; BP06], and has also been investigated in a form adjusted to use sequences of correction terms obtained from CCSD, CCSDT, and CCSDTQ calculations [FPC06]. However, there appears to be a limited consensus that the corrections provided by the technique are too unreliable for production use [Fel+03; FD03; Boe+04], at least in the CCSD(T) version.

4.2. Conventional composite methods

The unidirectional extrapolation approaches discussed above combine by definition sets of calculations that vary in only one of the two refinement types provided by a Pople-diagram grid: either an adjustment of the basis set, or by modification of the form of the correlation energy approximation. The concept of varying *both* refinement types was explored at least in [Pop+83]. There, total energies were first approximated by HF-based calculations using the 6-311G** basis set [Kri+80] and including an extrapolated Møller-Plesset term as per (4.4). These were then adjusted by three distinct correction terms, each formed as the difference between the original reference 6-311G** total energy and another total energy computed similarly,[1] but using one of three slightly different basis sets. These latter were constructed by extension from 6-311G** to be in some sense "orthogonal to each other" [Pop+83, p. 314], in the hope that the reductions in total basis-set related error that they produced in comparison to 6-311G** would be additive, approximately if not exactly. A similar construction is described in [Pop+85], although apparently without the use of an extrapolative treatment of the correlation energy. These schemes can be considered as early examples of *composite methods* [GKC17].

[1] It is not completely clear to this author's reading whether the correction terms in [Pop+83] were also calculated using the MPn extrapolative formula.

We will consider here only four particularly well-known and representative families of composite methods, all of which provide "fixed recipes" [GKC17; FD18] for the calculation of thermochemical values. We make general reference to the reviews of Helgaker et al. [HKT08], Peterson et al. [PFD12], Raghavachari and Saha [RS15], and Karton [Kar16], as well as [Jen17]. We mention also a more recent review again by Karton [Kar22]; see comments in Section 4.3 below. We leave full details, motivations, etc., to the sources, and limit ourselves basically to recalling the forms of their working energy equations. When we speak of "total energies" here without further qualification, we mean (approximate) ground-state solutions to the electronic Schrödinger equation (2.1), always adjusted by a post-inclusion of the nuclear repulsion energy, as in (2.6). It should be stressed that all of the following methods are intended in their original formulations to approximate true physical energies at 0 K, and thus include approximate corrections both for errors introduced by the Born-Oppenheimer approximation and also for the non-relativistic formulation of the (electronic) Schrödinger equation. For simplicity, we avoid this detail here, and so the energy equations given below should be considered adapted from the originals.

4.2.1. The Gaussian-n family (Gn)

The Gn family of composite methods [Cur+90; Cur+91; Cur+98; Cur+99b; Cur+99a; Cur+01; CRR07a; CRR07b] represent, in short, a series of incremental developments and improvements to the original *Gaussian-1* (G1) method [Pop+89]. We focus here explicitly only on the G4(MP2) model [CRR07b]. After the molecular system under study is subjected to an initial DFT geometry optimisation, a base approximation $E^{\mathrm{CCSD(T),FC}}_{\text{6-31G*}}$ to its total energy is calculated via, as suggested by the sub- and superscripts, a frozen-core CCSD(T) calculation according to the 6-31G* basis set [DHP71; HDP72; HP73]. Two additive corrections are defined:

$$\Delta E^{\mathrm{MP2,FC}}_{\text{G3(MP2)LargeXP}} = E^{\mathrm{MP2,FC}}_{\text{G3(MP2)LargeXP}} - E^{\mathrm{MP2,FC}}_{\text{6-31G*}}, \tag{4.5}$$

$$\Delta E^{\mathrm{HF}}_{\infty} = E^{\mathrm{HF}}_{\infty} - E^{\mathrm{HF}}_{\text{G3(MP2)LargeXP}}. \tag{4.6}$$

The former expression attempts to estimate the benefit that would have been obtained, had the base CCSD(T) calculation been performed using a more detailed basis set than 6-31G*, specifically the customised G3(MP2)LargeXP basis set [Cur+98; CRR07a; CRR07b]. The latter expression further compensates for the discretisation error in the Hartree-Fock component of that secondary calculation, and, implicitly, that of the original. Here, E^{HF}_{∞} is an extrapolated approximation to the CBS-limit Hartree-Fock energy, as per the two-point formula obtained from (4.1) for fixed $\alpha = 1.63$, presumably following [Hal+99]; the energies extrapolated from are calculated with adjusted variants of the aug-cc-pVTZ and aug-cc-pVQZ basis sets.

The G4(MP2) total energy is, then, adapting from [CRR07b, (1)],

$$E^{\mathrm{G4(MP2)}} = E^{\mathrm{CCSD(T),FC}}_{\text{6-31G*}} + \Delta E^{\mathrm{MP2,FC}}_{\mathrm{G3(MP2)LargeXP}} + \Delta E^{\mathrm{HF}}_{\infty} + \Delta E^{\mathrm{HLC}}. \tag{4.7}$$

The trailing ΔE^{HLC} term is a *higher-level correction* (HLC). For full details, see [CRR07a; CRR07b], but it suffices to say that this is given by a parametrised expression, the precise parameters of which minimise an error quantity over a particular training dataset; see further, e.g., [Cur+97; CRR05], and also discussion in [DCW06]. The total G4(MP2) energy at 0 K follows from (4.7) by the addition of further correction terms, as mentioned above. We omit the details here.

As independently benchmarked, the G4(MP2) method achieves accuracy levels at or slightly above the conventionally-set $1\,\mathrm{kcal\,mol^{-1}}$ cutoff of chemical accuracy [KSM17]. The transferability of the G4(MP2) HLC term to molecular systems that are substantially different to those on which it was originally parametrised has also been recently investigated [DCR21], and at least partially confirmed. The G4(MP2) protocol has seen wide use in practical applications, such as the provision of accurate training values for use in quantum machine learning [Ram+14; Zas+18; Nar+19; Bar+21].

4.2.2. The correlation-consistent Composite Approach (ccCA)

The correlation-consistent Composite Approach [*sic*] (*ccCA*) [DCW06] is closely and explicitly related to the G*n* methods, differing most notably in the deliberate omission of any term like, e.g., the G*n*-family HLCs. As in the G*n* case, there are a number of alternative formulations of the ccCA [DCW06; DeY+09]. We consider here the ccCA-PS3 form as described in [DeY+09], which also presupposes an equilibrium geometry calculated via a particular DFT calculation. A base total energy is derived as an extrapolated approximation of the CBS frozen-core MP2 total energy, $E^{\mathrm{MP2,FC}}_{\infty}$. This approximation is formed from individual extrapolations of the Hartree-Fock total energy and MP2 correlation energy. The Hartree-Fock energy is extrapolated according to (4.1) with $\alpha = 1.63$, as per [Hal+99]. The frozen-core MP2 correlation energy is taken as the mean of two separate extrapolations, one according to (4.2), and another according to (4.3). All extrapolations are performed using results obtained with the aug-cc-pVDZ, aug-cc-pVTZ, and aug-cc-pVQZ basis sets.

The complete ccCA total energy is then, adapting from [DeY+09, (1.6)],

$$E^{\mathrm{ccCA}} = E^{\mathrm{MP2,FC}}_{\infty} + \Delta E^{\mathrm{CCSD(T),FC}}_{\text{cc-pVTZ}} + \Delta E^{\mathrm{MP2,FC1}}_{\text{aug-cc-pCVTZ}}, \tag{4.8}$$

where the first correction term attempts to approximate the error reduction that would be provided by a full CCSD(T) calculation rather than MP2 [DeY+09, (1.3)],

$$\Delta E^{\mathrm{CCSD(T),FC}}_{\text{cc-pVTZ}} = E^{\mathrm{CCSD(T),FC}}_{\text{cc-pVTZ}} - E^{\mathrm{MP2,FC}}_{\text{cc-pVTZ}}, \tag{4.9}$$

and the second correction term partially addresses the use of the frozen-core approximation in the above [DeY+09, (1.4)],

$$\Delta E^{\mathrm{MP2,FC1}}_{\text{aug-cc-pCVTZ}} = E^{\mathrm{MP2,FC1}}_{\text{aug-cc-pCVTZ}} - E^{\mathrm{MP2,FC}}_{\text{aug-cc-pVTZ}}. \tag{4.10}$$

For our purposes, "FC1" can be considered to indicate an all-electron calculation, although some orbitals would be frozen for atoms living in higher rows of the periodic table than those we will explicitly consider; see [DeY+09, p. 1109]. Additional corrective terms are applied to push the ccCA energy E^{ccCA} towards the true (i.e., relativistic) total energy; we do not discuss these here.

When tested with reference to a dataset of thermochemical properties that was itself constructed to benchmark certain Gn models [CRR05], the ccCA methods, and ccCA-PS3 in particular, produced a mean accuracy very close to $1\,\mathrm{kcal\,mol^{-1}}$ [DeY+09]. A more recent benchmark using the W4-17 dataset [KSM17] found the ccCA-PS3 method to perform slightly better than the G4 and G4(MP2) methods.

4.2.3. The Weizmann family (Wn)

The Weizmann Wn family [MO99; Boe+04; Kar+06] of composite methods aim to produce much more accurate thermochemical values than do the Gn or ccCA methods. In their paper describing the original W1 and W2 methods, Martin and de Oliveira [MO99] state their target to be *calibration accuracy*, which they define at the dataset level to be a mean absolute error of $1\,\mathrm{kJ\,mol^{-1}}$ ($\approx 0.24\,\mathrm{kcal\,mol^{-1}}$), "with the additional constraint that no individual error be larger than the chemical accuracy goal of $1\,\mathrm{kcal\,mol^{-1}}$" [MO99, p. 1843]. This target can be broadly reached at least by the subsequently-developed W4 method; see, e.g., [KDM11; KSM17].

We sketch only a rough outline of W4 here, following [Kar+06; KSM17]. The details are complex; for a full description, we refer to the sources, but the total W4 energy can basically be written as[2]

$$\begin{aligned} E^{\mathrm{W4}} := E^{\mathrm{HF}}_{\infty} + \Delta E^{\mathrm{CCSD}}_{\infty} + \Delta E^{\mathrm{CCSD(T)}}_{\infty} + \Delta E^{\mathrm{CCSDT}}_{\infty} \\ + \Delta E^{\mathrm{CCSDT(Q)}}_{\text{cc-pVTZ}} + \Delta E^{\mathrm{CCSDTQ}}_{\text{cc-pVDZ}} + \Delta E^{\mathrm{CCSDTQP}}_{\mathrm{DZ}} + \Delta E^{\mathrm{CV}}. \end{aligned} \tag{4.11}$$

We will briefly outline the various correction terms below. As for the Gn and ccCA methods, the total W4 energy at 0 K includes further corrections, again left here undescribed.

In the following, when we write W4-pVnZ, we refer to a composite basis set that is, for our purposes, just aug-cc-pVnZ, but with cc-pVnZ used for hydrogen atoms; cf., however, [Kar+06; KDM11]. The first two terms on the RHS of (4.11), E^{HF}_{∞} and $\Delta E^{\mathrm{CCSD}}_{\infty}$, are two-point CBS extrapolations, calculated using formulae [KM06; Kar+06] which we

[2] Actually, omitting an additional term relating to technical differences between two CCSD(T) implementations used in the original W4 implementation [Kar+06].

did not give above, from W4-pV5Z and W4-pV6Z values of the total HF energy and the CCSD correlation energy respectively. The calculation of $\Delta E_\infty^{\mathrm{CCSD}}$ requires some additional work; see [Kar+06]. All of the remaining terms but the last estimate the improvements available from progressively better-quality coupled cluster treatments. For instance, the term $\Delta E_\infty^{\mathrm{CCSD(T)}}$ approximates the "true" difference $E_\infty^{\mathrm{CCSD(T)}} - E_\infty^{\mathrm{CCSD}}$ as a two-point extrapolation of those differences as approximated using W4-pVQZ and W4-pV5Z; $\Delta E_\infty^{\mathrm{CCSDT}}$ is similar, but uses only the standard cc-pVDZ and cc-pVTZ basis sets. The terms $\Delta E_{\text{cc-pVTZ}}^{\mathrm{CCSDT(Q)}}$ and $\Delta E_{\text{cc-pVDZ}}^{\mathrm{CCSDTQ}}$ are scaled differences; for instance, $\Delta E_{\text{cc-pVTZ}}^{\mathrm{CCSDT(Q)}} = 1.1(E_{\text{cc-pVTZ}}^{\mathrm{CCSDT(Q)}} - E_{\text{cc-pVTZ}}^{\mathrm{CCSDT}})$ [Kar+06, (1)]. In $\Delta E_{\mathrm{DZ}}^{\mathrm{CCSDTQP}} = E_{\mathrm{DZ}}^{\mathrm{CCSDTQP}} - E_{\mathrm{DZ}}^{\mathrm{CCSDTQ}}$, "DZ" indicates for our purposes the unpolarised double-zeta basis set of Dunning and Hay [DH77]. Since all of the preceeding calculations apply the frozen-core approximation [KSM17], the final correction term ΔE^{CV} includes an extrapolated estimate of the resulting systematic error from the differences of all-electron CCSD(T) correlation energies calculated with the aug-cc-pwCVTZ and aug-cc-pwCVQZ basis sets, and the frozen-core versions of the same.

Several variants of W4 are available [Kar+06; KSM17]. In particular, the cheaper W4Lite variant of (4.11) replaces $\Delta E_{\text{cc-pVTZ}}^{\mathrm{CCSDT(Q)}}$ with an equivalently-defined term $\Delta E_{\text{cc-pVDZ}}^{\mathrm{CCSDT(Q)}}$, and omits $\Delta E_{\text{cc-pVDZ}}^{\mathrm{CCSDTQ}}$ and $\Delta E_{\mathrm{DZ}}^{\mathrm{CCSDTQP}}$ entirely. Alternatively, the more expensive W4.2 variant calculates ΔE^{CV} using CCSDT. The calculation of a W4.2-approximated energy is only feasible for molecular systems "with up to 2–3 non-hydrogen atoms" [KSM17, p. 2064]. W4Lite, by contrast, can handle "up to eight" [KSM17, p. 2065].

4.2.4. High-accuracy Extrapolated *ab initio* Thermochemistry (HEAT)

The final composite method family that we shall outline are the *High-Accuracy Extrapolated* ab initio *Thermochemistry* (HEAT) methods [Taj+04; Bom+06; Har+08; Tho+19]. Like the Wn methods, the HEAT methods target accuracies of $\mathrm{kJ\,mol^{-1}}$ or better [Taj+04]. The total energy for the HEAT-456QP method [Bom+06; Har+08], which is the most comprehensive of the family in terms of its treatment of the correlation energy and thus also the most expensive, is, cf. e.g. [Bom+06, (1)]

$$E^{\text{HEAT-456QP}} := E_\infty^{\mathrm{HF}} + \Delta E_\infty^{\mathrm{CCSD(T)}} + \Delta E_\infty^{\mathrm{CCSDT}} + \Delta E_{\text{cc-pVDZ}}^{\mathrm{CCSDTQP}}. \tag{4.12}$$

Obtaining the first two terms here requires calculation of $E_{\text{aug-cc-pCVQZ}}^{\mathrm{CCSD(T)}}$, $E_{\text{aug-cc-pCV5Z}}^{\mathrm{CCSD(T)}}$, and $E_{\text{aug-cc-pCV6Z}}^{\mathrm{CCSD(T)}}$. The Hartree-Fock energy components of these are then extrapolated to E_∞^{HF} via (4.1), and the correlation energy components of the latter two provide $\Delta E_\infty^{\mathrm{CCSD(T)}}$ as per (4.3). These are all-electron calculations, and the HEAT methods do not usually include an explicit core-valence correction term [Taj+04; Har+08]; cf., however, the variants described in [Tho+19]. The next term, $\Delta E_\infty^{\mathrm{CCSDT}}$, is obtained as the difference between

CCSD(T) and CCSDT correlation energies, each of which is individually extrapolated from the results of frozen-core cc-pVTZ and cc-pVQZ calculations, again using (4.3). The final correction, $\Delta E^{\text{CCSDTQP}}_{\text{cc-pVDZ}}$, is just the difference between now-unextrapolated frozen-core CCSDTQP/cc-pVDZ and CCSDT/cc-pVDZ correlation energies. Once more, the full energy at 0 K also includes correction terms for, e.g., relativistic effects, and once more, we leave these undescribed.

Several variants of the HEAT-456QP method are obtained by modifying either one or both of two particular computational details [Bom+06; Har+08]. The first possible modification is to make use of the aug-cc-pCVTZ, QZ, and 5Z basis sets for calculation of E^{HF}_{∞} and $\Delta E^{\text{CCSD(T)}}_{\infty}$, rather than the aug-cc-pCVQZ, 5Z, and 6Z basis sets as described above. The second is to replace the use of the roughly $\mathcal{O}(K^{12})$-scaling CCSDTQP in $\Delta E^{\text{CCSDTQP}}_{\text{cc-pVDZ}}$ with a cheaper treatment, either CCSDT(Q) or CCSDTQ [Bom+06].[3] The resulting schemes are named accordingly, so, e.g., HEAT-345(Q) denotes the cheapest HEAT method, which uses basis sets up to aug-cc-pV5Z for extrapolation and requires coupled cluster calculations only up to CCSDT(Q).

As tested by the original authors, the HEAT methods produce results that reliably agree with high-quality experimental values to within 1 kJ mol^{-1} [Taj+04; Bom+06; Har+08]. We are unaware of any detailed critical benchmarking that has been performed in order to explicitly compare the results of the various Wn methods and the HEAT methods; see, however, an indirect comparison in [Kar16].

4.3. A generalised composite method

There is a certain similarity between the general formulations of the composite methods outlined above and that of the standard combination technique. This similarity has been previously noted at by Chinnamsetty et al. [CGH18] and particularly Zaspel et al. [Zas+18], whose work underlies the following and to which we shall return shortly.

To make the underlying similarity obvious to the reader with some experience of the combination technique, consider first the expression for the total G4(MP2) energy given in (4.7). By explicitly inserting the expressions (4.5) and (4.6), then eliding the HLC term and reformatting, we obtain

$$\begin{aligned} E^{\text{G4(MP2)}} \approx {} & +E^{\text{CCSD(T)}}_{\text{6-31G*}} \\ & -E^{\text{MP2}}_{\text{6-31G*}} \quad +E^{\text{MP2}}_{\text{G3(MP2)LargeXP}} \\ & \qquad -E^{\text{HF}}_{\text{G3(MP2)LargeXP}} +E^{\text{HF}}_{\infty} . \end{aligned} \tag{4.13}$$

The involved values are arrayed roughly according to an implicit underlying Pople-diagram grid; cf. here [Jen17, Fig. 5.4] for an explicit plot of the component values required by

[3] The use of CCSDTQ(P) appears to have not been explicitly considered in any HEAT-related work.

the full G4 method. A similar rewriting of the ccCA-PS3 total energy equation (4.8) provides

$$\begin{aligned} E^{\text{ccCA-PS3}} = &+E^{\text{CCSD(T)}}_{\text{cc-VTZ}} \\ &\quad +E^{\text{MP2,FC1}}_{\text{aug-cc-pCVTZ}} \\ &-E^{\text{MP2,FC}}_{\text{cc-pVTZ}} \; -E^{\text{MP2,FC}}_{\text{aug-cc-pVTZ}} \; +E^{\text{MP2,FC}}_{\infty}. \end{aligned} \tag{4.14}$$

The W4 energy equation (4.11) is significantly more complicated, but can also be rewritten as a weighted sum:

$$\begin{aligned} E^{\text{W4}} = & \\ &+E^{\text{CCSDTQP}}_{\text{DZ}} \\ &-E^{\text{CCSDTQ}}_{\text{DZ}} \; +1.1E^{\text{CCSDTQ}}_{\text{cc-pVDZ}} \\ &\quad -1.1E^{\text{CCSDT(Q)}}_{\text{cc-pVDZ}} \; +1.1E^{\text{CCSDT(Q)}}_{\text{cc-pVTZ}} \\ &\quad +(\cdots)E^{\text{CCSDT}}_{\text{cc-pVDZ}} \; +(\cdots)E^{\text{CCSDT}}_{\text{cc-pVTZ}} \\ &\quad +(\cdots)E^{\text{CCSD(T)}}_{\text{cc-pVDZ}} +(\cdots)E^{\text{CCSD(T)}}_{\text{cc-pVTZ}} +(\cdots)E^{\text{CCSD(T)}}_{\text{W4-pVQZ}} +(\cdots)E^{\text{CCSD(T)}}_{\text{W4-pV5Z}} \\ &\qquad\qquad +(\cdots)E^{\text{CCSD}}_{\text{W4-pV5Z}} +(\cdots)E^{\text{CCSD}}_{\text{W4-pV6Z}} \\ &\qquad\qquad +(\cdots)E^{\text{HF}}_{\text{W4-pV5Z}} +(\cdots)E^{\text{HF}}_{\text{W4-pV6Z}} \\ &+\Delta E^{\text{CV}}, \end{aligned} \tag{4.15}$$

where we use the notation $(\cdots)$ to indicate scalar factors that are fixed and in principle calculable. These factors arise from the use of the terms which they precede in one or more extrapolation formulae. For an explicit Pople-style plot of the W4 components, cf. again [Jen17], this time Figure 5.5. As the extrapolations used in the various HEAT methods do not produce expressions which are linear in the values extrapolated from, the presentation of an equivalently-formatted weighted sum is not possible; however, it is not hard to persuade oneself via inspection of (4.12) that the overall component structure of a HEAT total energy is fundamentally similar. Consistent with [Zas+18], the key point to take away from the above is, firstly, that all of these composite methods involve sets of values that can be organised in a way corresponding very loosely to index sets used in applications of the standard combination technique, and secondly, that the sign patterns of the involved sums are reminiscent of those encountered in, for example in the case of G4(MP2), the combination sum (3.6) taken over the two-dimensional index set I_L for $L = 2$.

The combination sum-like structure of such composite methods, and particularly that of the G2 method, led Zaspel et al. [Zas+18] to construct their multilevel *CQML* scheme for the machine-learning of molecular energetic properties. Basing their work on the standard formulation of the combination technique, they consider what are in our

terminology combination sums S_{I_L} over poset grids $\mathbb{N}^d$ with both $d = 2$ and $d = 3$. Both grids include an axis counting an increasing number of training samples. The second axis of the two-dimensional grid indexes a combined level of theory, capturing both an *ab initio* method and a pre-chosen, fixed basis set: HF/STO-3G, then MP2/6-31G, then CCSD(T)/cc-pVDZ; see the modified Pople diagram in [Zas+18, Fig. 1]. The three-dimensional grid splits this axis into two, one each for *ab initio* theory (HF, MP2, CCSD(T)) and for basis-set theory (STO-3G, 6-31G, cc-pVDZ). Each point on each grid corresponds to a machine-learning setup, trained on a dataset of appropriately-calculated atomisation energies; each model function represents an evaluated prediction by that trained setup.

Following very closely the example of the CQML method, we will investigate an adapted and more conventional application of the combination technique here. Specifically, without applying machine-learning techniques but considering instead extended and more regular hierarchies of both *an initio* and basis set theory, we investigate whether the combination technique can be used to calculate approximations to FCI/CBS energetic properties of a molecular system at a similar ratio of accuracy to cost as is provided by the conventional composite methods which we have discussed above, but in a more systematically-refinable way. Although the use of such larger axes in the CQML setting is explicitly suggested in [Zas+18] as an avenue of future investigation, this idea has not yet to our knowledge been thoroughly explored. For example, a very recent application of the CQML method is still limited in particular to only basis sets of up to triple-zeta quality [Vin+23]. We are also particularly interested in investigating to what extent the application of adaptivity might be helpful, and we are laying the foundations here for work in Chapter 7 to come.

We will refer to the following construction just as the *generalised composite method* (GCM). In this application of the order-theoretic combination technique, the target function is taken to be the (true) Born-Oppenheimer ground-state potential energy function, V^{BO}, as defined in Section 2.1. Each model function is an approximation $V^{\mathrm{BO}}_{(m,n,p,q)} \approx V^{\mathrm{BO}}$, where the four indices m, n, p, q specify levels of theory used in the approximation, to be outlined below. Each index is drawn from a contiguous subset of $\mathbb{N}$, all of which are either truly or practically finite, so the model functions form a family $\mathcal{F}_\Pi$ according to a poset grid Π composed of four finite chain posets. This setup thus exercises little of the poset-grid machinery introduced in Chapter 3.

The first poset axis, corresponding to the indices m, specifies a generally post-HF *ab initio* method as per (2.26). The lowest-indexed $m = 0$ indicates a "treatment", or rather a neglection of electron correlation as given by the Hartree-Fock method. An initial approximation to the correlation energy is then provided by the total MP2 energy ($m = 1$), followed by coupled cluster methods using cluster operators in (2.24) truncated at increasing orders. The truncated coupled cluster method for each excitation order is considered first in its standard form, and then with the addition of a perturbative correction for next-order correlation effects: CCSD ($m = 2$), then CCSD(T) ($m = 3$),

then CCSDT ($m = 4$), CCSDT(Q) ($m = 5$), and so on. The highest available value of m is a completely untruncated coupled cluster treatment, which is effectively equivalent to an FCI calculation.

The second index, n, allows a choice of discretising basis set. Here, we consider only correlation-consistent sets drawn from the (aug)-cc-p(C)VnZ families; a similar hierarchy was used in the ML-BOSSANOVA construction of [CGH18]. The first available index, $n = 2$, specifies the use of a double-zeta basis set, such as cc-pVDZ; $n = 3$, a triple-zeta basis set, e.g., cc-pVTZ, and so on. At least in principle, basis sets from this family could be constructed for arbitrary values of $n \geq 2$ by following the protocol outlined in [Dun89], so this poset axis is notionally unbounded. In practice, however, we have access to constructed basis sets only in the range $2 \leq n \leq 8$; see Section A.2 for details of the basis sets used, as well as appropriate citations.

The third and fourth indices, $p \in \{0, 1\}$ and $q \in \{0, 1\}$, are binary "fine-tuning" parameters. The index p controls the introduction of extra diffuse functions into the basis set. If $p = 0$, no diffuse functions are used, and the basis set is cc-p(C)VnZ; if $p = 1$, then aug-cc-p(C)VnZ. When $q = 0$, the frozen-core approximation is applied; for $q = 1$, an all-electron calculation is performed. The choice of q also has an impact on the precise basis set used: for frozen-core approximations, we use (aug)-cc-pVnZ, and for all-electron calculations, we use (aug)-cc-pCVnZ. Technically, the frozen core approximation has meaning only in the context of correlated calculations, i.e., for $m \geq 1$. However, for Hartree-Fock calculations ($m = 0$), there is still a difference in quality of results between $q = 0$ and $q = 1$, as the use of the core-valence basis sets increases the number of basis functions relative to, e.g., valence-only cc-pVnZ, and therefore also the variational scope of the solution space.

For the calculation of total energies, the evaluation functional $\mathcal{L}$ is simply point evaluation for a particular fixed nuclear configuration $\{X_A = (R_A, Z_A)\}_{A=1}^{M}$, following [CGH18]:

$$\mathcal{L}[V : (\mathbb{R}^3 \times \mathbb{N})^M \to \mathbb{R}] = V(X_1, \ldots, X_M). \tag{4.16}$$

We shall discuss the extension of $\mathcal{L}$ to the evaluation of total atomisation energies below. Although we shall not explicitly do so here, we note that we could also choose $\mathcal{L}$ to be point evaluation of the nuclear gradient ∇V with respect to $\{R_A\}_{A=1}^{M}$.

We mention here, before we begin, a recent review and comparitive discussion of composite methods by Karton [Kar22], which suggests a classification of various composite methods along "rungs" [Kar22, p. 15] of a notional *Jacob's ladder*, here echoing a name which has previously been used to taxonomise approximate DFT functionals [Goe+17]. We became aware of [Kar22] only late in the production of this thesis, and will not engage with it in detail. We note with interest, however, that Karton's classification (which considers many more composite methods than those we have mentioned here) also hinges upon the introduction of what we would call a one-dimensional hierarchical decomposition of the total energy in terms of a hierarchy of computational methods [Kar22, Tab. 2]

that is basically identical to that which we use here. This is also related directly to Pople-style diagrams [Kar22, Figs. 1 and 4]. Indeed, the generalised formulae [Kar22, (5), (6), and (7)] used to express the different rungs of the ladder can also be rewritten, with substitution as per [Kar22, Tabs. 3 and 4], in roughly triangular forms similar to those given above. This provides yet another motivation for the investigation of the GCM.

4.4. Case study: water (H_2O)

For an initial investigation of the generalised composite method, we consider the water monomer, H_2O, which has historically provided a simple quantum-chemical "test-bed" [Fel92, p. 6104]; a visualisation is given in Figure 4.4 below. We will benchmark the performance of the GCM as used for the calculation of both the total energy and total atomisation energies of water. We begin with the total energy.

4.4.1. Total energy

We draw a reference estimate for the true non-relativistic Born-Oppenheimer total energy of water as $E_\infty^{\mathrm{FCI}} = -76.4390(4)\,E_\mathrm{h}$ from a study by Bytautas and Ruedenberg [BR06]. This value was obtained by repeated application of their CEEIS extrapolation scheme (see Section 4.1.2), and then further extrapolation of those results using basically (4.1) and (4.3), plus an additional core-valence correction; for full details, see [BR06] and references within.

We performed a base set of single-point total energy calculations on H_2O at the reference geometry given in [Hel+97], which was that used in [BR06]. We then used these to derive total energy estimates according to a variety of extrapolation schemes and standard composite methods. Let us note that our focus here will be on the performance of the GCM, and these standard values, as well as similar which will be discussed throughout the remainder of this work, are used only to provide necessary context. We do not present their calculation as novel work in and of itself; in particular, some very similar calculations for a different geometry of H_2O are discussed in the context of the HEAT methods in [Har+08; Tho+21].

Specifically, we performed RHF calculations [SO89] using the cc-pVnZ, cc-pCVnZ, aug-cc-pVnZ, and aug-cc-pCVnZ basis sets for all $2 \leq n \leq 8$; full citations for these and all other used basis sets are given in Section A.2. We also attempted to calculate correlation energies using the MP2, CCSD, CCSD(T), CCSDT, CCSDT(Q), CCSDTQ, CCSDTQ(P), and CCSDTQP levels of theory [MP34; Číž66; PB82; NB87; Rag+89; SO89; KB92; KS01; Bom+05; KG05; KG08] for the same collection of basis sets. The frozen core approximation was applied for calculations performed using the (aug)-cc-pVnZ basis sets, while those with the (aug)-cc-pCVnZ basis sets were complete all-electron calculations. We write "attempted", since not all of these calculations proved to be computationally feasible in practice. CCSDT and higher calculations were performed

Method	E (E_h)	Abstract cost
G4(MP2), no HLC	−76.339 020	8.475×10^9
G4(MP2), HLC	−76.376 908	8.475×10^9
ccCA-PS3	−76.434 304	1.473×10^{11}
HEAT-345(Q)	−76.439 812	1.439×10^{14}
HEAT-345Q	−76.439 783	1.441×10^{14}
HEAT-345QP	−76.439 798	1.576×10^{14}
HEAT-456(Q)	−76.438 861	2.330×10^{14}
HEAT-456Q	−76.438 832	2.331×10^{14}
HEAT-456QP	−76.438 848	2.467×10^{14}
CCSD(T) extrap. (aug-cc-pCV8Z)	−76.438 192	3.234×10^{15}
CCSDT extrap. (cc-pCV6Z)	−76.438 549	2.279×10^{16}
Reference [BR06]	−76.439 0(4)	–

Table 4.1.: Total energies of H_2O according to a selection of composite methods and extrapolative procedures, along with their abstract costs of calculation. Note that the CCSDT extrapolation (and cost) includes the same Hartree-Fock total energy extrapolation at the aug-cc-pCV8Z level as used for the CCSD(T) extrapolation.

using MRCC [Kál+20; MRCC], and the remainder with PySCF [Sun15; Sun+17; Sun+20]. For full calculation details, see Appendix A, and in particular, Sections A.1, A.2, and A.4.

The set of calculations which completed successfully included all required single-point values necessary to calculate total Born-Oppenheimer energies according to the ccCA-PS3, HEAT-345(Q), HEAT-345Q, HEAT-345QP, HEAT-456(Q), HEAT-456Q, and HEAT-456QP composite methods, as described above. So that we could also calculate the G4(MP2) total Born-Oppenheimer energy, we also performed appropriate RHF, MP2, and CCSD(T) total energy calculations according to the 6-31G*, G3(MP2)LargeXP, and G4(MP2)-specialised basis sets mentioned above using PySCF; see again Sections A.1, A.2, and A.5. Using CBS extrapolation according to either Feller's exponential formula (4.1) (for Hartree-Fock total energies), or the two-point formula (4.3) of Helgaker et al. (for correlation energies), our best available values lead to the estimates $E_\infty^{\mathrm{CCSD(T)}} \approx -76.438\,192\,E_\mathrm{h}$ and $E_\infty^{\mathrm{CCSDT}} \approx -76.438\,549\,E_\mathrm{h}$. These follow from an extrapolation for E^{HF} using basis sets up to aug-cc-pCV8Z, and extrapolations for $\Delta E^{\mathrm{CCSD(T)}}$ and $\Delta E^{\mathrm{CCSDT}}$ using up to aug-cc-pCV8Z and cc-pCV6Z respectively. The latter value is very nearly within the tolerance of the reference estimate of [BR06]. We are unable to give a confident estimate of the total CCSDT(Q)/CBS level by direct extrapolation, as we were unable to calculate any all-electron data points at the 6Z level.

These CBS-extrapolated values are contrasted in Table 4.1 with the total energies of H_2O as calculated according to the G4(MP2), ccCA-PS3, and various HEAT-family

composite method formulae, as given earlier in the chapter. For computational details, see Section A.5. The total cost of each calculation according to the abstract cost model outlined in Chapter 2 is also provided; these values are simply the sums of the abstract costs for all required component calculations. It is important to clarify here that these total energies are not entirely faithful to the composite methods in their full definitions. In particular, all of these composite methods technically assume calculations at equilibrium geometries as per different levels of theory. This is particularly problematic for G4(MP2), since the HLC is fitted in terms of just such equilibrium-geometry calculations [CRR07b]. For comparison, we provide values for the G4(MP2) total energy of H_2O both with and without the inclusion of the ΔE_{HLC} term in (4.7); see and cf., e.g., [CRR07a; DCR21]. In this context, it is also slightly interesting to see that our HEAT-456QP energy is exactly that given in [Har+08, Tab. IV], up to the precision of that source, despite the presumably different geometry used.

Table 4.1 does not contain a total energy calculated using any Wn-family composite method. In particular, a W4-level calculation would have been desirable for comparison relative to the HEAT-family calculations. This omission is due to the fact that none of the quantum chemistry codes to which we had access could split the CCSD energy into the components required for the W4 extrapolation of $\Delta E^{\mathrm{CCSD(T)}}$; see [Kar+06; KSM17]. We considered replacing this extrapolation with one of the variants described in Section 4.1.1; however, we judged that the resulting approximation could be misleading, especially since "true" W4 calculations (as calculated by other authors) will be used later in this chapter.

The G4(MP2) total energies, both with and without the HLC, match the reference value to only approximately $0.1\,E_{\mathrm{h}}$. In the HLC-free case, we remember that (4.7) otherwise considers only calculations as per the frozen-core approximation. An estimate for the concomitant intrinsic error is collated by Bytautas and Ruedenburg as $-0.0628\,E_{\mathrm{h}}$; see [BR06, Sec. IV.D] and references therein. The inaccuracy of the full HLC-inclusive expression (4.7) is more interesting. An experimentally-derived relativistic total energy of H_2O (specifically $-76.352\,56\,E_{\mathrm{h}}$) is, in fact, included in the G2/97 dataset [Cur+97], an expanded variant of which [CRR05] was used to fit the parameters of the G4(MP2) HLC term [CRR07b]. It may or may not be meaningful that the calculated G4(MP2) total energy for our reference geometry is closer to this value (with a difference of approximately $0.02\,E_{\mathrm{h}}$) than to our reference (difference approximately $0.06\,E_{\mathrm{h}}$). However, a post-inclusion of the FC error term from [BR06] narrows this latter difference down to only approximately $0.0007\,E_{\mathrm{h}} \approx 0.4\,\mathrm{kcal\,mol^{-1}}$.

By contrast, the ccCA-PS3 energy agrees with the FCI/CBS reference to within approximately $0.005\,E_{\mathrm{h}} \approx 2.9\,\mathrm{kcal\,mol^{-1}}$. The various HEAT methods, although significantly more expensive than the G4(MP2) and ccCA-PS3 methods, are also much more accurate. The 345-family HEAT total energies all match the FCI/CBS reference to roughly $0.0008\,E_{\mathrm{h}} \approx 0.5\,\mathrm{kcal\,mol^{-1}}$, and the 456-family HEAT total energies are no more than approximately $0.0002\,E_{\mathrm{h}} \approx 0.1\,\mathrm{kcal\,mol^{-1}}$ from the reference; let us note that [Har+08] also compares HEAT results for H_2O with the same reference and comments on their

accuracy. The latter value, which is well within the $1\,\mathrm{kJ\,mol^{-1}}$ threshold of calibration accuracy, must, however, be considered in light of the relatively loose error bounds on the reference value, which are themselves on the same order of magnitude. The HEAT methods clearly outperform the conventionally extrapolated CCSD(T) and CCSDT results, which are one and two orders of magnitude more expensive again than the most expensive HEAT method respectively.

We consider now approximations of the total energy of H_2O according to the generalised composite method, obtained via the adaptive algorithm for index set selection described in Chapter 3. Since this algorithm is iterative, its execution produces a sequence of progressively-refined combination sums according to the chosen adaptive strategy; we investigated the relationship between abstract cost and accuracy of the combination sums at each iteration.

As well as the complete four-dimensional family of model functions $\mathcal{F}_\Pi$ defined in Section 4.3 above, we also considered the subfamily of $\mathcal{F}_\Pi$ consisting of those model functions indexed with $p = 0$ and $q = 1$. This corresponds to a two-dimensional "subgrid" of calculations, all using the cc-pCVnZ basis sets and without applying the frozen-core approximation; let us stress that this is even closer to the idea of the CQML method of [Zas+18]. The intention here was to investigate the impact of the fine-tuning parameters on the quality of the results.

For both grids (complete 4D and restricted 2D cc-pCVnZ-only), we report executions of the adaptive algorithm using both the All and Best adaptive strategies. The former produces a sequence of index sets and accompanying combination sums that is equivalent to marching out the parameter L in the standard combination-technique combination sums S_{I_L} as defined in (3.4). In each case, we consider as many iterations as possible given the complete set of precalculated total-energy calculations described above.[4] For completeness, we mention that the various reductions involved in the adaptive algorithm were calculated using high-precision floating-point arithmetic [Joh17] rather than standard double-precision arithmetic (see Appendix A), although we have no reason to believe this has any meaningful impact on the results.

The absolute errors relative to the FCI/CBS reference value of each combination sum obtained at each iteration are plotted in Figure 4.1, measured against the total abstract cost of their respective index sets. The absolute errors and costs of the various composite method total energies given in Table 4.1 are also marked on the plot for the purpose of comparison, as are horizontal lines marking the $1\,\mathrm{kcal\,mol^{-1}}$ and $1\,\mathrm{kJ\,mol^{-1}}$ thresholds of chemical accuracy and calibration accuracy respectively. Only one result is plotted for each of the HEAT-345 and HEAT-456 families, as the costs and accuracies of the members of each family are very similar.

We have also plotted the sequence of results obtained by simultaneously increasing

[4] Since neither calculation applies an explicit termination criteria, the All calculations are not adaptive in the strictest sense of the word.

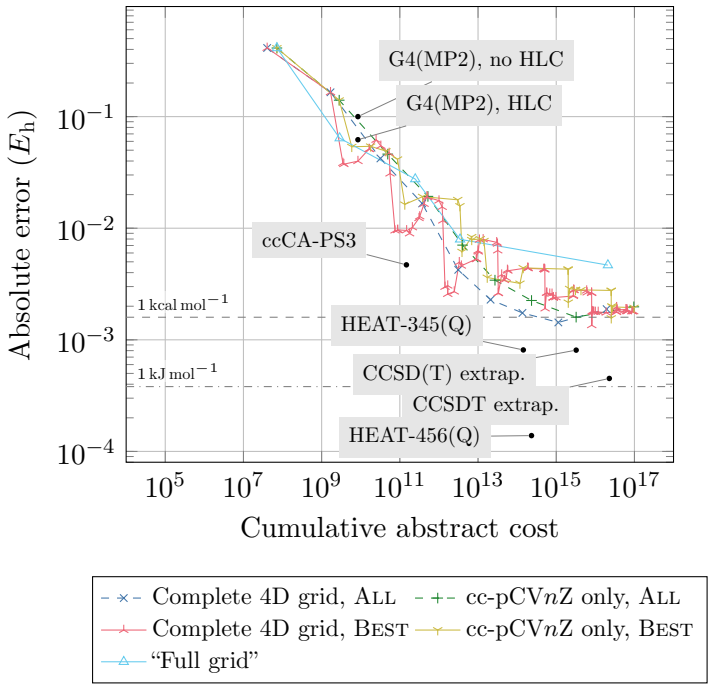

Figure 4.1.: Performance of the generalised composite method (GCM), in terms of absolute error relative to the FCI/CBS reference value of $E = -76.4390\,E_h$ for the total energy of H_2O, as a function of total abstract cost. Each coloured marker indicates a per-iteration result for a particular iterative adaptive calculation, with iterations connected by lines which move here from left to right. The dashed grey horizontal line indicates chemical accuracy ($1\,\mathrm{kcal\,mol^{-1}}$); the dash-dotted grey horizontal line indicates calibration accuracy ($1\,\mathrm{kJ\,mol^{-1}}$). The pinned black points indicate the cost and accuracy of the conventional composite method results given in Table 4.1.

both the basis set quality and the treatment of electron correlation for all-electron calculations, i.e., HF/cc-pCVDZ, MP2/cc-pCVTZ, CCSD(T)/cc-pCVQZ, etc. This is labelled in Figure 4.1 as "Full grid", following standard terminology from the sparse grids and combination technique literature; see, e.g., [GSZ92; BG04].

The total energy results obtained according to the conventional composite methods appear to describe a roughly algebraic trend in accuracy versus cost; although the HEAT-456(Q) result appears substantially "better" in terms of accuracy, we remember that the error bounds on the reference are approximately equivalent to the calibration accuracy threshold.

The per-iteration results from the four GCM calculations also follow a roughly algebraic trend. Index-set refinement according to the Best (i.e., greedy) adaptive strategy produces results which for the most part oscillate around or are less accurate than the All results,

before "settling" to about the same levels of accuracy in their final iterations. The ALL calculations are at best narrowly better than chemical accuracy, although they appear to grow less accurate in their own respective final iterations. We will return to this effect shortly.

There appears to be a slight but noticeable advantage in both cost and accuracy attached to the use of the fine-tuning parameters in the complete four-dimensional GCM grid, particularly for the ALL results. However, this advantage does not change the overall behaviour of the per-iteration results.

The "full grid" results, those obtained by simultaneously increasing the basis set quality and electron correlation treatment, are approximately as good as those obtained via the GCM for the first four iterations. The fourth iteration, however, produces only a slight increase in quality for a significant increase in cost. This is consistent with expectation taken from conventional applications of the combination technique, where full grid solutions may appear to perform equally well or better than combination-technique solutions for small index sets, but then scale much less favourably once a pre-asymptotic regime is exceeded. We observe here that the next such result after the sequence plotted, namely the total energy according to a CCSDT(Q)/cc-pCV7Z calculation, is completely computationally infeasible.

For reasons of legibility, we have not explicitly plotted here the adaptive error indicators produced by the various GCM index sets as described in Section 3.5.5. These follow the same overall trends as do the true absolute errors, although they tend towards an underestimation of the error. This is particularly true in the final iterations, where the the error indicators underestimate the true error by a half to a full order of magnitude.

Although the standard composite methods produce total energies which are almost always more accurate than the GCM index sets at a given cost, the latter do indeed seem to have similar overall accuracy/cost behaviour to the former. In this comparison, we remember that all of the standard composite methods make explicit use of extrapolated energies, which are expected to significantly increase delivered accuracy for a given cost. The particular extrapolations used are the result of careful calibration, while the GCM provides instead a systematically-refinable approximation.

Previously, we motivated the application of the combination technique in the form of the GCM by a rather handwaving analogy between the structure of the total energy expressions of the conventional composite methods and that of the standard combination technique in two dimensions. We can improve somewhat on this justification by an *a posteriori* analysis of the data produced by the GCM calculations.

We recall from Chapter 3, as well as the discussion in e.g. [TW18], that favourable performance of the combination technique may be expected if the involved benefit/cost ratios $\mathcal{L}[\tilde{f}_{\mathbf{m}}]/\mathcal{C}(\mathbf{m})$ decay in a certain bounded and predictable way in the multiindices $\mathbf{m}$. Here, we are unable to place any kind of generally-rigorous bounds on the benefit terms $\mathcal{L}[\tilde{V}^{\mathrm{BO}}_{(m,n,p,q)}]$ in particular. However, we can explicitly evaluate these benefit/cost ratios

for the functions $V^{\mathrm{BO}}_{(m,n,p,q)}$ that are evaluated in the course of our adaptive calculations.

A Pople-style plot of the true benefit/cost values for all points in the final index set over the complete four-dimensional grid for the Best adaptive calculation is given in Figure 4.2. Each of the four subplots displays the benefit/cost ratios for a particular fixed pair of model fine-tuning indices $p \in \{\pm 1\}$, $q \in \{\pm 1\}$. To interpret this figure, consider an execution of the adaptive algorithm performed using the All strategy. The first iteration of such an execution of the adaptive index-set selection algorithm produces an index set containing only the bottom left-hand point in the bottom-left hand subplot, i.e., the point (HF, DZ) in subplot (a), corresponding to the model function $V^{\mathrm{BO}}_{(m=0,n=2,p=0,q=0)}$, representing the Hartree-Fock total energy using the cc-pVDZ basis set, without diffuse functions and with a notional application of the frozen-core approximation.[5] The next refinement adds the points (MP2, DZ) and (HF, TZ) in the same subplot, so MP2/cc-pVDZ and HF/cc-pVTZ FC calculations, and also the points (HF, DZ) in subplots (a) and (d), so HF/cc-pCVDZ and HF/aug-cc-pVDZ calculations. The third refinement adds (HF, DZ) in subplot (b), both (MP2, DZ) and (HF, TZ) in subplots (a) and (d), and (CCSD, DZ), (MP2, TZ), and (HF, QZ) in subplot (a). Further refinements follow the general pattern of a regular "lower triangular" simplex in each subplot, with a boundary in subplot (a) that extends one grid cell further than those in subplots (b) and (c), which are themselves one grid cell further out than that in subplot (d).

To a first glance, the benefit/cost ratios display a fairly even pattern of decay as $\|(m, n, p, q)\|_1$ increases. This suggests one reason why the truly adaptive Best strategy does not perform significantly differently to the All strategy, since all strategies lead to selection of the same indices in more or less the same order. However, some imbalance does begin to become visible towards the frontiers of the plotted index set. Higher-order coupled cluster calculations using cheaper basis sets offer less benefit per cost than do CCSD and CCSD(T) calculations with more expensive basis sets.

The benefit/cost plots of Figure 4.2 also provide a visual hint as to the behaviour of the GCM calculations in their final iterations, as remarked on above. There, the rates of error decay slow, flatten, and then seem to turn upwards. This behaviour is likely due to incompleteness in the available data, as expressed in the finite poset grid axes. Specifically: the poset axis corresponding to the treatment of electron correlation contains nine elements, but there are only seven elements in the axis expressing basis set size (i.e., from DZ up to 8Z). This means that from the eighth iteration onwards, the progressively-refined index sets are no longer being augmented with points distributed equally along a simplex, and, in particular, no information from basis sets with quality higher than 8Z enters the approximations. Unfortunately, suitable basis sets of higher quality than 8Z are not available for further experimentation.

[5] "Notional" since the frozen-core approximation is meaningful only in the context of a correlated, i.e., post-Hartree-Fock calculation.

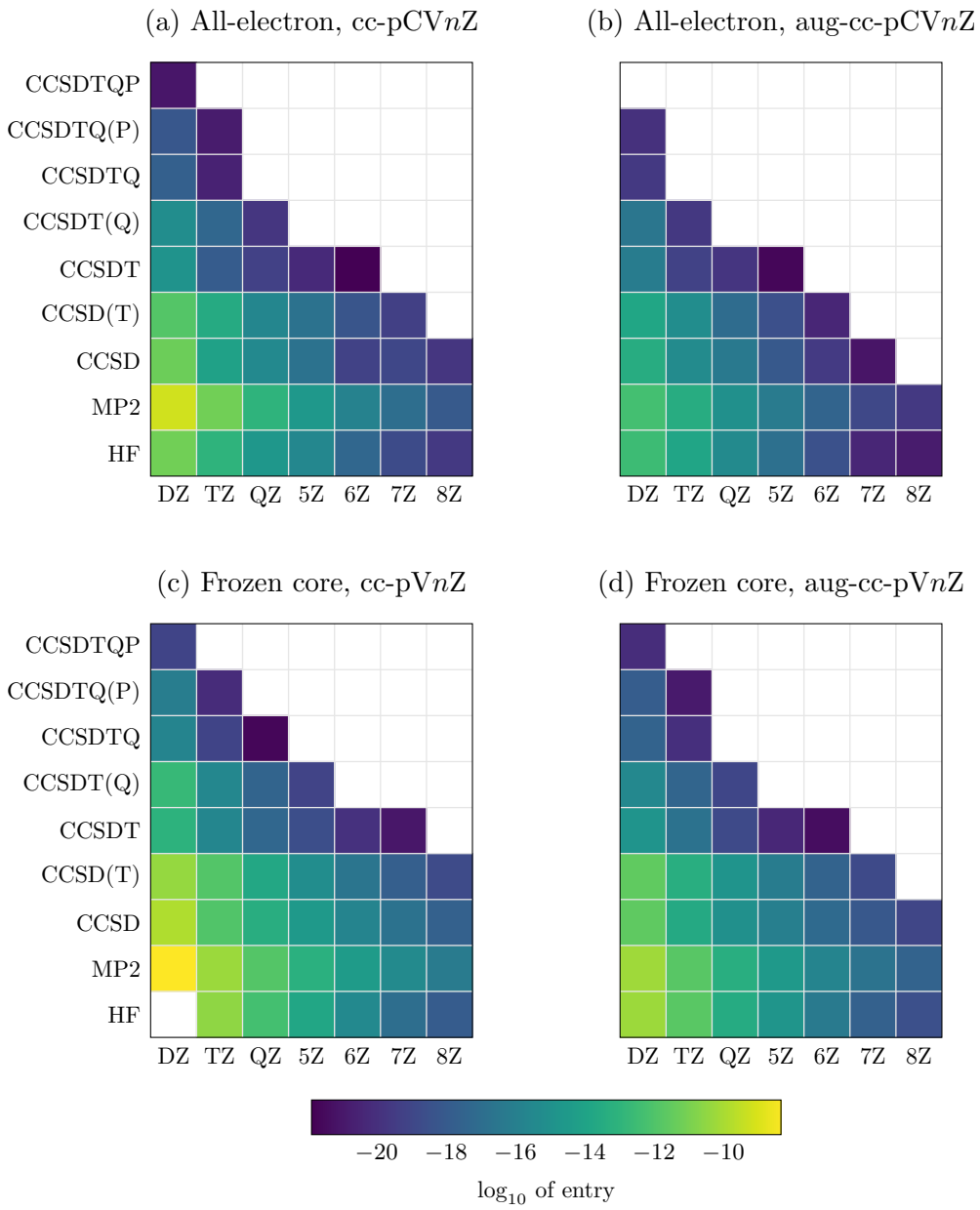

Figure 4.2.: Benefit/cost ratios for all index set elements involved in the GCM estimations of the total energy of H_2O given in Figure 4.1. Each coloured grid square displays the base-10 logarithm of the magnitude of the benefit/cost ratio $\mathcal{L}[\tilde{V}^{\mathrm{BO}}_{(m,n,p,q)}]/\mathcal{C}(m,n,p,q)$ for a particular element with indices m, n, p, q. The value for Hartree-Fock at cc-pVDZ is several orders of magnitude larger than surrounding points (log-magnitude approximately -5.73) and is omitted in order to avoid compressing the colour representation of the remaining data. The four subplots here represent a single set of benefit/cost ratios in terms of the complete four-dimensional GCM grid; refer to the main text for interpretation.

4.4.2. Total atomisation energy

We turn now to the calculation of the total atomisation energy of the H_2O monomer. As discussed in Section 2.3, the total (Born-Oppenheimer, non-relativistic) atomisation energy of a molecular system composed of M nuclei is a derived quantity, defined as (with a slight notational adjustment of (2.27))

$$E_{\text{atom}} := \left(\sum_{i=1}^{M} E^{(i)}\right) - E, \tag{4.17}$$

where E is the total energy of the full system, and $E^{(i)}$ is the total energy of the ith-indexed atom treated as a standalone system in its own right.

Our consideration of the atomisation energy of H_2O serves two purposes. Firstly, it is interesting to investigate whether our GCM can be usefully applied to the approximation of derived energetic quantities and specifically energy differences, particularly since these are more interesting in practice than are raw total energies [Taj+04], and also since most of the standard composite methods are directly targeted at the same [Kar16]. Secondly, high-quality reference estimates for total energies such as those used in the previous section are uncommon in the literature, while reference estimates for atomisation energies are comparitively easier to come by. We will follow this case study by investigating the application of the GCM to two additional and more complicated molecular systems, for which we shall only have atomisation energies for reference. Thus, the atomisation energy of water provides us with an informal control subject.

A reference estimate of the FCI/CBS total atomisation energy of H_2O as $E_{\text{atom}} \approx 0.3716\,E_{\text{h}}$ can be derived from a value given in the study of Bytautas and Ruedenberg [BR06, (40)] by discounting a term modeling relativistic effects; see also [BR06, Tab. V]. Unlike their FCI/CBS estimate for the total energy of water used in the previous section, this value comes without an explicit quantification of error. We note a newer empirical reference value given in [Tho+21, Tab. II] as $E_{\text{atom}} = 974.91(6)\,\text{kJ}\,\text{mol}^{-1} \approx 0.371\,32(2)\,E_{\text{h}}$; see [Tho+21] for details. For consistency with the previous section, however, we take the value derived from [BR06] as our reference here. In any case, the two values are still within $1\,\text{kJ}\,\text{mol}^{-1}$ of each other.

We derived a collection of atomisation energies for H_2O from the set of total energy calculations discussed in previous section. This required an equivalent collection of monoatomic total energy calculations for both H and O. For more details on these monoatomic calculations, refer to Section A.3 in Appendix A. In particular, to avoid any ambiguity regarding the order of operations, we note that we consistently performed CBS extrapolations using HF and correlation energy values, and then combined the resulting values into atomisation energies.

A collection of atomisation energies similarly derived from standard composite method total energies (both monoatomic and for H_2O) is given in Table 4.2, along with their

Method	E_{atom} (E_{h})	Abstract cost
G4(MP2), no HLC	0.360 651	8.475×10^{9}
G4(MP2), HLC	0.370 597	8.475×10^{9}
ccCA-PS3	0.373 560	1.473×10^{11}
HEAT-345(Q)	0.371 568	1.439×10^{14}
HEAT-345Q	0.371 528	1.441×10^{14}
HEAT-345QP	0.371 541	1.576×10^{14}
HEAT-456(Q)	0.371 427	2.330×10^{14}
HEAT-456Q	0.371 388	2.331×10^{14}
HEAT-456QP	0.371 400	2.467×10^{14}
W4.2	0.371 277 [a]	3.636×10^{14}
CCSD(T) extrap. (aug-cc-pCV8Z)	0.371 395	3.141×10^{15}
CCSDT extrap. (cc-pCV6Z)	0.371 192	2.011×10^{14}
Reference, derived from [BR06]	0.371 6	–

[a] At the W4-17 dataset geometry [KSM17; W4-17].

Table 4.2.: Atomisation energies (E_{h}) of H_2O calculated according to a selection of composite methods and extrapolative procedures, and their abstract costs of calculation. A reference energy is provided for comparison.

abstract costs, equivalently to those in Table 4.1. As the monoatomic total energies involved in each atomisation energy calculation do not depend on the structure of the full system, they may be precalculated and reused repeatedly. For this reason, we do not explicitly include the abstract costs of the monoatomic calculations when reckoning the abstract cost of an atomisation energy calculation, either here or elsewhere.

In addition to the G4(MP2), ccCA-PS3, and HEAT-family results, Table 4.2 also contains an atomisation energy calculated according to the W4.2 composite method. This value is taken from the W4-17 dataset of Karton et al. [KSM17; W4-17]. We will discuss this dataset in more detail in Section 4.5 below, but for now, we note that the provided value was obtained using a complete W4.2 calculation, including a CCSD(T)/cc-pV(Q+d)Z geometry optimisation as per [Kar+06; KSM17]. The geometry so obtained is close but not identical to the reference geometry of H_2O of [Hel+97]: the H-O bond lengths and H-O-H bond angles of each are approximately 0.958 Å and 104.120° (derived from geometry of [KSM17; W4-17]) versus 0.957 Å and 104.520° ([Hel+97]), placing the hydrogen atoms in each geometry 1.511 Å and 1.514 Å apart respectively. The total and atomisation energies of the two geometries may therefore be expected to be close to one another, but we shall not attempt to quantify precisely how close. As such, any direct comparison of the W4.2 value against the remaining values should be made with caution, and the value is included only for the sake of interest and completeness.

Compared to the total energy calculations above, the accuracy of the G4(MP2) composite method is now much more competitive. This is particularly so when the HLC is included as intended; the resulting G4(MP2) atomisation energy agrees with the chosen reference value to within approximately $0.6\,\mathrm{kcal\,mol^{-1}}$. Again, we stress that this comparison is not a completely fair test of the G4(MP2) approach, as the values here still disregard certain components of the experimental energies against which the HLC was fitted; even so, this accuracy is better than the $1.2\,\mathrm{kcal\,mol^{-1}}$ provided by the significantly more expensive ccCA-PS3 method. Note that the comments made in the previous section regarding the applicability of the HLC to this particular geometry also still hold.

The remaining, higher-quality composite methods all perform extremely well, delivering results within $1\,\mathrm{kJ\,mol^{-1}}$ of the reference. This is also true of the W4.2 value, even considering the associated caveats regarding geometry. The composite-method result with closest agreement to the reference value is provided by the relatively cheapest HEAT-345(Q) method.[6] It has been noted in the literature that the HEAT-345(Q) method operates somehow "better than it ought to" [Har+08, p. 10], but also that all of the HEAT-345 methods are prone to overestimating experimental atomisation energies [Bom+06]. Both observations may be relevant here. However, the reference result is probably insufficiently precise to support any deeper assessment of the relative quality of the HEAT methods.

The extension of the GCM to the calculation of atomisation energies is straightforward. For a fixed nuclear conformation, and a fixed set of monoatomic energies $E^{(i)}$, the evaluation of the total atomisation energy can be written as a suitable linear functional, itself defined in terms of that in (4.16), i.e.:

$$\mathcal{L}_{\mathrm{atom}}[V] = \left(\sum_{i=1}^{M} E^{(i)}\right) - \mathcal{L}[V]. \tag{4.18}$$

In order to maximise any available cancellation of errors, however, the monoatomic energies should be calculated at the same level of theory as that used to approximate V^{BO}. Achieving this requires only a slight tweak to the definition of (4.18).[7]

A plot of per-iteration results for All and Best adaptive refinements over both the complete four-dimensional grid Π and the cc-pCVnZ restriction is given in Figure 4.3, equivalently as for the total energy case above. The expected error-cancellation property that leads to an improvement in absolute accuracies of atomisation energies in comparison

[6] We observe in passing that our HEAT-345(Q) value matches the equivalent value for the atomisation energy of H_2O given in [Tho+21, Tab. V], up to the precision of that source. This is interesting, since our extrapolated values do not seem to agree to the same level of precision with the data in [Tho+21, Tab. IV]; we suspect that this relates to the use of a slightly different geometry of H_2O in [Tho+21].

[7] We could also construct a combination-sum approximation of each monoatomic energy according to the same index set as that used to approximate the full-system energy, and then combine these in turn according to (4.17).

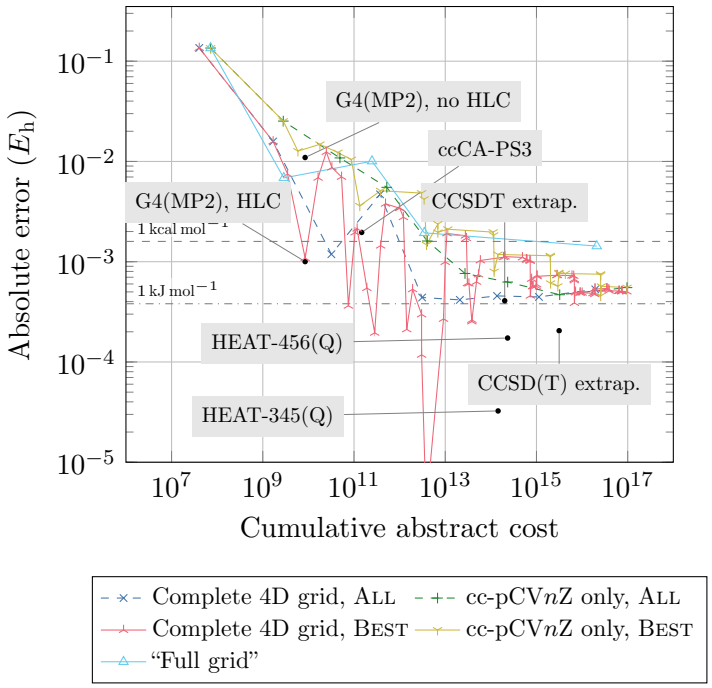

Figure 4.3.: Performance of the generalised composite method, in terms of absolute error relative to the reference value of $E_{atom} = 0.3716\,E_h$ for the atomisation energy of H_2O, as a function of abstract cost. Plot format is as for Figure 4.1.

to that of total energies is visible: now all four sets of results easily pass below chemical accuracy. The greedy-style Best strategy again does not seem to produce an improvement on the conventional All strategy; for the complete four-dimensional grid in particular, although many of the Best results are highly accurate (at least up to the precision of the reference value), they still oscillate around the All results and converge in the end towards a similar value.

Using the All strategy, both the complete 4D grid and the reduced cc-pCVnZ grid show similar behaviour relative to each other as previously seen for total energies. The complete grid appears to produce higher accuracy at lower cost, with a faster rate of increase in accuracy, but refined index sets over both grids also both plateau to a result that is slightly more than 1 kJ mol^{-1} away from the reference value. This plateau effect is again most likely due to some combination of the limitations of the available basis sets, and the inherently limited precision of the reference value.

As for the total energy case, the "full grid" sequence of solutions obtained by simultaneously increasing both the basis set quality and the treatment of electron correlation perform on par with the GCM results for the first four iterations, before failing to produce a significant error increase in the fifth iteration. This is consistent with both the total

energy results and with expectations from applications of the standard combination technique.

The All calculation over the complete grid shows a particularly accurate result at the third iteration, with a cost slightly more than 10^{10}, before returning in the fourth iteration to a less accurate approximation. The changes in these points appears to correspond to a sign flip in the true error, and we conclude that the high accuracy of the third iteration result is likely just numerical accident.

Overall, and allowing for the fact that all points plotted with accuracy less than approximately $1\,\mathrm{kJ\,mol^{-1}}$ cannot reliably be distinguished as being truly more or less accurate than each other, we conclude that again, the GCM produces systematically-improvable series of results that broadly match the cost/accuracy trend of the various standard composite methods.

4.5. Case study: ozone (O_3) and β-lactim (C_3H_5NO), total atomisation energies

Although the GCM appears to function reasonably well in the simple case of the water monomer, it is prudent to also consider its performance when faced with less straightforward high-accuracy calculations. We conclude the chapter by investigating the application of the GCM to two further molecular systems: the ozone monomer O_3, and β-lactim (C_3H_5NO).

Approximations of the atomisation energies of both ozone and β-lactim are included in the W4-17 dataset of Karton et al. [KSM17; W4-17]. This dataset contains atomisation energies for 200 small molecular systems, all of which have been calculated using variants of the W4 family of composite methods. The calculations contained in the dataset are explicitly intended to approximate physically-measurable atomisation energies, and so are corrected for, e.g., relativistic effects and the Born-Oppenheimer approximation; however, uncorrected values are also included in the dataset, matching our restricted problem formation. The values in the W4-17 dataset are expected by the authors to be physically accurate to within a full-dataset 3σ confidence interval of at least $1\,\mathrm{kJ\,mol^{-1}}$, and are intended as high-quality benchmark references; indeed, these values are used in [KSM17] to assess the quality of some of the composite methods we consider here. It should be noted that we understand the confidence interval stated in [KSM17] to apply to the atomisation energies in the W4-17 dataset intended for experimental comparison, and it is not clear to our reading whether the nonrelativistic clamped-nuclei values are as reliable; nevertheless, we will operate here on the general assumption that this is so.

Visualisations of both ozone and β-lactim are shown in Figure 4.4, using the respective geometries as obtained from [W4-17]. Although both molecules are still small by the usual standards of quantum chemistry, each still presents a significantly more difficult computational problem for high-accuracy calculation than does the water monomer.

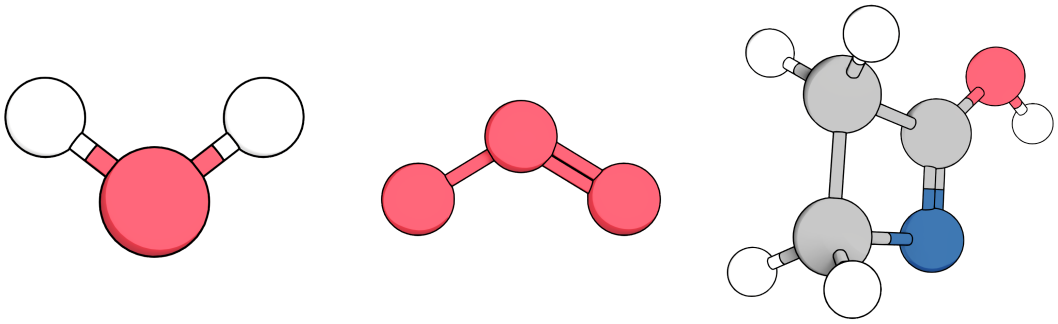

Figure 4.4.: Ball-and-stick visualisations of H_2O (left), ozone monomer (O_3, middle), and β-lactim (C_3H_5NO, right). Geometries are drawn from [Hel+97] (H_2O) and [KSM17; W4-17] (ozone, β-lactim), as described in the main text.

The conformation of ozone is superficially similar to that of H_2O. Ozone involves more than twice as many electrons as does water — 10 for H_2O, 24 for O_3 — and, more importantly, has also been called a "pathological system" [Fog+12, p. 2]. Ozone is a *multireference system* [KSM17], in the sense that it can be problematic to treat via *ab initio* calculations which involve only a single reference Slater determinant; see e.g. [HG11].[8] Karton et al. [KSM17] expect that such systems may not be well approximated at the CCSD(T) level of theory; here, they implicitly include also composite methods that involve at most CCSD(T) calculations, such as G4(MP2) and ccCA-PS3.

By contrast, Karton et al. [KSM17] classify β-lactim as a non-multireference system, suggesting that the incorporation of CCSDT calculations or higher is comparatively less likely to be necessary to achieve chemical or calibration accuracy than would be so in the case of ozone. Instead, it is simply the size of β-lactim that makes it here a difficult problem. β-lactim is one of the larger molecules included in the W4-17 dataset, consisting of 38 electrons orbiting ten nuclei. As a result, the W4-17 reference value is calculated using only the W4Lite scheme, as described at the end of Section 4.2.3 above.

We performed equivalent sets of single-point and extrapolated total energy calculations on both ozone and β-lactim as those described for water in the previous section, subject again to the limits of practical computational feasibility. Only a very limited subset of these calculations proved practically possible for β-lactim. For example, a CCSDT(Q) calculation was only achievable using the cheapest basis set considered (cc-pVDZ), even under the frozen-core approximation, and no all-electron correlation energy treatment of any kind proved feasible under any 7Z or 8Z basis set.

A summary of the atomisation energies of both ozone and β-lactim as predicted by various standard composite methods, and their attendant costs, is given in Table 4.3.

[8] Alternatives include, e.g., MCSCF methods [Jen17], which we do not consider here.

Method	Ozone (O_3)		β-lactim (C_3H_5NO)	
	E_{atom} (E_h)	Abstract cost	E_{atom} (E_h)	Abstract cost
G4(MP2), no HLC	0.219 436	1.309×10^{11}	1.542 215	2.285×10^{12}
G4(MP2), HLC	0.233 548	1.309×10^{11}	1.582 537	2.285×10^{12}
ccCA-PS3	0.232 680	1.072×10^{12}	1.585 637	9.002×10^{13}
HEAT-345(Q)	0.234 654	7.500×10^{15}	–	3.700×10^{18}
HEAT-345Q	0.233 187	7.571×10^{15}	–	3.811×10^{18}
HEAT-345QP	0.233 852	2.687×10^{16}	–	1.218×10^{20}
HEAT-456(Q)	0.234 779	9.708×10^{15}	–	3.932×10^{18}
HEAT-456Q	0.233 312	9.779×10^{15}	–	4.043×10^{18}
HEAT-456QP	0.233 977	2.908×10^{16}	–	1.221×10^{20}
W4Lite	–	4.318×10^{14}	1.583 825[b]	1.630×10^{17}
W4.2	0.234 945[a]	1.197×10^{17}	–	2.278×10^{19}
Best extrap.	0.230 327[c]	1.416×10^{16}	1.585 033[d]	5.843×10^{17}

[a] Reference for O_3 [KSM17; W4-17].

[b] Reference for C_3H_5NO [KSM17; W4-17].

[c] Hartree-Fock extrap. to aug-cc-pCV8Z, CCSD(T) to aug-cc-pCV7Z.

[d] Hartree-Fock extrap. to cc-pCV8Z, CCSD(T) to cc-pCV6Z.

Table 4.3.: Atomisation energies (E_h) of ozone (O_3) and β-lactim (C_3H_5NO), calculated according to a selection of composite methods and extrapolative procedures, and their abstract costs of calculation. Both reference values are drawn from [KSM17; W4-17], and are here given converted into E_h. Missing values indicate unavailability, either in the W4-17 dataset (for Wn methods), or due to practical computational infeasibility (for HEAT methods for β-lactim).

A sample approximation assembled from CBS extrapolations of Hartree-Fock total and CCSD(T) correlation energies is also given for each system; we could not obtain CCSDT results at a sufficiently high basis set level that their extrapolations would merit comparison. The values in the W4-17 dataset (as explicitly obtained from [W4-17]) are provided to three decimal places and given in units of $\text{kcal}\,\text{mol}^{-1}$, so we list the W4 reference values to an equivalent level of precision in E_h. It should still be kept in mind that these values may not necessarily be reliable past the $1\,\text{kJ}\,\text{mol}^{-1} \approx 0.0004\,E_\text{h}$ level, if that.

For ozone, the G4(MP2) method provides an impressive level of accuracy considering its low cost and the anticipated difficulty of the problem: the G4(MP2) atomisation energy is within $1\,\text{kcal}\,\text{mol}^{-1}$ of the reference solution. However, like water, O_3 is also a member of the G2 test set [Cur+97], and thus transitively of the G3/05 test set on which the G4(MP2) HLC is trained [CRR05; CRR07b]. Therefore, this result may again be unrepresentative of the quality of the G4(MP2) method when applied to arbitrary

systems; cf. again [DCR21]. By contrast, the ccCA-PS3 prediction is not even close to chemical accuracy with respect to the W4-17 reference.

The HEAT variant that provides an estimate closest to the reference atomisation energy of ozone is HEAT-456(Q), with an absolute error of approximately $0.1\,\mathrm{kcal\,mol^{-1}} \approx 0.4\,\mathrm{kJ\,mol^{-1}}$. This is followed closely by HEAT-345(Q)'s error of approximately $0.8\,\mathrm{kJ\,mol^{-1}}$. The other HEAT variants are undercompetitive; those two using an iterative treatment of quadruple excitations both fail to come even within $1\,\mathrm{kcal\,mol^{-1}}$ of the reference solution. The energies produced by the two HEAT variants that involve both quadruple and pentuple excitations are still no closer than approximately $0.6\,\mathrm{kcal\,mol^{-1}}$ to the W4.2 reference, despite having access to roughly the same quality of correlation information as that built into the W4.2 method itself. (And, indeed, these methods are roughly an order of magnitude more expensive than W4.2.)

The practical unaffordability of high-accuracy composite calculations over β-lactim is clear: since we were unable to perform the CCSDT/cc-pVQZ calculation required for the construction of $\Delta E_\infty^{\mathrm{CCSDT}}$ in (4.12), we can provide no HEAT-family atomisation energies in this case.[9] Again, the G4(MP2) HLC-inclusive atomisation energy for β-lactim is more accurate than the ccCA-PS3 atomisation energy. As far as we aware, the definition of the G4(MP2) HLC does not involve β-lactim, so this may be a better test of its quality and transferability than either water or ozone.

Again following the same approach as in the previous sections, we also consider the GCM applied to ozone over both the complete 4D Pople grid, as well as the restricted two dimensional cc-pCVnZ hierarchy. As above, we consider adaptive calculations according to both the All and Best strategies, iterated as far as possible given the available calculation data. Per-iteration results are plotted in Figure 4.5, with an equivalent format as used for the results for water given above. Here, however, we have not plotted a sequence of "full grid" results, since one fewer result is available than in the previous case of H_2O and the resulting data are inconclusive.

The eye is drawn first to the second-iteration results of both All- and Best-obtained index sets using the restricted cc-pCVnZ grid, which are roughly as accurate at the G4(MP2) result, but even cheaper. Again, we see no reason to believe that this indicates anything more than numerical accident.

The index sets refined used the All strategy again seem to increase in accuracy more smoothly and also slightly more cheaply than do those according to the Best strategy. The complete four-dimensional grid outperforms the cc-pCVnZ grid, but up to the last possible iteration of each, only in terms of a cost benefit of at most an order of magnitude. At this last iteration, the four-dimensional All index set produces a result that is within calibration accuracy of the W4.2 reference. The final possible cc-pCVnZ All-derived index set is not quite good to chemical accuracy, and is indeed less accurate than the

[9] Given the very high number of involved triple excitations, we consider it unlikely that this calculation would be feasible on any currently-available hardware.

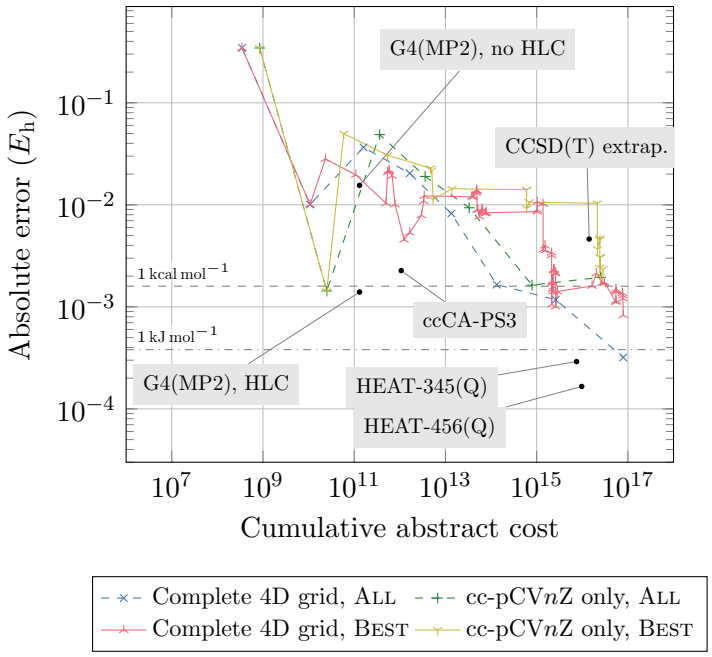

Figure 4.5.: Performance of the generalised composite method, in terms of absolute error relative to the reference value of $E_{\text{atom}} = 0.234\,945\,E_\text{h}$ for the atomisation energy of O_3, as a function of abstract cost. Plot format is as for Figure 4.1.

penultimate such result. Once more, although the low precision of the reference data hampers the drawing of strong conclusions, it seems reasonable to say that the GCM functions here again as a systematically refinable method to approach the true atomisation energy of ozone, although not as efficiently as the tuned and extrapolation-based standard composite methods.

The unaffordability of high-quality β-lactim calculations proved limiting for a similar investigation of the GCM in the case of that molecule. In particular, we were unable to perform frozen-core CCSDT(Q) or higher calculations with any basis set of more than double-zeta quality, frozen-core CCSDT calculations with better than triple-zeta quality, or any all-electron CCSDT or greater calculation at more than double-zeta quality. As a result, GCM All-strategy results were only possible up to the sixth iteration for the complete four-dimensional poset grid, and the fifth iteration for the cc-pCVnZ-only subgrid. Best-strategy results were similarly limited.

The major difficulty in the preparation of results stems from the prohibitive jump in cost from CCSD(T) calculations to CCSDT calculations. We were still able to perform CCSD(T) calculations for β-lactim up to reasonably high-quality basis sets (6Z, 7Z, or for frozen-core, cc-pV8Z). As discussed above, we might expect a CCSD(T) treatment of correlation to be sufficient to provide at least chemical accuracy for a non-multireference

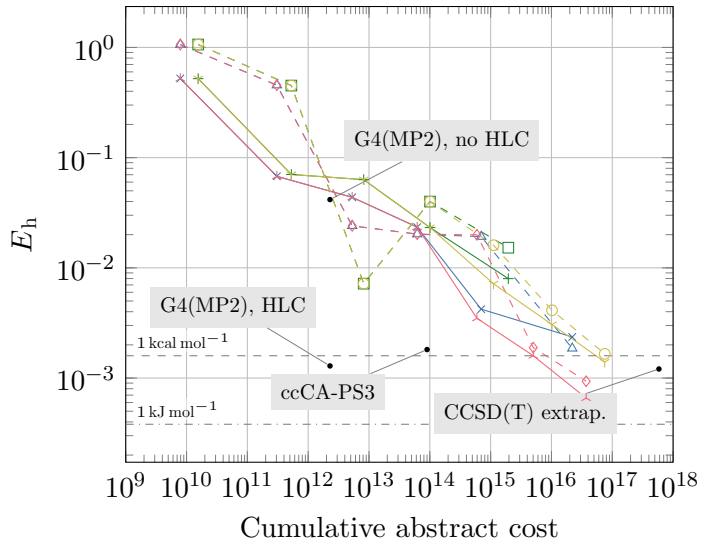

Figure 4.6.: Performance of the generalised composite method, in terms of absolute error relative to the reference value of $E_{\text{atom}} = 1.583\,825\,E_h$ for the atomisation energy of C_3H_5NO, as a function of abstract cost.

system such as β-lactim.

Based on this idea, we also investigated a further restriction of the GCM poset grid. Here, the treatment of correlation was restricted to go no higher than CCSD(T), corresponding to m-index values only in the range $0 \leq m \leq 3$. The restricted poset grid is technically unbounded in only one of its four axes, that corresponding to basis set size. It follows that an index set and corresponding combination sum over this grid is better viewed as producing an approximation, not of the full FCI/CBS atomisation energy, but instead only of the CCSD(T)/CBS atomisation energy; cf., e.g., Karton's classification of ccCA-PS3 and G4(MP2) as CCSD(T)/CBS approximations in [Kar16], and again comments in [Zas+18].

Per-iteration results for both the standard 4D and cc-pVnZ only poset grids are plotted in Figure 4.6, along with results for the CCSD(T)-limited 4D and cc-pCVnZ only grids. Here, we show only results obtained using the All strategy, but have also plotted the per-iteration error indicators for each of the four sets of results. When interpreting this plot, it is helpful to realise that results for the 4D grid both with and without the

restriction on the index m are necessarily identical for the first four iterations, as are the results for the cc-pCVnZ grid both with and without the restriction on m.

The per-iteration error indicators, which are plotted using dashed lines in Figure 4.6, are generally quite close in magnitude to the true absolute errors to which they correspond. The greatest difference between a true error and an error indicator is seen in the results at the third iteration of the cc-pCVnZ-only index sets, where the difference is still less than an order of magnitude.

All sets of per-iteration results show the same roughly-algebraic trend in increasing (true) accuracy and cost as has been consistently observed for the cases previously studied in this chapter, at least as far as we can judge given the limited availability of single-point total energies. This is also true for the results which use no higher than a CCSD(T) treatment of correlation; these two sets of results both obtain chemical accuracy in their final iterations.

Overall, we see once again that the GCM results display the expected and desired patterns of systematic improvement. Although a restriction of the correlation treatment to CCSD(T) technically removes the ability of the GCM to approximate true solutions to the Schrödinger equation, the best-quality resulting approximations are still slightly cheaper than the best-available standard CBS extrapolation at the CCSD(T) level of theory. However, the G4(MP2) and ccCA-PS3 composite methods provide atomisation energies at or close to chemical accuracy, at orders of magnitude less cost.

In summary, the generalised composite method which we have discussed in this chapter appears to offer a well-founded mechanism for the relatively efficient approximation of solutions to the Schrödinger equation, as well as for derived energetic quantities such as atomisation energies. It seems reasonable to think that the understanding of the additivity of error relied on both implicitly and explicitly by the composite methods we have considered here may benefit from deeper analysis from the established viewpoint of the combination technique.

In practice, the accuracy offered by the GCM at a given cost seems at best equivalent to and generally somewhat worse than that offered by the standard composite methods by which it is inspired. This is unsurprising: as should be clear from the above, those methods are mature and well-developed, and benefit in particular from the use of CBS extrapolation. The GCM, however, remains flexible, and the rather rudimentary poset grid on which it is based can be itself further extended. Such an extension will be the work of the remainder of this thesis.

5. Adaptive many-body expansions

Although the composite methods discussed in the previous chapter may well offer the reasonable possibility of a chemically-accurate approximation to the true FCI/CBS total energy of a small molecular system, they all require post-HF estimations of the correlation energy. Therefore, the fact that even Hartree-Fock calculations scale at best cubically in the number of basis functions — and so, in the number of atoms — places larger systems well beyond their practical reach [RS15].

As mentioned in Section 2.4.3 above, the complexity reduction for Hartree-Fock calculations over sufficiently large molecular systems from formally quartic down to practically cubic occurs because the $H^1(\mathbb{R}^3)$-discretising basis functions are spatially localised around the nuclear centres on which they are placed [Hea96; EA07]. Similar and related manifestations of locality form the basis of a number of efforts to construct approximate solution techniques for the electronic problem that scale more sympathetically, and in the best case linearly, in the number of atoms in a molecular system; see, e.g., [Mas+98; Goe99; Gor+11; BM12; BBR13; BCK16].

It was in the context of such *reduced-scaling methods* that Kohn first introduced his well-known principle of the *nearsightedness of electronic matter* [Koh96; PK05], which limits the response of local electronic properties to remote changes in an external potential. Kohn was careful to highlight that his principle is not always applicable, and in particular can only be expected to hold for many-particle systems [Koh96].

For the remainder of this thesis, we consider a class of possibly [Her19] reduced-scaling protocols that rely fundamentally, although often rather fuzzily, on Kohn's principle [MR11; RH13; CGH18]. These approaches build, augment, or approximate full-system electronic properties via one or more calculations which are performed with particular reference to smaller regions of the complete system. Thus, we will refer collectively to these methods as *subsystem techniques*; see similar language used in e.g. [MR11; Her19].

We will be particularly interested in those subsystem techniques which formally decompose the complete set $[M]$ of nuclear indices of a molecular system into a family of *fragment* subsets $F_i \subseteq [M]$, and through that decomposition also the total energy of the system [Gor+11; CB15; RS15; Her19]. Most such *energy-based fragmentation methods* either explicitly are or at some conceptual level can be linked to the well-known *many-body expansion* (MBE); see, e.g., [Fis64; HMS70; MR12; RH12; GHH14; CGH18; LH19]. Although they could be considered fragmentation methods in the above sense, we keep our scope manageable and explicitly exclude from this category those approximation

techniques that apply the related and also nearsighted *atomic decomposition ansatz*, e.g., [Bar+10; Bar+17; Fis18].

Our central contention, here and in the following chapters, is that the energy expressions produced by fragmentation methods based on or related to the MBE can be usefully viewed as truncations of particular combination sums in the sense of the order-theoretic combination technique presented in Chapter 3. We are particularly interested in obtaining a new perspective on the BOSSANOVA [Heb14; GHH14] and ML-BOSSANOVA [CGH18] techniques; although we mention these only obliquely in this chapter, we still draw significant general influence here from [GHH14; CGH18] in particular. For now, we will lay the necessary groundwork for this new viewpoint. We collect and relate a number of basic ideas and existing definitions and observations from the copious literature on subsystem techniques and fragmentation methods. We will consider the construction of the MBE at length, and contrast an order-theoretic perspective on the MBE with two other perspectives more commonly seen in the literature. Although historically well-known [Kle86; DFS04], this viewpoint seems to have been mostly forgotten in contemporary discussions of fragmentation techniques, and its reapplication provides a formal connection between several modern constructions and ideas. After briefly summarising some known issues related to the cost and numerical stability of MBE truncations, we then apply the adaptive algorithm from Chapter 3 to the problem of calculating quasi-optimal MBE truncations, leading to an *adaptive many-body expansion*.

5.1. Subsystem techniques

We will consider subsystem techniques of two loosely-divided types. As mentioned above, *fragmentation* methods somehow disintegrate the complete system under study into plural *fragment* subsystems; see, e.g., [Fis64; HMS70; Gor+11; RH12; Her19]. Full-system electronic properties are then obtained by some recombination of those properties as calculated for the individual fragments. By contrast, *embedding* techniques single out one or more subsystems as being more deserving of careful treatment [WL76; Chu+15]. Calculations over these *embedding regions* are performed using a high-quality and therefore expensive level of theory, and somehow coupled to or merged with a lower-quality but more affordable full-system calculation. The two types are not disjoint, conceptually or in practice. Many modern fragmentation methods embed their fragments in some coarse model of the complete system; see, e.g., [DT06; IWT13; LH16; Jon+20]. Similarly, embedding techniques can be viewed as separating the system into at least two fragments, namely the embedding region (or regions) and the embracing *environment region*; see, e.g., [BT96; Bat+11; Chu+15; SC16; Jon+20].

However they are categorised, subsystem techniques are a major subfield of modern computational chemistry, supported by a considerable body of literature. We will provide here only sufficient background to contextualise our investigation of additive

decompositions related to the MBE. For classical embedding techniques and QM/MM methods, we refer to the reviews of Lin and Truhlar [LT06], Senn and Theil [ST09], and Chung et al. [Chu+15], also [Jen17, Sec. 2.12]. For quantum embedding theories, we refer to the reviews of Sun and Chan [SC16] and Jones et al. [Jon+20]. For fragmentation methods, we refer to the reviews of Gordon et al. [Gor+11], Collins and Bettens [CB15], Raghavachari and Saha [RS15], and Herbert [Her19].

5.1.1. Embedding schemes

The development of modern embedding techniques is entwined with the field of *quantum mechanics/molecular mechanics* (QM/MM) [WL76; FBK90; BT96; LT06; ST09; Chu+15]. Here, a molecular system is split into an environment region, called in context the *MM region*, and an embedding, *QM region*. A full-system *effective Hamiltonian* is written as [FBK90; WCC08; Jen17; Kir+21]

$$H_{\text{eff}} = H_{\text{QM}} + H_{\text{QM/MM}} + H_{\text{MM}}. \tag{5.1}$$

The first term, H_{QM}, is the electronic Hamiltonian (2.3), defined in restricted terms of the nuclei and electrons in the QM region. The final term, H_{MM}, describes the energy of the MM region as per a classical molecular mechanics model, and is effectively a scalar constant, like the nuclear repulsion term in (2.6). The middle term, $H_{\text{QM/MM}}$, is a *coupling Hamiltonian* which represents cross-region interactions. The total electronic energy of the system decomposes additively according to the terms of the effective Hamiltonian, as $E = \langle \Psi, H_{\text{eff}} \Psi \rangle = E^{\text{QM}} + E^{\text{QM/MM}} + E^{\text{MM}}$; in addition to the above citations, see and cf., e.g., [BT96; ST09; Chu+15].

For details on different forms of the coupling Hamiltonian, see, e.g., [BT96; LT06; ST09; Kir+21]. In the particular case of an *electrostatic embedding* scheme, however, $H_{\text{QM/MM}}$ is basically constructed like H_{MM}, but includes a term explicitly imposing the electrostatic configuration of the MM region on the QM-region wavefunction. That is, as in, e.g., [FBK90; BT96; WCC08],

$$H_{\text{QM/MM}} = \sum_{A=1}^{M_{\text{QM}}} \sum_{B=1}^{M_{\text{MM}}} \frac{Z_A q_B}{\|R_A - R_B\|} - \sum_{i=1}^{N_{\text{QM}}} \sum_{B=1}^{M_{\text{MM}}} \frac{q_B}{\|r_i - R_B\|} + \cdots \tag{5.2}$$

Here, M_{MM} counts the atoms in the MM region, and M_{QM} and N_{QM} the QM region nuclei and electrons respectively. Each q_B is a point charge associated with the atom indexed by B, usually a partial charge chosen with reference to the molecular mechanics model; see, e.g., [WCC08] and discussion in [LT06]. The first summation tallies pairwise Coulomb interactions between MM atoms and QM nuclei, and the second sets an external potential on the electrons in the QM region [BT96; Kir+21]. The trailing ellipsis in (5.2) includes any remaining non-electrostatic cross-region interactions that may be specified by the molecular mechanics model; see, e.g., the van der Waals terms in [FBK90, (9);

WCC08, (4)]. The external potential term in (5.2) can be subsumed into the QM-region Hamiltonian H_{QM} [FBK90; BT96; Vre+06]. Obtaining an approximate solution to the electronic problem posed in terms of the so-augmented H_{QM} requires minor modifications to the solver; this is supported in particular by the PySCF package [Sun15; Sun+17; Sun+20], which we use in this chapter.

There is another class of alternatively-formulated embedding techniques which also seek to emphasise the theoretical treatment of a particular region, or regions, of a full system. These are commonly referred to as *subtractive* [Chu+15], to distinguish them from the additive-style QM/MM approach outlined above. The canonical subtractive embedding formulation is the ONIOM (*Our own N-layered Integrated molecular Orbital and molecular Mechanics*) [Sve+96] method, which is constructed in terms of a nested hierarchy of N subsystems of the full system. The outermost, lowest-level region is just the full system itself, referred to in context as the *real* system. Each interior higher-level region is referred to as a *model* system, and receives a notionally higher level of theoretical treatment than those below it. The full ONIOM energy equation for a nested hierarchy of N systems approximated with N matching levels of theory is given by [Sve+96, (4)]

$$E_{\mathrm{ONIOM}(N)} := \sum_{i=1}^{N} E^{\mathrm{level}(i)}_{\mathrm{model}(n+1-i)} - \sum_{i=2}^{N} E^{\mathrm{level}(i)}_{\mathrm{model}(n+2-i)}, \tag{5.3}$$

where we slightly adjust the notation of the source and write $E^{\mathrm{level}(i)}_{\mathrm{model}(j)}$ to indicate the total energy of the jth-nested model region calculated using the ith-ranked level of theory. It is interesting to observe that this is, up to notation, exactly the expression (3.6) for the two-dimensional standard combination technique sum S_{I_L} of level L, and ONIOM is directly presented in [Sve+96] as an extrapolation of a set of cheaper values towards $E^{\mathrm{level}(N)}_{\mathrm{model}(N)}$; cf. [Sve+96, Fig. 1], and see also comments in [CGH18]. Explicit connections have also been made between the ONIOM formulation and composite methods [Chu+15] such as those which we contrasted with the standard combination technique in the previous chapter.

In the implementation, the additive and subtractive embedding formulations outlined above reduce to performing *a postereori* arithmetic on sets of energies — or, although not discussed here explicitly, nuclear gradients; see, e.g., [MM95; WCC08] — that are provided by distinct calculations. We mention briefly the existence also of *quantum embedding theories*, which can be viewed as somehow partitioning either the full-system wavefunction or an equivalent property such as the electron density and then performing basically one single calculation [SC16; Jon+20]. We refer to those reviews for more information, but it will be slightly relevant below that these include *DFT-in-DFT* approaches, like the *frozen density embedding* (FDE) method of Wesolowski and Warshel [WW93]; also *WFT-in-DFT* techniques, where "WFT" stands for "wavefunction theory", like the *embedded correlated wavefunction* (ECW) approach of Carter and co-workers, see, e.g., [GWC99;

Yu+17]; finally, full *WFT-in-WFT* embedding approaches, see, e.g., work by Hégely et al. [Hég+16].

In the classical additive and subtractive approaches, particular difficulties are encountered when a covalent bond exists between two atoms, one inside and one outside an embedding region. The absence of one half of a bonded pair from a subsystem considered in isolation is said to leave behind a *dangling bond* [AR96; VM03; ST09]. Such a subsystem will be "chemically unrealistic" [Dap+99, p. 2], and, left untreated, the electron density near the dangling bond will not resemble that in the complete system [FBK90]. Although more complicated and intrusive methods exist [AR96; Gao+98], a simple and widely-applied fix is to adjoin a *link atom* to terminate any dangling bond [FBK90]. It is generally anticipated, either implicitly or explicitly, that the embedding structure is chosen such that only single bonds will be cut, and so a link atom must be monovalent [AR96]; hydrogen atoms are the most common choice [LT06; ST09; Chu+15; RS15], and their precise placement is an implementation detail. Although it is technically possible to use divalent link atoms to cap dangling double bonds [Dap+99], we are aware of only one study in which this is actually performed [MMM16]. Certain additional subtleties must be considered when introducing link atoms in an electrostatic embedding context [LT06]; see, e.g., the discussion of charge balancing and redistribution in [WT10].

The handling of embedding/environment interfaces in quantum embedding theories is not important for our purposes. We mention, however, the DFT-in-DFT *embedded mean-field theory* (EMFT) of Fornace et al. [For+15], which uses the association of the underlying LCAO basis functions with their nuclear centres to decompose the DFT density matrix. Here, although the concept of a dangling bond is not meaningful in the sense above, certain irregularities can still occur in the electron density local to the boundary between the embedding and environment regions [MMM16; Lee+17; HNK18].

5.1.2. Fragmentation methods and the many-body expansion

There exist a number of fragmentation-style methods phrased explicitly in terms of densities and density matrices [Gor+11; CB15]; see, e.g., [Yan91; YL95; LYY96]. However, we will focus in this section, and in this thesis more generally, only on *energy-based fragmentation methods*, hereafter just "fragmentation methods" where there is no chance of confusion. We make in this section general reference to reviews in [SDS09; Gor+11; RH12; RLH14; CB15; RS15; Her19]. Let us recommend here also a body of important work by Herbert and co-workers which we found influential and interesting; see generally [RH12; Jac+13; RLH13; RH13; RLH14; Lao+16; LH16; LH17; LH19].

We follow the taxonomies in, e.g., [SDS09; RH12; RS15] and consider two ways of deconstructing the complete molecular system under study: into sets of fragments that are either *disjoint* or *overlapping*. Mathematically, the former is equivalent to a partition, in the strict sense, of the set of nuclear indices $[M]$; that is, a family of K non-empty sets $\{F_i\}_{i=1}^{K}$ which are pairwise disjoint, i.e., $F_i \cap F_j = \emptyset$ for $i \neq j$, and such that

$\bigcup_{i=1}^{K} F_i = [M]$. The latter corresponds to a family of non-empty sets $\{F'_i\}_{i=1}^{K}$ that still recover $[M]$ in their union, but which are not necessarily pairwise disjoint. Since we are mostly interested in the forms of the final energy equations of fragmentation methods, we will not linger here on the technical details of how these families are actually decided. We will often conflate as *fragments* both sets of indices $F_i \subseteq [M]$ and the notional molecular subsystems of physical atoms to which those indices correspond.

We summarise for the moment only a small selection of methods that begin with disjoint fragments. We will come to some methods that start with sets of potentially-overlapping fragments in the following section, and we defer a discussion of multilevel fragmentation methods until Chapter 7. In general, we mostly try to follow the overall notational style of the sources, although we make modifications for clarity and consistency, usually without explicit comment.

We start with and will focus most closely on the well-known concept of the *many-body expansion* (MBE), which permeates the literature [SDS09; Gor+11; Fed+14; CB15; Her19]. Following, e.g., [CGH18], we will consider the MBE as an additive decomposition of the Born-Oppenheimer potential,

$$V^{\mathrm{BO}}(X_1, \ldots, X_M) = \sum_{A=1}^{M} \tilde{V}^{(1)}(X_A) + \sum_{A<B}^{M} \tilde{V}^{(2)}(X_A, X_B) + \cdots + \tilde{V}^{(M)}(X_1, \ldots, X_M). \quad (5.4)$$

Here, each $X_A = (R_A, Z_A)$ is a nuclear variable as usual. Each $\tilde{V}^{(k)}$ is a *k-body potential energy function*, to be defined shortly.

The notation used to express the MBE in the literature varies considerably from work to work; ours is inspired primarily by that in [CGH18], also [HMS70; Heb14], but is modified for consistency with the remainder of this thesis. In practice, such an explicitly nuclear form of the MBE as (5.4) is not common, and each X_A is more usually a composite variable collecting the nuclei/atoms in a fragment F_i; see, e.g., [HMS70] for an early example. We will formalise this more precisely below. It is, however, common, particularly in the more modern literature, to treat the MBE instead as a decomposition of a scalar energy value for a fixed nuclear conformation; see, e.g., [MR12; RH12; Her19]. But up to these and other small details, the k-body potentials are usually constructed recursively [Fis64; HMS70; CGH18; Her19]. To make this even slightly rigorous, we require an extended family of Born-Oppenheimer potential functions $\{V^{\mathrm{BO},k} : (\mathbb{R}^3 \times \mathbb{N})^k \to \mathbb{R}\}_{k=1}^{M}$; the definitions are equivalent to that of the "full" $V^{\mathrm{BO}} = V^{\mathrm{BO},M}$ but for molecular systems containing fewer nuclei, and the superscripts k will usually be clear from context and so omitted. We ignore and will continue to ignore the technical question of the well-definition of each $V^{\mathrm{BO},k}$ on $(\mathbb{R}^3 \times \mathbb{N})^k$; see, however, related discussion in [CGH18].

Noting that V^{BO} is symmetric in its variables $\{X_A\}_{A=1}^{M}$, we consider sets $\mathbf{u} \subseteq [M]$ and write

$$\tilde{V}_{\mathbf{u}} :=: \tilde{V}_{\mathbf{u}}(\ldots, X_{A\in\mathbf{u}}, \ldots) := \tilde{V}^{(|\mathbf{u}|)}(\ldots, X_{A\in\mathbf{u}}, \ldots), \quad (5.5)$$

and similarly for $V_{\mathbf{u}}^{\mathrm{BO}}$, etc. Then the usual recursive definition of the k-body potentials is

$$\tilde{V}^{(k)}(X_1, \ldots, X_k) := V^{\mathrm{BO}} - \sum_{\substack{\mathbf{u} \subset [k] \\ \mathbf{u} \neq \emptyset}} \tilde{V}_{\mathbf{u}}. \tag{5.6}$$

The MBE given by

$$V^{\mathrm{BO}}(X_1, \ldots, X_M) = \sum_{\emptyset \subset \mathbf{u} \subseteq [M]} \tilde{V}_{\mathbf{u}} \tag{5.7}$$

is thus formally exact by construction. Again following [CGH18], we will refer to each $\tilde{V}_{\mathbf{u}}$ as the *contribution* of $\mathbf{u} \subseteq [M]$, and, indirectly, of the molecular subsystem so indexed.

As is well noted in the literature, see, e.g., [GHH14; CGH18; Her19], the MBE makes practical sense only after the elision of certain terms in (5.4). The simplest standard approach is to make a truncation of (5.4) after all terms $\tilde{V}^{(k)}$ for $k \leq n$ for some particular $1 \leq n \leq M$; the resulting expression is referred to as an *n-body expansion*. Critical here is the anticipation of a swift pointwise decay in $|\tilde{V}^{(k)}|$ for higher k, and thus an acceptably small divergence from V^{BO} in truncation even for very low values of n [CGH18]. Since those terms which are left remaining are expected to be individually cheap to evaluate, their calculation and summation may, under some conditions, be markedly more affordable than a complete full-system evaluation of V^{BO}. It has also been very thoroughly noted that these calculations can be performed completely in parallel; see, e.g., [CB15; CGH18], but cf. cautionary comments in [Her19].

Although it seems *a priori* obvious that the error intrinsic to an evaluated n-body truncation should shrink with increasing n, certain practical aspects mean that this may not necessarily be so, and particularly not for larger systems [RLH14; Lao+16; LH17]. We shall return to this topic briefly in Section 5.3 below. But it is especially problematic given that the "convergence" of n-body truncations of the MBE, insofar as that term makes sense in the context of a finite expansion, may also not be as quick or as smooth as historically believed [OCB14].

For a full discussion on the cost of fragmentation methods and particularly the MBE, we refer to [Her19], also, e.g., [Lao+16; LH17; LH19]. It suffices to say that, since the number of k-body terms in (5.4) is $\binom{M}{k}$, the total number of the same involved in an n-body expansion scales roughly as $\mathcal{O}(M^n)$ in M [LH19]. If the evaluation cost of any k-body potential is bounded above by a constant (see, e.g., [LH19] and cf. [Her19, (1)]), this implies an equivalent overall cost scaling in M. Note here that k-body terms of order $k = 5$ or possibly even higher may be required in order to obtain an accurate result in some cases [OCB14].

Such unfavourable scaling can be mitigated to some extent when a second kind of decay in the magnitudes $|\tilde{V}^{(k)}|$ is exploited, namely one in the distance(s) between the nuclear spatial variables [OB16; LH17; LH19]. This decay, which can be viewed as a manifestation of Kohn's nearsightedness principle [OB16; CGH18], is relied on either

implicitly or explicitly by many fragmentation methods which build their sets of fragments using distance- or connectivity-based arguments, e.g. [DC05; Gan+06; LLJ07; WHM10; MR12; RH12; KC16; CGH18]. We shall have more to say on this later in this chapter, and particularly in the next. For now, we leave the idea with the reader and move on to quickly reviewing some other fragmentation methods.

As the name suggests, the *electrostatically-embedded many-body expansion* (EE-MB) of Dahlke and Truhlar [DT06; DT07b] is based upon a fragment-based formulation of the MBE, given in terms of scalar energy values. Here, each k-body energy calculation is performed using an augmented Hamiltonian as for an electrostatically-embedded QM/MM calculation, with one point charge term in that Hamiltonian for each atom outside the k-body subsystem. Dahlke and Truhlar are not prescriptive as to the source of the point charges; indeed, one interesting extension to the EE-MB suggests computing the point charges themselves as standard partial charges using a three-body expansion [Lev+12].

In the simplest version of the *kernel energy method* (KEM) of Huang et al. [HMK05], a chain-like molecule is split into K fragments, which are non-overlapping and indexed such that fragments i and $i+1$ are directly adjacent in the chain. These fragments are called in context *kernels*. Any pair of adjoining fragments form together a *double kernel*, with energy $E_{i,i+1}$. The total KEM energy of such a molecule is then [HMK05, (1)]

$$E^{\mathrm{KEM}} := \sum_{i=1}^{K-1} E_{i,i+1} - \sum_{i=2}^{K-2} E_i. \tag{5.8}$$

It has been noted that when the adjacency requirement on a double kernel is dropped, as also done in [HMK05], the resulting energy expression is just a standard two-body expansion [SDS09]; cf, e.g., [Gor+11, (52)]. More interesting, and less standard, is a generalised variant of the KEM (GKEM) due to Weiss et al. [WHM10]. The GKEM considers the kernels as vertices of an undirected graph, and produces a sum of terms, one for each of the connected induced subgraphs of that graph with size up to some fixed n. This sum is itself converted into a weighted sum of energy values for systems formed as combinations of up to n kernels. In [WHM10], $n \leq 4$, and a generalisation of the GKEM energy equations to higher values of n seems non-trivial.

The *systematic fragmentation method* (SFM) of Deev and Collins [DC05; CD06] partitions a molecule or nonconducting crystal [NC07] into disjoint fragments referred to as *functional groups*. These are "defined in the usual way" [NC07, p. 2], but can in principle also be chosen more flexibly [CD06]. The complete system is then split into K potentially-overlapping subsystems basically by dissolving pairs of bonds separated by some particular number of functional groups, which determines the *level* of the resulting fragmentation; see the appendices of [CD06] for details. The total SFM energy is [CD06, (A7)]

$$E^{\mathrm{SFM}} = \sum_{k=1}^{K} \mathrm{sign}(k) E_k; \tag{5.9}$$

where E_k is the energy of the kth subsystem and sign(k) is a weight assigned during the splitting process. We foreshadow the next section by mentioning that the precise values of sign(k) can be understood as correcting for somehow overcounted interactions caused by functional groups that appear in multiple subsystems [RH12; RS15].

The *DCMB* scheme of Wu and Xu [WX12] is a slightly unusual approach that also bases upon a disjoint partition of the complete set of nuclear indices. Here, a complete full-system *mixed basis* DFT calculation specialises each fragment; in each, the nuclei indexed by the fragment, as well as those in a narrow surrounding buffer region, are equipped with a *large* basis set, such as 6-31G* [DHP71; HDP72; HP73], while the remaining nuclei are equipped with a *small* basis set, such as STO-3G [HSP69] or 3-21G [BPH80]. The DCMB scheme focuses on the calculation of nuclear gradients, rather than energies; the former for the complete system are not summed, but are instead simply collated componentwise. Obtaining the latter requires extra effort.

The *fragment combination range* (FCR) approach of König and Christiansen [KC16] was introduced in the context of a *double-incremental expansion* of V^{BO}.[1] They start from a standard MBE, which is formulated with reference to a set of N disjoint fragments, each represented by a composite variable z_i. This MBE is truncated to include only terms corresponding to certain subsets $\mathbf{f}_l \subseteq [N]$. The FCR provides particularly for a non-recursive expression, which we present based on both [KC16] and also [HK21, (9), (10), (11)]:

$$E(z_1, z_2, \ldots, z_N) \approx \sum_{\mathbf{f}_l \in \{\mathrm{FCR}\}} p^{\mathrm{FCR}}_{\mathbf{f}_l} E_{\mathbf{f}_l}(\{z\}_{\mathbf{f}_l}), \tag{5.10}$$

where {FCR} is the set of all subsets considered in the truncation; this set must be downwards-closed. Each such subset has an associated coefficient

$$p^{\mathrm{FCR}}_{\mathbf{f}_l} = \sum_{\substack{\mathbf{f}_{l'} \supseteq \mathbf{f}_l \\ \mathbf{f}_{l'} \in \{\mathrm{FCR}\}}} (-1)^{l'-l}. \tag{5.11}$$

This expression is not given explicitly for the FCR in the original [KC16], but rather for the more general VCR (*variable combination range*) upon which the FCR is based. The authors of [KC16] are, however, very clearly aware of the equivalence.

Finally, we mention the *Bond-Order diSSection ANOVA* (BOSSANOVA) approach, due to Heber [Heb14] and co-workers [GHH14]. The precise formal presentation of BOSSANOVA differs between the two sources just given, and a third version again can be found in the later [CGH18], but all three presentations involve exact decompositions of the full-system Born-Oppenheimer potential which are formally equivalent and based on the theory of approximation of high-dimensional functions. From one perspective,

[1] Getting somewhat ahead of ourselves, we remark in passing that the double-incremental expansion can also be formalised as a particular application of the order-theoretic combination technique. The specific poset grid under consideration is $\Pi = B_M \times B_M$. We leave an investigation of this setup for future work.

BOSSANOVA truncates a standard MBE, but one constructed in terms of modified potentials $V^{(k)}$ which are defined in such a way that many of the resulting k-body potentials $\tilde{V}^{(k)}$ vanish. From another perspective, the covalent bond graph of the target molecule is pulled apart into generally non-disjoint connected induced subgraphs, and per-fragment potentials are calculated and recombined according to those subgraphs and their own subgraph structures. We shall discuss BOSSANOVA in detail in Chapter 6.

5.2. Mathematical viewpoints on the many-body expansion

We discuss now three mathematical viewpoints on the many-body expansion. Specifically, we shall first review the use of counting arguments in explications of the energy expressions of certain fragmentation methods related to the MBE, and then consider the MBE as an example of a particular kind of general decomposition of high-dimensional functions. Finally, we will discuss the MBE from a combinatorial and order-theoretic perspective.

5.2.1. Counting arguments

The literature describing and contrasting fragmentation methods and the MBE contains a number of counting-style arguments, carried out to varying levels of rigour. Many of these rely on or reduce to an application of the *principle of inclusion/exclusion* (PIE) [MR11; MR12; RH12; RH13; RS15]. Stanley gives a very general expression of the PIE as [Sta12, Thm. 2.1.1]; this is directly recoverable as a special case of Möbius inversion, applied to a boolean algebra B_n. We are aware of only one place in the fragmentation method-related literature (specifically [WHM10, Supp. info]) where such a general form of the PIE is mentioned explicitly. There, it is used to derive expressions for k-body terms in a certain generalisation of the MBE. We will come back to Möbius inversion in Section 5.2.3 below.

A less general but much more widely-known variant of the PIE has been used both to motivate and compare a number of fragmentation methods that are defined in terms of families of potentially-overlapping fragments $\{F_i\}_{i=1}^K$ [Gan+06; MR11; MR12; RH12; CB15; RS15; Her19]. As well as our general references for fragmentation methods, the following summary is particularly informed by the comparison of the MOBE and the GMBE in [RH13].

For the convenience of the reader, we begin with an explicit statement of this set-cardinality version of the PIE, given without proof. We refer generally here to [Sta12] for detailed background, but this is of course a standard result; the precise form we give here is actually closer to that given in, e.g., [RH12].

Theorem 5.2.1 (Principle of inclusion/exclusion [Sta12])**.** *Let X be a finite set, and $\{X_i \subseteq X\}_{i=1}^n$ a family of subsets of X. Then*

$$|X_1 \cup \cdots \cup X_n| = \sum_{k=1}^{n} (-1)^{k+1} \sum_{1 \leq i_1 < \cdots < i_k \leq n} |X_{i_1} \cap \cdots \cap X_{i_k}|. \tag{5.12}$$

We now recite the energy equation of the *cardinality-guided molecular tailoring approach* (CG-MTA) [Gan+06], which "preserves the count of atoms and bonds in the parent system" [Gan+06, p. 3]. This equation, which is explicitly and directly related to (5.12), is given as [Gan+06, (4)]

$$E^{\text{CG-MTA}} = \sum_{i=1}^{K} E_{F_i} - \sum_{i<j} E_{F_i \cap F_j} + \cdots + (-1)^{K+1} \sum_{i<j<\cdots<k} E_{F_i \cap F_j \cap \cdots \cap F_k}, \tag{5.13}$$

where we slightly adjust and clarify the notational style of the source. Here, for example, we write $E_{F_i \cap F_j}$ for the notional total energy of the subsystem $F_i \cap F_j$.

Mayhall and Raghavachari [MR12] carried the same argument further in the derivation of their *many-overlapping-body expansion* (MOBE). Here, an extended set $\{F'_i\}_{i=1}^{K'}$ of $K' \leq 2^K - 1$ of *derivative subsystems* or just *monomers* is constructed as the original set of fragments $\{F_i\}_{i=1}^{K}$, along with all of the k-fold intersections of those sets for every $1 \leq k \leq K$ [RH13]. The inequality for K' holds because the intersections are not guaranteed to be elementwise distinct. The MOBE is then [MR12, (5)]

$$E^{\text{MOBE}} := \sum_{i=1}^{K'} c_i E_{F'_i} + \sum_{i<j}^{K'} c_i c_j \Delta E_{F'_i \cup F'_j} + \sum_{i<j<k}^{K'} c_i c_j c_k \Delta E_{F'_i \cup F'_j \cup F'_k} + \cdots, \tag{5.14}$$

by extension of (5.4). Each coefficient c_i is the sum of the ± 1 terms which would be attached in a calculation of $|F_1 \cup \cdots \cup F_K|$ using (5.12) to the cardinalities of the intersections of original sets $F_j \cap \cdots \cap F_k = F'_i$ that lead to the monomer F'_i; as well as the original [MR12], cf. on this point also [RH13; CB15]. The scaled terms $\Delta E_{F'_i \cup F'_j}$ and $\Delta E_{F'_i \cup F'_j \cup F'_k}$ substitute for the normal recursive two- and three-body MBE terms; see [MR12, (6) and (7)] for their full definitions, which also involve a counting argument. Note that, in particular and very deliberately, the first sum in (5.14) is just the RHS of (5.13).

Richard and Herbert [RH12] have also suggested a competing many-body generalisation of sums such as (5.13). Their *generalized many-body expansion* (GMBE) is phrased in terms of unions of elements of the original set of potentially-overlapping fragments $\{F_i\}_{i=1}^{K}$, rather than intersections as in the case of the MOBE [RH13]. As per the equivalent protocols given in [RH12; RH13] and with slight reference to [CB15], to obtain a truncation of the GMBE after order $1 \leq n \leq K$, a family of *n-mers* is first built to be the $K' = \binom{K}{n}$ precisely n-fold unions of distinct fragments, $\{F'_i\}_{i=1}^{K'}$. The order-n energy is then written, as per [RH13, (1.9), (2.4), (2.5)] up to notation,

$$E^{\text{GMBE}}_{(n)} = \sum_{i=1}^{K'} E_{F'_i} - \sum_{i<j}^{K'} E_{F'_i \cap F'_j} + \sum_{i<j<k}^{K'} E_{F'_i \cap F'_j \cap F'_k} + \cdots + (-1)^{K'+1} E_{F'_1 \cap \cdots \cap F'_{K'}}, \tag{5.15}$$

again directly constructed by reference to the cardinality PIE.

Wondering about the legitimacy of the counting arguments involved in the derivations of the above expressions, Richard and Herbert suggested that the cardinality PIE can be better understood in context as a tool "to keep track of" [RH12, p. 5] the two-particle interactions embodied in the full-system electronic Hamiltonian (2.3) [RH12]. This viewpoint is formalised in [RH13], with reference to a family of Hamiltonians $\{\hat{H}[S]\}_{S\subseteq[M]}$. The terms of each such Hamiltonian are restricted to "all pairwise interactions among the particles" [RH13, p. 1409] indexed by S, so $\hat{H}[[M]]$ is just (2.3). Then, given a K'-set of n-mers as per the construction of the GMBE, an inductive argument is used to show that [RH13, (2.1)]

$$\hat{H}[[M]] = \sum_{i=1}^{K'} \hat{H}\big[F'_i\big] - \sum_{i<j}^{K'} \hat{H}\big[F'_i \cap F'_j\big] + \cdots + (-1)^{K'+1} \hat{H}\big[F'_1 \cap \cdots \cap F'_{K'}\big]. \tag{5.16}$$

Expanding the RHS of $E_0 = \langle \Psi_0, \hat{H}[[M]]\Psi_0 \rangle$ via the bilinearity of the inner product,[2] where Ψ_0 and E_0 are a true eigenpair for the electronic problem, leads to an expression [RH13, (2.2)] that can be approximately matched termwise with (5.15) [RH13].

We mention one last counting-style argument, which seems to have been first applied by Li et al. [LLJ07] in the context of the *generalised energy-based fragmentation* method (GEBF); cf. also [IWT13; LH16]. Here, the terms in an energy equation that is fundamentally just (5.13), see [MR11; RH12], are evaluated using electrostatic-embedding Hamiltonians, much like, e.g., the EE-MB approach described above. These are defined in terms of a complete set of point charges $\{q_A\}_{A=1}^{M}$, one for each each atom in the molecular system, and each calculation explicitly includes Coulomb interactions between the partial charges outside the involved subsystem. That is, given for example some fragment F_i, the term E_{F_i} includes the value

$$\sum_{1\leq A\notin F_i\leq M} \sum_{A<B\notin F_i\leq M} \frac{q_A q_B}{\|R_A - R_B\|}. \tag{5.17}$$

To avoid considering these and other pairwise Coulomb interactions too many times, a counting argument is used to justify the form of, in the terminology of [IWT13], an overcounting correction in the GEBF energy equation [LLJ07; IWT13]. This is, in full, following [IWT13, (13)] and using notation adjusted for consistency with the expression of the CG-MTA energy given previously,

$$E^{\mathrm{GEBF}} = \sum_{i=1}^{K} E_{F_i} - \sum_{i<j} E_{F_i\cap F_j} + \cdots + (-1)^{K+1} \sum_{i<j<\cdots<k} E_{F_i\cap F_j\cap\cdots\cap F_k} - \left[\left(\sum_{i=1}^{K} c_i\right) - 1\right] \sum_{A=1}^{M}\sum_{A<B}^{M} \frac{q_A q_B}{\|R_A - R_B\|}. \tag{5.18}$$

[2] Here assuming, unlike in Chapter 2, that Ψ_0 is real-valued; see, e.g., [Can+03].

The additional term on the right-hand side compared to (5.13) is the correction; the sum in parentheses runs over coefficients c_i for each of the K' intersections in the inclusion/exclusion sum, defined equivalently as for the MOBE above.

5.2.2. ANOVA-like decompositions

It has been previously and repeatedly observed that MBEs such as (5.4) are reminiscent of certain decompositions used in the approximation and analysis of high-dimensional functions; see, e.g., [Gri06; Heb14; GHH14; Pas14; KC16; CGH18; Fis18]. We will recall and consider a general construction for the latter given in [Kuo+09]; the notation we use in this section mostly matches that of [Kuo+09], and the reader is warned that it may conflict slightly with the notation we use elsewhere. We omit some definitional details in the interest of brevity.

The setting of [Kuo+09] is a vector space $\mathcal{F}$ of d-dimensional functions $f(x_1, \ldots, x_d) : D \subseteq \mathbb{R}^d \to \mathbb{R}$. We require a family of commuting projections $\{P_i : \mathcal{F} \to \mathcal{F}\}_{i=1}^d$ such that, for any choice of $1 \leq i \leq d$ and $f \in \mathcal{F}$, then $P_i f$ is invariant in x_i, and moreover, $P_i f = f$ whenever f is invariant in x_i [Kuo+09, (2.1)]. Then any $f \in \mathcal{F}$ can be decomposed as

$$f = \sum_{\mathbf{u} \subseteq [d]} f_{\mathbf{u}}, \tag{5.19}$$

in terms of functions $f_{\mathbf{u}}$ invariant in those variables $\{x_i\}_{i \in \mathbf{u}}$ [Kuo+09, Thm. 2.1]. Here, each $f_{\mathbf{u}}$ can be defined equivalently as [Kuo+09, Thm. 2.1(b)]

$$f_{\mathbf{u}} := \sum_{\mathbf{v} \subseteq \mathbf{u}} (-1)^{|\mathbf{u}|-|\mathbf{v}|} P_{[d]-\mathbf{v}} f, \tag{5.20}$$

where $P_{\mathbf{u}} := \prod_{i \in \mathbf{u}} P_i$, or recursively [Kuo+09, Thm. 2.1(a)],

$$f_{\mathbf{u}} := P_{[d]-\mathbf{v}} f - \sum_{\mathbf{v} \subset \mathbf{u}} f_{\mathbf{v}}, \tag{5.21}$$

with $f_\emptyset := P_{[d]} f$ providing the base case. Regardless of the precise choice of commuting projections P_i, the decomposition (5.19) is minimal, in the sense that if some decomposition $f = \sum_{\mathbf{u} \in [d]} t_{\mathbf{u}}$ exists such that $t_{\mathbf{v}} = 0$ for all $\mathbf{v} \supseteq \mathbf{z}$ for some $\mathbf{z} \subseteq [d]$, then also $f_{\mathbf{v}} = 0$ for all $\mathbf{v} \supseteq \mathbf{z}$ [Kuo+09, Thm. 3.1].

If $\mathcal{F}$ is taken to be $L^2([0,1]^d)$, and the projections P_i are chosen such that

$$P_i f := \int_{[0,1]} f(x_1, \ldots, x_d) \, \mathrm{d}x_i, \tag{5.22}$$

then (5.19) becomes the *ANalysis Of VAriance* (ANOVA) decomposition; see [Kuo+09, Ex. 2.2], and also [Gri06; Feu10] and references therein. For this reason, following also [Gri06], we will call decompositions of the form (5.19) to be *ANOVA-like*.

There is a relationship here to the concept of *effective dimension* [CMO97]; see also [Gri06; Feu10]. We leave a formal definition for [CMO97], but it has been suggested that low-order n-body expansions function in practice because the Born-Oppenheimer potential energy function possesses something akin to low effective dimension [Gri06; Pas14; CGH18]. Also noteworthy in this context is the *high-dimensional model representation* (HDMR) of Rabitz and Aliş [RA99]. It has been explicitly observed by those authors that "many-body expansions can be viewed as a special case of an HDMR" [RA99, p. 199].

To illustrate, we give a sketch of an explicit construction of the MBE as an ANOVA-like decomposition (5.19) in the style of [Kuo+09], and motivated by the termwise decomposition (5.16) of the Hamiltonian discussed in Section 5.2.1; cf. here the very closely related constructions of MBEs in [GHH14; CGH18]. The simplest choice of underlying vector space $\mathcal{F}$ is just the set of functions $f : (\mathbb{R}^3 \times \mathbb{N})^M \to \mathbb{R}$, which we consider as taking M nuclear variables $\{X_A = (R_A \in \mathbb{R}^3, Z_A \in \mathbb{N})\}_{A=1}^M$ consistent with our setting. The use of such four-dimensional composite variables means that the results of [Kuo+09] are not directly applicable here. We claim without proof, however, that Theorems 2.1 and 3.1 of [Kuo+09] generalise trivially to this setting.

We define the family of operators $\{P_A : \mathcal{F} \to \mathcal{F}\}_{A=1}^M$ to be such that

$$(P_A f)(X_1, \ldots, X_M) = f(\ldots, X_{A-1}, (R_A, 0), X_{A+1}, \ldots). \tag{5.23}$$

that is, the charge of the Ath variable of $P_A f$ is constrained to be zero. Clearly, $P_A^2 = P_A$, so P_A is a projection; it also easy to see that $P_A P_B f = P_B P_A f$ for $1 \leq A,\, B \leq M$, and more generally that the family $\{P_A\}_{A=1}^M$ is as required by the preconditions of [Kuo+09, Thm. 2.1]. We thus obtain a minimal decomposition of any $f \in \mathcal{F}$. This is just a very slight variation of the anchored-ANOVA decomposition; see [Gri06; Kuo+09], and cf. basically equivalent applications of the cut-HDMR [HLR05] such as, e.g., [KC16].

We recall the definition of the Born-Oppenheimer potential function from (2.6) above:

$$V^{\mathrm{BO}}(X_1, \ldots, X_M) := \sum_{1 \leq A < B \leq M} \frac{Z_A Z_B}{\|R_A - R_B\|} + \inf_{\substack{\Psi \in \mathcal{V} \\ \langle \Psi, \Psi \rangle = 1}} \big\langle \Psi, H\big[\{X_A\}_{A=1}^M\big]\Psi \big\rangle. \tag{5.24}$$

Assuming for simplicity a net zero total charge, the solution space over which the infimum is taken is implicitly defined for wavefunctions in terms of only $\sum_{A=1}^M Z_A$ electrons. Since any term in the Hamiltonian (2.3) involving $X_A = (R_A, Z_A)$ vanishes when $Z_A = 0$, it is clear that the projected potential $P_{[M]-\mathbf{u}} V^{\mathrm{BO}}$ loses any dependence on the variables $\{X_A\}_{A \in [M]-\mathbf{u}}$. Thus, with some abuse of notation, $P_{[M]-\mathbf{u}} V^{\mathrm{BO}} = V^{\mathrm{BO},|\mathbf{u}|}$, where the latter is just the standard Born-Oppenheimer potential function $V^{\mathrm{BO},|\mathbf{u}|}$ for the charge-neutral subsystem described by the index set $\mathbf{u}$. So the minimal decomposition of V^{BO} of the form (5.19) promised by [Kuo+09] is effectively just the MBE in (5.4).

The implementations of the MBE as applied in practice by the various fragmentation methods discussed in Section 5.1.2 above do not lend themselves quite so easily to a

projector-based construction. The analysis is complicated by the introduction of link atoms, as well as the use of embeddings. The latter is particularly problematic, since a potential function for some subsystem indexed by some $\mathbf{u}$ must in such a case still retain some explicit dependence on all the nuclear variables. Also, the guarantee of minimalism is not especially useful in context, since although the higher-order terms of the MBE may be small, they are not expected to be exactly zero. A more thorough analysis of particular variants of the MBE from the perspective of related work in high-dimensional function approximation is sure to be of value, but is beyond the scope of this work.

5.2.3. An order-theoretic perspective

We now consider two, more carefully-defined versions of the MBE (5.4), built in the particular context of order theory. Our primary intention is to bring the MBE, and MBE-like sums, within the scope of the adaptive order-theoretic combination technique described in Chapter 3. In so doing, we shall also collect some straightforward results that will help us to understand some MBE-like fragmentation techniques both mentioned in the sections above, and to be discussed in the pages to come.

The phrasing of chemical problems in terms of posets is well-established [KB97], and the basic idea of the following is certainly not new. Although it has long been known that a non-recursive expression for the k-body terms in (5.6) can be derived using Möbius inversion [DFS04], the technique still goes mostly unmentioned in the modern literature on fragmentation methods; cf., however, [SDS09]. But the MBE is also deeply connected to various *cluster expansion methods* found in statistical physics [Mar75; DFS04], where the explicit application of Möbius inversion is also well-established [Dom74] and more common. For example, the use of Möbius inversion to obtain a different perspective on the *cluster variation method* [An88; Mor94] inspired in turn an order-theoretic redevelopment [LC05] of the *lattice fundamental-measure theory* (LFMT) of Lafuente and Cuesta [LC04]. We mention also the *chemical graph-theoretic cluster expansion* of Klein [Kle86], an extremely general formulation which explicitly allows for but is not restricted to the construction of decompositions in terms of Möbius functions and Möbius inversion. We shall return briefly to the CGTCE in the following chapter, but remark now without elaboration that it provides as a special case a construction that is formally equivalent to that which we now give; see, e.g., [Kle86, App. A].

We begin by precisely defining, or redefining, some terminology which we have up to this point used either informally or in a less-than-general way. Again, the core ideas here are well-known, up to terminology and notation. We are most strongly influenced at the outset here by the development in [CGH18, Sec. 3.2]. Although this does not explicitly involve Möbius inversion, it does mention powersets and posets, and also downward-closed subsets of the same; see also comments below. Select a family of functions $\{V_{\mathbf{u}} : (\mathbb{R}^3 \times \mathbb{N})^M \to \mathbb{R}\}_{\mathbf{u}\subseteq[M]}$, each depending on M variables $\{X_A = (R_A \in \mathbb{R}^3, Z_A \in \mathbb{N})\}_{A=1}^M$. We call each $V_{\mathbf{u}}$ the *subproblem potential* for the index

subset $\mathbf{u}$, and we will sometimes in this context call $\mathbf{u}$ a *subproblem*. For each $\mathbf{u} \subseteq [M]$, we also define a *contribution potential* by

$$\tilde{V}_{\mathbf{u}} := \sum_{\mathbf{v} \subseteq \mathbf{u}} (-1)^{|\mathbf{u}-\mathbf{v}|} V_{\mathbf{v}}, \tag{5.25}$$

as in [CGH18, (8)]. In what follows, it will always be the case that $V_{[M]} := V^{\mathrm{BO}}$, the Born-Oppenheimer potential for the complete system assuming a net-zero total charge. Each $V_{\mathbf{u}}$ for $\emptyset \subset \mathbf{u} \subset [M]$ will be a potential that somehow accentuates the set of atoms indexed by $\mathbf{u}$. We will discuss the particular choice and role of $V_{\emptyset}$ below, but note for now that it may but need not be identically zero. We depart from the prequel in that each subproblem potential is now deliberately $4M$-dimensional, to allow for embedding-style potentials which depend somehow on the complete set of nuclear variables.

As covered in Chapter 3, the powerset of $[M]$ ordered by set inclusion is the boolean algebra B_M, which has the Möbius function

$$\mu(\mathbf{u}, \mathbf{v}) = \begin{cases} (-1)^{|\mathbf{v}-\mathbf{u}|} & \text{if } \mathbf{u} \subseteq \mathbf{v}, \\ 0 & \text{otherwise.} \end{cases} \tag{5.26}$$

As in [CGH18], fix some arbitrary nuclear configuration $\{X_A\}_{A=1}^{M}$, and define a point-evaluation functional by

$$\mathcal{L}[V : (\mathbb{R}^3 \times \mathbb{N})^M \to \mathbb{R}] = V(X_1, \ldots, X_M). \tag{5.27}$$

Then Möbius inversion of

$$\mathcal{L}[\tilde{V}_{\mathbf{u}}] = \sum_{\mathbf{v} \subseteq \mathbf{u}} (-1)^{|\mathbf{v}-\mathbf{u}|} \mathcal{L}[V_{\mathbf{v}}] = \sum_{\mathbf{v} \subseteq \mathbf{u}} \mu(\mathbf{v}, \mathbf{u}) \mathcal{L}[V_{\mathbf{v}}] \tag{5.28}$$

as per Theorem 3.3.4 provides that, for any $\mathbf{u} \subseteq [M]$,

$$V_{\mathbf{u}} = \sum_{\mathbf{v} \subseteq \mathbf{u}} \tilde{V}_{\mathbf{v}}, \tag{5.29}$$

hence,

$$\tilde{V}_{\mathbf{u}} = V_{\mathbf{u}} - \sum_{\mathbf{v} \subset \mathbf{u}} \tilde{V}_{\mathbf{v}}. \tag{5.30}$$

This is in no way novel, although it is more usually done in the other direction. That is, we could equally well have started with (5.30) and obtained (5.25); this is explicitly referenced in [SDS09], cf. [DFS04], and the derivation of [KC16, (14)] can be understood similarly.[3]

[3] We mention here also a discussion by others on the Physics Stack Exchange forum on the subject of the MBE. One response [Kor22] provides a very interesting and quite technical and general derivation of what is essentially (5.25), starting from essentially (5.29). Without going into details, the derivation involves what is recognisably a special case of the linear-algebraic construction of the Möbius function that we mentioned in Section 3.5.3; cf. in particular [NW78, Chap. 26]. Hence, although the term is not explicitly used there, the derivation in [Kor22] can also be viewed as an application of Möbius inversion.

This equivalence of definition is also a key feature of the ANOVA-like decompositions of [Kuo+09] discussed just previously. It is also made clear in [CGH18], which simply states the equivalence of contribution potentials defined as in (5.25) and (5.30) above. In this last, reference is made to the PIE, rather than explicitly to Möbius inversion, but as mentioned previously, the former can be understood as just a special case of the latter [Sta12].

No matter which definitional direction we choose, equation (5.29) provides an exact decomposition of any $V_{\mathbf{u}}$, and of $V_{[M]}$ in particular. We will call this decomposition the *nuclear many-body expansion* of $V_{[M]}$ according to the family of (nuclear) subproblem potentials $\{V_{\mathbf{u}}\}_{\mathbf{u}\subseteq[M]}$:

$$V_{[M]} = \sum_{\mathbf{u}\subseteq[M]} \tilde{V}_{\mathbf{u}}. \tag{5.31}$$

Now let $I \in J(B_M)$ be an order ideal of B_M. Then, completely consistently with definitions in Chapter 3 that were made in the context of the order-theoretic combination technique, we call the sum

$$S_I := \sum_{\mathbf{u}\in I} \tilde{V}_{\mathbf{u}} = \sum_{\mathbf{u}\in I} D^{(I)}_{\mathbf{u}} V_{\mathbf{u}} \tag{5.32}$$

the *I-truncation*, or just a *truncation*, of the nuclear MBE of $V_{[M]}$, where the *combination coefficient* $D^{(I)}_{\mathbf{u}}$ for each $\mathbf{u} \in B_M$ is

$$D^{(I)}_{\mathbf{u}} = \sum_{\substack{\mathbf{v}\in I\\ \mathbf{v}\supseteq\mathbf{u}}} \mu(\mathbf{u},\mathbf{v}) = \sum_{\substack{\mathbf{v}\in I\\ \mathbf{v}\supseteq\mathbf{u}}} (-1)^{|\mathbf{v}-\mathbf{u}|}. \tag{5.33}$$

It is to be stressed that here, the general order-theoretic construction leads only to a specific form which is already known; in particular, note that the preceeding expression for $D^{(I)}_{\mathbf{u}}$ is just that used in the FCR of König and Christiansen [KC16], for an order ideal I identical to the downward-closed set which they write as {FCR}. As we remarked in Example 3.3.13, the derivation of this expression in [KC16] can be viewed as an independent reformulation of Möbius inversion, limited to B_M. Indeed, most of the earlier versions of the identities that we rederived in Example 3.3.13, e.g., those in [KC06; RH12; RH13; KC16], were explicitly constructed in the setting of truncated MBEs; the only exception, that of [Kuo+09], was in the context of the ANOVA-like decompositions just mentioned.

The construction of the nuclear MBE given above is easy to extend to match the more common disjoint-fragment setup. Specifically, let $\{F_i \subseteq [M]\}_{i=1}^K$ be a partition of $[M]$, that is, a family of K non-empty sets such that $F_i \cap F_j = \emptyset$ for $1 \leq i \neq j \leq M$, and also that $\bigcup_{i=1}^K F_i = [M]$. Then we call each F_i a *fragment* of the complete index set $[M]$, and the partition $\{F_i\}_{i=1}^K$ is called a *fragmentation* of $[M]$.

Let $F = \{F_{\mathbf{u}} := \bigcup_{i \in \mathbf{u}} F_i \mid \mathbf{u} \subseteq [K]\}$ be the set of all possible k-fold unions of fragments for $0 \leq k \leq K$. Note in particular that $F_{\emptyset} = \emptyset$ and $F_{[K]} = [M]$. We recall from [Sta12] that a *subposet* of some poset P is a poset $Q \subseteq P$ where, given $s, t \in Q$, then $s \leq_Q t$ implies $s \leq_P t$, and also vice versa. Ordered by set inclusion, F is then a subposet of B_M. We will call F the *fragmentation poset* associated with the particular fragmentation of $[M]$ under consideration. F is easily seen to be isomorphic to B_K, the boolean algebra of rank K, with the required bijection provided by $\phi(\mathbf{u} \in B_K) := F_{\mathbf{u}} \in F$. Moreover, the Möbius function of F is equivalent to that of B_K:

$$\mu_F(F_{\mathbf{v}}, F_{\mathbf{u}}) = \mu_{B_K}\big(\phi^{-1}(F_{\mathbf{v}}), \phi^{-1}(F_{\mathbf{u}})\big) = \mu_{B_K}(\mathbf{v}, \mathbf{u}) = (-1)^{|\mathbf{u}-\mathbf{v}|}. \tag{5.34}$$

It will be relevant in what follows that F, B_M, and B_K are examples of particular kinds of poset. We recall some more basic definitions here; see again [Sta12], also, e.g., [AK16]. Given a poset P, if two elements $s, t \in P$ have a greatest lower bound in P, it is named their *meet* and written $s \wedge t$. Their least upper bound is their *join*, written $s \vee t$. If $s \wedge t$ is well-defined for any s, $t \in P$, then P is a *meet semilattice*; a *join semilattice* is defined similarly. If P is both a meet and a join semilattice, then it is a *lattice*. If P is finite and has a $\hat{1}$, that is, a unique maximal element, then if it is a meet semilattice, it must also be a lattice [Sta12, Prop. 3.3.1]. Any boolean algebra B_n is a lattice, with meets and joins provided by intersections and unions respectively. Since obviously $\mathbf{u} \cap \mathbf{v} = \mathbf{w}$ for $\mathbf{u}, \mathbf{v}, \mathbf{w} \in B_K$ if and only if $F_{\mathbf{u}} \cap F_{\mathbf{v}} = F_{\mathbf{w}}$, it follows in particular that F is a meet semilattice, with meets provided by intersections.

Now, given a family of fragment subproblem potentials $\{V_{F_{\mathbf{u}}} : (\mathbb{R}^3 \times \mathbb{N})^M \to \mathbb{R}\}_{\mathbf{u} \subseteq [K]}$, we can obtain a corresponding family of fragment contribution potentials $\{\tilde{V}_{F_{\mathbf{u}}} : (\mathbb{R}^3 \times \mathbb{N})^M \to \mathbb{R}\}_{\mathbf{u} \subseteq [K]}$ via a definition equivalent to any one of (5.25), (5.29), or (5.30), just as in the nuclear case. Then

$$V_{F_{[K]}} = \sum_{F_{\mathbf{u}} \in F} \tilde{V}_{F_{\mathbf{u}}} = \sum_{\mathbf{u} \subseteq [K]} \tilde{V}_{F_{\mathbf{u}}}, \tag{5.35}$$

which, since $V_{[M]} = V_{F_{[K]}}$ by definition, we call the *fragment many-body expansion* of $V_{[M]}$ according to the family of fragment subproblem potentials.

It is in order to compare nuclear and fragment MBEs that we have not taken the more usual approach of defining the fragment potentials, either subproblem or contribution, to be functions of composite fragment variables; see and cf., e.g., [HMS70; RH12; KC16; HK21; Kor22]. To avoid confusion in what follows, we will use circumflexes to distinguish between the truncations and combination coefficients S_I and $D_{\mathbf{u}}^{(I)}$ of a nuclear MBE and those $\hat{S}_I$ and $\hat{D}_{F_{\mathbf{u}}}^{(I)}$ of a fragment MBE, which can be defined in terms of order ideals $I \in J(F)$ equivalently as in the nuclear case.

The combination coefficients of any truncation of a fragment MBE in terms of a molecular system can be precisely related to those of a truncation of the corresponding nuclear MBE. From an order-theoretic perspective, this relationship is essentially a special case of one previously observed by Lafuente and Cuesta [LC05] in the context

of the LFMT. A little explanation is required to make the connection clear in our setting. Like those authors, we use a fundamental theorem due to Rota [Rot64], which we recite verbatim in the form given by Stanley [Sta12]. Here, the meet of some subset $X = \{t_1, \ldots, t_n\}$ of a finite lattice L is obviously $\bigwedge X = \bigwedge_{i=1}^{n} t_i = t_1 \wedge \cdots \wedge t_n$, with $\bigwedge \emptyset$ vacuously $\hat{1}$; see, e.g., [AK16].

Theorem 5.2.2 (Crosscut theorem; verbatim from [Sta12])**.** *Let L be a finite lattice, and let X be a subset of L such that (a) $\hat{1} \notin X$, and (b) if $s \in L$ and $s \neq \hat{1}$, then $s \leq t$ for some $t \in X$. Then*

$$\mu(\hat{0}, \hat{1}) = \sum_{k} (-1)^k N_k, \tag{5.36}$$

where N_k is the number of k-subsets of X whose meet is $\hat{0}$.

Proof. See [Sta12, Cor. 3.9.4]. □

The following corollary is also used in [LC05]; it can be attributed to Hall [Hal36; Gre82] but we recite it also verbatim from [Sta12]. Here, the *coatoms* of a finite lattice L are the members of the set $\{p \in L \mid p \prec \hat{1}\}$.

Corollary 5.2.3 (Verbatim from [Sta12])**.** *If L is a finite lattice for which $\hat{0}$ is not a meet of coatoms, then $\mu(\hat{0}, \hat{1}) = 0$.*

Proof. See [Sta12, Cor. 3.9.5]. □

In short and omitting much interesting detail, Lafuente and Cuesta [LC05] consider a cluster expansion of a density functional defined for a multicomponent lattice model. Here, "lattice" refers to a point lattice such as, e.g., $\mathbb{Z}$ or $\mathbb{Z}^2$, rather than the order-theoretic construction. Using essentially their notation, given a generally infinite point lattice $\mathcal{L}$, each of multiple *clusters* are defined as $\boldsymbol{\mathcal{C}} = (\mathcal{C}_1, \mathcal{C}_2, \ldots, \mathcal{C}_p)$, where p is the number of particle types considered by the model and each $\mathcal{C}_i \subseteq \mathcal{L}$; also, notationally, $\boldsymbol{\mathcal{L}} = (\mathcal{L}, \mathcal{L}, \ldots, \mathcal{L})$. The poset of all clusters is partially ordered by componentwise set inclusion.

As per [LC05, Thm 2], given a subposet of clusters $\mathcal{W}$ with $\boldsymbol{\mathcal{L}} \in \mathcal{W}$, an application of Möbius inversion is used to approximately decompose a full-lattice density functional $\mathcal{F}_{\mathcal{L}}[\boldsymbol{\rho}]$ in terms of functionals $\mathcal{F}_{\mathcal{C}}[\boldsymbol{\rho}]$ defined with respect to individual clusters. Specifically, adapting slightly from [LC05, (3.7) and (3.18)],

$$\mathcal{F}_{\boldsymbol{\mathcal{L}}}[\boldsymbol{\rho}] \approx \sum_{\boldsymbol{\mathcal{C}} \in \mathcal{W} - \{\boldsymbol{\mathcal{L}}\}} [-\mu_{\mathcal{W}}(\boldsymbol{\mathcal{C}}, \boldsymbol{\mathcal{L}})] \mathcal{F}_{\boldsymbol{\mathcal{C}}}[\boldsymbol{\rho}]. \tag{5.37}$$

The key link to our setting is that the RHS is just, in our terminology, an I-truncation (3.22) of a decomposition of $\mathcal{F}_{\boldsymbol{\mathcal{L}}}[\boldsymbol{\rho}]$ in terms of $\mathcal{W}$, with the truncation taken according to the order ideal $I = \mathcal{W} - \{\boldsymbol{\mathcal{L}}\}$. This follows immediately by noting that, for each $\boldsymbol{\mathcal{C}} \in \mathcal{W} - \{\boldsymbol{\mathcal{L}}\}$,

we have $D_{\boldsymbol{\mathcal{C}}}^{(I)} = \sum_{\boldsymbol{\mathcal{C}} \leq \boldsymbol{\mathcal{C}}' \neq \boldsymbol{\mathcal{L}}} \mu_{\mathcal{W}}(\boldsymbol{\mathcal{C}}, \boldsymbol{\mathcal{C}}') = -\mu_{\mathcal{W}}(\boldsymbol{\mathcal{C}}, \boldsymbol{\mathcal{L}})$ directly by the definition of the Möbius function, (3.14). This viewpoint is very convenient and we shall make repeated use of it below.[4]

As constructed in [LC05], the set $\mathcal{W} = \{\boldsymbol{\mathcal{L}}\} \cup \{\bigcap S \,|\, \emptyset \subset S \subseteq W_{\max}\}$, where intersection is defined componentwise and $\mathcal{W}_{\max}$ is a set of somehow maximal clusters. The meaning of "maximal" here corresponds to the specific lattice model under study, but formally, $\mathcal{W}_{\max}$ could be any arbitrary antichain of the full cluster poset. Considering also a poset which we write in generator notation (see here Section 3.5.5 and transitively [Sta12]) as $\mathcal{V} = \langle \mathcal{W}_{\max} \rangle \cup \{\boldsymbol{\mathcal{L}}\}$, where the order ideal is generated in the full poset of clusters, it is observed in [LC05, App.] that $[\boldsymbol{\mathcal{C}}, \boldsymbol{\mathcal{L}}]_{\mathcal{V}}$ is an order-theoretic lattice for any $\boldsymbol{\mathcal{C}} \in \mathcal{V}$. Since Theorem 5.2.2 and Corollary 5.2.3 then provide that $\mu_{\mathcal{V}}(\boldsymbol{\mathcal{C}}, \boldsymbol{\mathcal{L}}) = \mu_{\mathcal{W}}(\boldsymbol{\mathcal{C}}, \boldsymbol{\mathcal{L}})$ if $\boldsymbol{\mathcal{C}} \in \mathcal{W}$, and $\mu_{\mathcal{V}}(\boldsymbol{\mathcal{C}}, \boldsymbol{\mathcal{L}}) = 0$ if not, respectively, it is further noted that the decomposition (5.37) could equally well have been formed in terms of $\langle \mathcal{W}_{\max} \rangle$, rather than just intersections of subsets of $\mathcal{W}_{\max}$.

Thus, although we state and explicitly prove the following proposition and subsequent corollary directly in the MBE setting, both the core observation and the ideas of proof are just those of [LC05]. Indeed, this can be viewed as a special case of their result, since the poset of all possible clusters considered in [LC05] is a direct product of p possibly-infinite boolean algebras, while we consider here only one finite boolean algebra. As mentioned in Footnote 4, the idea of rephrasing the combination coefficients by adjoining a $\hat{1}$ to an order ideal I is basically that on [Sta12, p. 265].

Proposition 5.2.4. *Let $[M]$ be some set of nuclear indices, and let $\{F_i\}_{i=1}^{K}$ be a fragmentation of $[M]$ with the associated fragment poset $F \cong B_K$. Further, let I' be an*

[4] This alternative expression for the combination coefficients emerges here simply because Lafuente and Cuesta take a slightly different formal approach in the application of Möbius inversion than that which we used to construct the order-theoretic combination technique in Chapter 3. We demonstrate in the more general notation of the order-theoretic combination technique, but still following the basic form of [LC05], and using an idea that can be found on [Sta12, p. 265] to which we will return shortly. Given some poset grid Π, a subposet $I \subseteq \Pi$ — here not necessarily an order ideal of Π — is essentially fixed *a priori*, and an explicit $\hat{1}$ is adjoined to it, even if I already has a unique maximal element. An additional model function $f_{\hat{1}}$ is then defined to be the target function f. Möbius inversion of $f = f_{\hat{1}} = \sum_{\mathbf{p} \in I \cup \{\hat{1}\}} \tilde{f}_{\mathbf{p}}$ provides for the case $\mathbf{p} = \hat{1}$ the expression $\tilde{f}_{\hat{1}} = \sum_{\mathbf{p} \leq \hat{1}} \mu(\mathbf{p}, \hat{1}) f_{\mathbf{p}}$, and thence $f_{\hat{1}} = \tilde{f}_{\hat{1}} + \sum_{\mathbf{p} < \hat{1}} [-\mu(\mathbf{p}, \hat{1})] f_{\mathbf{p}}$. The term $\tilde{f}_{\hat{1}}$ is treated as an error quantity. In some cases considered in [LC05], properties of the model functions $f_{\mathbf{p}}$ and the particular choice of I can be used to show that this term vanishes and the expansion is exact. In general, this is not so, and omitting $\tilde{f}_{\hat{1}}$ makes (5.37) an approximation.

Alternatively, let us reimpose in the above that I is an order ideal with combination sum S_I, and define instead $f_{\hat{1}} = 0$. Then Möbius inversion leads firstly to $\tilde{f}_{\hat{1}} = -S_I$, and secondly to $\tilde{f}_{\hat{1}} = \sum_{s < \hat{1}} \mu(\mathbf{p}, \hat{1}) f_{\mathbf{p}}$. This also serves to justify the alternative expression for the combination coefficients.

Finally, we mention that the idea from [Sta12] referenced above appears in discussion relating Möbius inversion to the cardinality PIE. There, effectively the RHS of (5.12) is written as a sum that is, of course up to setting, basically the same shape as (5.37). This is very closely related to the construction of the combination coefficients for an arbitrary meet semilattice in [HGC07, Sec. 3.1], which is founded upon a different expression of the PIE again.

arbitrary order ideal of F. Then there exists an order ideal I of B_M such that

$$D^{(I)}_{\mathbf{u}\in B_M} = \begin{cases} \hat{D}^{(I')}_{\mathbf{u}\in F} & \text{if } \mathbf{u}\in F\text{, i.e., if there exists } \mathbf{v}\in B_K \text{ such that } \mathbf{u}=F_{\mathbf{v}}, \\ 0 & \text{otherwise.} \end{cases} \tag{5.38}$$

Proof. Fix some I' as in the statement of the claim, and let A' be the set of maximal elements of I'. Each $\mathbf{a}\in A'\subseteq F$ is also a member of B_M, so let $I=\langle A'\rangle_{B_M}$ be the order ideal generated by those elements in B_M. Define J to be I with an explicit additional element $\hat{1}_J$ adjoined, such that $\hat{1}_J >_J \mathbf{u}$ for all $\mathbf{u}\in I$, and define J' similarly for I'. Clearly, both J and J' are lattices. Also, by (3.14) as noted above, $D^{(I)}_{\mathbf{u}} = -\mu_J(\mathbf{u},\hat{1}_J)$ for arbitrary $\mathbf{u}\in I$, and similarly, $\hat{D}^{(I')}_{F_{\mathbf{u}}} = -\mu_{J'}(F_{\mathbf{u}},\hat{1}_{J'})$ for arbitrary $F_{\mathbf{u}}\in I'$.

Fix an arbitrary $\mathbf{u}\in I$, and note, as in [LC05], that the interval $[\mathbf{u},\hat{1}_J]$ in J is itself a lattice, with $\mathbf{u}$ as $\hat{0}$. Since the elements of A' are all in F, the setwise meet of any subset of them is also in F. So, if $\mathbf{u}$ is a meet of coatoms of J, then also $\mathbf{u}\in F$, and furthermore, since $[\mathbf{u},\hat{1}_{J'}]$ in J' is a lattice, $D^{(I)}_{\mathbf{u}} = \hat{D}^{(I')}_{\mathbf{u}}$, by Theorem 5.2.2 with $X=A'$. If $\mathbf{u}$ is not a meet of coatoms of J, then $D^{(I)}_{\mathbf{u}}=0$, by Corollary 5.2.3. If in this case $\mathbf{u}\in F$, then also $\hat{D}^{(I')}_{\mathbf{u}}=0$, which is sufficient to show (5.38). □

Corollary 5.2.5. *Let $[M]$, $\{F_i\}_{i=1}^K$, and F be as for Proposition 5.2.4. Further, let $\{V_{\mathbf{u}}\}_{\mathbf{u}\subseteq[M]}$ be a family of nuclear subproblem potentials for the nuclear subsets $\mathbf{u}\subseteq[M]$, and let $\{V'_{F_{\mathbf{u}}} := V_{F_{\mathbf{u}}}\}_{\mathbf{u}\subseteq[K]}$ be a family of fragment subproblem potentials defined in terms of that nuclear family. Then for every order ideal I' of F, there exists some order ideal I of B_M such that $\hat{S}_{I'}=S_I$.*

Proof. Fix an order ideal I' of F, and let I be any corresponding order ideal of B_M as provided by Proposition 5.2.4. Then

$$S_I = \sum_{\mathbf{u}\in I} D^{(I)}_{\mathbf{u}} V_{\mathbf{u}} = \sum_{\mathbf{u}\in I'} \hat{D}^{(I')}_{\mathbf{u}} V_{\mathbf{u}} = \sum_{F_{\mathbf{u}}\in I'} \hat{D}^{(I')}_{F_{\mathbf{u}}} V'_{F_{\mathbf{u}}} = \hat{S}_{I'}, \tag{5.39}$$

by definition, by Proposition 5.2.4, by construction, and by definition, respectively. □

In effect, Corollary 5.2.5 allows any truncation of any fragment MBE to be identified as a truncation of an underlying nuclear MBE, if the family of nuclear subproblem potentials used to construct that nuclear MBE is an extension of the family of fragment subproblem potentials. From an intuitive perspective, this is of course completely unsurprising.

To see one practical application of this, consider in inverse a family of nuclear subproblem potentials which are always well-defined in theory, but some of which are somehow problematic in practice. For example, if the evaluation of some $V_{\mathbf{u}}$ requires the introduction of multiple link atoms to terminate dangling bonds, and if some or all of those link atoms are spatially close to each other, then the value $V_{\mathbf{u}}$ might carry an unrepresentative bias due to their interaction, and that bias would carry into any nuclear truncation

S_I over an order ideal I where $D_{\mathbf{u}}^{(I)} \neq 0$. This difficulty has been anticipated in the construction of existing subsystem methods; see, e.g., discussion in [Das+02; DC05, Sec. II.D.1; CD06, Sec. II.B; MR12; CB15; See+22]. The general response suggested by these methods is to somehow just remove problematic fragments from consideration. So, consistent with this idea and with a particular eye to an adaptive approach, if we can construct a fragmentation of the nuclear indices such that the evaluation of any $V_{F_{\mathbf{u}}}$ either does not require link atoms, or only introduces them separated by a sufficiently large distance, then the set $\{\hat{S}_I \mid I \in J(F)\}$ of truncations of the fragment MBE can be viewed as a "safe" subset of the set of nuclear truncations $\{S_I \mid I \in J(B_M)\}$, since any such bias is automatically excluded.

Moreover, from a theoretical level, since such a fragment truncation will still be a truncation of the underlying nuclear MBE, any properties of that nuclear MBE that we have either assumed or that we have been able to prove will still hold. For example, this can be used to view any truncation of an appropriate fragment MBE as a truncation of the explicitly nuclear-focused Hamiltonian-based ANOVA-like decomposition of V^{BO} sketched in the previous section, without the need for any further definition. This may be helpful in deeper analysis of particular MBEs.

Since this equivalency of nuclear and fragment truncations occurs because each nuclear combination coefficient $D_{\mathbf{u}}^{(I)}$ is either zero or equal to the corresponding fragment combination coefficient $\hat{D}_{\mathbf{u}}^{(I')}$, we will say that such a fragment MBE is *combination-consistent* with the underlying nuclear MBE. If we use this terminology, Lafuente and Cuesta noted in [LC05], in effect, that their decomposition (5.37), defined in terms of the order ideal $\mathcal{W} - \{\boldsymbol{\mathcal{L}}\}$ of the subposet $\mathcal{W}$ produced by the effectively arbitrary $\mathcal{W}_{\max}$, is combination-consistent with one defined in terms of the order ideal $\langle \mathcal{W}_{\max} \rangle$ of the full poset of clusters. Since this property is important for the order-theoretic combination technique well beyond the standard MBE (or indeed LFMT) setting, we will fix a very general definition of combination-consistency, and extend Proposition 5.2.4 accordingly.

Definition 5.2.6 (Combination-consistency)**.** Let P be a locally finite poset, and let Q be a subposet of P. For any finite order ideal I of P and any $s \in I$, let

$$D_s^{(I)} = \sum_{\substack{t \in I \\ t \geq s}} \mu_P(s,t) \tag{5.40}$$

be the combination coefficient of s in I.[5] For any finite order ideal I' of Q and $s \in Q$, define $\hat{D}_s^{(I')}$ equivalently. For any two such finite order ideals I of P and I' of Q, if

$$D_s^{(I)} = \begin{cases} \hat{D}_s^{(I')} & \text{if } s \in Q, \\ 0 & \text{otherwise} \end{cases} \tag{5.41}$$

[5]Note that this is a slight generalisation of Definition 3.3.7, which requires P to have a $\hat{0}$.

for every $s \in P$, then I and I' are *combination-consistent*. If such an I exists for every finite order ideal I' of Q, then Q is a *combination-consistent* subposet of P.

We first confirm, without explicitly using the crosscut theorem or requiring either P or Q to be a lattice, that given some finite order ideal $I' \in J_f(Q)$, if there exists any order ideal $I \in J_f(P)$ which is combination-consistent with I', then it must be that generated in P by the maximal elements of I'.

Lemma 5.2.7. *Let P be a locally finite poset, let Q be a combination-consistent subposet of P, and let $I' \subseteq Q$ be an arbitrary finite order ideal of Q. If I is a finite order ideal of P which is combination-consistent with I', then $I = \langle A' \rangle_P$, where $A' \subseteq I'$ is the antichain of all maximal elements of I'.*

Proof. Let P, Q, I', and A' be as in the statement of the claim. Further, let I be some finite order ideal of P which is combination-consistent with I'. Since I is finite, it is generated by some antichain $A \subseteq I$.

The proof is by contradiction. Suppose that $A \neq A'$. Then there is either some $a \in A$ such that $a \notin A'$, or some $a' \in A'$ such that $a' \notin A$. Begin by supposing the former. Since a is maximal in I, we have $D_a^{(I)} = \mu_P(a, a) = 1$ by definition. If $a \notin Q$, then by (5.41), $D_a^{(I)} = 0$, a contradiction. So $a \in Q$. It must be that $a \in I'$, for otherwise, $\hat{D}_a^{(I')} = 0$, again a contradiction. Since $a \notin A'$, there exists some $a^\dagger \in A'$ such that $a^\dagger > a$. Just as above, $\hat{D}_{a^\dagger}^{(I')} = 1$, but since a is maximal in I, we know that $a^\dagger$ cannot also be an element of I, so also $D_{a^\dagger}^{(I)} = 0$, a contradiction.

It must then be that there exists $a' \in A'$ such that $a' \notin A$. Just as above, $\hat{D}_{a'}^{(I')} = 1$. If $a' \notin I$, then $D_{a'}^{(I)} = 0$, a contradiction, so $a' \in I$. Then there exists some $a^\dagger \in A$ such that $a^\dagger > a'$, and $D_{a^\dagger}^{(I)} = 1$. Since this is nonzero, it must be that $a^\dagger \in Q$, by (5.41). But since $\hat{D}_{a^\dagger}^{(I')}$ is nonzero only when $a^\dagger \in I'$, we have that a' is not maximal in I', a contradiction. □

We need now one more definition. Let P be a meet semilattice, and let Q be a subposet of P such that Q is closed under $\wedge_P$, that is, if $t, t' \in Q$, then also $t \wedge_P t' \in Q$. Then we call Q a *meet subsemilattice* of P. Although [Aig97; Sta12] explicitly give definitions only for a *sublattice*, the more restricted version is also standard in the literature.

It is noted in [LC05] that the set $\mathcal{W}$ used in (5.37) is closed by direct construction under componentwise intersection; thus, $\mathcal{W}$ is a meet subsemilattice of the full poset of clusters. Similarly, $\mathcal{W}$ is a meet subsemilattice of $\langle \mathcal{W}_{\max} \rangle \cup \{\boldsymbol{\mathcal{L}}\}$. In the MBE setting, the fragmentation poset F is also a meet subsemilattice of B_M, again effectively by direct construction.

Meet semilattices provide a natural environment for combination techniques; see, in particular, the construction in [HGC07, Sec. 3.1], and discussion of [Won16] to follow below. We will show that, in the general case when P is a meet semilattice with a $\hat{0}$, in

order for some $Q \subseteq P$ to be fully combination-consistent with P, it is both necessary and sufficient that it be a meet subsemilattice of P. Although necessity in particular is not to our reading recognised explicitly in [LC05], this remains only a mechanical and slight extension of their work. Note also here that, from a purely order-theoretic perspective, this is an immediate and not especially interesting result, although we do not believe we have seen it stated explicitly elsewhere. We give it as a theorem only for its importance to the order-theoretic combination technique.

Theorem 5.2.8 (Combination-consistent subposets of certain meet semilattices)**.** *Let P be a locally finite meet semilattice with a $\hat{0}$, and let Q be a subposet of P. Then Q is combination-consistent with P if and only if Q is a meet subsemilattice of P.*

Proof. The proof of ($\Leftarrow$) is almost identical to that of Proposition 5.2.4 above; we repeat ourselves for completeness, again noting the previous work in [LC05; Sta12]. Let P be a locally finite meet semilattice with a $\hat{0}$, and let $Q \subseteq P$ be a meet subsemilattice of P. Let I' be an arbitrary finite order ideal of Q with generating antichain A', and let $I = \langle A' \rangle_P$ be the (necessarily finite) order ideal of P generated by A' in P.

Define $J := I \cup \{\hat{1}_J\}$ and $J' := I' \cup \{\hat{1}_{J'}\}$, such that $\hat{1}_J >_J t$ for any $t \in I$, and $\hat{1}_{J'} >_{J'} t'$ for any $t' \in I'$. Now fix an arbitrary $s \in I$. Consider $D_s^{(I)} = -\mu_J(s, \hat{1})$. If s is a meet of coatoms of J, then $D_s^{(I)} = -\mu_J(s, \hat{1}_J) = -\mu_{J'}(s, \hat{1}_{J'}) = \hat{D}_s^{(I')}$, by Theorem 5.2.2, with $X = A'$. If not, then $D_s^{(I)} = 0$ by Corollary 5.2.3; if here also $s \in Q$, then similarly $\hat{D}_s^{(I')} = 0$, and we are done.

In the other direction, ($\Rightarrow$), the proof is by contradiction. Let P be a locally finite meet semilattice with a $\hat{0}$, and let Q be a combination-consistent subposet of P. Suppose that Q is, however, not a meet subsemilattice of P. Then there exist two distinct elements $t, t' \in Q$ such that $t \wedge_P t' \notin Q$. Let $I' = \langle t, t' \rangle_Q$ be the finite order ideal of Q generated by those two elements.

Since Q is combination-consistent with P, there exists a finite order ideal I of P that is combination-consistent with I'. By Lemma 5.2.7, this $I = \langle t, t' \rangle_P$. Form $J := I \cup \{\hat{1}_J\}$, as above. Since the only subset of $\{t, t'\}$ whose meet is $t \wedge_P t'$ is $\{t, t'\}$ itself, it follows from Theorem 5.2.2 that $D^{(I)}_{t \wedge_P t'} = -\mu_J(t \wedge_P t', \hat{1}_J) = 1$. But, since I is combination-consistent with I', and since $t \wedge_P t' \notin Q$, also $D^{(I)}_{t \wedge_P t'} = 0$, a contradiction. □

The ramifications of this result in the broader order-theoretic combination technique setting are basically those of Proposition 5.2.4 as applied to MBEs. As mentioned above, the most natural choices for individual poset axes seem to often be lattices [Heg03], and thus meet semilattices; cf. again [HGC07, Sec. 3.1]. It is easy to see that any direct product of meet semilattices is also a meet semilattice. Suppose, then, that one constructs a theoretically well-behaved and -understood poset model hierarchy over a meet-semilattice grid Π that nevertheless exhibits undesirable computational characteristics for particular models, or is perhaps too sprawling for an adaptive algorithm to explore

efficiently. Theorem 5.2.8 provides a precise description of exactly when the order-theoretic combination technique can be applied instead to a smaller subposet of the original poset grid, safe in the knowledge that any resulting I-truncation will be exactly the same as one which might have been encountered in the original. Note that while this viewpoint is very closely related to that of [LC05], the work there is instead more focused on obtaining a single exact decomposition that is, effectively, provably minimal in the sense of [Kuo+09].

We take a brief detour here to draw a connection with a generalised construction of the combination coefficients for the standard combination technique as given by Wong [Won16, Chap. 3], which we mentioned briefly in Section 3.2 above. We give an abbreviated and slightly adapted sketch for the purposes of comparison. In the original, Wong considers finite sets C of particular finite grids $g_{\mathbf{m}} \subset [0,1]^d$, each of which latter is indexed by an element $\mathbf{m} \in \mathbb{N}^d$. For simplicity and consistency with our setting, we speak in adapted terms rather of finite subsets $C \subset \mathbb{N}^d$. The involved ideas translate directly, cf., e.g., [Won16, Assumption 3.2.4 and Prop. 3.2.6], and so we still follow Wong's development very closely and reuse much of his notation with only slight modification, if any. Each set C must be closed under meets, that is, $\mathbf{m} \wedge \mathbf{n} = (\ldots, \min(m_i, n_i), \ldots) \in C$ for $\mathbf{m}$, $\mathbf{n} \in C$, and is thus a meet subsemilattice of $\mathbb{N}^d$; cf. [Won16, Def. 3.2.7]. Fixing now some particular such C, the set ∂C is defined to contain exactly those *meet-irreducible* elements $\mathbf{m} \in C$ such that $\mathbf{m} = \mathbf{n} \wedge \mathbf{n}'$ for $\mathbf{n}$, $\mathbf{n}' \in C$ only when either $\mathbf{m} = \mathbf{n}$ or $\mathbf{m} = \mathbf{n}'$, consistent with [Won16, Def. 3.2.19]. After defining $\mathfrak{F}(\mathbf{m} \in C) := \{F \subseteq \partial C \mid \bigwedge F = \mathbf{m}\}$ as per [Won16, Def. 3.2.22], the *inclusion/exclusion coefficient* for each $\mathbf{m} \in C$ is, from [Won16, Def. 3.2.25],

$$c_{\mathbf{m}} := \sum_{F \in \mathfrak{F}(\mathbf{m})} (-1)^{|F|+1}. \tag{5.42}$$

As is suggested by the name, this construction is motivated by an inclusion/exclusion argument, explicitly connected to the cardinality PIE; see [Won16, Sec. 3.2.2]. Wong shows in [Won16, Thm. 3.5.4] that these combination coefficients match those of the standard combination technique, as given in, e.g., Section 3.1 above.

To see the connection to our order-theoretic combination technique, consider some non-empty and finite meet subsemilattice $\emptyset \neq C \subset \mathbb{N}^d$, and then write $I = \langle C \rangle = \{\mathbf{m} \in \mathbb{N}^d \mid \exists \mathbf{n} \in C \text{ s.t. } \mathbf{m} \leq \mathbf{n}\}$, by a slight abuse of our usual notation for an antichain-generated order ideal. Clearly, since I is an order ideal, it is also a meet subsemilattice, since $\mathbf{m} \wedge \mathbf{n} \leq \mathbf{m}$ and $\mathbf{m} \wedge \mathbf{n} \leq \mathbf{n}$ for any $\mathbf{m}, \mathbf{n} \in \mathbb{N}^d$. Further, it has a zero, $\hat{0} = \mathbf{0}$. Adjoin an explicit one to I as above, $J := I \cup \{\hat{1}_J\}$. Now fix some $\mathbf{m} \in C$, and write $\partial_{\mathbf{m}} C := \{\mathbf{n} \in \partial C \mid \mathbf{n} \geq \mathbf{m}\}$. Every $F \in \mathfrak{F}(\mathbf{m})$ is then clearly a subset of $\partial_{\mathbf{m}} C$. In fact, $\partial_{\mathbf{m}} C$ is a subset $X \subset [\mathbf{m}, \hat{1}_J]_J$ of the form required by the preconditions of Theorem 5.2.2, and the definition of the inclusion/exclusion coefficient $c_{\mathbf{m}}$ in (5.42) is, up to trivial manipulation, precisely how we would calculate $-\mu_J(\mathbf{m}, \hat{1}_J) = D^{(I)}_{\mathbf{m}}$ according to (5.36).

Again, cf. here the discussion relating Möbius inversion to the PIE on [Sta12, p. 265], and also the construction in [HGC07, Sec. 3.1].

Wong's argument can be easily extended to construct inclusion/exclusion coefficients in terms of any finite meet subsemilattice C of an arbitrary poset grid Π which is a meet semilattice, and not only for those of $\mathbb{N}^d$. Moreover, an equivalent appeal to the crosscut theorem shows that those coefficients so obtained will match those obtained via a Möbius function-based definition for the order ideal $I = \langle C \rangle_\Pi$. Although this viewpoint may be helpful for future theoretical analysis of combination sums in particular problem settings, it should be clear from the discussion of combination-consistency above that there is no real difference in terms of the available combination sums themselves.

We return now to the MBE setting, and to the counting arguments in Section 5.2.1. Here, we can revisit some previous observations in the fragmentation-method literature in the context of the order-theoretic construction and in light of the above discussion.

The basic observation that the energy equations of many fragmentation methods reduce to particularly-truncated MBEs goes back at least to Suárez et al. [SDS09]. In [CB15], Collins and Bettens state that energy equations like those in Section 5.2.1 above "can be constructed by conventional many-body expansion methods through the inclusion and exclusion of various higher order many-body interaction terms" [CB15, p. 5622]. As evidence, they give the equivalence of the MOBE-like energy expression provided by the *combined fragment method* (CFM) [Le+12] with a particular truncated MBE involving terms of up to fourth order. A similar observation can also be found in [RS15].

Although such an equivalence is heavily implicit in the work of Herbert and co-workers regarding the GMBE, e.g., [RH12; RH13; LH16], it does not seem to our reading to be stated there explicitly. König and Christiansen argue in [KC16] that the terms of any GMBE energy equation should emerge as, in their terminology, a particular FCR summation, or in ours, a suitably-chosen I-truncation of an MBE. Further practical demonstration of this is given in [HK21]. We are now equipped to confirm (or at least reconfirm) this with full rigour in the general case. Let us be clear that, from a combinatorial perspective, we apply here what reduces fundamentally to just a differently-organised version of the same argument that Stanley uses on [Sta12, p. 265] to connect Möbius inversion with the cardinality PIE, and we use in particular an effectively identical poset construction as there.

We build this poset in our notation and setting as follows. Consider some family of distinct, non-empty, and now potentially-overlapping fragments $\{F_i \subseteq [M]\}_{i=1}^K$. Define $F_{\cap \mathbf{u}} := \bigcap_{i \in \mathbf{u}} F_i$ for each non-empty $\mathbf{u} \subseteq [K]$, and then let $\hat{F} := \{F_{\cap \mathbf{u}} \,|\, \emptyset \subset \mathbf{u} \subseteq [K]\} \cup \{\emptyset\}$; that is, denote by $\hat{F}$ the set of all possible k-fold intersections of fragments F_i for $1 \leq k \leq K$, with the empty set always explicitly included. By construction, $\hat{F}$ is a subposet of B_M that is closed under intersection.[6] So is the subposet $\hat{F}^+ := \hat{F} \cup \{[M]\}$

[6] And indeed, if the set of fragments $\{F_i \subseteq [M]\}_{i=1}^K$ is an antichain, it is completely formally equivalent to the set $\mathcal{W}_{\max}$ of [LC05]; either way, the decomposition (5.43) is clearly analogous to (5.37).

formed by adjoining $[M]$ to $\hat{F}$ if it is not already present. An exact decomposition

$$V^{\mathrm{BO}} = \sum_{\hat{\mathbf{u}} \in \hat{F}^+} \tilde{V}_{\mathbf{u}}^{(\hat{F}^+)}, \tag{5.43}$$

can then be defined in terms of some family of subproblem potentials $\{V_{\hat{\mathbf{u}}}^{(\hat{F}^+)}\}_{\hat{\mathbf{u}} \in \hat{F}^+}$ as usual, either by slight extension of the definitions given earlier in this section or by directly using the equivalent machinery in Chapter 3, along with I-truncations of this decomposition and their involved combination coefficients. Since $\hat{F}$ is trivially an order ideal of $\hat{F}^+$, there exists some order ideal of B_M that is combination-consistent with $\hat{F}$.

Consider now the energy equation (5.15) for the GMBE; for notational ease, we use K rather than K' for the total number of n-mers. This equation can be rewritten

$$E_{(n)}^{\mathrm{GMBE}} = \sum_{\emptyset \subset \mathbf{u} \subseteq [K]} (-1)^{|\mathbf{u}|+1} E_{F_{\cap \mathbf{u}}} = \sum_{\substack{\hat{\mathbf{u}} \in \hat{F} \\ \hat{\mathbf{u}} \neq \emptyset}} d_{\hat{\mathbf{u}}} E_{\hat{\mathbf{u}}}; \tag{5.44}$$

this is just a rephrasing of [LH16, (14)], up to an additional correction term to which we will return shortly. Here, we consider $\hat{F}$ to be defined in terms of the set of GMBE n-mers, and define each $d_{\hat{\mathbf{u}}}$ to be the sum of the ± 1 coefficients in front of each term $(-1)^{|\mathbf{u}|+1} E_{F_{\cap \mathbf{u}}}$ for all $\mathbf{u}$ such that $F_{\cap \mathbf{u}} = \hat{\mathbf{u}}$, consistent with [RH12; RH13; LH16].

We claim that each $d_{\hat{\mathbf{u}}}$ in (5.44) is just and exactly the combination coefficient of $\hat{\mathbf{u}}$ in the $\hat{F}$-truncation of (5.43), which we remind the reader is not necessarily either a standard nuclear or fragment MBE. To see this, begin by fixing some non-empty $\hat{\mathbf{u}} \in \hat{F}$. There must be some maximal $1 \leq k_{\hat{\mathbf{u}}} \leq K$ such that $\hat{\mathbf{u}}$ can be written as a $k_{\hat{\mathbf{u}}}$-fold intersection of fragments F_i; that is, there exists some $\mathbf{u} \subseteq [K]$ such that $\hat{\mathbf{u}} = F_{\cap \mathbf{u}}$, and $|\mathbf{u}| =: k_{\hat{\mathbf{u}}}$ is as large as possible. This $\mathbf{u}$ must be unique, for were it not, then we could obtain $\hat{\mathbf{u}}$ as the intersection of some $k' > k_{\hat{\mathbf{u}}}$ fragments, contradicting the maximality of $k_{\hat{\mathbf{u}}}$.

Now, each non-empty subset $\mathbf{v} \subseteq \mathbf{u}$ corresponds to some $\hat{\mathbf{v}} = F_{\cap \mathbf{v}} \in \hat{F}$. Clearly, $\hat{\mathbf{v}} \geq_{\hat{F}} \hat{\mathbf{u}}$, for $\hat{\mathbf{u}}$ is a subset of each F_i for $i \in \mathbf{u}$ by construction. Moreover, every $\hat{\mathbf{v}}' \geq_{\hat{F}} \hat{\mathbf{u}}$ must appear as $F_{\cap \mathbf{v}'}$ for at least one non-empty $\mathbf{v}' \subseteq [k_{\hat{\mathbf{u}}}]$. Thus, the leading coefficient in each term $(-1)^{\mathbf{v}+1} E_{F_{\cap \mathbf{v}}}$ must be counted in exactly one $d_{\hat{\mathbf{v}}}$ for some $\hat{\mathbf{v}} \geq_{\hat{F}} \hat{\mathbf{u}}$. It follows that

$$\sum_{\hat{\mathbf{v}} \geq_{\hat{F}} \hat{\mathbf{u}}} d_{\hat{\mathbf{v}}} = \sum_{j=1}^{k_{\hat{\mathbf{u}}}} (-1)^{j+1} \binom{k_{\hat{\mathbf{u}}}}{j} = 1 - \sum_{j=0}^{k_{\hat{\mathbf{u}}}} (-1)^{j} \binom{k_{\hat{\mathbf{u}}}}{j} = 1, \tag{5.45}$$

where the first equality is obtained by separating non-empty subsets $\mathbf{v} \subseteq \mathbf{u}$ by cardinality, and the last by application of (3.26).

We introduce now another standard idea from order theory: the poset P^* derived from another poset P by reversal of order, that is, such that $s \leq_P t$ if and only if $s \geq_{P^*} t$,

is called the *dual* of the original poset P [Rot64; Aig97; Sta12]. It can be shown that $\mu_P(s,t) = \mu_{P^*}(t,s)$ for all $s,t \in P$ [Sta12]. Möbius inversion of (5.45) in terms of $\hat{F}^*$ provides, then, that

$$d_{\hat{\mathbf{u}}} = \sum_{\hat{\mathbf{v}} \geq_{\hat{F}^*} \mathbf{u}} \mu_{\hat{F}^*}(\hat{\mathbf{v}}, \hat{\mathbf{u}}) = \sum_{\hat{\mathbf{v}} \geq_{\hat{F}} \hat{\mathbf{u}}} \mu_{\hat{F}}(\hat{\mathbf{u}}, \hat{\mathbf{v}}), \tag{5.46}$$

which is precisely how we would define the combination coefficient of $\hat{\mathbf{u}}$ in the $\hat{F}$-truncation of (5.43) according to (3.23). So the claim holds. As a result, and assuming for the moment that the subproblem potential $V_{\emptyset}^{(\hat{F}^+)}$ is everywhere zero, the GMBE energy equation (5.15) can be viewed as being the $\hat{F}$-truncation of (5.43), and also and simultaneously a truncation of both a standard nuclear MBE in terms of the order ideal generated by the maximal elements of $\hat{F}$ in B_M, or of a fragment MBE in terms of a similarly-generated order ideal of B_K, just as anticipated in [KC16].

Echoing an observation made at least by Mayhall and Raghavachari in the context of the MOBE [MR12], and also one in [CB15], we point out that this equivalence also works inversely. Since any order ideal I of B_M is trivially seen to be closed under intersection, it can be used directly as the set of potentially-overlapping fragments that are fed into (5.15). In this case, $\hat{F} = I$, and the above confirms the obviously desirable property that this GMBE energy corresponds exactly to the I-truncation of an MBE using appropriately-chosen contribution potentials.

We highlight here one subtle difference between the FCR constructions in [KC16; HK21] and the order-theoretic one just used. Mixing our terminology with theirs, the FCR approach begins by considering a full MBE-style expansion of V^{BO}, truncated in terms of an order ideal $I = \{\mathrm{FCR}\} \subseteq B_M$. As combination coefficients in (5.11) are established to be zero, the relevant subsets $\mathbf{f}_l$ are discarded from {FCR}, leading to an *effective FCR*; however, coefficients must still be calculated for all terms in {FCR}, even if some subproblem potential evaluations can be subsequently avoided. By contrast, we construct an ANOVA-like expansion of V^{BO} directly in terms of some potentially much smaller subposet of B_M, pick an order ideal of that poset, and rely on Theorem 5.2.8 to guarantee that the other terms in the relevant order ideal have zero coefficients. In the following chapters, we will use similar subposets to supply our adaptive algorithm with what is effectively a smaller search space than the full boolean algebra B_M.

In any case, the MBE/GMBE equivalence just demonstrated also applies effectively unchanged to, e.g., the CG-MTA energy expression (5.13). Equivalence to the overcounting-corrected GEBF equation (5.18), or to [LH16, (14)], requires slightly more attention to be paid to $V_{\emptyset}$. Setting $V_{\emptyset} = 0$ is obviously the natural choice for a family of subproblem potentials that model, e.g., the total energies of subsystems $\mathbf{u} \subseteq [M]$ when treated in complete isolation. Liu and Herbert [LH16] considered instead the situation of what we would call a family of electrostatic-embedding subproblem potentials, defined as for the GEBF in Section 5.2.1. Again by appeal to a PIE-style decomposition of the full-system Hamiltonian, they point out that the correct definition for the empty-set potential should

be, in our notation,

$$V_\emptyset = \sum_{A \leq B} \frac{q_A q_B}{\|R_A - R_B\|}. \tag{5.47}$$

Considering the terms that we wrote as $d_{\hat{\mathbf{u}}}$ in (5.44), and reasoning that "the sum of all coefficients in the PIE equals unity" [LH16, p. 575], Liu and Herbert conclude that, again in our notation, $d_\emptyset = 1 - \sum_{\emptyset \neq \hat{\mathbf{u}} \in \hat{F}} d_{\hat{\mathbf{u}}}$, which suffices to recover (5.18) as the $\hat{F}$-truncation of (5.43) for electrostatic-embedding subproblem potentials.

A slight generalisation of Liu and Herbert's expression for $d_\emptyset$ is found in a "top-down" recursive expression for the FCR combination coefficients introduced in [HK21]. In our notation, this reads $D^{(I)}_{\mathbf{u}} = 1 - \sum_{\mathbf{u} \subset \mathbf{v} \in I} D^{(I)}_{\mathbf{v}}$, where I is any order ideal of B_M and $\mathbf{u} \in I$. This expression is justified in [HK21] by an inductive-style counting argument. A significantly more general version again, $D^{(I)}_s = 1 - \sum_{s < t \in I} D^{(I)}_t$ for any combination coefficient $D^{(I)}_s$ for any element s in any finite order ideal I of any poset P, is trivially obtained by dual-form Möbius inversion of (5.40); here, observe the connection to (5.45) and (5.46), and cf. [Won16, Prop. 3.2.20] in the specific case of the standard combination technique.

Thus, although the argument of [LH16] just mentioned was for the specific case of electrostatic-embedding subproblem potentials used in the GMBE setting, it also informally motivates the following idea. If any family of embedding-style subproblem potentials $V_{\mathbf{u}}$ is constructed over any meet subsemilattice of B_M, such that each subproblem potential $V_{\mathbf{u}}$ improves on a coarser treatment of the full-system energy for the particular set of atoms $\mathbf{u}$ than is embodied in $V_\emptyset$, then any I-truncation includes an automatic "overcounting correction" in the form of $D^{(I)}_\emptyset$. These subproblem potentials could, for example, apply some form of quantum embedding rather than a simple electrostatic-embedding approach. We will come back very briefly in Section 5.4 to some existing works applying such potentials, but we note without deeper exploration that the formal treatment of overcounting in these seems to us generally somewhat different, see, e.g., explicit discussion in [BAM12; SJ20].

5.3. Numerical condition of the many-body expansion

The accuracy and numerical stability of several versions and variations of the MBE has been thoroughly assessed in a series of papers by Herbert and co-workers [RLH14; Lao+16; LH17]; we highlight here one particular aspect of their study. In [RLH14], Richard et al. performed effectively an informal assessment of the numerical condition of n-body truncations of an MBE using a propagation-of-errors analysis. Their results show that while small propagated uncertainties compound quite gently for two-body MBE truncations as the number of fragments increases, those for third- and higher-order truncations become progressively and prohibitively more superlinear in the same;

see [RLH14, Fig. 5(a)]. This "calls into question the assumption that the n-body expansion is systematically improvable as a function of n" [RLH14, pp. 7–8].

Using essentially the notation of [RLH14], if f is a d-dimensional function in a family of uncorrelated variables $\{x_i\}_{i=1}^{d}$, each of which is certain only to within some $\pm \mathrm{d}x_i$, then the overall uncertainty of the function is given by [RLH14, (4.2)]

$$\mathrm{d}f = \sqrt{\sum_{i=1}^{d} \left(\frac{\mathrm{d}f}{\mathrm{d}x_i} \right)^2 (\,\mathrm{d}x_i)^2}. \tag{5.48}$$

In the context of the MBE, and mixing in now our notation, the dimensionality d is the total number of terms in a truncation S_I, with each "variable" x_i being an evaluated subproblem potential, $x_i = V_{\mathbf{u}} \in \mathbb{R}$ for some $\mathbf{u} \in I$. The uncertainties $\mathrm{d}x_i$ stem in particular from the fact that *ab initio* methods generally involve an iterative component, in which some property is refined until a representative error estimate shrinks below some convergence threshold; see again [RLH14]. The derivative terms $\mathrm{d}f / \mathrm{d}x_i$ are just the combination coefficients $D_{\mathbf{u}}^{(I)}$; here, Richard et al. use the equivalent expression to (3.37) that we mentioned in Example 3.3.13.

The structure of our adaptive combination technique algorithm proves to be useful in this context. Recall from Chapter 3 that our algorithm does not calculate individual contribution values directly; rather, it manipulates sparse tensors, which store individual Möbius function evaluations and their sums. As a result, at every stage of the adaptive refinement process, we know — explicitly and by design — the combination coefficients for every term of the candidate combination sum. Thus, if the results for individual calculation results can be ascribed an inherent uncertainty, then we can also evaluate the total propagated uncertainty of that combination sum according to (5.48), with no more effort than is required to evaluate the combination sum itself. This technique is not limited to MBE-style combinations, and applies without modification to any truncation of a combination sum taken over an arbitrary poset grid.

The impact of compounding uncertainty on the calculation of S_I is found in [RLH14] to be more pressing than that of raw floating-point error. Nevertheless, those authors note that double-precision arithmetic might not always suffice in the K-fragment MBE case for larger K; cf. [RLH14, Fig. 7]. In the implementation of the adaptive order-theoretic combination technique algorithm used throughout this thesis, the summation of S_I is handled using arbitrary-precision arithmetic [Joh17], using 100 bits of intermediate precision; see Appendix A.8. This latter number carries no inherent meaning in the MBE context. But since it is only a little less than the 113 bits [Mul+18, Tab 3.1] that would be provided were we to use the quadruple-precision arithmetic suggested as prophylaxis in [RLH14], we will assume in the case studies that follow that possible floating-point error can be effectively disregarded.

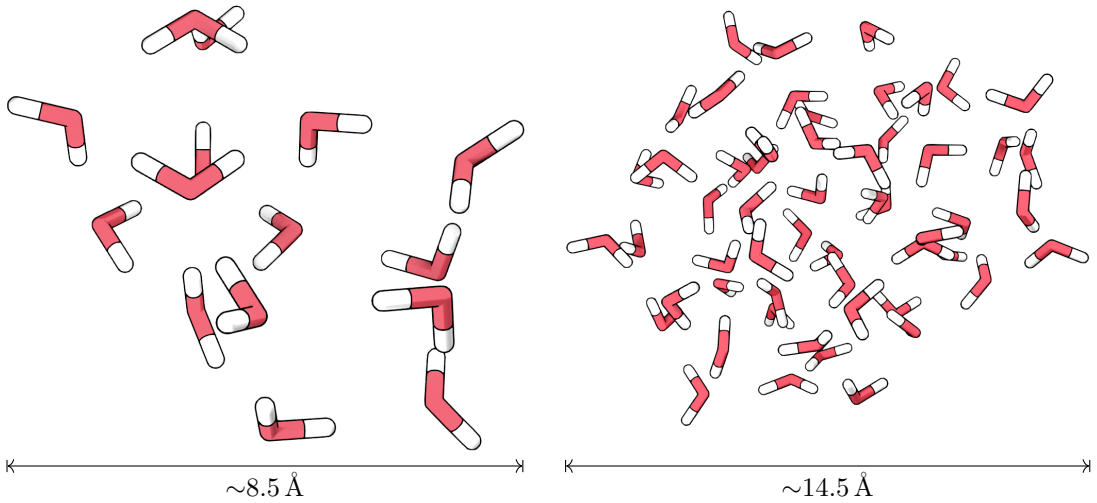

Figure 5.1.: Stick-model visualisations of water clusters $(H_2O)_{15}$ and $(H_2O)_{55}$. Geometries are drawn from [KT13], as described in the main text.

5.4. Case study: water clusters

At the beginning of the chapter, we stated that we can use our order-theoretic combination technique to calculate truncations of various MBEs and MBE-like sums. We will now make this application concrete, and investigate the behaviour of the resulting *adaptive many-body expansion* when used to approximate the total energies of two *water clusters*, that is, collections of water molecules. Water clusters have a long history of study from the perspectives of both MBE-type approximations and subsystem techniques in general [HMS70; Xan94; DT06; Pru+12; PL15; Cis+16; Che+17; HX20]. Our purpose here is not to investigate or benchmark the standard MBE as applied to water clusters *per se*; this has been done authoritatively by Herbert and co-workers [RLH14; Lao+16; LH17; LH19]. Instead, guided at a high level by their work, we ask whether the application of adaptivity may be helpful, particularly in light of the numerical and cost-related issues mentioned in Section 5.3 above.

We have previously mentioned that both the adaptive algorithm we use and the order-theoretic combination technique more generally have their origin in the ML-BOSSANOVA approach [Heb14; GHH14; CGH18]. In fact, the adaptive formulation in [CGH18] is explicitly given in terms of a powerset $2^{[M]}$, and so can be viewed formally also as an adaptive MBE, albeit one where — unlike here — most of the contribution potentials vanish exactly. We shall elaborate on this in the following chapter, and it will become clear why we have returned to a more traditional MBE form first rather than trying to

apply the BOSSANOVA approach immediately. Techniques also exist for explicit *a priori* selection of included MBE terms based on particular screening criteria [LH17; LH19], and as we shall discuss in the following chapters, various connectivity-based fragmentation methods can be understood as implicit examples of the same. In particular, the method of Iyengar and co-workers [RHI18; RI18; KI19] has been referred to as providing an "adaptive-many-body-expansion" [KI19, inset graphic, p. 5769], but their approach is only adaptive in the sense that the set of potentially-overlapping fragments they use is a direct function of the spatial layout of the system, up to some adjustable parameters; see discussion in Section 7.1 below. Notions of adaptivity are also applied in a construction by Artiukhin et al. [Art+20] that is related to the FCR method; here, the adaptivity is in terms of the set of sampling coordinates used to assemble a fitted approximation to the Born-Oppenheimer potential energy surface. But beyond ML-BOSSANOVA, we are unaware of any other adaptive schemes for selecting terms from a standard MBE in a quasi-optimal way.

We consider the two water clusters $(H_2O)_{15}$ and $(H_2O)_{55}$. Approximately equilibrium-state geometries for these clusters were obtained from [KT13]; specifically, we used the `w15[AMOEBA]` and `w55[AMOEBA]` structures given in the supplementary material of [KT13]. Visualisations of these clusters are given in Figure 5.1. For each cluster, we obtained a reference total energy via an all-electron MP2 calculation [MP34; SO89] using the cc-pCVTZ [Dun89; WD95] basis set. These calculations were performed with NWChem [Apr+20], with an RHF iterative convergence threshold set to $10^{-9}\,E_\mathrm{h}$ and ERI prescreening thresholds for both RHF and MP2 set to $10^{-14}\,E_\mathrm{h}$. This represents only a lightweight treatment of electron correlation, particular in view of the extremely high quality treatments considered in the previous chapter. However, these are not entirely trivial calculations, particularly in the case of $(H_2O)_{55}$, which when equipped according to cc-pCVTZ presents a problem involving 3905 contracted atomic orbitals.

In the terminology of the order-theoretic combination technique, the target function f which we want to approximate is the charge-neutral MP2/cc-pCVTZ Born-Oppenheimer potential energy function $V^{\mathrm{BO}}(X_1, \ldots, X_M)$, where M is the total number of atoms in each cluster, so $M = 45$ and $M = 165$ for $(H_2O)_{15}$ and $(H_2O)_{55}$ respectively. We use a single-axis poset grid $\Pi = B_{M'}$, where M' is the number of water molecules in each cluster, so $M' = 15$ and $M' = 55$. The implementation of the poset axis interface required to use this axis is given in Section B.2.

We construct the obvious fragmentation $\{F_i\}_{i=1}^{M'}$ of the nuclear indices, where each F_i corresponds to the indices of one of the M' water molecules in each cluster. The model hierarchy is then taken to be a family of subproblem potentials, $\{V_{F_\mathbf{u}}\}_{\mathbf{u}\in B_{M'}}$, where $F_\mathbf{u} = \bigcup_{i\in\mathbf{u}} F_i$. The property evaluation functional $\mathcal{L}$ is as usual point evaluation as per the cluster geometries. The arguments made in this chapter make it clear that, given any order ideal $I \subseteq B_{M'}$, the resulting combination sum S_I is an I-truncation of the fragment MBE of $V^{\mathrm{BO}} = V_{[M']}$ in terms of the fragmentation $\{F_i\}_{i=1}^{M'}$.

We consider four different families of subproblem potentials. The members of the first

family, $\{V^{\text{vac}}_{F_\mathbf{u}}\}_{\mathbf{u}\subseteq[M']}$, are simply standard MP2/cc-pCVTZ Born-Oppenheimer potential energy functions, and so deliver the total energy of the isolated system specified by the subfamily of the nuclear variables $\{X_A\}_{A\in F_\mathbf{u}}$. Although no explicit embedding scheme is applied here, we will refer to members of this family as *vacuum embedding* subproblem potentials, following terminology in [Hég+16]. We set $V^{\text{vac}}_\emptyset$ to be identically zero.

The second and third families of subproblem potentials, written $\{V^{\text{EE-xTB}}_{F_\mathbf{u}}\}_{\mathbf{u}\subseteq[M']}$ and $\{V^{\text{EE-IAO}}_{F_\mathbf{u}}\}_{\mathbf{u}\subseteq[M']}$, also calculate the MP2/cc-pCVTZ total energies of charge-neutral systems formed from subfamilies of the nuclear variables, but use electrostatic-embedding Hamiltonians defined for subsets of certain sets of point charges $\{q^{\text{xTB}}\}_{A=1}^M$ and $\{q^{\text{IAO}}\}_{A=1}^M$ respectively, following the EE-MB idea of [DT06; DT07b]. For reasons that will shortly become clear, we will call members of these families *xTB electrostatic embedding* potentials and *IAO electrostatic embedding* potentials respectively. The Coulomb interaction energy (5.17) of all used point charges is explicitly included in each such potential, and the empty-fragment potentials are taken to be the complete sum of all distinct Coulomb interactions between the point charges, that is,

$$V^{\text{EE-xTB}}_\emptyset = \sum_{A<B}^{M} \frac{q^{\text{xTB}}_A q^{\text{xTB}}_B}{\|R_A - R_B\|}, \tag{5.49}$$

and equivalently for $V^{\text{EE-IAO}}_\emptyset$. This is consistent with discussion in Section 5.2.3 above, and also, e.g., [LH16]. We note that this goes against the advice of [RLH14]; nevertheless, the point charges used for all subproblem potential evaluations were equivalent up to full double precision.

Each q^{xTB}_A is the Mulliken partial charge [Mul55] for atom A provided by an appropriate full-system calculation using the GFN2-xTB semi-empirical method [BEG19; Ban+20]. Similarly, each q^{IAO}_A is a partial charge calculated according to the intrinsic atomic orbitals of Knizia [Kni13], obtained from a reference full-system RHF wavefunction according to the cc-pVTZ basis set [Dun89] and calculated using PySCF [Sun15; Sun+17; Sun+20]. We will not describe the GFN2-xTB model here, observing only that it carries a formal scaling as $\mathcal{O}(M^3)$ in the number of atoms [Ban+20]. In practice, however, this scaling carries a dramatically smaller prefactor than, e.g., a Hartree-Fock calculation, so the xTB charges can be viewed as a relatively affordable estimate of the electrostatic environment that might feasibly be obtained for any medium-size system. The calculation of the IAO charges, by contrast, requires a full-system Hartree-Fock calculation, and the use of at least a triple-zeta basis set seems to be suggested by [Kni13, Tab. 1]. We do not suggest that IAO charges should be used in general; we consider them in order to investigate results obtained using electrostatic embedding under a "best-case" scenario, since the semi-empirical nature of GFN2-xTB in conjunction with potential issues with Mulliken partial charges [Her19] suggests caution in the use of the xTB charges; cf., however, comments in [RLH14].

The final family of subproblem potentials that we consider, $\{V^{\mathrm{mix}}_{F_{\mathbf{u}}}\}_{\mathbf{u}\subseteq[M']}$, are *mixed-basis embedding* potentials. For a deeper discussion of mixed-basis embeddings, see, e.g., [HNK18], but the idea is straightforward and well-known [JG91]. The evaluation of any $V^{\mathrm{mix}}_{F_{\mathbf{u}}}$ delivers a total energy obtained by a complete full-system MP2 calculation. When performing this calculation, the atoms indexed by $F_{\mathbf{u}}$ are equipped with atomic orbitals according to cc-pCVTZ, which in context we call the *embedding basis*. The remaining atoms are equipped only with the Dunning-Hay DZ basis set [DH77], called the *environment basis*.[7] The empty-fragment potential $V^{\mathrm{mix}}_{\emptyset}$ thus delivers the total energy according to a standard full-system MP2 calculation using the DZ basis set. Such an approach is closely related to the DCMB method [WX12] and the EMFT [For+15]; indeed, a mixed-basis DFT calculation can be viewed as a special case of EMFT [Lee+17]. We note here that there is a body of previous work investigating the use of quantum embeddings and related techniques to obtain terms of MBEs; see, e.g., [Man+12; BAM12; GW12; SJ20; SJ21], and particularly [VLH19] for an approach explicitly applying the EMFT. Although we will not attempt to engage with or compare against this work for reasons of space, we observe that this work generally considers expansions truncated after only second- or third-order terms,[8] and adaptivity is not used. Moreover, to our knowledge, simple mixed-basis embeddings treating the complete system with a correlated wavefunction theory such as MP2 have not been previously investigated in the MBE context.

We use our standard abstract cost model to represent the cost of all subproblem potential evaluations. For calculation details on all subproblem potentials, see Section A.7. For completeness, we mention a practical aspect of the adaptive calculations described here and throughout the remainder of this thesis (with the particular exception of that described in Section 6.8.2). As discussed in Section A.7, we carefully cached and reused the results of subproblem potential evaluations between adaptive calculations. This allowed us to repeat and extend adaptive calculations in a resilient manner, making efficient use of available computational resources and backed by progressively larger caches of evaluated subproblem potentials. Thus, although we will consider the adaptive calculations discussed below as being notionally standalone, each draws underlying evaluated subproblem potentials from what is effectively a precalculated dataset. Since the adaptive algorithm is fully deterministic, and since we consider only the abstract costs

[7] The reader may wonder why we have chosen DZ for the environment basis set, rather than, say, a minimal and thus cheaper basis set such as STO-2G [HSP69]. During the preparation of this work, we did indeed experiment with the use of minimal environment basis sets, but found that they performed poorly. We can only speculate informally about this behaviour. Mixed-basis calculations are known to lead to irregularities in the electronic density [Lee+17; HNK18]. Since the constructions of the STO-nG and cc-pCVnZ monoatomic basis functions are quite different, we hypothesise that this leads to particularly pronounced irregularities. However, the DZ set is effectively an unpolarised variant of cc-pVDZ, so the effect of increasing basis-set quality in particular regions should be correspondingly less disruptive. This would be consistent with comments and observations in, e.g., [Lee+17; HNK18].

[8] Up to four-body in [SJ20; SJ21], but there only for small systems.

of evaluations and not their real-world wall-time costs, the use of precached evaluations has no impact on the final results as given here.

We did not explicitly modify the cost model to account for the use of electrostatic-embedding Hamiltonians, either for the calculation of the point charges themselves or their inclusion in the one-electron term. Thus, in what follows, each electrostatic-embedding potential "costs" exactly the same as an equivalent vacuum embedding potential. We justify this decision for reasons of simplicity, and by our anecdotal observation that the true costs of evaluating the vacuum potentials and the electrostatic-embedding potentials are in these cases indeed comparable, although this would not necessarily be expected to hold were the sizes of the systems and thus the numbers of point charges to grow significantly. The real-world computational costs of each full-system evaluation of the xTB point charges were here negligible; the costs of each evaluation of the IAO point charges were not.

The evaluations of the mixed-basis subproblem potentials are by nature much more expensive than those for either the vacuum or electrostatic-embedding potentials, since their costs scale in the full-system size rather than the size of the fragment. We have no reason to believe that a cost benefit will be obtained relative to the other potential types; indeed, we expect mixed-basis potentials to rapidly become infeasible for large systems, and consider them only out of interest with respect to accuracy.

For each type of subproblem potential, we report per-iteration results for two adaptive calculations, one each using the All and Threshold strategies. Results are given up to termination thresholds defined in terms of cumulative abstract cost. For calculations over $(H_2O)_{15}$, the termination threshold was chosen to be the abstract cost of the reference calculation multiplied by 100; for $(H_2O)_{55}$, the abstract cost of the reference calculation multipled by 10. This latter was further restricted for All calculations with mixed-basis potentials for $(H_2O)_{55}$; see below.

The nth "adaptive" refinement according to the All strategy is, of course, not truly adaptive, but adds $\binom{M'}{n}$ elements to the index set and produces a result which is exactly equivalent to an n-body truncation of a fragment MBE. By contrast, index set growth according to the Threshold strategy will depend on the precise value of the threshold parameter α. For simplicity, and since we are interested at first in knowing whether true adaptivity can be of any value at all, we sidestep the question of how one might best choose such a value *a priori*. Instead, we consider results only for the particular value $\alpha = 0.5$, which seems to introduce a relatively granular but usually non-trivial amount of new work at each adaptive iteration. A deeper analysis of behaviour obtained with different values of α is certainly important, but is left for future work.

As well as per-iteration approximations to the total energy, we also calculated at each iteration a propagated uncertainty, as per [RLH14] and as described in Section 5.3 above. Here, an uncertainty of $10^{-8}\,E_\mathrm{h}$ was assigned to the result of each subproblem potential evaluation, consistent with the RHF iterative convergence thresholds used for the set of underlying single-point calculations.

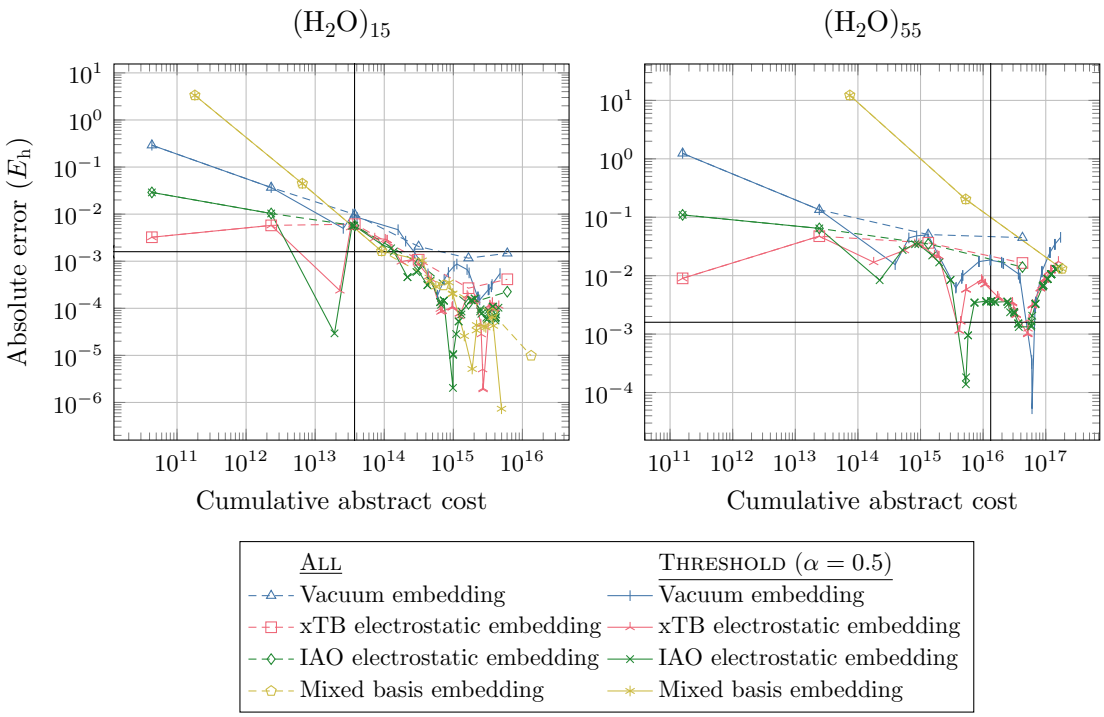

Figure 5.2.: Per-iteration absolute errors for progressively-refined adaptive fragment-MBE calculations over the water clusters $(H_2O)_{15}$ and $(H_2O)_{55}$. Errors are measured relative to reference MP2/cc-pCVTZ calculations, as described in the main text. The horizontal axis represents the total cumulative cost expended by the algorithm to calculate the index set at each iteration, according to our abstract cost model. Vertical and horizontal lines indicate the abstract cost of the reference calculation and chemical accuracy ($\approx 0.0016\,E_h$) respectively.

The absolute error obtained by adaptively-refined index sets relative to the reference calculations is plotted against their cost in Figure 5.2. The left-hand plot shows the results for $(H_2O)_{15}$, the right-hand for $(H_2O)_{55}$. Vertical and horizontal lines are shown indicating the reference cost and chemical accuracy respectively. The latter is not inherently meaningful here, since it is quite clear from Chapter 4 than an MP2/cc-pCVTZ calculation has no chance of being within chemical accuracy of the true FCI/CBS total energy; nevertheless, we include it in its traditional role as an accuracy benchmark. These lines allow an easy informal assessment of the quality of any adaptively-obtained approximation: if it offers an accurate approximation to the reference result at cheaper cost, then it must be plotted in the lower left-hand "quadrant".

Beginning with $(H_2O)_{15}$, we see first that all adaptive calculations considered here eventually produce approximations which are within chemical accuracy of the reference.

However, this does not reliably occur until the cost of each respective approximation is well past that of the reference calculation. The third iterations of both Threshold calculations for electrostatic-embedding potentials are both below chemical accuracy, but immediately return to errors of approximately $10^{-1}\,E_\mathrm{h}$ in the subsequent iteration. The fact that the respective All calculations do not show the same dip, and also that a smaller although less pronounced dip is also visible in the Threshold calculation using vacuum embedding potentials, suggests that this is probably not of any great interest.

The overall error/cost scaling of the various approximations do not dramatically differ from each other. The vacuum embedding potential calculations, both All and Threshold, appear to perform the worst, but the Threshold calculation is not far enough from the remaining group of results that we can conclude any inherent superiority of the latter over the former. Similarly, although the remaining non-vacuum potentials do produce results that venture far below chemical accuracy, these excursions are oscillatory and unpredictable. In all cases, the Threshold calculations generally outperform their All counterparts, suggesting that the adaptive algorithm successfully finds and incorporates more important contribution potentials; again, however, the scaling differences between the two are not so significant as to be remarkable.

The mixed-basis embedding approximations perform well, and are competitive in both accuracy and cost with the remainder, despite initial costs and errors that are much higher. Indeed, of all of the All results, it is those for mixed-basis potentials which appear to perform the best, and show the most consistent reduction of error at high cost. Nevertheless, all results for mixed-basis embeddings with costs that do not exceed the reference cost threshold carry errors well above $10^{-1}\,E_\mathrm{h}$.

We turn now to the results for $(H_2O)_{55}$, which are concerning. Although the calculations using vacuum and electrostatic-embedding potentials again pass briefly below chemical accuracy, they do not reliably stay there. Worse, the errors of all such calculations appear to reach minima between approximately $10^{-2}\,E_\mathrm{h}$ and $10^{-3}\,E_\mathrm{h}$, before deteriorating as they approach the point at which the calculations were terminated. This does not suggest optimism for the reliability of the MBE for larger systems, even with the application of adaptivity. Furthermore, although the electrostatic-embedding potentials do outperform the vacuum-embedding potentials, particularly in the Threshold cases, their results are still far from reliable. In short: both vacuum and electrostatic-embedding adaptive MBE calculations fail to provide useful approximations to the total energy of $(H_2O)_{55}$ at any measured cost.

The mixed-basis calculations for $(H_2O)_{55}$ do appear to be trending reliably downwards, but their evaluation was so expensive that we could only calculate them up to the inclusion of two-body terms. As a result, the Threshold and All calculations have no chance to differentiate themselves, and we are unable to make any clear further judgement. This highlights a general issue which we encountered during informal experiments applying adaptive MBE-type calculations to large systems. Specifically, zero- and one-body calculations produce large contribution potentials, but are inherently unable to capture

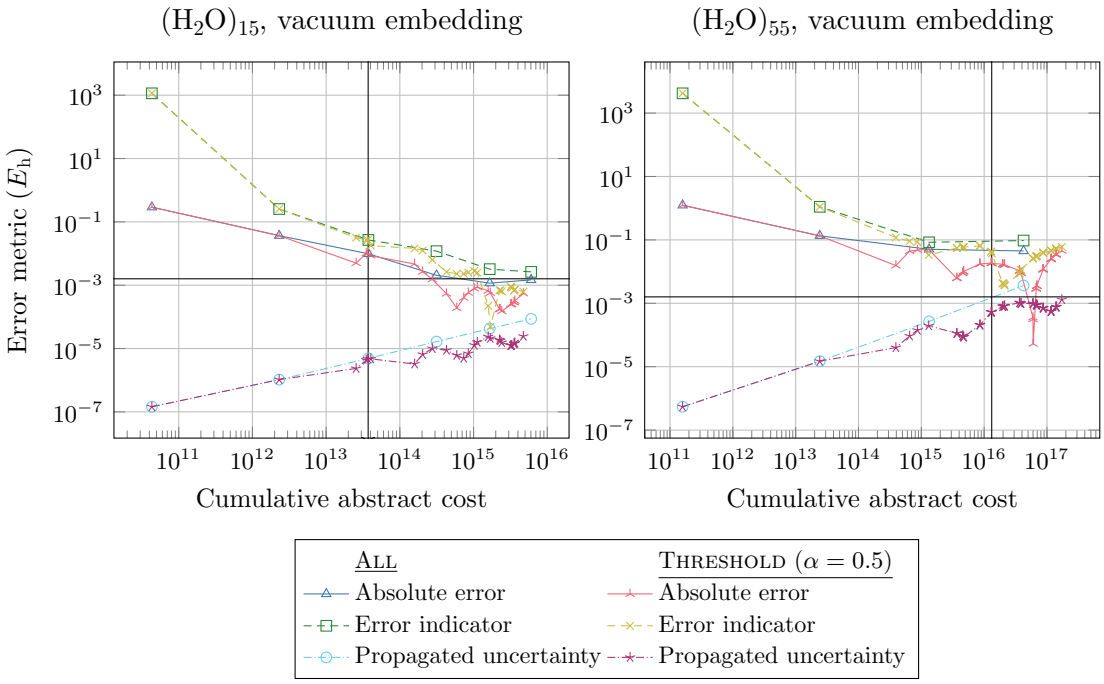

Figure 5.3.: Absolute errors, error indicators, and propagated uncertainties for progressively-refined adaptive fragment-MBE calculations over the water clusters $(H_2O)_{15}$ and $(H_2O)_{55}$ using vacuum embedding subproblem potentials. Absolute errors and costs are calculated as in Figure 5.2. Error indicators are calculated as described in Section 3.5.5. Propagated uncertainties are calculated as described in Section 5.3, with an assumed uncertainty of $10^{-8}\,E_h$ ascribed to the value of each individual subproblem potential. Vertical and horizontal lines indicate the abstract cost of the reference calculation and chemical accuracy ($\approx 0.0016\,E_h$) respectively, also as in Figure 5.2.

any meaningful spatial information from the system. As a result and in practice, at least all two-body terms must be calculated before an adaptive algorithm can have sufficient information for a meaningful exploration of the remaining space of possible terms.

We consider now the behaviour of the two error metrics available to the adaptive calculations: the adaptive error indicator outlined in Section 3.5.5, and the propagation-of-uncertainty analysis discussed in Section 5.3. For simplicity, we consider only the calculations involving vacuum embedding potentials. The plots in Figure 5.3 show again the absolute errors for the relevant All and Threshold calculations for both considered water clusters. Also plotted in each case are the values of the error indicator, and the propagated uncertainty in the final result, according to (5.48), calculated with an assumed per-result uncertainty of $10^{-8}\,E_h$, which corresponds to the SCF convergence threshold

settings used for the subproblem potential calculations.

A positive first observation is that the error indicator is generally reliable in both calculations and for both types of adaptive refinement, consistently overestimating the error of the approximation by at most an order of magnitude. In both cases, the error estimators for the Threshold calculations do once dip below the true error, only to return to reliability quickly. These dips are attributable in both cases to a sign change of the underlying combination sum of evaluated subproblem potentials; since the error indicator updates by only a small amount at each iteration, it passes the true error closely during the transition between signs.

The propagated uncertainties of both calculations grow, which is as anticipated given the analysis in [RLH14]. The uncertainty for Threshold calculations is consistently below that of the All calculations; this is likely due to the fact that the combination coefficients of some subproblem potentials vanish, unlike those for All, which are always non-zero, as mentioned in Example 3.3.13. In the case of $(H_2O)_{55}$, there is a suggestion that the growth of the propagated uncertainty for the Threshold calculation may be slightly slower than that for the All calculation; however, both approach the chemical accuracy threshold as the cost of their respective calculations passes that of the reference calculation. It seems reasonable to conclude that this uncertainty may provide at least a partial explanation for the failure of the true errors of the $(H_2O)_{55}$ calculations to come near chemical accuracy, let alone fall below it.

We close our study of the adaptive MBE by a brief empirical investigation of the expected decays in the sizes of the evaluated contribution potentials $\mathcal{L}[\tilde{V}_{F_\mathbf{u}}]$, both in terms of the size of the subsystem $|F_\mathbf{u}|$, and in terms of the spatial organisations of the atoms in those subsystem. Similar and more detailed studies can be found in the literature. For example, Heindel and Xantheas considered the magnitudes of many-body contributions for several water clusters in [HX20], and basically concluded that the contribution potential terms become generally negligible beyond fourth order. Their work focused particularly on many-body corrections for a phenomenon called *basis set superposition error*, which we do not consider in this thesis for reasons of simplicity. We aim here not to compare with or extend on their work or similar, but rather only to obtain a quick and informal picture for a single molecular system of the relationship between the spatial dispersion of involved subsystems, the magnitudes of their resulting contribution potential, and the individual computational costs associated with them.

Here, we calculated all 32 768 possible contribution potentials $\mathcal{L}[\tilde{V}_{F_\mathbf{u}}]$ for the vacuum-embedding case of $(H_2O)_{15}$. We grouped the results by the total number of H_2O monomers involved in each subsystem, $k = |\mathbf{u}|$. The single contribution potential for $k = 0$ is zero by definition, and those for $k = 1$ are just the total energies of individual H_2O monomers. Two-dimensional histograms of some of the remaining values are plotted in Figure 5.4. Each histogram groups all values for some particular $k = |\mathbf{u}|$ with $2 \leq k \leq 10$. The vertical axis of each histogram corresponds to the base-10 logarithm of $\mathcal{L}[\tilde{V}_{F_\mathbf{u}}]$. The horizontal axis classifies the distance between each of the k H_2O monomers involved in

each $F_{\mathbf{u}}$, measured in terms of the central points calculated as $1/3(R_A + R_B + R_C)$ for the three nuclear spatial variables R_A, R_B, and R_C involved in each monomer. Naturally, we cannot compress the full geometric conformation of each subsystem into a single scalar value. Instead, we calculated both the mean and the maximum of the pairwise distances between the fragments, and then took the mean of these two values again. The intention here was to obtain a value that represents, very roughly, how "elongated" each collection of monomers is, while not ignoring its overall shape.

The bin counts in the histogram are normalised such that all bins for every k sum to unity. Each bin is coloured according to the base-10 logarithm of its normalised count; the same colour scale is used for all plots for all values k. We do not provide histograms for collections of values $11 \leq k \leq 15$ for reasons of space. However, the values for these do not deviate strongly from those for $k = 10$, and indeed seem to continue the pattern of "compression" into a small region that is visible in the data that is plotted.

Decays in the size of the terms $\mathcal{L}[\tilde{V}_{F_{\mathbf{u}}}]$ are certainly visible, both in terms of increasing $\mathbf{u}$ and increasing distance. These are strongest for the first few values of k plotted, although it is interesting to note that even some four-body terms still have magnitudes greater than chemical accuracy; this would seem to confirm previously-mentioned statements in the literature about the importance of some four- and five-body terms, although clearly not all of them. Once k passes six or seven, the decay in the largest of the evaluated terms seems to slow.

Perhaps more significantly, the number of terms with magnitudes that could be considered "negligible", in that they are less than the applied convergence tolerance threshold of $10^{-8}\,E_\mathrm{h}$, also decreases: the magnitudes seem to cluster between $10^{-6}\,E_\mathrm{h}$ and $10^{-8}\,E_\mathrm{h}$. We remark that this might also be understandable in some cases via a propagation-of-errors analysis, since the explicit calculation of each $\mathcal{L}[\tilde{V}_{F_{\mathbf{u}}}]$ requires the summation of $2^{\mathbf{u}}$ evaluated subproblem potential terms $\mathcal{L}[V_{F_{\mathbf{u}}}]$. Since the weights in this sum are always ± 1, it is easy to see that the propagated uncertainty will be, in the example case of a system with $k = 9$, just $\sqrt{2^9 \times 10^{-16}\,E_\mathrm{h}} \approx 2 \times 10^{-7}\,E_\mathrm{h}$. Of course, since these sums are not performed explicitly in the evaluation of most truncated MBE sums — at least, not when using our tensor-based formulation — this effect will have no direct impact on the accuracy of the same.

A decay in terms of distance is also obvious. This would seem to be clear and further confirmation that the importance of individual k-body terms can be related to the locality properties of the subsystems they represent, as has been widely remarked on and exploited in the literature mentioned earlier in the chapter. However, we note that there is still usually a large spread in the magnitudes of the plotted terms for any given choice of k and at any particular distance. It is difficult to characterise this spread further without a more detailed analysis, since the *ad hoc* distance metric we use here is not very descriptive. Nevertheless, it still seems reasonable to hope that some kind of adaptive approach should be able to locate structure here.

In closing, we remember that the adaptive index-set algorithm given in Chapter 3.5

implicitly requires, or at least hopes, that the sum of the ratios of the evaluated model functions to their calculation costs should also decay rapidly. A histogram of these benefit/cost ratios $|\mathcal{L}[\tilde{V}_{F_\mathbf{u}}]|/\mathcal{C}(\mathbf{u})$ in this case is given in Figure 5.5, with an equivalent format to that of Figure 5.4. The overall shape of the plotted distributions is basically the same as in the earlier figure; this is unsurprising, since the costs $\mathcal{C}(\mathbf{u})$ for all terms with $\mathbf{u} = k$ involve exactly the same number and species of atoms, and will be distinguished in our cost model only by the number of non-neglible ERIs that must be evaluated. Thus, this data can be viewed as basically a rescaling of the earlier-shown data.

However, the decays for these terms is in fact slightly stronger than that seen previously, particularly as k grows. Although we cannot do anything more rigorous with these values, this is, at least in principle, exactly what we would hope to see. Although the adaptive MBE approach we have tried here is clearly not of any real practical value in the form given, this is obviously just an expression of the same numerical issues that were anticipated in [RLH14; Lao+16]; put differently, even an adaptive index set must simply grow too much in order to gain too little. In the following chapter, we will consider how we might adjust the MBE to better exploit the locality effect on the contribution potentials that is so visible in Figures 5.4 and 5.5.

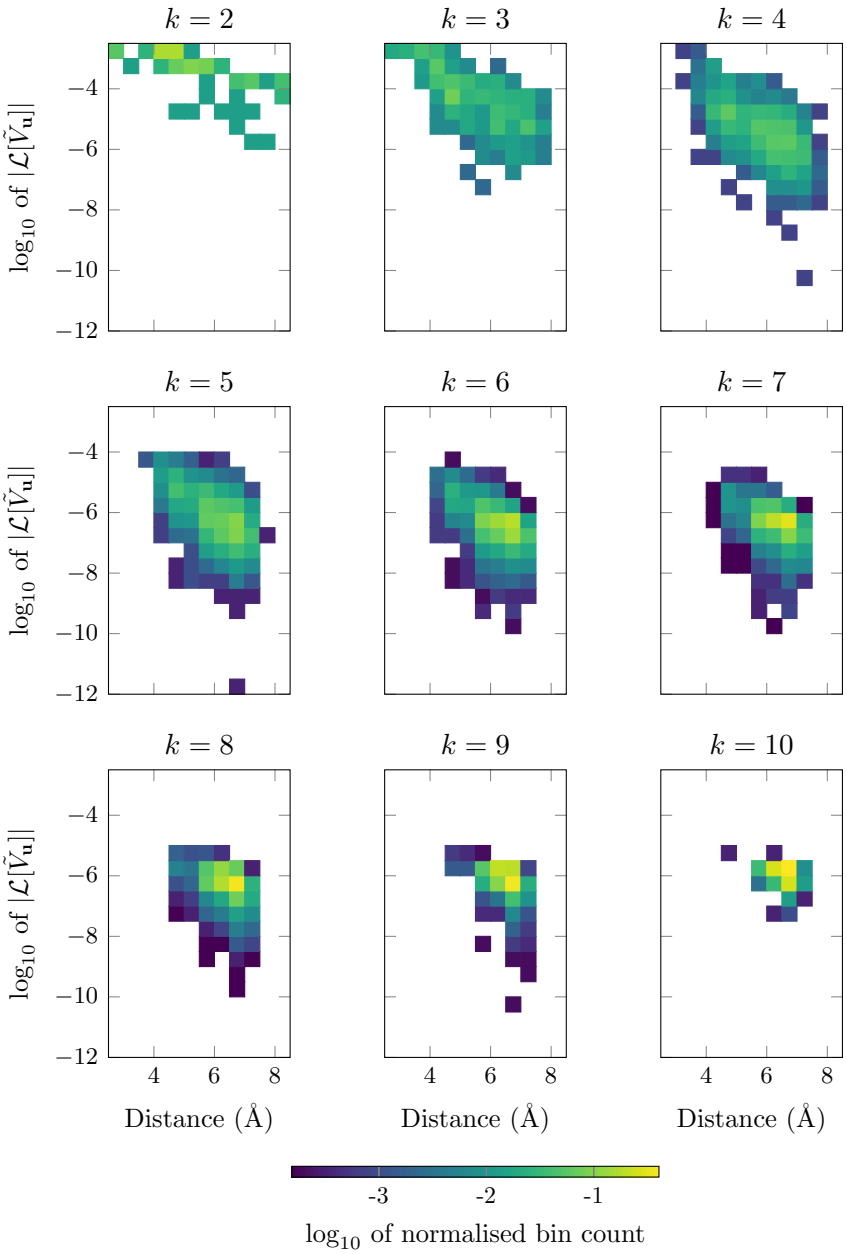

Figure 5.4.: Histograms of magnitudes of evaluated vacuum-embedding MP2/cc-pCVTZ fragment-MBE contribution potentials $|\mathcal{L}[\tilde{V}_{F_{\mathbf{u}}}]|$ for the water cluster $(H_2O)_{15}$. Each histogram corresponds to values with $\mathbf{u} = k$. See the main text for further details.

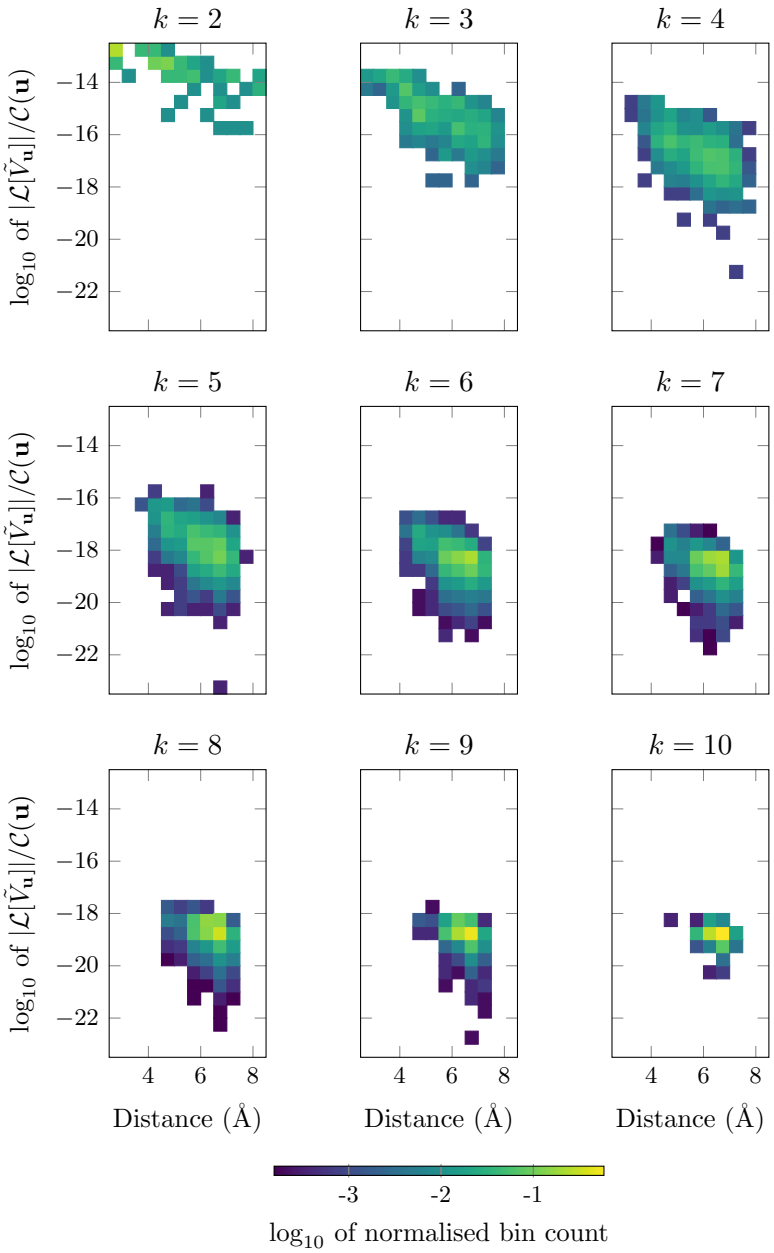

Figure 5.5.: Histograms of the benefit/cost ratios of evaluated vacuum-embedding MP2/cc-pCVTZ fragment-MBE contribution potentials $|\mathcal{L}[\tilde{V}_{F_{\mathbf{u}}}]|/\mathcal{C}(\mathbf{u})$ for the water cluster $(H_2O)_{15}$. Each histogram corresponds to values with $\mathbf{u} = k$. See the main text for further details.

6. Graph-based ANOVA-like decompositions

Since it is quite clear from the previous chapter that many-body expansions as usually considered cannot extend gracefully in application to even moderately substantial molecular systems, we consider now the question of how the basic shape of the MBE might be adjusted or reworked in such a way as to ameliorate issues both of cost and of compounding numerical instability. However we do so, it is desirable to retain some kind of systematic improvability, and for this, we must also retain the formal exactness of the expansion. The switch from a nuclear MBE to a fragment MBE is actually already a step in this direction, since, as we have seen, any truncation of a fragment MBE is provably equivalent to a truncation of an underlying nuclear MBE.

Although it is not technically required, fragment MBEs are basically always obtained by grouping together collections of atoms that are spatially local to each other [Gor+11; CB15; RS15; Her19]. Both the literature discussed and the data analysed in the previous chapter suggest a relationship between the magnitudes of the contributions $|\tilde{V}_{\mathbf{u}}|$ and the distances between the individual atoms indexed by $\mathbf{u}$. It seems therefore *a priori* sensible to discard these terms if the relevant sets of atoms are somehow spatially non-local. This idea underlies a variety of *distance-based thresholding* [LH17] schemes used in the construction of a number of fragmentation methods; see, e.g., [CLJ06; Gan+06; DT07a; HHL10; MR11; QLT13; Fis18] and discussion in [OB16; LH17]. In general, however, simply discarding contribution potential terms will destroy the formal exactness of an MBE. It has also been recognised that some such terms may contribute more strongly to the resulting sum than might be suggested by the spatial density of the underlying atoms [OB16].

More generally, fragmentation methods can and have been defined in terms of some abstract concept of the *connectivity* between pairs of atoms or fragments. Such connectivity is usually defined by the presence of chemical bonds (covalent or otherwise) between atoms or fragments [DC05; CD06; LLJ07; GHH14; Heb14; CB15; RS15; Her19], sometimes augmented or filtered according to spatial separation [Gan+06; MR11; LH17; See+22]. The application of connectivity hinges upon the expectation that collections of atoms or fragments interact most strongly when they are closely connected to each other, thus providing another mechanism with which to attempt to select *a priori* the "important" terms in an MBE-like sum.

Imbuing a molecular system with connectivity then leads very naturally to an explicit treatment using graph theory. This can be used to select a fragmentation, as in, e.g., [Le+12; See+22], or directly to arrive at a decomposition or approximation of the

total energy, as in, e.g., [Kle86; WHM10; GHH14; Heb14; KDI21; ZI22]. Depending on method, the distinction is not always clear. In this chapter, hypothesising some full-system *interaction graph*, we extend the order-theoretic development of MBE-like sums given in the prequel to a class of exact decompositions of V^{BO} defined with reference to particular subcollections of the induced subgraphs of that graph.

6.1. Interaction graphs and SUPANOVA decompositions

We use here and throughout standard graph-theoretic notation and definitions consistent with the textbook of Diestel [Die17]; we make also some general reference to the textbooks of Harary [Har72] and of Cormen et al. [Cor+22]. We consider here only finite undirected graphs $G = (V, E)$; although we are already making heavy notational use of V to represent subproblem/contribution potentials and E to represent energetic quantities, the distinction will be clear in context. We prefer to use the term *connected component* rather than just *component*. Following [Har72], we write $C_n = ([n], E = \{\{1,2\},\{2,3\},\ldots,\{n-1,n\},\{n,1\}\})$ to mean the canonical *cycle graph* of n vertices, and K_n to mean the *complete graph* of n vertices. For clarity, if we say that a graph is *cyclic*, we mean only that it contains at least one cycle.

We now introduce a semi-formal definition for an *interaction graph* in our context. Our naming is based on usage of "interaction graph" in [CGH18]; the term also appears in [GHH14, Sec. 3.1], although in context the meaning is more specific than here. We are very far from the first to introduce such a definition. In the fragmentation method setting, for example, something similar is explicitly used to define the GKEM of Weiss et al. [WHM10], which we shall discuss further in Section 6.4; see also and cf. [Kle86]. Similar graphs also appear in works by Iyengar and co-workers, e.g., [KDI21; ZI22]; there, edges also embody pairwise interactions in a quite general sense.

Definition 6.1.1 (Interaction graph)**.** Given a set of nuclear indices $[M]$, a *nuclear interaction graph* is an undirected graph $G = ([M], E)$, where every vertex $i \in [M]$ is a nuclear index. Similarly, given some fragmentation $F = \{F_i\}_{i=1}^{K}$ of $[M]$, a *fragment interaction graph* is an undirected graph $G = ([K], E)$, where every vertex $i \in [K]$ is the index of some fragment $F_i \in F$. In both cases, each edge $\{i, j\} \in E$ is called a *direct interaction* between the atoms or fragments indexed by i and j, which are said to *interact directly* according to G.

If the choice between the nuclear or fragment setting is clear in context, we will usually speak only of an interaction graph. The simplest possible (here nuclear, but also equivalently fragment) interaction graph is the edge-free interaction graph $([M], \emptyset)$, which represents a system of M non-interacting atoms. Similarly, an M-atom system wherein every atom interacts directly with every other atom is represented by the complete interaction graph $K_M = ([M], \{\{i, j\} \mid 1 \leq i < j \leq M\})$. Since the definition is not

prescriptive as to precisely what an "interaction" between two atoms entails, an interaction graph can be constructed with an edge set capturing effectively any pairwise definition of connectivity between atoms in a system. The prototypical interaction graph, is, however, the *covalent bond graph*, such that atoms i and j interact directly (or alternatively, the vertices i and j are adjacent) whenever they share a covalent bond. This graph is handled, either implicitly or explicitly, by basically all of the overlapping-fragment methods mentioned and referenced in the previous chapter. There are various mechanisms for obtaining the covalent bond structure and thus graph of a given molecule, and in practice, this information is often provided alongside or directly recoverable from a molecular conformation in standard storage formats [OBo+11; Zha+12; Heb14; Bar+21].

We explicitly introduce one particular graph-theoretic idea which we will find useful. The following definition is a slight rephrasing of that given in [GL78, Sec. 4.2], but this is a standard concept and other equivalent versions can be readily found elsewhere in the literature.

Definition 6.1.2 (Quotient graph [GL78])**.** Let $G = (V, E)$ be an undirected graph, and let $F = (F_i \subseteq V)_{i=1}^{K}$ be a partition of V. The *quotient graph* $G/F = (F, E')$ of F in G is defined such that, if $\{i', j'\} \in E$, and there exists $i \neq j$ such that $i' \in F_i$ and $j' \in F_j$, then $\{F_i, F_j\} \in E'$.

In our setting, it should be clear that any quotient graph of a nuclear interaction graph is a fragment interaction graph. In practice and in the interest of simplicity, we shall abuse the definition by assuming that any quotient graph of a fragment interaction graph is also a fragment interaction graph. To make this precise, we would need to complicate the definition by specifying that each vertex of the quotient graph of a fragment interaction graph is not a set of fragments, but instead the union of those fragments. Since we will use quotient graphs only as a practical tool, and will not rely on the definition explicitly in proof, we trust the reader will forgive our sloppiness.

As is the case for any graph, a nuclear (equivalently fragment) interaction graph G can be broken down into subgraphs. In particular, if $\mathbf{u} \subseteq [M]$ is a subset of the relevant set of nuclear indices, then the induced subgraph $G[\mathbf{u}]$ can itself be considered as an interaction graph, and preserves by definition the direct interactions between the atoms indexed by elements of $\mathbf{u}$. Thus, any subset of the set of all induced subgraphs of an interaction graph can be considered as a poset of interaction-preserving subgraphs, with a natural ordering provided by set inclusion calculated on the inducing vertex sets of those subgraphs; that is, $G[\mathbf{u}] \leq G[\mathbf{v}]$ whenever $\mathbf{u} \subseteq \mathbf{v}$; see [Nie80] and cf. [Kle86]. The poset of all induced subgraphs of G is therefore isomorphic to the boolean algebra B_M, and any subposet P of the poset of all induced subgraphs is isomorphic to one of B_M.

We consider a general type of ANOVA-like decomposition of the Born-Oppenheimer potential energy V^{BO} of some M-atom molecule which is equipped with a nuclear interaction graph G. A formally equivalent, and indeed markedly more general rendition of the following construction is given in [Kle86]; more details below. Slightly adjusting

notation, suppose that P is an arbitrary subposet of B_M corresponding to some particular family of induced subgraphs $\{G[\mathbf{u}]\}_{\mathbf{u}\in P}$, rather than the poset of such induced subgraphs themselves. Define subproblem potentials $\{V_\mathbf{u}\}_{\mathbf{u}\in P}$ with $V_{[M]} = V^{\mathrm{BO}}$ and contribution potentials

$$\tilde{V}_\mathbf{u} = \sum_{\mathbf{v}\leq_P \mathbf{u}} \mu_P(\mathbf{v},\mathbf{u}) V_\mathbf{v}. \tag{6.1}$$

In context, we will then call the exact decomposition

$$V^{\mathrm{BO}} = \sum_{\mathbf{u}\in P} \tilde{V}_\mathbf{u}. \tag{6.2}$$

a *SUPANOVA decomposition*, for *SUbgraph Poset ANOVA*. Of course, since the definition is technically in terms of a subposet of B_M, the entire machinery of the previous chapter applies without modification and this definition just mirrors those given before. The distinction is only one of viewpoint. It is to stress that we choose to see P as an ordered collection of induced subgraphs rather than one of subsets that we write the order relation as $\leq_P$ rather than $\subseteq$; where there can be no confusion, we may simply conflate posets of induced subgraphs and isomorphic posets of their inducing vertex sets. The generalisation of (6.2) to induced subgraphs of a fragment interaction graph is immediate, and we will not give it explicitly.

Such an expansion class is not inherently novel. In particular, the SUPANOVA decomposition (6.2) emerges as an immediate special case of the *chemical graph-theoretic cluster expansion* (CGTCE) of Klein [Kle86], which we mentioned in the previous chapter. Klein's formalism is extremely general, and the construction of a CGTCE using Möbius functions and total-energy potentials is only one of multiple possibilities mentioned in [Kle86]. We take the liberty of introducing an explicit new name for the decomposition (6.2) for two reasons. Firstly, we find it a convenient way to denote the precise choice of, in Klein's terminology, cluster function and f transform in the particular context of energy-based fragmentation methods. Secondly, our investigation in this chapter is based upon the foundation laid by the BOSSANOVA decomposition [GHH14; Heb14], and we structure our development so as to converge in the next chapter with the multilevel ML-BOSSANOVA decomposition [CGH18]. This latter represents a further generalisation that is not, to our reading, explicitly anticipated by the work of Klein. From this perspective, we could also expand SUPANOVA as *SUPerset of bossANOVA*.

In any case, Klein's original work suggests truncations of the CGTCE according to subgraph size, analogously to order-n truncations of an MBE. We mention in connection with the previous chapter that, if one considers a Möbius function-based CGTCE expansion over all induced subgraphs of a complete interaction graph, then a standard MBE form is recovered; essentially just this is obtained by the Bethe-Goldstone hierarchy discussed in [Kle86, App. A]. There is also a strong connection between the CGTCE formulation and the *linear combination of graph invariants* (LCGI) scheme described

in e.g. [GK73; Ess+77]; we note in particular the use of Möbius inversion and Möbius functions in [Ess+77]. We are not aware of any treatments of either the CGTCE or the LCGI which consider truncations in terms of arbitrary order ideals, or which explicitly apply adaptivity as we do here.

6.2. BOSSANOVA

The original development of the BOSSANOVA (*Bond Order diSSection ANOVA*) decomposition is described most completely in the German-language doctoral thesis of Heber [Heb14];[1] see also [GHH14; CGH18] for alternative and relatively concise English-language presentations.[2] To contextualise further discussion, we will give a brief and in places interpretative sketch of the construction of this decomposition, following at first [Heb14] but later also [GHH14; CGH18]. We use our notation rather than that of the sources where possible and appropriate. We omit much detail, particularly including but not limited to the handling of short- and long-range contributions in [Heb14].

The construction as given in [Heb14] is based on an ANOVA-like decomposition of the negative-semidefinite component N of a generally-indefinite Hermitean matrix $H \in \mathbb{C}^{n\times n}$; for full details, see [Heb14, Chap. 6]. The eigenvectors u_i of H are assumed to be sparse, in a sense which we will not make precise here; see [Heb14, Defs. 28 and 29] and surrounding discussion. Each eigenvector is matched with a set $\mathcal{I}_i \subseteq [n]$ which indexes its non-negligible entries. The set

$$\mathcal{J} = \{\mathcal{I}_i \cap \mathcal{I}_j \mid 1 \leq i, j \leq n\}. \tag{6.3}$$

is named a „minimalüberlappende Zerlegung“ [*minimal overlapping decomposition*] [Heb14, Def. 33] of $[n]$. Such a decomposition is a special case of what Heber calls a „hierarchische Zerlegung“ [*hierarchical decomposition*] [Heb14, Def. 30] of $[n]$:[3] a family of potentially-overlapping subsets $\mathcal{I}' = \{\mathcal{I}'_i \subseteq [n]\}_{i=1}^{m}$ such that $\bigcup_i \mathcal{I}'_i = [n]$,[4] and such that the family is closed under intersection. It is shown by a casewise argument that N can be decomposed as [Heb14, Lem. 12]

$$N = \sum_{\mathcal{I}_i \in \mathcal{J}} \tilde{N}_i, \qquad \tilde{N}_i = N_i - \sum_{\mathcal{I}_j \subset \mathcal{I}_i} \tilde{N}_j, \tag{6.4}$$

[1] This author does not claim fluency in German, let alone mathematical German, and apologises in advance for any resulting mistranslation and/or misunderstanding of this source.

[2] Let us note the existence of an extended preprint [GHH08] of [GHH14], to which we also make some general reference.

[3] Note that this is not the same as our equivalently-named definition in Chapter 3.

[4] The term „(überlappende) Zerlegung“ [*(overlapping) decomposition*] [Heb14, p. 80] used in [Heb14, Def. 30] is not, to our reading, entirely precisely defined. In context, we understand „überlappende“ to mean “not necessarily pairwise disjoint”; thus, we write “potentially-overlapping” in the main text, for consistency with the remainder of this work.

where each $N_i = \sum_{j \in \mathcal{I}_i, \lambda_j < 0} \lambda_j u_j u_j^\dagger$, with λ_j the eigenvalue of u_j. It is conjectured [Heb14, Behauptung 1] that an equivalent decomposition also exists for any hierarchical decomposition $\mathcal{J}'$ which contains $\mathcal{J}$; we understand this last in the sense of [Heb14, Def. 20].

The decomposition (6.4) is applied to the restricted Hartree-Fock density matrix P of some molecular system [Heb14, Chap. 7], which we recall from [SO89] is defined elementwise by $P_{\mu\nu} = 2\sum_{a=1}^{N/2} c_{\mu a} c_{\nu a}^*$ for $1 \leq \mu, \nu \leq K$ and in terms of values $c_{\mu a}$ drawn from the Hartree-Fock coefficient matrix C. Since the total RHF energy is then given by $E^{\mathrm{HF}} = \frac{1}{2}\operatorname{Tr}[PF]$ where F is the Fock matrix [SO89], we obtain, as in [Heb14, (7.2.4)] but with slight formulaic differences,

$$E^{\mathrm{HF}} = \frac{1}{2}\operatorname{Tr}\left[\sum_{\mathcal{I}_i \in \mathcal{J}} \tilde{P}_i F\right] \tag{6.5}$$

$$\approx \sum_{\mathcal{I}_i \in \mathcal{J}} \frac{1}{2}\operatorname{Tr}[\tilde{P}_i F_i]. \tag{6.6}$$

Here, each $\tilde{P}_i$ is defined basically just as in (6.4), and each F_i is the principal submatrix of F according to $\mathcal{I}_i$. For details on the legitimacy and conditions of this approximation, see [Heb14, Chap. 7].

From a practical perspective, the approximation (6.6) can only be evaluated as written if the precise structure of the minimal overlapping decomposition $\mathcal{J}$ is known. In general, it is not, so a technique for choosing a plausible alternative decomposition $\mathcal{J}'$ is required [Heb14, Sec. 7.5]. One is built by taking each $\mathcal{I}'_i$ to collect the basis function indices associated with all of the nuclear indices specified by some particular $\mathbf{u}'_i \subseteq [M]$. The term on the RHS of (6.6) for each $\mathcal{I}'_i$ can then be interpreted in our notation and terminology just as an evaluated contribution potential, $\mathcal{L}[\tilde{V}_{\mathbf{u}'_i}]$. In particular, if $\mathcal{J}'$ is chosen to consist of all possible k-fold unions of the sets of indices of the basis functions located at the various nuclear centres, then (6.6) becomes just a standard nuclear MBE of an equivalent form to (5.31); this more familiar approach is used as a starting point in both of [GHH14; CGH18], although differently in each.

Here enters the covalent bond graph G [Heb14, Sec. 7.5; GHH14; CGH18]. Since any observed sparsity of the density matrix is interpretable as an manifestation of spatial locality, as is covalent bonding, the intention is to use the information contained in the covalent bond graph in order to winnow down the number of terms in the nuclear MBE, or, equivalently, the size of the decomposition $\mathcal{I}'$.

At this point, the BOSSANOVA construction makes a critical assumption, which we phrase in our notation following [CGH18]; cf. the ultimately equivalent phrasings in [GHH14, Lem. 1; Heb14, Lem. 13]. For any non-empty subset $\mathbf{u} \subseteq [M]$ of the nuclear indices, the induced subgraph $G[\mathbf{u}]$ of the covalent bond graph can clearly be separated into some number $1 \leq N_{\mathrm{comp}}^{(\mathbf{u})} \leq M$ of connected components. The subproblem potential

for any $\mathbf{u}$ such that $G[\mathbf{u}]$ is disconnected is then assumed to be exactly additive in the subproblem potentials corresponding to those connected components:

$$V_{\mathbf{u}} = V_{\mathbf{v}_1} + V_{\mathbf{v}_2} + \cdots + V_{\mathbf{v}_{N_{\text{comp}}^{(\mathbf{u})}}}, \tag{6.7}$$

where the sum runs over the subsets of $\mathbf{u}$ which induce the connected components of $G[\mathbf{u}]$. Under this assumption, an inductive argument [CGH18, Lem. A.1] shows that if $G[\mathbf{u}]$ has two or more connected components, then the contribution potential $\tilde{V}_{\mathbf{u}} = 0$. This justifies the omission of any such $\tilde{V}_{\mathbf{u}}$ from the nuclear MBE, and leads, finally, to the BOSSANOVA decomposition of V^{BO}. Again in our notation, this reads

$$V^{\text{BO}} = V_{[M]} = \tilde{V}_{\emptyset} + \sum_{\substack{\mathbf{u}\in\text{conn}\,[G]\\|\mathbf{u}|=1}} \tilde{V}_{\mathbf{u}} + \sum_{\substack{\mathbf{u}\in\text{conn}\,[G]\\|\mathbf{u}|=2}} \tilde{V}_{\mathbf{u}} + \cdots + \sum_{\substack{\mathbf{u}\in\text{conn}[G]\\|\mathbf{u}|=N}} \tilde{V}_{\mathbf{u}}, \tag{6.8}$$

where we introduce $\text{conn}[G]$ to indicate the set of all subsets $\mathbf{u} \subseteq [M]$ that induce a connected subgraph $G[\mathbf{u}]$ of G. We will always consider the empty graph to be vacuously connected. The contribution potentials for each $\mathbf{u} \in \text{conn}\,[G]$ have the recursive definition

$$\tilde{V}_{\mathbf{u}} = V_{\mathbf{u}} - \sum_{\substack{\mathbf{v}\in\text{conn}[G]\\\mathbf{v}\subset\mathbf{u}}} \tilde{V}_{\mathbf{v}}. \tag{6.9}$$

Using the terminology and notation we have developed in this and the preceeding chapter, we can view (6.8) as being a SUPANOVA decomposition of $V_{[M]}$ in terms of the poset of the connected induced subgraphs of G ordered by set inclusion of their inducing vertex sets.

Although the development of BOSSANOVA is more rigorous and complicated than is common for fragment-based energy methods, once it is arrived at, equation (6.8) and its applications are closely related to earlier constructions. In particular, an n-body truncation of the BOSSANOVA decomposition (6.8) after the terms for connected subgraphs of orders $1 \leq |\mathbf{u}| \leq n$ appears,[5] in the case of chain-like molecules, equivalent to the level-n SFM method [DC05], see Section 5.1.2 in the previous chapter, if an SFM "functional group" is taken to mean "atom" and "bond" is taken to mean "any covalent bond".[6] Here, we observe also, again without explicit proof, that a two-body BOSSANOVA truncation is equivalent to the original KEM double-kernel total energy (5.8) as described in [HMK05]. The FCR method can here also deliver an equivalent, non-recursive formula [KC16; Art+20].

From a technical perspective, the subproblem potentials used in the BOSSANOVA method are for the most part standard. The subproblem potential $V_{\emptyset}$ is taken to

[5] A rigorous proof of this is complicated by the relatively informal presentations of the SFM in, e.g., [DC05; CD06].

[6] References [DC05; CD06] are mentioned in passing in [GHH14].

be identically zero. No electrostatic embedding is used, although an extension which separately approximates the long-range exchange component of the Fock matrix is described in [Heb14, Sec. 7.3]. Dangling bonds are treated using hydrogen link atoms. One novelty is found in the handling of dangling double bonds, which are not treated with a divalent link atom as suggested in Section 5.1.1 above. Instead, a dangling double bond is saturated with two hydrogen atoms arrayed according to the geometry of the original bond and of any other bonds emanating from the link atom host; see, e.g., [Heb14, Fig. 9.31].

Although BOSSANOVA is constructed in terms of the nuclear covalent bond graph G, this graph is preprocessed before its connected subgraphs are used to express (6.8) [GHH14; Heb14, Sec. 9.1.2]. We will refer to the adjusted version as the *dehydrogenated bond graph*. For our purposes, we define it to be that which is obtained by first building a (disjoint) fragmentation $F = \{F_i = \{j, k_1, k_2, \ldots\}\}_{i=1}^{K}$ of $[M]$, where each j indexes a non-hydrogen atom and k_1, k_2, ... index the possibly several hydrogen atoms which interact directly with j in G, and then taking the quotient graph G/F. The BOSSANOVA decomposition is therefore in practice a variation on a fragment MBE (5.35) rather than on the nuclear MBE (5.31). The implicit assumption that this underlying fragment MBE is consistent with the original nuclear MBE appears to pass unremarked in the original BOSSANOVA constructions, but is of course unproblematic, as per Proposition 5.2.4 in the previous chapter.

6.3. On the connected induced subgraphs of cyclic graphs

Although the BOSSANOVA method produces good results when applied to simple linear chain-like molecules, it is known to perform poorly when confronted with systems for which the covalent bond graphs contain chordless cycles [Heb14; GHH14]. Such cyclical structures are often referred to in the chemical context as *rings*.

BOSSANOVA approximations of the total energy of the aromatic hydrocarbon benzene (C_6H_6) are reported in [Heb14, Fig. 9.30], where the subproblem potentials represent HF calculations using the 6-311G* basis set [Kri+80]. As there remarked, the n-body truncations of the BOSSANOVA decomposition (6.8) after all $|\mathbf{u}|$-body terms for $1 \leq |\mathbf{u}| \leq n \leq 6$ produce a series of approximations of the total Hartree-Fock energy that gain steadily in accuracy up to $n = 3$, but then lose accuracy for $n = 4$ and $n = 5$. Basically the same is further and consistently observed in equivalent sets of results for other small biomolecules which contain chordless cycles in their bond graphs [Heb14, Sec. 9.4.5]. Heber advances a physical explanation for this behaviour. Specifically, it is suggested that the „Krümmung“ [curvature] [Heb14, p. 174] of larger ring subfragments leads increasingly to straining forces and steric effects, which are only offset by electron delocalisation in the finally-closed aromatic ring; see also [GHH14].[7]

[7] It is suggested that this issue might be ameliorated by „gesonderter Berücksichtigung aller Zyklen im

Although this observation seems sound, a more fundamental underlying problem is in fact at least partially at work. This is due to the structure of the poset of connected induced subgraphs over which the BOSSANOVA decomposition is defined, and is explicable via Theorem 5.2.8 of the previous chapter.

We will demonstrate the problem by a series of model calculations of the total energy of the linear alkane hexane (C_6H_{14}), which will help to underscore the importance of that theorem. We choose hexane for two reasons. Firstly, it is clearly established [GHH14; Heb14; CGH18] that the BOSSANOVA technique is productive for linear alkanes; indeed, they seem to be the class of molecule for which BOSSANOVA performs best. Secondly, hexane has the same number of non-hydrogen atoms as benzene, so the dehydrogenated bond graphs of both admit connected subgraphs of orders $0 \leq n \leq 6$. We can thus contrast results for hexane with those which we also obtain for benzene, which serves the additional purpose of demonstrating that our method is consistent with that used to obtain the results given in [Heb14] and [GHH14].

Briefly summarising our method: we used molecular geometries that were obtained from the ChemSpider database [CSHex; CSBenz], and then optimised to plausible equilibrium geometries according to B3LYP/cc-pVDZ KS-DFT calculations [KS65; Dun89; Ste+94]; see Section A.6 for details. Calculations were performed using the same implementation of the order-theoretic combination technique as elsewhere in this work; as in Section 5.4, we used a poset grid $\Pi = P$ composed of only a single poset axis P. Here, however, P was taken to represent the poset $\mathrm{conn}[G]$ of vertex subsets inducing connected induced subgraphs of an interaction graph; for a description of the necessary implementation of the poset axis interface, see Appendix B.3. The subproblem potentials used correspond to Hartree-Fock calculations with the 6-311G* basis set, also chosen to match [Heb14; GHH14]; these are similar to the vacuum embedding potentials described in Section 5.4, but also include a link-atom treatment of cut covalent bonds. Such cut bonds were treated using single or pair hydrogen link atoms according to the strength of the cut bond, in a manner that should be consistent with the original BOSSANOVA implementation [Heb17].[8]

For these calculations, explicit adaptivity was not used; rather, we performed a complete series of progressive refinements using the All strategy. The intermediate results here correspond exactly to n-body truncations of (6.8) after the terms for connected subgraphs of orders $0 \leq k \leq n \leq 6$, and so should be consistent with results reported in [Heb14; GHH14], up to differences in the exact molecular geometries, and the additional initial approximation for $n = 0$, which is here always zero. As a reference, we also performed fragment MBE calculations for hexane and benzene, taking the same carbon/hydrogen

Bindungsgraphen“ [separate consideration of all cycles in bond graphs] [Heb14, p. 182]. However, since the precise details of this treatment are to our reading slightly ambiguously described in [Heb14], we do not engage with it here.

[8] We note here that the distances between link atoms and their parent atoms differs from our implementation to the original. We do not believe that this meaningfully impacts the results given in this section.

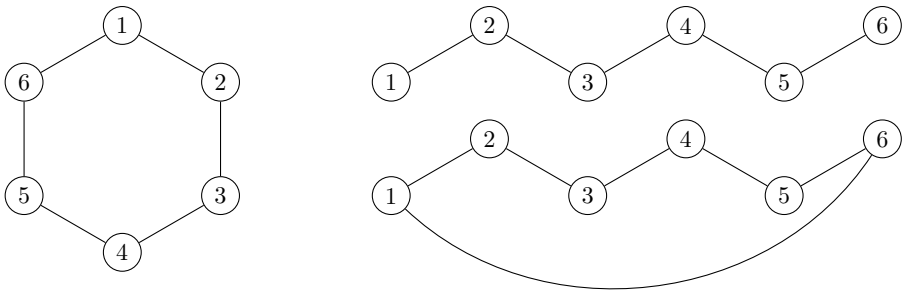

Figure 6.1.: Dehydrogenated covalent bond interaction graphs for benzene (C_6H_6, left) and hexane (C_6H_{14}, right). Each vertex in a graph corresponds to a carbon atom, as well as all hydrogen atoms which are directly bonded to it. On the right-hand side of the plot, two graphs are shown stacked. The top graph is the standard dehydrogenated bond graph of hexane; the bottom is the same graph, but augmented with an additional direct interaction edge between vertices 1 and 6. Note that the augmented hexane graph is actually identical to the standard benzene graph, but with vertices drawn in different relative locations.

groups as in the BOSSANOVA case for the sets of fragments.

For BOSSANOVA calculations on benzene, we considered only the dehydrogenated covalent bond graph. For hexane, we considered the *standard* dehydrogenated bond graph, and also an *augmented* variant, which includes an extra edge between vertices 1 and 6, i.e., between the terminal groups capping the alkane chain. Visualisations of these graphs are given in Figure 6.1. One might consider the augmented graph as modelling an additional direct interaction between the terminal groups, in addition to the indirect interaction provided by the standard bond graph; we would reasonably expect this interaction to be fairly "small", and that its inclusion would only slightly improve the quality of any truncation of the resulting BOSSANOVA-like decomposition. However, the fact that we have constructed the augmented interaction graph of hexane so as to be identical to the standard bond graph of benzene is not a coincidence.

Plots of the absolute errors of the n-body BOSSANOVA total-energy approximations for all three cases are shown in Figure 6.2, along with the equivalent absolute errors for the n-body fragment-MBE calculations. The observed errors for hexane and benzene are generally consistent with, although not entirely equivalent to those plotted in Figures 9.24 and 9.30 of [Heb14] respectively, even after adjusting for the fact that our graphs are plotted against absolute rather than relative error. In particular, our implementation seems to achieve slightly better accuracy for higher-order truncations over hexane than the original BOSSANOVA implementation, as reported in [Heb14]. We suspect that the original results were generated by naïvely summing individually-calculated contribution potentials, which implies uncertainty-related error as outlined in the preceeding chapter.

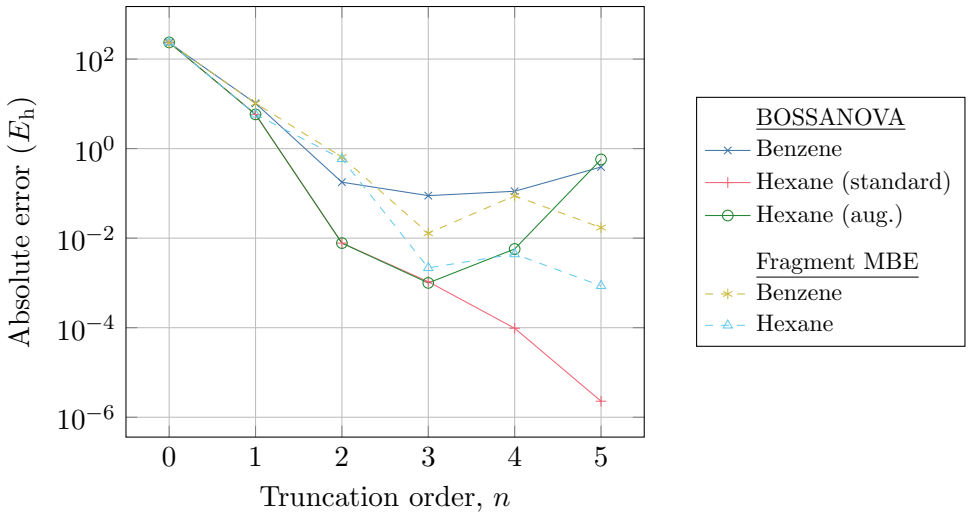

Figure 6.2.: Absolute errors of n-body BOSSANOVA truncations for benzene (C_6H_6) and hexane (C_6H_{14}) as described in the main text, contrasted with absolute errors for n-body fragment-MBE truncations in terms of the same systems/fragmentations. Since the six-body truncations are exact by definition for both BOSSANOVA and the fragment MBE, their associated absolute errors are zero, and are not plotted.

It is interesting to contrast the standard BOSSANOVA results with the behaviour of the n-body fragment-MBE approximations for both hexane and benzene. Both sets of MBE results display error "bumps" in the four-body case, and the five-body cases are only slightly better or approximately equivalent to the three-body cases. Also, the standard BOSSANOVA hexane calculations significantly outperform the equivalent MBE calculations. This is somewhat surprising, given that the BOSSANOVA decomposition is obtained by excluding information from the fragment MBE, and so would seem *a priori* guaranteed to be less accurate for the same truncation order. One potential explanation for this behaviour is the interaction between link atoms introduced by fragment calculations for index sets which induce disconnected subgraphs of the complete bond graph. If the connected components of such subgraphs map to molecular subsystems which are not far from each other, the steric interactions of these link atoms may be non-negligible, thus presenting a source of error just as suggested in Section 5.2.3. Since BOSSANOVA avoids such subproblem calculations by nature, these errors would then be absent.

The most striking aspect of Figure 6.2, however, is the behaviour of the BOSSANOVA approximation driven by the augmented hexane graph. These results are equivalent to or very slightly more accurate than the standard-graph calculations for truncation orders $0 \leq n \leq 3$. Past this point, however, the four- and five-body approximations decrease markedly in accuracy, to the point that the five-body approximation is almost two orders of magnitude less accurate than the two-body approximation. We stress here

that the only difference to the standard hexane calculations is an additional edge in the underlying interaction graph. That this small change can cause such a dramatic decrease in approximation quality is both unexpected and concerning, and this decrease cannot be convincingly explained away by either electron delocalisation or steric effects.

To locate the source of the errors in the augmented bond graph case, we remember that each subproblem potential used on the right-hand side of the recursive definition (6.9) of the BOSSANOVA contribution potentials can also be expanded exactly using a full fragment many-body expansion as in (5.35). Insertion of such an expression into (6.9), followed by some tedious manipulation, leads to an explicit representation of any n-body BOSSANOVA approximation $V^{(n)}_{\text{BOSSANOVA}}$ in terms not of BOSSANOVA contribution potentials, but instead in terms of fragment MBE contribution potentials; cf., e.g., [GHH08, (3.8)].

Performing this manipulation in the case of the standard hexane bond graph, and indeed for any system modelled by fragments with a chain-like interaction graph of length 6, we find that the n-body BOSSANOVA approximations $V^{(n)}_{\text{BOSSANOVA}}$ for $1 \leq n \leq 4$ are given by

$$V^{(1)}_{\text{BOSSANOVA}} = \tilde{V}_{\emptyset} + \tilde{V}_{\{1\}} + \tilde{V}_{\{2\}} + \tilde{V}_{\{3\}} + \tilde{V}_{\{4\}} + \tilde{V}_{\{5\}} + \tilde{V}_{\{6\}}, \tag{6.10}$$

$$V^{(2)}_{\text{BOSSANOVA}} = V^{(1)}_{\text{BOSSANOVA}} + \tilde{V}_{\{1,2\}} + \tilde{V}_{\{2,3\}} + \tilde{V}_{\{3,4\}} + \tilde{V}_{\{4,5\}} + \tilde{V}_{\{5,6\}}, \tag{6.11}$$

$$\begin{aligned} V^{(3)}_{\text{BOSSANOVA}} = V^{(2)}_{\text{BOSSANOVA}} &+ \tilde{V}_{\{1,3\}} + \tilde{V}_{\{2,4\}} + \tilde{V}_{\{3,5\}} + \tilde{V}_{\{4,6\}} \\ &+ \tilde{V}_{\{1,2,3\}} + \tilde{V}_{\{2,3,4\}} + \tilde{V}_{\{3,4,5\}} + \tilde{V}_{\{4,5,6\}}, \end{aligned} \tag{6.12}$$

$$\begin{aligned} V^{(4)}_{\text{BOSSANOVA}} = V^{(3)}_{\text{BOSSANOVA}} &+ \tilde{V}_{\{1,4\}} + \tilde{V}_{\{2,5\}} + \tilde{V}_{\{3,6\}} \\ &+ \tilde{V}_{\{1,2,4\}} + \tilde{V}_{\{1,3,4\}} + \tilde{V}_{\{2,3,5\}} + \tilde{V}_{\{2,4,5\}} \\ &+ \tilde{V}_{\{1,2,3,4\}} + \tilde{V}_{\{2,3,4,5\}} + \tilde{V}_{\{3,4,5,6\}}. \end{aligned} \tag{6.13}$$

We remind the reader that the contribution potentials $\tilde{V}_{\mathbf{u}}$ here are those as defined for a full fragment MBE, not for BOSSANOVA; we choose not to explicitly indicate this by e.g. placing a superscript on the contribution potential terms in order to minimise the visual noise of the equations.

At each successive truncation order n, along with the expected contributions for the connected subgraphs of size n, some number of lower-order MBE contribution potentials also enter the BOSSANOVA sum. These terms can be understood as representing all pairwise, triple-wise, etc. subfragments of those connected subgraphs, which are not themselves connected and are thus not directly calculated, but are instead provided implicitly by some relevant $V_{\mathbf{u}}$ for $\mathbf{u} \in \text{conn}[G]$ with $|\mathbf{u}| = n$. The overall effect is just as intended: each $V^{(n)}_{\text{BOSSANOVA}}$ is a "sub-truncation" of the full n-body MBE sum, guided by the connectivity information encoded in the covalent bond graph.

We consider now the equivalent n-body BOSSANOVA truncations in terms of connected subgraphs of a cycle interaction graph of size 6, as in the case of the augmented hexane

graph, rather than in terms of a chain graph like the standard hexane graph. For $1 \leq n \leq 3$,

$$V^{(1)}_{\text{BOSSANOVA}} = \tilde{V}_{\emptyset} + \tilde{V}_{\{1\}} + \tilde{V}_{\{2\}} + \tilde{V}_{\{3\}} + \tilde{V}_{\{4\}} + \tilde{V}_{\{5\}} + \tilde{V}_{\{6\}}, \tag{6.14}$$

$$V^{(2)}_{\text{BOSSANOVA}} = V^{(1)}_{\text{BOSSANOVA}} + \tilde{V}_{\{1,2\}} + +\tilde{V}_{\{1,6\}} + \tilde{V}_{\{2,3\}} + \tilde{V}_{\{3,4\}} + \tilde{V}_{\{4,5\}} + \tilde{V}_{\{5,6\}}, \tag{6.15}$$

$$\begin{aligned} V^{(3)}_{\text{BOSSANOVA}} = V^{(2)}_{\text{BOSSANOVA}} &+ \tilde{V}_{\{1,3\}} + \tilde{V}_{\{1,5\}} + \tilde{V}_{\{2,4\}} + \tilde{V}_{\{2,6\}} + \tilde{V}_{\{3,5\}} + \tilde{V}_{\{4,6\}} \\ &+ \tilde{V}_{\{1,2,3\}} + \tilde{V}_{\{1,2,6\}} + \tilde{V}_{\{1,5,6\}} + \tilde{V}_{\{2,3,4\}} + \tilde{V}_{\{3,4,5\}} + \tilde{V}_{\{4,5,6\}}. \end{aligned} \tag{6.16}$$

The results again seem to be as intended; the only difference to the chain case is the inclusion of additional disconnected lower-order contribution potentials, which come implicitly from the additional connected induced subgraphs of order n that include the "bridging" edge between vertices 1 and 6. However, for $n = 4$, we find that

$$\begin{aligned} V^{(4)}_{\text{BOSSANOVA}} = V^{(3)}_{\text{BOSSANOVA}} &+ 2\tilde{V}_{\{1,4\}} + 2\tilde{V}_{\{2,5\}} + 2\tilde{V}_{\{3,6\}} \\ &+ \tilde{V}_{\{1,2,4\}} + \tilde{V}_{\{1,2,5\}} + \tilde{V}_{\{1,3,4\}} + \tilde{V}_{\{1,3,6\}} + \tilde{V}_{\{1,4,5\}} + \tilde{V}_{\{1,4,6\}} \\ &+ \tilde{V}_{\{2,3,5\}} + \tilde{V}_{\{2,3,6\}} + \tilde{V}_{\{2,4,5\}} + \tilde{V}_{\{2,5,6\}} + \tilde{V}_{\{3,4,6\}} + \tilde{V}_{\{3,5,6\}} \\ &+ \tilde{V}_{\{1,2,3,4\}} + \tilde{V}_{\{1,2,3,6\}} + \tilde{V}_{\{1,2,5,6\}} + \tilde{V}_{\{1,4,5,6\}} + \tilde{V}_{\{2,3,4,5\}} + \tilde{V}_{\{3,4,5,6\}}. \end{aligned} \tag{6.17}$$

Three MBE contribution potentials of order two are introduced in the first line, each with a non-unit coefficient. As a result, the four-body BOSSANOVA approximation of any system represented by a cycle interaction graph of size 6 is not combination-consistent with any truncation of a fragment MBE over the same system. The five-body BOSSANOVA truncation suffers from the same problem, in fact even more so: nine order-two contribution potentials are double-counted, as are twelve order-three contribution potentials, and eight order-four contribution potentials. Additionally, two order-three contribution potentials ($\tilde{V}_{\{1,3,5\}}$ and $\tilde{V}_{\{2,4,6\}}$) are triple-counted.

To understand why this happens, we could try to investigate how the overcounted terms enter the sum by counting the involved subsets. For example, the subset $\{1, 4\}$ occurs as a subset of $\{1, 2, 3, 4\}$ and of $\{1, 4, 5, 6\}$, and so we might suspect that $\tilde{V}_{\{1,4\}}$ is counted twice as a result. However, $\{2, 3\}$ occurs as a subset of three of the size-4 subsets relevant to (6.17), but there is no triple-counting of $\tilde{V}_{\{2,3\}}$. This kind of analysis is not helpful without considering the contextual properties of the subsets. The critical difference is that $\{2, 3\}$ induces a connected subgraph, while $\{1, 4\}$ does not, so the recursive definition of the BOSSANOVA contribution potentials simply does not involve the latter.

Structurally, $\{1, 4\}$ appears as the intersection of two sets which induce connected subgraphs of the cycle bond graph, namely $\{1, 2, 3, 4\}$ and $\{1, 4, 5, 6\}$. In the standard hexane case, it is easy to persuade oneself visually by studying Figure 6.1 that there is no

way to choose two connected subgraphs such that the intersection of their inducing sets does not itself induce a connected subgraph. This is effectively a restatement of the fact that the poset of index sets which induce connected subgraphs for the standard hexane case is closed under intersection, so the poset is a meet subsemilattice of the underlying boolean algebra. Clearly, however, the equivalent poset in the cycle bond graph case is not closed under intersection, so the poset is not a meet subsemilattice. By Theorem 5.2.8, this guarantees that there will exist at least one order ideal of conn[G] such that the resulting truncation is not combination-consistent with the underlying nuclear MBE. The four-body and five-body truncations of the BOSSANOVA decomposition in this case involve just such order ideals.

The issue in the construction of the BOSSANOVA method as described in the previous section is the assumption of additivity of subproblem potentials of connected components according to (6.7). Since the contribution potentials for any disconnected induced subgraph are shown on this basis to be zero, then the overcounted terms above would be expected to simply vanish without ill effect. However, the additivity assumption (which is justified in [Heb14] by an appeal to size-consistency) does not generally hold, certainly not to chemical accuracy or better, and the contribution potentials for disconnected subgraphs cannot be so easily neglected. Also, since some sets of vertex subsets which induce connected subgraphs for systems with cyclic bond graphs such as benzene do not fulfill the closure-under-intersection property required of the hierarchical decompositions defined in [Heb14], neither will decompositions $\mathcal{I}'$ formed by expanding those vertex sets into sets of underlying basis function indices. We suspect that this may cause trouble related to the original matrix decomposition theorem [Heb14, Lem. 12, Behauptung 1].

For hexane, for example, the actual value of the MBE contribution potential for $\{1, 4\}$ produced by our calculations is $\tilde{V}_{\{1,4\}} \approx 1.948 \times 10^{-3}\,E_\mathrm{h}$, and for benzene, $\tilde{V}_{\{1,4\}} \approx 1.216 \times 10^{-2}\,E_\mathrm{h}$. Thus, the four-body BOSSANOVA truncation of the total energy of hexane using the augmented bond graph is afflicted by an inherent error on the order of $6 \times 10^{-3}\,E_\mathrm{h}$ due to the double-counting of the three diametric-pair contribution potentials; this is directly visible in the results plotted in Figure 6.2. The observed error for the five-body BOSSANOVA truncation for hexane shown in Figure 6.2 admits a similar explanation. Such inherent errors also afflict the the four- and five-body BOSSANOVA approximations of the total energy of benzene, although there do also appear to be additional issues in these cases which may well be related to steric effects and delocalisation as originally suggested.

We conclude, then, that the BOSSANOVA decomposition "works" in the case of the standard hexane graph despite rather than because of the assumption of additivity in the subproblem potentials. It seems that the selection of subproblem potentials which are explicitly evaluated here is fortuitous, leading to relatively error-free truncations of a hypothesised and pure underlying nuclear MBE which do not incur, for example, issues related to link atom placement. However, in the augmented case for hexane, as well as the standard case for benzene, the structure of the underlying poset of connected

induced subgraphs leads to higher-order truncations which fundamentally cannot be good approximations of the true total energy.

We end the section by addressing the obvious next question: does this problem occur for the BOSSANOVA decomposition in terms of any interaction graph containing chordless cycles, or is the particular graph we have considered here just a special and unfortunate case? The answers are, basically, yes and no respectively.

The conditions under which the connected induced subgraphs of a connected graph form a lattice were considered at least by Leclerc [Lec76], and seem to have been first settled in the general case by Nieminen [Nie80]. In the latter, Nieminen considers a particular class of graphs, which he calls *tree structures*; this class includes but is not limited to trees. For our purposes, it is more helpful to use a characterisation in terms of certain *forbidden subgraphs*[9] that is also used in proof by Nieminen. The following is a partial restatement of some of his results rephrased in our terminology; the proof we give serves only to justify this rephrasing and contains nothing novel.

Theorem 6.3.1 ([Nie80])**.** *Let $G = ([M], E)$ be a connected graph, with $M \geq 0$. The poset* $\mathrm{conn}[G]$ *is a meet subsemilattice of B_M if and only if there exists no $\mathbf{u} \subseteq [M]$ such that $G[\mathbf{u}]$ is isomorphic to either the graph $C_4' = ([4], \{\{1,2\},\{1,4\},\{2,3\},\{3,4\},\{2,4\}\})$, that is, the cycle graph C_4 with an additional edge connecting two diametric vertices,*[10] *or to the cycle graph C_n with $n \geq 4$.*

Proof. For $M \leq 3$, the claim is equivalent to $\mathrm{conn}[G]$ always being a meet subsemilattice of B_M, and this is mechanical to verify for all such connected graphs. So assume that $M \geq 4$. We use the definition of a tree structure given in [Nie80] without repeating it here. For ($\Rightarrow$), suppose $\mathrm{conn}[G]$ is a meet subsemilattice of B_M. Since $G[\emptyset]$ is connected, as is G itself, $\mathrm{conn}[G]$ is a lattice. The presence of either forbidden subgraph would contradict this, as reasoned in the proof of [Nie80, Lem. 1]. For ($\Leftarrow$), the condition of the claim is only possible if G is a tree structure; the contrapositive of this is used in the proof of [Nie80, Lem. 1]. Then, as noted in the first few lines of the proof of [Nie80, Lem. 2], for every pair of vertex subsets $\mathbf{u}$, $\mathbf{v}$ inducing connected $G[\mathbf{u}]$, $G[\mathbf{v}]$, it holds that $G[\mathbf{u} \cap \mathbf{v}]$ is connected, so $\wedge_{\mathrm{conn}[G]}$ is well-defined on $\mathrm{conn}[G]$. □

In particular, note that if G is a tree, then $\mathrm{conn}[G]$ is a meet subsemilattice of B_M; see and cf. Leclerc [Lec76], who cites earlier works by Boulaye which we have been unfortunately unable to access. For other order-theoretic properties of $\mathrm{conn}[G]$, see also [JKS95; KS96].

Taken together, Theorems 5.2.8 and 6.3.1 imply that if G contains one of the forbidden induced subgraphs C_4' or C_n for $n \geq 4$, then there will be at least one order ideal of $\mathrm{conn}[G]$ which leads to a summation of BOSSANOVA contribution potentials which is inconsistent with the underlying nuclear MBE. But these results do not directly guarantee

[9] A standard term in graph theory; see, e.g., [Har72].
[10] See [Nie80, Fig. 1] for a visual representation of the original graph.

that any of the order-n BOSSANOVA truncations will be "broken" in this way, although this clearly can happen in practice.

A general characterisation of exactly which order ideals do possess the inconsistency property would require some extra work, as would a characterisation of the manner in which the resulting truncation sums are inconsistent (e.g. which MBE contribution potentials are overcounted, and how many times). Such further investigation might be of some minor interest from an order-theoretic perspective. Practically and informally, however, we observe that inconsistent truncations seem to follow at least from any order ideal which is itself not closed under intersection, and the multiple-counting issue appears to grow worse as the number of "missing" intersections between elements in the order ideal increases. In particular, if a large interaction graph contains induced chordless cycles of some bounded size, say size 6, then the issue does not seem to vanish for the n-body BOSSANOVA truncations for $n \geq 6$, and usually seems to become worse, not better.

6.4. On amending the poset of connected induced subgraphs

Many non-trivial molecules of interest, and most proteins and large biomolecules in particular, contain cyclic structures within their covalent bond graphs. Most commonly, these are five- and six-membered, but larger chordless cycles can also occur. The unreliability of the BOSSANOVA approach when applied to the bond graphs of these systems is therefore a significant limitation to its general applicability.

However, it is clear that n-body truncations of the BOSSANOVA decomposition can provide good approximation quality when applied to some systems with acyclic covalent bond graphs, indeed sometimes better quality than the equivalent n-body truncations of the traditional fragment MBE on which the BOSSANOVA decomposition is based. It is therefore natural to wonder if we might somehow modify the poset of connected induced subgraphs in such a way as to ensure that truncations of the resulting SUPANOVA decomposition are guaranteed to retain consistency with the underlying nuclear MBE.

Before investigating this idea further, we observe firstly that effectively the same problem that afflicts the BOSSANOVA decomposition when applied to cyclic bond graphs has been identified and addressed in other connectivity-based fragmentation schemes. We have already mentioned that an n-body BOSSANOVA decomposition seems to be related to the level-n SFM scheme [DC05]. It was quickly recognised that chordless cycles presented issues for SFM calculations, which were initially ascribed to steric effects due to close-range interactions between introduced hydrogen link atoms [CD06]. The resulting *ring repair* modification to the SFM algorithm is difficult to phrase in our terminology — see [CD06, App. B] for details — but it seems to essentially remove the contribution potentials of any subfragment of a ring from explicit consideration.

In later work by Collins [Col12], the SFM was used as the basis for the *systematic*

molecular fragmentation by annihilation (SMFA) protocol. We omit detail, but the SMFA method can be basically viewed as performing a recursive top-down splitting of the original interaction graph into a series of connected subgraphs. It is observed in [Col12] that splitting chordless cycles can lead to the overcounting of certain "non-bonded interactions" [Col12, p. 7745], and so biasing the weighted summation of per-fragment energies used to deliver the full SMFA energy; thus, when a chordless cycle is encountered during the top-down splitting, it is simply left whole.

An alternative approach is taken by the *generalized kernel energy method* (GKEM) of Weiss et al. [WHM10], which we mentioned briefly in Sections 5.1.2 and 6.1 above. The GKEM is an explicitly graph-based algorithm, phrased in terms of graphs that match in structure and purpose what we call here interaction graphs. Very briefly, using our notation and omitting much detail, one begins with an fragment interaction graph G in terms of some fragmentation. For some truncation order n, all subsets $\mathbf{u} \in \mathrm{conn}[G]$ of size $1 \leq k \leq n$ are enumerated. Then, the sum of all of the fragment-MBE contribution potentials $\tilde{V}_{\mathbf{u}} = V_{\mathbf{u}} - \sum_{\mathbf{v} \subset \mathbf{u}} \tilde{V}_{\mathbf{v}}$ for such subsets $\mathbf{u}$ is manipulated into a weighted sum of subproblem potentials; this sum may involve subproblems $\mathbf{v}$ such that $G[\mathbf{v}]$ is not connected.[11]

The GKEM energy expressions are given in [WHM10] explicitly only up to $n = 4$; the derivations are complicated and the extension to $n \geq 5$ is not immediate. Nevertheless, to our best although informal and unproven understanding, the order-n GKEM total energy is identical to the order-n BOSSANOVA decomposition for tree interaction graphs, but also provides combination-consistent truncations of the underlying fragment MBE for cyclic interaction graphs, while BOSSANOVA generally does not.

These examples suggest two possible approaches to amending the poset of connected induced subgraphs. Given $\mathrm{conn}[G]$, we could seek to add additional disconnected induced subgraphs $G[\mathbf{v}]$ in such a way as to make that poset closed under intersection, which seems to be effectively although not explicitly done by the GKEM in cases where summed contribution terms $V_{\mathbf{v}}$ do not cancel out. Alternatively, we could seek to remove certain connected induced subgraphs from the poset, following the example of the SMFA.

Both approaches are problematic from the perspective of our adaptive algorithm. Whichever adjusted poset $P \subseteq B_M$ we end up using, we need its cover relation to be somehow locally enumerable, in the sense that given some induced subgraph $G[\mathbf{u}]$ and its inducing vertex set $\mathbf{u} \in P$, we need to be able to discover all elements $\mathbf{v} \in P$ such that $\mathbf{v} \prec_P \mathbf{u}$ or $\mathbf{v} \succ_P \mathbf{u}$. The cost of such discovery should ideally scale in a way dependent only on $\mathbf{u}$, $\mathbf{v}$, and possibly the underlying graph G, but should not require global knowledge of the complete poset P. This is true for $\mathrm{conn}[G]$; see Section B.3. We could in theory calculate, enumerate, and record the entirety of P before beginning a calculation, but this is unlikely to be computationally feasible for even moderately-sized

[11] It is in the derivation of the four-body terms in [WHM10] that we encounter the only explicit invocation of the general form of the PIE in the fragment literature that we are aware of.

full-system interaction graphs.[12] Nor is it easily feasible to somehow augment or prune the poset during the adaptive process, since adding or removing a term $s \in P$ can potentially change the value of the Möbius function $\mu_P(s,t)$ for any $t \geq s$.

6.5. Graph convexities and convex subgraphs

We consider now a particular subposet of the connected induced subgraphs which is guaranteed by definition to be closed under intersection: that of the *convex subgraphs* of G, or isomorphically, the subposet of vertex subsets of G which induce convex subgraphs. We defer a definition of "convex subgraph" for the moment, but mention in anticipation that the cover relations $\mathbf{u} \prec \mathbf{v}$ and $\mathbf{u} \succ \mathbf{v}$ of this subposet can also be explored in a way that is more-or-less local to some $\mathbf{u}$, although not as straightforwardly as in the case of $\mathrm{conn}[G]$.

Once more, we are not the first to consider decompositions of chemical properties in terms of convex subgraphs. In particular, Klein suggests the use of the set of convex subgraphs of a graph in the context of the CGTCE [Kle86]. In the words of Klein, "we believe connected convex subgraphs form an especially well-behaved class of subgraphs in some cluster expansion schemes to be discussed" [Kle86, p. 156]. By the latter, he refers to a coupled cluster-like expansion. A full consideration of this formulation and the role which the convex subgraphs play in it is beyond the scope of this work. To our reading, however, precisely why and in what sense the convex subgraphs are "well-behaved" is not elaborated upon.

The convex subgraphs of a graph have been considered when applying the CGTCE in the context of, e.g., Heisenberg model Hamiltonians [PSK89]. Although it is possible to recover the SUPANOVA decomposition form (6.2) in terms of the convex subgraphs of some graph G from Klein's very general class of expansions, it does not appear that this has ever been fully and explicitly done in a setting like ours. It turns out, however, that a particular subset of the convex subgraphs — specifically, the complete induced subgraphs — provide the building blocks for a multilevel graph-based fragmentation scheme developed more recently by Iyengar and co-workers [RHI18; RI18; KI19; RI20; Zha+21], although the convexity of these subgraphs in the sense we consider here does not seem to have been noticed by those authors.[13] Although the derivation we give here is not based directly on their work and we present it distinctly, the resulting *convex SUPANOVA* expansion provides a natural extension to their formulation and recovers it as a special case. We will discuss their scheme in detail and describe the connections between it and that which we now give in Sections 7.1 and 7.2 in the following chapter.

[12] The SMFA encounters effectively just this issue during its top-down splitting, and attempts to overcome it by a *compression* technique in the fragmentation algorithm [Col12].

[13] A different kind of convexity is mentioned in [RI20; Zha+21]. See Footnote 6 on page 195 in Section 7.2 for more details.

To make the connection with our order-theoretic construction clear, we will introduce the convex subgraphs via the well-established theory of *abstract convexity*. This development is not novel, and we shall only scratch the surface of the subject here. The reader interested in a deeper treatment could begin with the survey by Edelman and Jamison [EJ85], the textbook of van de Vel [Vel93], and the monograph of Pelayo [Pel13]. Precise nomenclature and notation differs across the literature; we pick and choose for consistency and according to taste.

We begin by recalling informally that a convex subset $X \subseteq \mathbb{R}^d$ is one such that every point z that lies on a straight line between two points x, $y \in X$ is itself in X. It follows immediately that, if X, $Y \subseteq \mathbb{R}^d$ are convex, then so too is their intersection. This leads to the following basic abstraction, which is defined equivalently in, e.g., [EJ85, Sec. 2; Vel93, Sec. 1.1; Pel13, Sec. 1.3]. Our notation partially follows that in [GO10; BO13].

Definition 6.5.1 (Convexities and convex structures [EJ85; Vel93; Pel13])**.** Let X be a finite set, and let $\mathcal{M} \subseteq 2^X$ be such that $\emptyset \in \mathcal{M}$, $X \in \mathcal{M}$, and if $A \in \mathcal{M}$ and $B \in \mathcal{M}$, then $A \cap B \in \mathcal{M}$. Then $\mathcal{M}$ is called a *convexity*, and the members of $\mathcal{M}$ are called *convex sets* or *convex subsets of* X. Collectively, $(X, \mathcal{M})$ is called a *convex structure*.

The additional requirement of closure under nested union specified in [Vel93; Pel13] is unnecessary in the finite case [CMS05]. From an order-theoretic perspective, note that any convexity $\mathcal{M}$ ordered by set inclusion is trivially isomorphic to a meet subsemilattice of $B_{|X|}$.

For the next definition, we refer again to, e.g., [EJ85, Sec. 2; Vel93, Sec. 1.1], and borrow some notation from [BO15].

Definition 6.5.2 (Convex hull [EJ85; Vel93])**.** Let $(X, \mathcal{M})$ be a convex structure, and let $S \subseteq X$ be an arbitrary subset of X. The *convex hull* of S, written $\mathrm{CH}[S]$, is defined as

$$\mathrm{CH}[S] := \bigcap \{A \in \mathcal{M} \mid S \subseteq A\}. \tag{6.18}$$

The convex hull of $S \subseteq X$ can be considered equivalently as the unique smallest convex subset of X containing S; see, e.g., [Vel93; Pel13]. Now following specifically [EJ85]:

Definition 6.5.3 (Extreme points of a set [EJ85])**.** Let $(X, \mathcal{M})$ be a convex structure. The *extreme points* of some $A \subseteq X$, written $\mathrm{ex}(A)$, are those points $p \in A$ such that $p \notin \mathrm{CH}[A - \{p\}]$.

In [EJ85, Thm. 2.1], Edelman and Jamison show the equivalence of several properties that may be possessed by a convexity. The following definition, which we consider an abridgement of that in [EJ85], is in terms of just one of those conditions. This simpler form is also given in, e.g., [FJ86; Pel13].

Definition 6.5.4 (Convex geometry [EJ85])**.** Let $(X, \mathcal{M})$ be a convex structure. If for every convex set $A \in \mathcal{M}$ it holds that $A = \mathrm{CH}[\mathrm{ex}(A)]$, then $(X, \mathcal{M})$ is called a *convex geometry*. If X is clear in context, we say informally just that $\mathcal{M}$ is a convex geometry.

The convexity of any convex geometry turns out to be a kind of lattice called *lower semidistributive* [EJ85, Thm. 4.1; Mon85]. It is particularly interesting from our perspective that a non-recursive expression has been given for the Möbius functions of these lattices [EJ85, Thm. 4.3].

The intersection of abstract convexity and graph theory has a long history of study [FJ86; FJ87; Duc88; BO13; Pel13]. Naturally, one can define a convexity of subsets of the vertex set of some graph G. When so doing, it is usual to include an additional connectedness requirement on both G and the subgraphs induced by members of the convexity. Partially following, e.g., [Duc88; Pel13]:

Definition 6.5.5 (Graph convexity and convex subgraph [Duc88; Pel13])**.** Let $G = (V, E)$ be a connected undirected graph. Let $\mathcal{M}$ be a convexity over the vertex set V of G. If every convex set of $\mathcal{M}$ induces a connected subgraph of G, i.e., if $\mathbf{u} \in \mathcal{M}$ implies that $G[\mathbf{u}]$ is connected, then every such $G[\mathbf{u}]$ is called a *convex subgraph* of G, and $\mathcal{M}$ is called a *graph convexity*.

We will sometimes write a graph convexity as $\mathcal{M}[G]$ to emphasise the particular graph involved, much like [GO10; BO13]. We note that the term "convex subgraph" is not very widely used in the graph-convexity literature (cf., however, e.g., [BC08]), and is in particular not used in [Duc88; Pel13]. But we find it simpler than writing, e.g., "subgraph of G induced by a convex set $\mathbf{u}$".

As usual, we define the shortest-path distance $d(u, v)$ between two vertices u and v in a connected graph G to be the minimum length of any path between them. Since G is connected, at least one such shortest path exists. Each such shortest path is called a $u-v$ *geodesic*, as in, e.g. [CZ99; GO10; Pel13, Sec. 1.2], leading to the following definitions. We cite here only [HLT93; CZ99; BC08; GO10; Pel13, Sec. 2.1] as exemplary references for particular names and forms, but there are many other equivalent versions to be found throughout the literature.

Definition 6.5.6 (Geodesic interval and geodesic closure [HLT93; CZ99; BC08; GO10; Pel13])**.** Let $G = (V, E)$ be a connected undirected graph, and let $u, v \in V$. The *geodesic interval* [GO10] between u and v is defined[14] to be the set of all vertices which lie along any shortest path between u and v in G, and is written $I_g(u, v)$. That is, following the explicit form given in [BC08, Sec. 1],

$$I_g(u, v) := \{w \in V \mid d(u, w) + d(w, v) = d(u, v)\}. \tag{6.19}$$

Similarly, the *geodesic closure* [HLT93; CZ99; Pel13, Sec. 2.1] is defined to be

$$I_g[\mathbf{u}] := \bigcup_{u,v \in \mathbf{u}} I_g(u, v); \tag{6.20}$$

[14] Both "geodesic" and "geodetic" can be found used interchangeably in this context in the literature. For consistency, we use "geodesic" throughout.

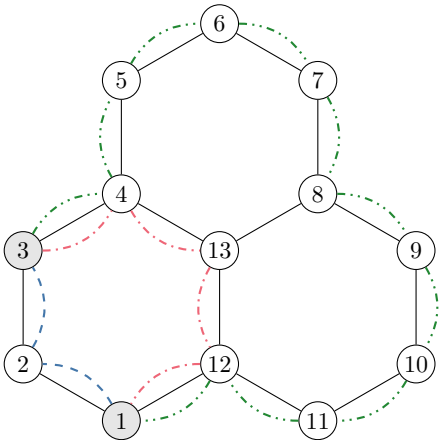

Figure 6.3.: All chordless paths between two vertices in the dehydrogenated covalent bond graph of phenalene ($C_{13}H_{10}$) [CSPhen]. The numbering is arbitrary. The coloured arcs between vertices indicate all three possible chordless paths, up to orientation, between vertices 1 and 3 (shaded grey). The path marked by blue dashed arcs ( ) is the unique shortest path between vertices 1 and 3, so the geodesic convex hull of $\{1, 3\}$ is $\mathrm{CH}_g[\{1, 3\}] = \{1, 2, 3\}$. Since the vertices along the paths marked by red dash-dotted arcs ( ) and green dash-double-dotted arcs ( ) are also included in the monophonic closure $I_m[\{1, 3\}]$, the monophonic convex hull of $\{1, 3\}$ is $\mathrm{CH}_m[\{1, 3\}] = [13]$, i.e., the complete vertex set of the graph.

that is, $I_g[\mathbf{u}]$ is defined to be the set of all vertices that lie along any shortest path between any pair of vertices $u, v \in \mathbf{u}$.

Clearly $I_g[\emptyset] = \emptyset$ and $I_g[V] = V$. Moreover, suppose that $\mathbf{u}, \mathbf{v} \subseteq V$ are such that $\mathbf{u} = I_g[\mathbf{u}]$ and $\mathbf{v} = I_g[\mathbf{v}]$. Then clearly also $\mathbf{u} \cap \mathbf{v} = I_g[\mathbf{u}] \cap I_g[\mathbf{v}] = I_g[\mathbf{u} \cap \mathbf{v}]$. As a result, the collection of sets which are fixed points of the geodesic closure [Dou+09; Pel13] is a graph convexity over G, called the *geodesic convexity* [FJ86; GO10; BO13; Pel13]:

$$\mathcal{M}_g[G] := \{\mathbf{u} \subseteq V \mid \mathbf{u} = I_g[\mathbf{u}]\}. \tag{6.21}$$

As well as being called *convex* by dint of membership in a convexity, any $\mathbf{u} \in \mathcal{M}_g[G]$ may also be called *geodesically convex*. Similarly, we will call each $G[\mathbf{u}]$ geodesically convex. The *geodesic convex hull* of any $\mathbf{v} \in V$, written in context $\mathrm{CH}_g[\mathbf{v}]$, is defined as above and can be found by repeated application of I_g to $\mathbf{v}$ until a fixed point is encountered [HN81; Dou+09; Pel13].

Although the geodesically convex subgraphs of a graph are usually those meant when "convex subgraph" is used unqualified, as in [Kle86], a variety of other path-based graph convexities can also be defined; see, e.g., [FJ86; Duc88; ASS13; Pel13; BO15; DS16]. For example, the *monophonic convexity* $\mathcal{M}_m$ is obtained by defining an interval between u and v to be all vertices along any chordless path between u and v, rather than just those

along any shortest path [FJ86; Pel13; BO13]. However, the geodesic convexity is the only graph convexity described in the literature which we consider suitable for our purpose here. We seek a hierarchy of subgraphs that is "balanced", in the sense that it allows our adaptive combination technique to explore small increments in cost and accuracy. For this, we ask informally that the convex hull of a set of vertices should be somehow "as small as possible" in order to cover those vertices. The monophonic convexity can produce much larger convex hulls for given sets of vertices than the geodesic convexity. For an illustration of this effect, see Figure 6.3, which shows the dehydrogenated bond graph of the polycyclic aromatic hydrocarbon phenalene ($C_{13}H_{10}$) [CSPhen]. The bond structure of this molecule consists of three joined benzene-like rings. The empty set, the singleton sets, and the sets of adjacent vertex pairs are convex sets in both the geodesic and monophonic graph convexities. However, the smallest monophonically convex set including $\{1, 3\}$ is $\mathrm{CH}_m[\{1, 3\}] = [13]$, i.e., the entire graph, while $\mathrm{CH}_g[\{1, 3\}] = \{1, 2, 3\}$.

In general, it would be desirable to work with a graph convexity $\mathcal{M}[G]$ that is a convex geometry. As well as the aforementioned non-recursive expression for the Möbius function, certain properties of a convex geometry would suit our adaptive algorithm well. In particular, it can be shown that each maximal chain in a convex geometry over X has length $|X|$ [EJ85, Thm. 2.2]. This would immediately imply that every cover $\mathbf{v} \succ \mathbf{u}$ of some $\mathbf{u} \in \mathcal{M}[G]$ could be obtained by the addition of a single additional vertex to $\mathbf{u}$; see also [EJ85, Thm. 2.1(b)]. Likewise, every $\mathbf{v} \prec \mathbf{u}$ could be obtained by the removal of a single extreme point of $\mathbf{u}$.

Consider, however, the following slightly rewritten version of [GO10, Thm. 24], which is itself a partial restatement of a key result on the geodesic convexity originally due to Farber and Jamison [FJ86].

Theorem 6.5.7 ([FJ86; GO10])**.** *Let $G = (V, E)$ be a connected undirected graph. Then $\mathcal{M}_g[G]$ is a convex geometry if and only if G contains no chordless cycles of length four or greater, and there does not exist $\{v_1, v_2, v_3, v_4, v_5\} \subseteq V$ such that each of $\{v_1, v_2\}$, $\{v_1, v_3\}$, $\{v_1, v_4\}$, $\{v_1, v_5\}$, $\{v_2, v_3\}$, $\{v_3, v_4\}$, $\{v_4, v_5\} \in E$.*[15]

Proof. See [FJ86, Thm. 4.1]. □

This is a strong condition, which is not fulfilled by precisely the kind of problematic interaction graphs which led us to this point, i.e., those containing five- and six-membered chordless cycles. However, it is not hard to see that $\mathcal{M}_g[G] = \mathrm{conn}[G]$ for any tree interaction graph, and here also $\mathcal{M}_g$ is a convex geometry. We remark that just this occurs in the known-good case for BOSSANOVA of linear alkanes.

In summary, although the geodesic convexity $\mathcal{M}_g[G]$ of an interaction graph is a meet subsemilattice and thus a combination-consistent subposet of B_M that we can safely use in a SUPANOVA decomposition, the presence of chordless cycles of length greater than

[15] In this case, both [FJ86; GO10] call the induced subgraph $G[\{v_1, v_2, v_3, v_4, v_5\}]$ a *3-fan*. For visual representations of the 3-fan, see either [FJ86, Fig. 1] or [GO10, Fig. 6].

three (or, in their absence, the substructure described in Theorem 6.5.7) in G will cause $\mathcal{M}_g[G]$ to be not as well-behaved as we would like. We mention in particular that we have so far been unable to construct a general formulation of the Möbius function for $\mathcal{M}_g$ that is usefully more efficient than the standard recursive definition; we suspect that this is likely to be a difficult problem.

6.6. Case study: heterocyclic molecules

Undeterred by the last, we will examine SUPANOVA decompositions of V^{BO} made using the poset $\mathcal{M}_g[G]$ of those vertex subsets which induce geodesically convex subgraphs of a connected interaction graph G. Here, the nuclear subproblem potentials are, in our standard notation,

$$V_{\mathbf{u}} = \sum_{\mathbf{v} \in \mathcal{M}_g[G[\mathbf{u}]]} \tilde{V}_{\mathbf{v}}. \tag{6.22}$$

As before, we relegate a description of the poset axis interface functionality to Section B.4 in Appendix B.

We consider now two molecules of non-trivial size: limonin ($C_{26}H_{30}O_8$) [CSLimo] and chignolin ($C_{48}H_{63}N_{11}O_{18}$) [Hon+04]. We choose these molecules because their properties allow us to demonstrate particular strengths and weaknesses of various SUPANOVA decompositions in terms of covalent bond graphs. An initial geometry for limonin was obtained from the ChemSpider database [CSLimo], and an initial geometry for chignolin from the Protein Data Bank [HY04, PDB key: 1UAO]. Both geometries were further optimised to plausible equilibria; see Section A.6. As in the case of the water clusters considered in the previous chapter, we then performed reference calculations on both systems at the all-electron MP2/cc-pCVTZ level of theory [MP34; SO89; Dun89; WD95]. Both calculations were also performed using NWChem [Apr+20], with the RHF iterative convergence threshold set to $10^{-9}\,E_{\mathrm{h}}$ and ERI prescreening thresholds for both RHF and MP2 set to $10^{-14}\,E_{\mathrm{h}}$. The reference total energy of limonin was calculated to be $E^{\mathrm{MP2}}_{\text{cc-pCVTZ}} \approx -1609.265\,066\,E_{\mathrm{h}}$, and that of chignolin, $E^{\mathrm{MP2}}_{\text{cc-pCVTZ}} \approx -3820.686\,852\,E_{\mathrm{h}}$. The reference energy for chignolin was somewhat non-trivial to obtain, as the involved MP2 calculation involved 4193 atomic orbital basis functions.

We begin with limonin, an abstract visualisation of which is provided in Figure 6.4. The structure of limonin is strongly heterocyclic, containing a total of six five- and six-membered rings of oxygen and carbon atoms. There is also a single three-member ring formed of two carbon atoms and an oxygen.

Although the original BOSSANOVA formulation allows the severing of double bonds, we choose to explicitly avoid this possibility here. We do so mostly to avoid the additional complexity introduced by the questions of how one should saturate a dangling double bond, and — should two hydrogen atoms be used — where precisely they should be placed. Thus, we must use a different fragmentation of the nuclear indices of the system.

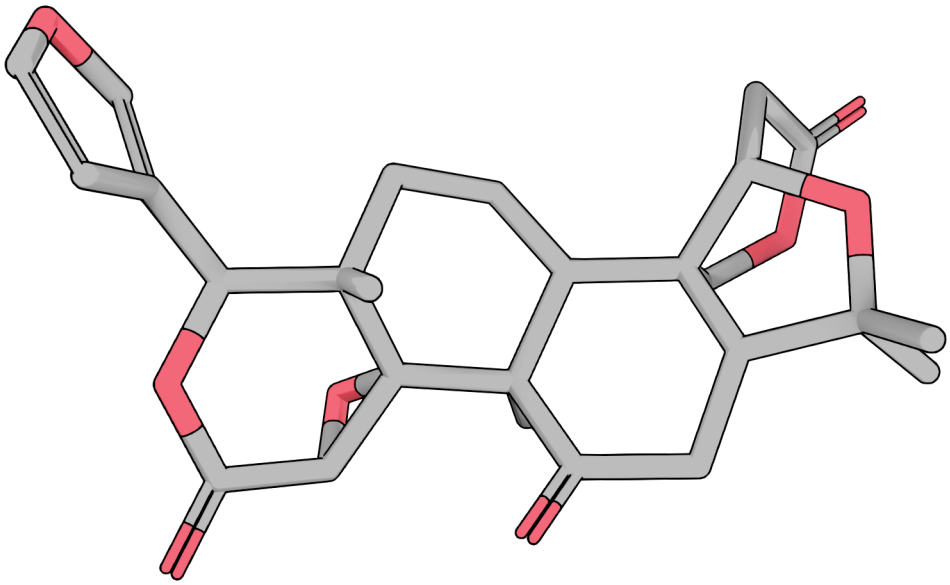

Figure 6.4.: Stick-model visualisation of limonin ($C_{26}H_{30}O_8$) [CSLimo], after geometry optimisation according to B3LYP/cc-pVDZ. Each stick represents a covalent bond between two non-hydrogen atoms; hydrogens are omitted in order to better display the main structure of the molecule. Fully grey sticks indicate bonds between carbon atoms; half-grey and half-red sticks indicate bonds between carbon and oxygen atoms. Single sticks indicate single bonds, and double sticks indicate double bonds.

There are, of course, very many ways of generating such a fragmentation, and a full comparison of the different approaches described in the existing literature is beyond the scope of this work. For the purposes of this study, we applied a simple heuristic algorithm based upon repeated refinement of a candidate fragmentation F'. We will describe this algorithm only informally. Very similar heuristics are used to produce the initial sets of functional groups used by both the SMFA [Col12] as well as the CFM (*combined fragmentation method*) [Le+12] of Le et al. The basic approach of the algorithm is also very similar to that given in [Le+12]. We note, however, that our approach here is intended to function for an arbitrary interaction graph, rather than the covalent bond graphs considered in [Col12; Le+12].

We begin by setting $F' := \{\{i\}\}_{i=1}^{M}$, that is, assigning each distinct atom to a singleton fragment, and then forming the quotient graph G'/F' of the base interaction graph G'. The algorithm then proceeds in two phases.

In the first phase, for each fragment F_i, we consider every adjacent fragment F_j in the quotient graph. If there exists a non-single covalent bond between a atom in F_i and one in F_j, the two fragments are combined. Similarly, if there exists a hydrogen atom in F_i

which is bonded to any atom in F_j, or vice versa, the fragments are combined into a single fragment in an updated version of F'. Once all fragments have been considered, the quotient graph G/F' is recalculated. This process is repeated until a fixed point is found, that is, until no fragment should be combined with an adjacent fragment according to the two rules given above. Although we do not offer a proof, this results in practice in a quotient graph that is effectively just the dehydrogenated covalent bond graph, with any two fragments in that graph that are linked by a non-single bond merged together. Exactly such a quotient graph is used as the initial group structure by the SMFA, as presented in [Col12]; the two conditions here are also rules H1 and H2 in [Le+12, Tab. 1].

The second phase of the heuristic algorithm aims to produce a quotient graph such that no vacuum embedding calculation in terms of any convex induced subgraph would result in the cutting of multiple bonds to any atom not indexed by the subgraph, and therefore require the placement of two or more link atoms in close spatial proximity. This can occur in two cases. Firstly, if a single atom in one fragment is bonded to two or more atoms in another fragment adjacent in the quotient graph. Secondly, if two atoms, one in each of two quotient graph-adjacent fragments F_i and F_j, are bonded, and there exists an atom bonded to either of the two original atoms in a third fragment F_k, that is itself adjacent in the quotient graph to both F_i and F_j. Similarly to the first phase, the second phase again iterates over all adjacent pairs of fragments F_i and F_j, testing whether either of these two conditions holds, and merging F_i and F_j if so. Once a fixed point is found, the algorithm terminates, and the current candidate F' is taken to be the final fragmentation F. The two heuristics used in this step are very similar to rule H3 in [Le+12, Tab. 1], although not identical; we do allow an atom to be bonded to multiple other atoms outside its containing fragment, provided that each of those atoms resides in a distinct fragment, and none of those fragments are adjacent in the quotient graph.[16]

Although this heuristic approach is observed to construct suitable fragmentations in practice, we offer no formal proof of its optimality in any sense. Indeed, the second phase in particular does not necessarily lead to a unique fragmentation, since the merging of fragments depends on the order in which the pairs are considered. A more rigorous development of this algorithm, as well as comparison with those described in, e.g., [Col12; Le+12; CCB14; See+22] may be of some slight interest in the future.

When applied to limonin, the heuristic approach produces a fragmentation F that is as follows. As expected, each non-hydrogen atom is found in the same fragment as are all hydrogen atoms to which that atom is covalently bonded. Additionally, any two atoms connected by a double bond are also found in the same fragment. The carbon and oxygen atoms involved in the three-member ring are likewise grouped into the same fragment,

[16] For example, consider the case of the covalent bond graph of any branched alkane containing a tertiary carbon atom. Our algorithm simply provides the dehydrogenated covalent bond graph, and the tertiary carbon atom will be placed alone in a fragment with its single bonded hydrogen atom. The CFM algorithm, to our understanding, will never allow this atom to exist in a group without the presence of at least two of its three bonded carbon atoms, since this would break rule H3 in [Le+12, Tab. 1].

and the oxygen atom in the five-member ring containing double bonds is grouped with the two carbon atoms involved in the double bond that is drawn further from the page in Figure 6.4. The fragmentation F consists of 26 fragments in total; seven of these are singleton sets, seven are pairs, five are triplets, and seven are quadruplets. The final quotient graph $G = G'/F$ contains three chordless cycles of length six, and two chordless cycles of length five.

We report several order-theoretic combination technique calculations using single-axis poset grids, with two types of SUPANOVA axis. The first type is in terms of the poset grid $\Pi = \mathrm{conn}[G]$, the poset of connected induced subgraphs of G. The second set involves $\Pi = \mathcal{M}_g[G]$, the poset of geodesically convex subgraphs of G. For comparison, we also consider fragment MBE calculations, using the single-axis boolean algebra grid $\Pi = B_{|G|}$ of subsets of the complete set of fragments.

Calculations were performed in terms of both vacuum embedding subproblem potentials and mixed-basis subproblem potentials, all at the MP2/cc-pCVTZ level of theory [MP34; SO89; Dun89; WD95] as calculated with PySCF [Sun15; Sun+17; Sun+20]. We refer again to Section A.7 for full details, remarking also again that the practical process of obtaining the final results given here and throughout this chapter (with the exception of those in Section 6.8.2) involved the use of repeatedly restarted adaptive calculations and precached subproblem potential results. Given that the various electrostatic embedding potentials in the MBE case discussed in the previous chapter did not demonstrate a persuasive benefit relative to vacuum embedding potentials, and in the interest of simplicity, we do not consider such potentials either here or anywhere in the remainder of this thesis. For vacuum embedding calculations, dangling bonds were treated using hydrogen link atoms, placed according to the scaled covalent radii [Cor+08] of the involved atoms as per [RH12, (9)]. The embedding basis set for the mixed-basis potentials was again chosen to be DZ.

As in the previous chapter, we report per-iteration results for adaptive calculations using both the Threshold and All strategies. For Threshold calculations, we give results here and for the remainder of the chapter only for a single threshold value, $\alpha = 0.1$, rather than the $\alpha = 0.5$ considered in the previous chapter. The reduction was intended to increase the amount of work available at each iteration; we note that the posets investigated here are significantly "sparser" than the full boolean algebra. Again, a full investigation of different values of α for such calculations is left for future work. For full-MBE calculations, however, the threshold is kept at $\alpha = 0.5$ for consistency. Calculations are considered to have terminated once the cumulative abstract cost of evaluation of all terms in an adaptively-obtained index set exceeds the reference cost by a factor of ten.

In what follows, we will refer to as "BOSSANOVA" those calculations which are made in terms of $\mathrm{conn}[G]$, and "SUPANOVA" or "convex SUPANOVA" those in terms of $\mathcal{M}_g[G]$. The former are not, strictly speaking, BOSSANOVA calculations exactly as described previously in the chapter, since some of the involved fragments are larger than those

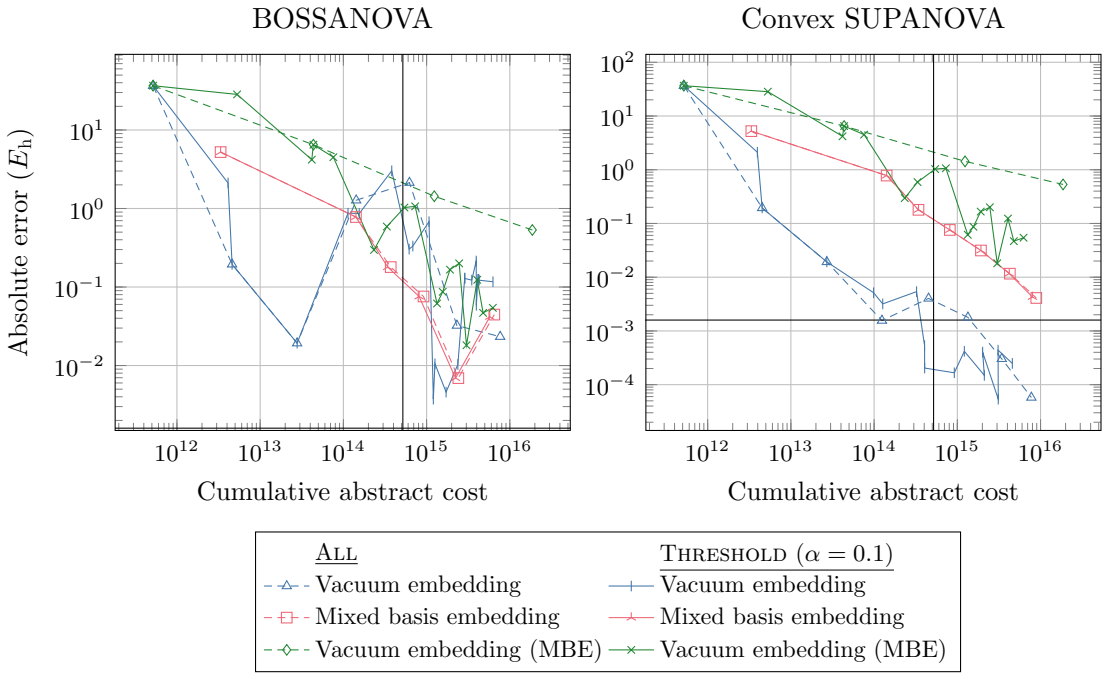

Figure 6.5.: Absolute errors for progressively-refined adaptive SUPANOVA calculations over limonin ($C_{26}H_{30}O_8$), performed using vacuum and mixed-basis embedding subproblem potentials. The fragmentation is as described in the main text. Also shown for comparison are errors for a vacuum-embedding fragment MBE performed over the same system with the same fragmentation; the Threshold calculations here used $\alpha = 0.5$. Errors are measured relative to a reference MP2/cc-pCVTZ calculation. Costs measure the total evaluation cost of the complete index set at each stage of the refinement. The black vertical and horizontal lines indicate the abstract cost of the reference calculation and chemical accuracy ($\approx 0.0016\,E_h$) respectively.

provided by the dehydrogenated covalent bond graph of limonin. Informal experimentation suggests that a truly faithful implementation of the original BOSSANOVA formulation performs no better, indeed significantly worse, perhaps due to issues related to both the treatment of double bonds and steric effects arising from too-close placement of link atoms.

Per-iteration plots of the absolute errors obtained during progressive refinement versus the cumulative cost expended are given in Figure 6.5. As expected, the BOSSANOVA calculations perform poorly. The progressive refinements produce a steady decrease in absolute error until three-body subgraph terms are included in the combination sum, at which point overcounting errors begin to accumulate, as discussed in Section 6.3 above. The errors of the subsequent approximations are oscillatory, but they do not fall

meaningfully and consistently below $5 \times 10^{-2}\,E_\mathrm{h}$ at best.

It is perhaps interesting to note that the adaptive MBE calculations produce errors roughly equivalent to the various BOSSANOVA calculations once their cumulative cost passes 10^{14}. It is likely that some of the MBE contribution potentials are also incurring significant errors due to proximally-placed link atoms. More interesting is that the Threshold-adaptive MBE calculation here does seem to clearly outperform the All calculation, which produces effectively a sequence of standard n-body fragment MBEs for $n = 1, 2, \ldots$ This might suggest that longer-range interactions contribute less strongly to the total energy of limonin than they do for the water clusters considered previously, and that the adaptive algorithm has more flexibility to ignore fragments separated by medium and long distances.

The error behaviour of the SUPANOVA expansions over the poset of convex subgraphs is much better behaved than that of the BOSSANOVA expansions. Although the errors for these calculations also oscillate from one iteration to the next, particularly those obtained according to Threshold adaptivity with vacuum-embedding potentials, they appear to decay algebraically against cost in the overall trend. The obtained error for a given cost is almost always better than that of the fragment-MBE calculations; for the vacuum-embedding calculations, this is usually by multiple orders of magnitude.

Even the adaptive results involving mixed-basis embedding potentials, which are full-system calculations and thus individually expensive, produce results which are still at least competitive with those of the Threshold fragment-MBE calculation in the convex subgraph case. Here, we observe that the distinction between Threshold and All calculations is very slight; indeed, for the first few iterations, it turns out here that the Threshold expansion simply selects all of the elements in the queue. It is unclear whether and to what extent this behaviour would continue for further iterations; it may be that a different and higher choice of α would produce more fine-grained iterations and be able to winnow out some of the less-important contribution potentials.

The behaviour of the error indicator and propagated uncertainty for the BOSSANOVA and convex SUPANOVA calculations using vacuum embedding potentials is displayed in Figure 6.6. The propagated uncertainties grow here much slower with calculation cost than they do for the fragment-MBE calculations discussed in the previous chapter, never surpassing $10^{-5}\,E_\mathrm{h}$. This is most likely not due to any inherent structural characteristics of the involved posets, but rather to the simple fact that far fewer subproblem potentials are explicitly involved in each combination sum than would be in a comparable fragment-MBE combination sum; this would be consistent with [RLH14; Lao+16; LH17]. Thus, we can effectively rule out numerical condition as a contributing factor to the poor performance of the BOSSANOVA calculations. The major problem remains most likely the overcounting issues inherent to combination sums over a combination-inconsistent subposet of the relevant boolean algebra.

The error indicator is again observed to behave reliably for the convex SUPANOVA calculations, both for Threshold and All adaptivity. The error indicator does at times

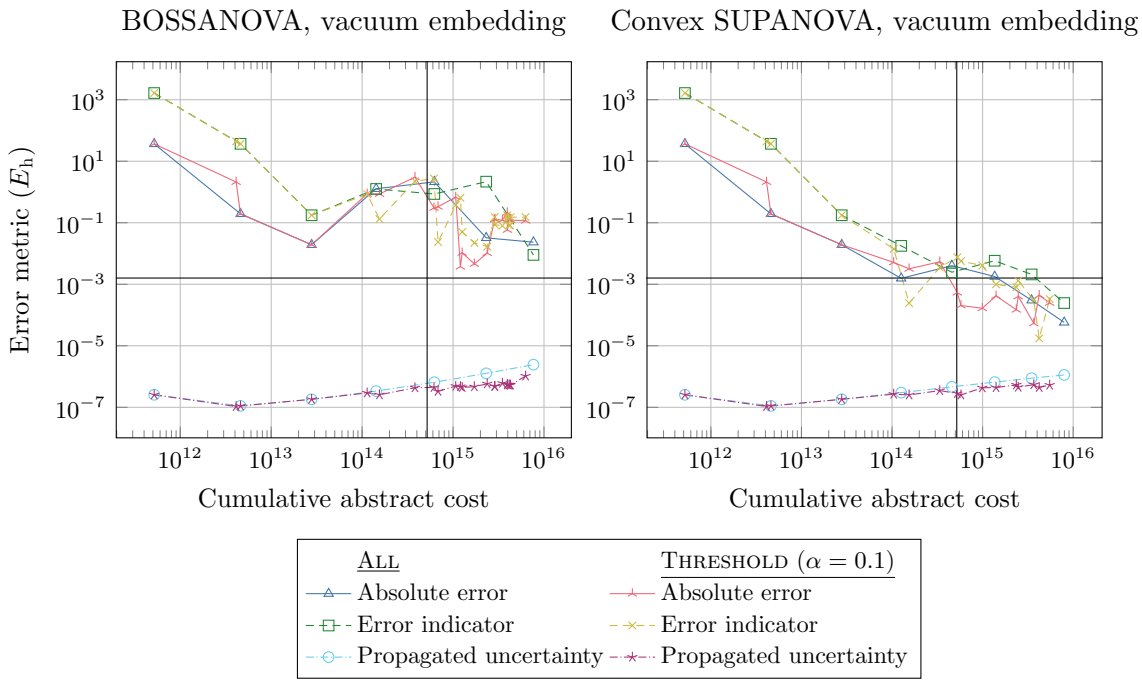

Figure 6.6.: Absolute errors, error indicators, and propagated uncertainties for progressively-refined adaptive BOSSANOVA and convex SUPANOVA calculations on limonin ($C_{26}H_{30}O_8$) using vacuum embedding subproblem potentials. Absolute errors and costs are calculated as in Figure 6.5. Error indicators are obtained as described in Section 3.5.5. Propagated uncertainties are calculated as described in Section 5.3, with an assumed uncertainty of $10^{-8}\,E_\mathrm{h}$ ascribed to the value of each individual subproblem potential. Vertical and horizontal lines indicate the abstract cost of the reference calculation and chemical accuracy ($\approx 0.0016\,E_\mathrm{h}$) respectively, also as in Figure 6.5.

fall below the true error for the Threshold calculations, but never by more than half an order of magnitude. For the BOSSANOVA cases, the error indicator appears less reliable, oscillating around the true error and in some cases underestimating it by an order of magnitude or more. Here, the same over-counting issues which afflict the full BOSSANOVA combination sums also afflict the partial combination sum used to produce the error indicator. This does not indicate a problem with the formulation of the error indicator itself, but it does emphasise that that the error indicator is deeply tied to the poset and index set which are used to derive the approximation that it measures. Thus, the error indicator does not and cannot provide an independent assessment of quality.

We turn now to chignolin. As the visualisation in Figure 6.7 shows, chignolin takes a *β-hairpin* conformation; see [Hon+04] and references within. This is held in place by

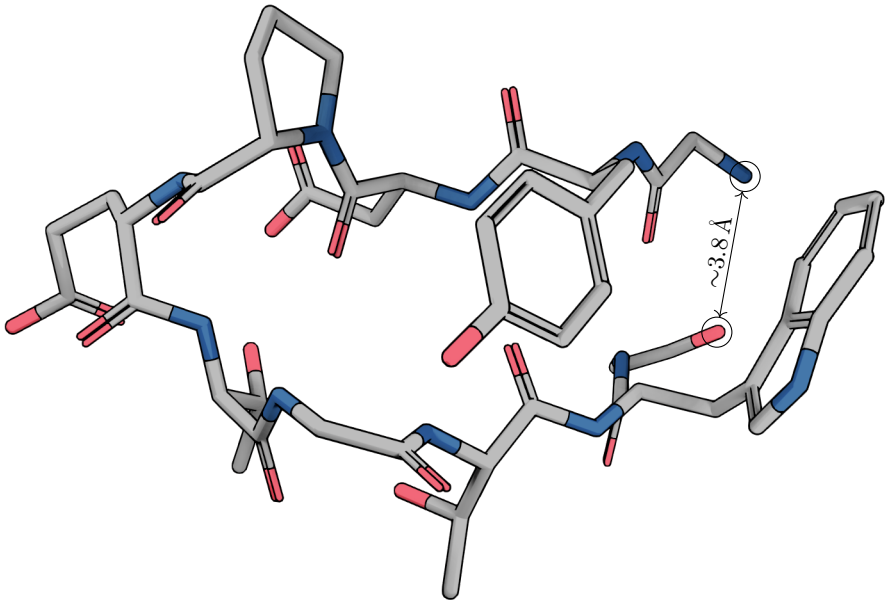

Figure 6.7.: Stick-model visualisation of chignolin ($C_{48}H_{63}N_{11}O_{18}$) [Hon+04; HY04, PDB key: 1UAO], after geometry optimisation according to B3LYP/cc-pVDZ. The visualisation format is that of Figure 6.4. Here, blue half-sticks indicate nitrogen atoms. The two circled atoms, one nitrogen and one oxygen, are separated in space by approximately 3.8 Å, as indicated. The shortest-path distance between these two atoms in the covalent bond graph of chignolin is 30.

non-covalent interactions, particularly by hydrogen bonds, between atoms on both sides of the hairpin; see again [Hon+04], particularly Figure 4.B. Chignolin contains three cyclic substructures. The first is a fused pair of aromatic rings. The remaining two are isolated rings, one aromatic and one non-aromatic.

We consider chignolin in order to investigate a situation in which the guiding assumptions underlying the use of an adaptively-truncated SUPANOVA decomposition over posets of connected induced subgraphs or convex subgraphs of covalent bond graphs may not hold. Many of the atoms in chignolin which are not close in the covalent bond graph are however close in terms of distance in space, and so may also produce strong non-covalent effects. For a particularly striking case, observe that the two circled atoms in the visualisation in Figure 6.7 are almost maximally distant from each other in the bond graph, being separated by a shortest path across 29 intervening atoms. Nevertheless, their spatial separation is only approximately 3.8 Å. It seems therefore doubtful that considering larger and larger connected or convex subgraphs of the bond graph will smoothly resolve all of the medium- to long-range effects involved in the total energy of chignolin. This is a known problem with graph-based fragmentation methods [See+22].

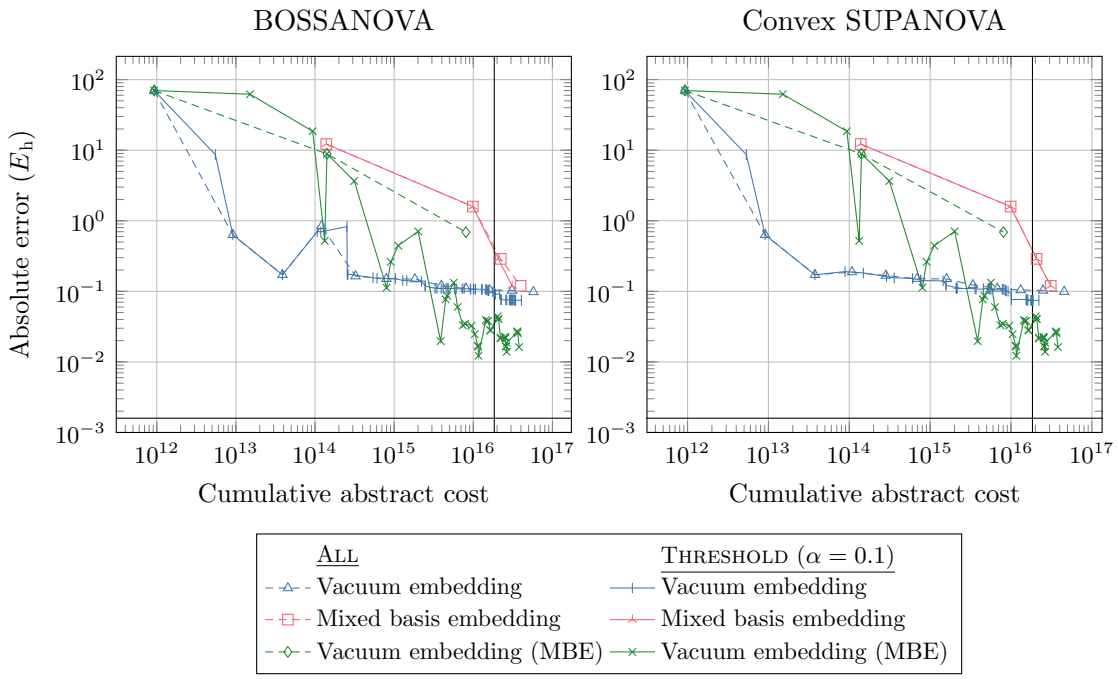

Figure 6.8.: Absolute errors for progressively-refined adaptive SUPANOVA calculations over chignolin ($C_{48}H_{63}N_{11}O_{18}$). Plot formats are as for Figure 6.5.

We report equivalent calculations on chignolin to those performed on limonin, using a fragmentation F and quotient interaction graph $G = G'/F$ of the covalent bond graph G' obtained using the same heuristic algorithm. This fragmentation F consists of 57 fragments: one singleton fragment, 35 pair fragments, 15 triplet fragments, and 6 quadruplet fragments.

Due to practical computational limits, we give adaptive calculations for chignolin up to a termination threshold set as the abstract cost of the reference calculation multiplied by a factor of two, rather than the tenfold factor used for limonin. The All vacuum-embedding fragment MBE calculation is given only up to the introduction of three-body terms; inclusion of the four-body terms would involve 395 010 subproblem potentials, and was not considered interesting enough to justify the significant additional computational expense.

Plots against cumulative cost of the per-iteration absolute errors for the BOSSANOVA and convex SUPANOVA calculations for chignolin are given in Figure 6.8, in the same format as those for limonin above. It is immediately clear that neither BOSSANOVA nor SUPANOVA truncations provide good approximations to the total energy when vacuum embeddings are used, at least not in the cost range considered. After an initially rapid decrease, the per-iteration errors produced by both All and Threshold calculations

stabilise around $10^{14}\,E_\mathrm{h}$, and do not improve even as the cost of the approximations increases by multiple orders of magnitude.

When mixed-basis embedding potentials are used, a steep decay in error is also observed from the second iteration onwards; recall that the first iteration provides in effect just a DZ-level MP2 calculation in the form of $V_\emptyset$. However, these errors also only reach approximately $10^{14}\,E_\mathrm{h}$ before the cost limit is exceeded. It is unclear from these plots whether the decay would continue at the same rate if more expensive approximations were considered. This would be interesting to know in principle, but it is not practically relevant in the face of a cheaper reference calculation.

It is interesting that the "control" fragment MBE calculations using the Threshold strategy provide the best result; not results, plural, since only one such calculation was performed, and plotted on each of the left- and right-hand side plots in Figure 6.8 for reference. We note that this is also the clearest indication that we have yet seen of a true adaptive algorithm outperforming a simple All strategy. For example, at a cumulative abstract cost of around 10^{16}, the former provides an approximation that is almost two orders of magnitude better than the latter. This suggests that many of the MBE terms in a larger system are indeed negligible and can be successfully screened by our adaptive approach. Nevertheless, even the Threshold MBE calculation does not approach chemical accuracy in the explored cost range, which is consistent with the results for fragment MBEs given in the previous chapter.

Taken together, these results, specifically those for the vacuum-embedding BOSSANOVA/SUPANOVA expansion and for the fragment MBE, seem to confirm our suspicion that the interaction graph underlying these decompositions does not suffice as a representation of the underlying system. The most likely explanation here is that those MBE terms that include components of the system on both sides of the hairpin structure are important, but are not being included in either the BOSSANOVA or SUPANOVA truncated summations, since they do not appear as connected induced subgraphs of the complete graph, nor as convex subgraphs.

In the absolute error and related error metrics for the vacuum-embedding case, shown in Figure 6.9, we find a more concerning aspect of the inability of the adaptive algorithm to locate more important contribution potentials. The error indicators for both BOSSANOVA and convex SUPANOVA calculations are far too optimistic. Although this might be ascribed in the BOSSANOVA case to the now-clear issues with that decomposition, this is not expected in the convex SUPANOVA case. Again, however, we realise that the error indicator is connected to the decomposition itself. All that the error indicator can truly tell us is that the sum of the contribution potentials in the antichains which generate the progressively-refined index sets is shrinking. This does not necessarily imply that the resulting approximation is accurate.

The differences in absolute errors between results obtained using BOSSANOVA and convex SUPANOVA appear to be much less pronounced here than for limonin. This can be understood as a consequence of the fragmentation scheme we have employed,

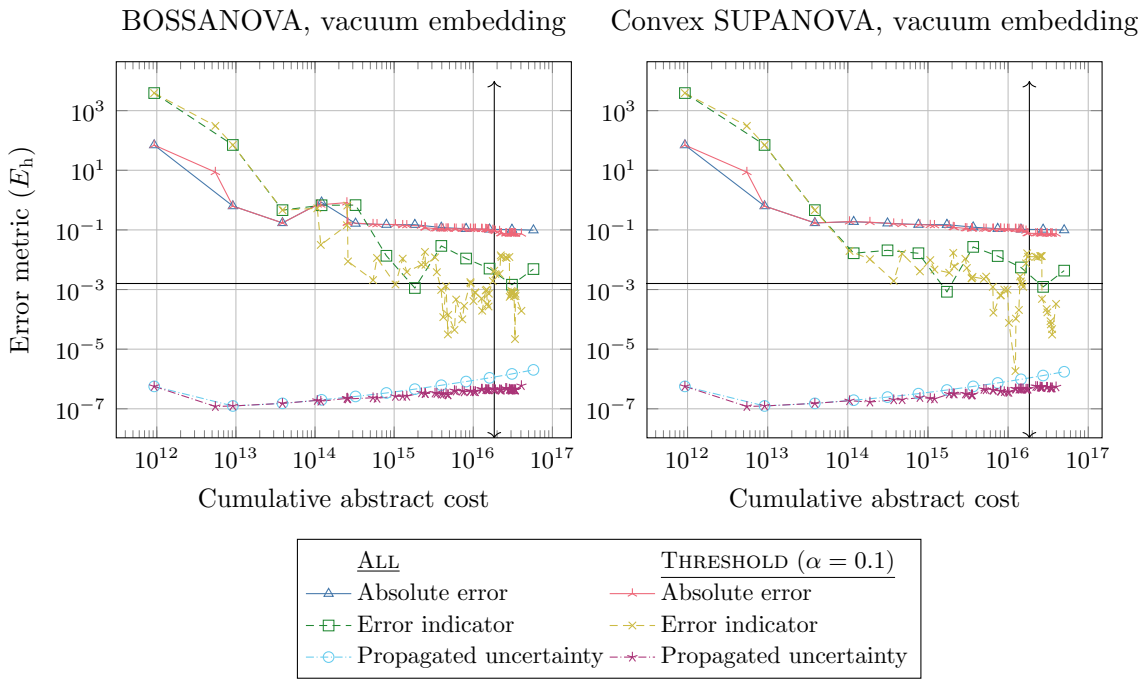

Figure 6.9.: Absolute errors, error indicators, and propagated uncertainties for progressively-refined adaptive BOSSANOVA and convex SUPANOVA calculations on chignolin ($C_{48}H_{63}N_{11}O_{18}$) using vacuum embedding subproblem potentials. Plot formats are equivalent to those in Figure 6.6.

specifically that the quotient graph G/F possesses fewer chordless cycles than does the original covalent bond graph G'. The quotient graph contains only three chordless cycles: two of length three, the fragments of which make up the two six-membered aromatic rings that involve double bonds, and one of length four, which includes the five-membered non-aromatic ring along the main line of the hairpin. The two length-three chordless cycles in the quotient graph do not lead to overcounting problems in the BOSSANOVA decomposition; only the length-four chordless cycle does. It may be the case that the resulting overcounted contribution potential values are sufficiently small in number and magnitude that their contribution falls below the larger inaccuracy caused by the inadequacy of the covalent bond graph.

It is tempting to wonder if this suggests a remedy for the problems with the BOSSANOVA decomposition that is simpler than the use of convex subgraphs. Given a covalent bond graph, we would need only construct a fragmentation such that the resulting quotient graph G contains none of the forbidden subgraphs discussed above. The poset of connected induced subgraphs would then be closed under intersection, and the significant additional complexity of the convex-subgraph approach could be avoided entirely. Put

differently, the geodesic convexity $\mathcal{M}_g[G]$ would be exactly equal to conn$[G]$; it would also be a convex geometry, and we would have a cleaner expression for its Möbius function.

There are several ways in which this could be manually achieved for chignolin. Noting that only a single carbon-nitrogen bond in the non-aromatic ring is a part of the main hairpin structure, we might group together in a fragment the three carbon atoms that lie off the hairpin. Alternatively, we might introduce an additional rule for the construction of the fragmentation, stating that any two bonded atoms which are not both carbon should be constrained to lie in the same fragment. Again, this would compress the non-aromatic ring to a triangle subgraph in the quotient graph.

Consider, however, the three-ringed structure of phenalene, as drawn in Figure 6.3. The same fragmentation scheme we have used in this section there produces a quotient graph composed of two square subgraphs and a triangle subgraph. Each of the square subgraphs is isomorphic to C_4, and thus problematic. Certainly this quotient graph and its corresponding fragmentation could be further compressed, but an algorithm which could also operate on arbitrary other similar molecular structures (graphenes, coronenes, fullerenes, etc.) is not immediate. Nor is it clear how one could perform a similar transformation for limonin, or for an arbitrary valid interaction graph in the general case.

6.7. Alternative interaction graphs

Although a SUPANOVA expansion of the total energy in terms of the convex subgraphs of the covalent bond graph of a molecule is combination-consistent with the underlying MBE, the results shown for chignolin in the previous section suggest that adaptive truncation of that expansion may not be well-behaved if the bond graph provides a poor representation of the spatial connectivity of that molecule.

Additionally, such a SUPANOVA decomposition is only formally exact if the complete covalent bond graph of the system is itself connected. This is trivially not so in the case of the water clusters considered in the previous chapter, nor for any cluster of non-covalently bonded molecules. It may also be not be true for other macromolecular systems. Consider, for example, the famous double-helix structure of DNA; the two helices are not covalently bonded, but are instead held coherent by hydrogen bonds between base pairs [Gue+00].

We consider briefly a way to treat the former issue, that bond graphs may underdescribe the spatial connectivity of a molecular system, and in so doing also treat the latter issue, that the full-system covalent bond graph may be disconnected. Specifically, we can use a different full-system interaction graph, which includes a superset of the direct interactions in the original covalent bond graph, and which can always be made to be connected.

For this purpose, we will consider what we will call *radial interaction graphs*, which have the edge set $E = \{\{i, j\} \subseteq [M] \mid i \neq j,\ \|R_i - R_j\| \leq r_{\text{cut}}\}$ for some choice of cutoff $r_{\text{cut}} > 0$. Here, each atom interacts directly only with those atoms which lie within a

surrounding sphere of radius r_{cut}. This is just the same idea as applied by effectively all of the distance-based thresholding methods mentioned above; see, for very non-exhaustive example, the original definition of the GEBF [LLJ07], and a study of the SFM method as applied to water clusters [Pru+12]. The idea is also well-known in other areas of computational chemistry and molecular dynamics, particularly in the construction of approximate potential functions [GKZ07; Bar+10; TM11; FT13; Bar+17; Jen17; GGG20]. Radial interaction graphs have been directly considered and used by other existing fragmentation methods, e.g., [Fis18; RI20; RKI20; KDI21; Zha+21], sometimes explicitly to garner non-bonded interactions. In fact, a slightly more sophisticated idea with species-dependent values of r_{cut} is used in, e.g., [CCB14; See+22] to construct a bond graph itself.

Since the precise choice of r_{cut} determines the graph, so will it also determine the behaviour and characteristics of truncations of the resulting SUPANOVA decomposition. If r_{cut} is taken to be greater or equal to the Euclidean length of the longest bond represented in the covalent bond graph, the edge set of the resulting radial interaction graph will be a superset of the edge set of the covalent bond graph. There must also exist some minimal r_{cut} such that the radial interaction graph is connected. This value is bounded above by the largest pairwise distance between atoms in the system, but will generally be smaller in magnitude. If a radial interaction graph G is constructed using an r_{cut} which exceeds this threshold, then it will be connected, and so any SUPANOVA decomposition in terms of the poset of convex subgraphs of G will be formally exact.

It seems reasonable to wish to choose an r_{cut} sufficiently large enough to enmesh any important direct interactions which are omitted from the covalent bond graph, but no larger, so as to retain as much sparsity in the graph as possible. If r_{cut} is taken sufficiently large, the according radial interaction graph will be complete, and every involved atom modelled as interacting directly with every other atom. In this case, every induced subgraph is convex, and so the convex SUPANOVA decomposition becomes nothing more than a standard nuclear or fragment MBE.

We return now not only to chignolin, but also to the 55-monomer water cluster $(H_2O)_{55}$ considered in the previous chapter. The structures and thus energies of both systems are known to be influenced by non-covalent bonding. As remarked above, chignolin contains hydrogen bonds across the gap formed by its β-hairpin conformation [Hon+04]. The conformations assumed by water clusters are also deeply influenced by hydrogen bonds [Xan94; RS15]; this is visible to some extent in Figure 5.1, where the individual water molecules are oriented such that their oxygen-hydrogen bonds tend to point towards the oxygen atoms in other molecules. Informally and roughly, hydrogen bonds such as these are around 2 Å in length, if one measures from the hydrogen atom to the heavy atom [HM99]; see also [Gue+00] for a discussion of hydrogen bond lengths in the DNA case. Thus, for each system, we equip each system with a radial interaction graph, constructed with a cutoff threshold of $r_{\text{cut}} = 2.5\,\text{Å}$; the intention here is to model at least all hydrogen bonds, allowing a reasonable threshold for similar interactions at a slightly

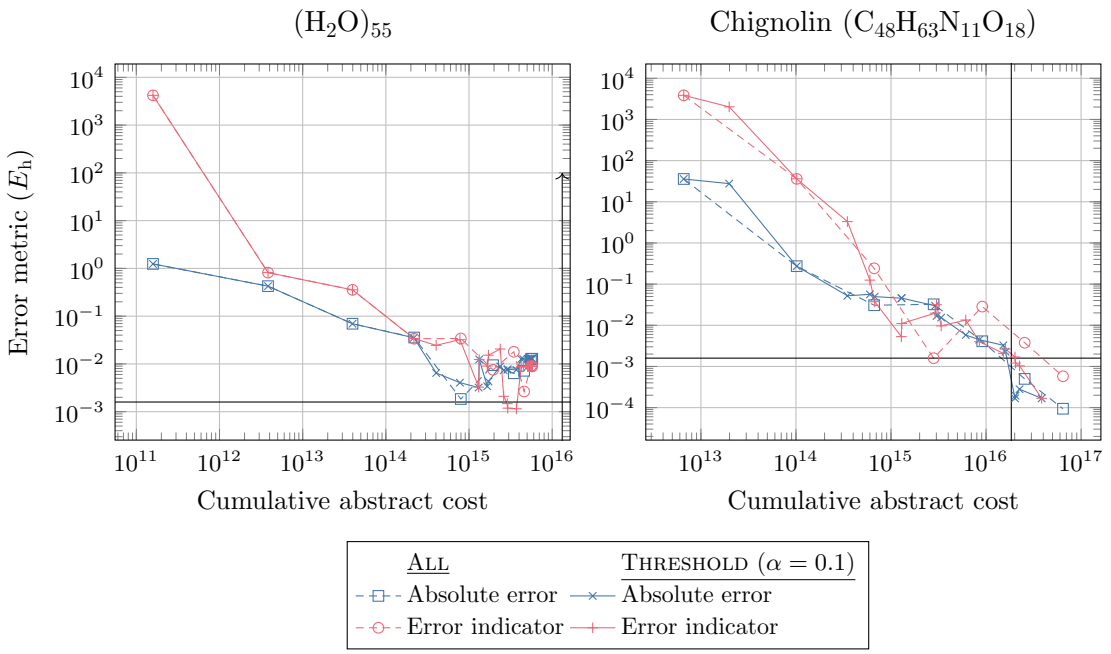

Figure 6.10.: Absolute errors and error indicators for progressively-refined adaptive convex SUPANOVA calculations over $(H_2O)_{55}$ and chignolin ($C_{48}H_{63}N_{11}O_{18}$), at the MP2/cc-pCVTZ level of theory. The calculations are in terms of posets of convex subgraphs of processed radial interaction graphs with $r_{\text{cut}} = 2.5\,\text{Å}$ as described in the main text. Horizontal and vertical lines display the threshold of chemical accuracy and the cost of the reference, respectively.

longer length scale.

Again, we consider approximate calculations of the MP2/cc-pCVTZ total energies of both systems using adaptively-refined poset-grid combination sums. The radial graphs used for both systems are highly cyclic, so we consider only calculations in terms of convex SUPANOVA decompositions, and for simplicity, only vacuum embedding potentials. Again, both All and Threshold ($\alpha = 0.1$) strategies are considered.

Per-iteration results for the calculations using radial interaction graphs are given in Figure 6.10. The calculations for $(H_2O)_{55}$, both All and Threshold, proved to be only feasible up to a total abstract cost of approximately 10^{16}. This was due to an interesting emergent property of the poset of convex subgraphs, which is related to the observation that $\mathcal{M}_g[G]$ cannot in this case be a convex geometry. Instead, $\mathcal{M}_g[G]$ contains saturated[17] chains between the empty induced subgraph and the full (sub)graph G which are of widely varying length. The longest saturated chain in the poset is of

[17] A saturated chain in a poset P is one where, informally, each element of that chain is covered by the next element. See [Sta12] for a more precise definition.

length 13, while the shortest is of length 6. In the latter case, the numbers of monomers in the involved fragments are 0, 1, 2, 3, 4, 5, and then suddenly 55; put differently, there is a five-monomer convex-subgraph fragment which is covered in $\mathcal{M}_g[G]$ by the complete system.

This is quite different to the poset of connected subgraphs $\mathrm{conn}[G]$, where every saturated chain of that poset has the same length, 56, and adds a single monomer at each element of that chain. This structure of $\mathcal{M}_g[G]$ means in practice that the expansion of a small subgraph in the adaptive index set algorithm can introduce a dramatically larger subgraph into the index set, and the cost associated with this larger subgraph can be several orders of magnitude greater than that of the predecessor which introduced it. Precisely this occurred during the calculations for $(H_2O)_{55}$: at some iteration of the adaptive process, a large jump in the individual cost of some or all of the required MP2 calculations occurred, and these calculations were not practically feasible with the PySCF solver on the particular hardware resources used.

This effect is a good illustration of the desirability of a SUPANOVA decomposition in terms of a convex geometry. A thorough analysis of this behaviour would surely be of interest, as would be the development of a mechanism by which a radial-type interaction graph might be structured so as to ameliorate its impact, but these are well beyond the scope of this thesis. There is likely to be a connection here to the *hull number* of a graph [ES85]: the size of the smallest subgraph G' of some graph G such that $\mathrm{CH}_g[G'] = G$. The hull number and related quantities have been well-studied for various classes of graph, see, e.g., [BO13; Dou+09; Ton09; JG12] and Chapter 2 of [Pel13]. Note that the five-monomer fragment mentioned above would then correspond to a hull graph of the full graph. Also interesting here is an empirical study of convexity in a variety of real-world networks [MŠ18].

The limited results for $(H_2O)_{55}$ that we do have are slightly better in comparison to those for the conventional MBE setting as plotted in Figure 5.2 in the previous chapter, and are at least more stable. Both All and Threshold calculations comfortably achieve a true accuracy of $10^{-2}\,E_\mathrm{h}$. Although these results are not particularly impressive, it is reassuring to see that the adaptive error indicator is again acceptably accurate, never over- or underestimating the true error by more than an order of magnitude.

The convex-subgraph poset $\mathcal{M}_g[G]$ for the equivalent quotient graph for chignolin seems to be better-behaved than that for $(H_2O)_{55}$, and the performance of the convex SUPANOVA method is in this case much more encouraging. Both All and Threshold calculations show a progressive and smooth decay in error as the cost increases. In the final iterations as plotted, the combination sums comfortably and convincingly pass chemical accuracy. The error indicators are also generally well-behaved and accurate, and consistently overestimate the true error, although there is one particular All iteration where the error indicator instead underestimates the true error by slightly more than an order of magnitude.

The results for chignolin seem then to suggest that our suspicions about the descriptive

quality of the covalent bond graph were well-founded, and that the use of a radial interaction graph can indeed produce a SUPANOVA decomposition that is systematically improvable and does not produce misleading error indicators. Despite this, however, the SUPANOVA calculations still do not produce a meaningful speedup when considered against the reference calculation. Presumably, if such is to materialise, we must consider even larger systems again.

6.8. Case study: proteins

We conclude the chapter by investigating the application of the convex SUPANOVA decomposition to two proteins, one relatively small, the other enormous.

6.8.1. Antifreeze protein

We consider first a model of a natural *antifreeze protein* [Sön+96; Sön+97, PDB key: 1KDF], which we obtained from the Protein Data Bank (PDB). We converted the representation of 1KDF from the original PDB format into MOL format and explicitly added hydrogen atoms using the OpenBabel toolkit [OBo+11]. The resulting structure has the empirical formula $C_{303}H_{512}N_{82}O_{89}S_3$, for a total of 991 atoms, 479 of which are non-hydrogen. The covalent bond graph of 1KDF is connected, and contains seven chordless cycles, all of length either five or six. An abstract visualisation of the non-hydrogenic bond connectivity structure of 1KDF is given in Figure 6.11, with those cyclic substructures highlighted. Unlike previous systems considered in this chapter, we performed no further optimisation on the geometry of 1KDF, and treated it exactly as obtained from the PDB, up to the addition of hydrogens.

This particular protein is briefly investigated in [Heb14, Sec. 9.6.2] in the context of the BOSSANOVA decomposition. There, a three-body BOSSANOVA truncation provides an approximate MP2/6-311G* total energy of 1KDF as $E^{\mathrm{MP2}}_{\text{6-311G*}} \approx -24\,996.98\,E_\mathrm{h}$. The accuracy of this approximation is not explicitly estimated. In the context of similar calculations reported in that work, however, we suggest that this approximation might have a relative error between 10^{-3} and 10^{-4}. Despite the presence of chordless cycles in the bond graph, this approximation will not be afflicted with the inconsistency-related errors discussed in Section 6.3, since only up to three-body terms are used and no chordless cycles of length four exist in the bond graph. However, such a relative error would lead to an absolute error at least on the order of $10\,E_\mathrm{h}$. Such an approximation is unlikely to be of any great use without further refinement, especially without validation of its reliability, and any further refinement would necessarily introduce consistency-related errors.

We now investigate whether an adaptive convex SUPANOVA approximation can provide a refineable estimate of the total energy of 1KDF. We attempted to perform reference MP2 total energy calculations on 1KDF at various levels of basis set theory,

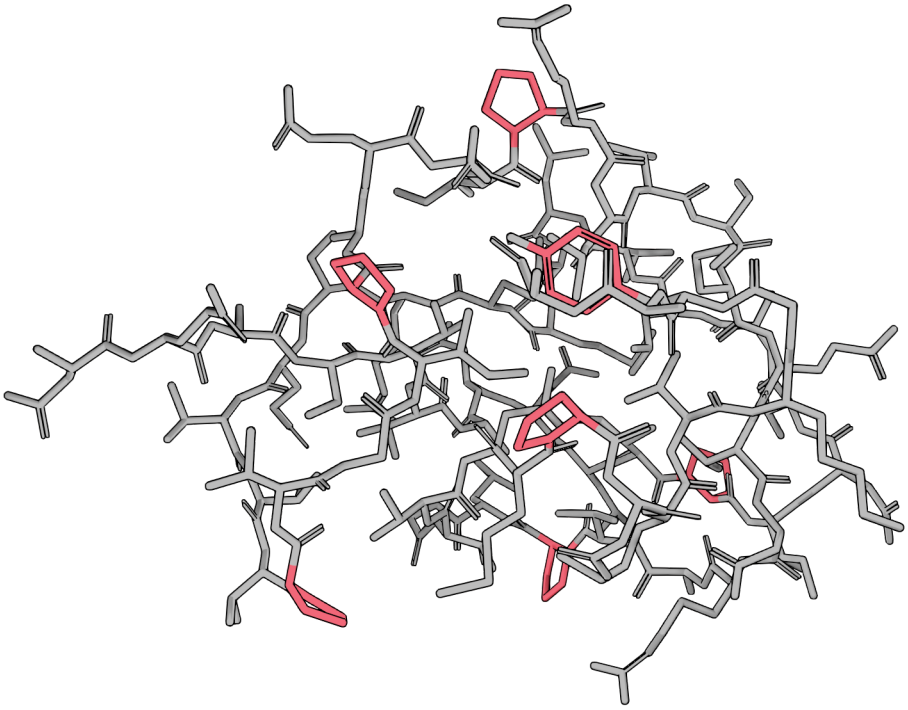

Figure 6.11.: Stick-model visualisation of the covalent bond structure of the antifreeze protein [Sön+96; Sön+97, PDB key: 1KDF]; bonds to hydrogen atoms are not shown. Chordless cycles in the bond graph are highlighted in red; all other elements are left universally grey.

including 6-311G*, but found these calculations to be computationally impractical for any non-trivial basis set. Instead, we consider here only the HF/cc-pVTZ [SO89; Dun89] total energy of 1KDF, $E^{\mathrm{HF}}_{\text{cc-pVTZ}} \approx -24\,896.350\,029\,E_{\mathrm{h}}$, which — requiring a calculation in terms of 21 558 contracted basis functions — was the best-available reference value that we could reasonably obtain.

The set of atoms composing the 1KDF system were split into a fragmentation F according to the heuristic algorithm outlined above. This fragmentation comprised 143 fragments, with fragment sizes ranging between two (four fragments) and 17 (one fragment), most (131 fragments) involving ten or fewer atoms. We then constructed the quotient graph $G = G'/F$ of the radial interaction graph G' of 1KDF with $r_{\text{cut}} = 2.5\,\text{Å}$, similarly to as in the previous section.

As above, we consider here two adaptive convex SUPANOVA calculations in terms of G, using the All and Threshold adaptive strategies respectively, the latter with $\alpha = 0.1$.

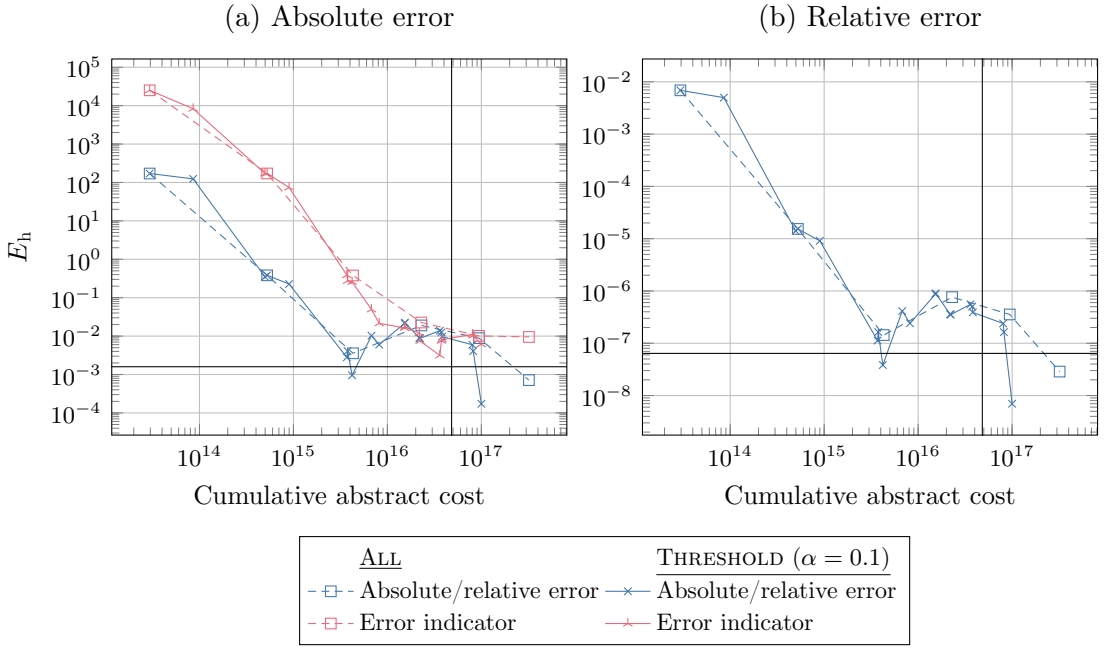

Figure 6.12.: Error metrics for adaptive SUPANOVA calculations on the antifreeze protein. Calculations are in terms of the poset $\mathcal{M}_g[G]$ of convex subgraphs of a radial interaction graph with $r_\mathrm{cut} = 2.5\,\text{Å}$, as described in the text. The left-hand plot (a) shows per-iteration absolute errors, measured against a reference value of $E^\mathrm{HF}_\text{cc-pVTZ} \approx -24\,896.350\,029\,E_\mathrm{h}$, as well as the absolute values of the error indicator. The right-hand plot (b) shows per-iteration relative errors. Solid black vertical and horizontal lines indicate reference abstract cost and chemical accuracy as usual; in plot (b), chemical accuracy is measured as the relative error of the difference of the reference value and $1\,\text{kcal}\,\text{mol}^{-1}$.

Only vacuum embedding subproblem potentials are considered. These calculations were terminated once the total cumulative abstract cost exceeded that of the reference calculation by a factor of two. Per-iteration results for the true absolute errors and calculated error indicators of these calculations versus the cumulative abstract costs are plotted in Figure 6.12. Here, we depart from the prequel in plotting not only the absolute error of the per-iteration results (in the left-hand plot), but also the corresponding relative errors (in the right-hand plot).

The results here are mixed. Most importantly, the per-iteration results for both calculations show that the convex SUPANOVA approximation does appear to approximate the HF/cc-pVTZ total energy more and more accurately. Also, the error indicators appear to track the true absolute error quite closely, at least once past an initial regime where they significantly overestimate it. From this, we conclude that the radial interaction

graph is again an appropriately comprehensive description of the system.

However, neither of the two adaptive calculations is quite able to convincingly obtain chemical accuracy, certainly not with any kind of speedup relative to the reference calculation. This is somewhat discouraging — 1KDF is by normal quantum chemical standards a large system, so even if speedup were only to be found for large systems, we would hope to see at least a glimmer of it here. However, it should still be noted that since the threshold of chemical accuracy is of course fixed, calculations over larger and larger systems will require more and more relative accuracy in order to exceed it.

6.8.2. SARS-CoV-2 spike glycoprotein

The infection process of the SARS-CoV-2 (*Severe Acute Respiratory Syndrome Coronavirus 2*) [Wal+20b] virus connected to the COVID-19 pandemic is driven by *spike glycoproteins* arrayed over its exterior [TV19; Wal+20b; Mur+21]. We consider here a model of the SARS-CoV-2 spike glycoprotein obtained via cryoelectron microscopy and published in the Protein Data Bank [Wal+20b; Wal+20a, PDB key: 6VXX]; see the visualisation in Figure 6.13, and also those in Figure 3 of [Wal+20b]. As for 1KDF above, we explicitly hydrogenated the original model of 6VXX using OpenBabel [OBo+11]. The resulting molecular system consists of 27 distinct non-covalently bonded subunits, ranging in size between 63 and 2690 non-hydrogen atoms. In total, these subunits involve 46 923 atoms, connected by 47 526 covalent bonds.

From an *ab initio* quantum chemical standpoint, this glycoprotein is prohibitively large. Equipped with the STO-2G minimal basis set [HSP69; Heh+70], a description of the 6VXX glycoprotein system would require 142 095 atomic orbitals, increasing to 1 036 422 AOs for the more moderate cc-pVTZ basis set, and 9 565 164 AOs for aug-cc-pCV6Z. These are problem formulations that are vastly beyond the capabilities of any conventional quantum-chemical solver. However, some research has been performed [Aki+21] using the *fragment molecular orbital* (FMO) method [Kit+99; FNK12] to investigate interaction energies between subunits of the spike glycoprotein, as well as between the glycoprotein and certain enzymes and antibodies possibly related to virus infection. These calculations involved post-HF calculations up to a simplified formulation of MP4, applying the 6-31G* and cc-pVDZ basis sets and considering only up to two-body fragment terms. Although the computational requirements for this study were substantial, involving well more than 100 000 CPU cores working in parallel, the authors were still prevented for reasons of feasibility from performing a more desirable cc-pVTZ calculation [Aki+21, Supp. info.].

As a speculative application of our SUPANOVA approach, we attempted to approximate the Hartree-Fock total energy of the modelled spike glycoprotein according to the cc-pVTZ basis set. We calculated the radial interaction graph G' for the spike protein with $r_{\text{cut}} = 2.5\,\text{Å}$. When applied to this graph, the heuristic fragmentation method outlined above produced a fragmentation F containing 7524 fragments, ranging in size from 93 monoatomic fragments up to seven fragments containing 17 atoms. Thus, we consider

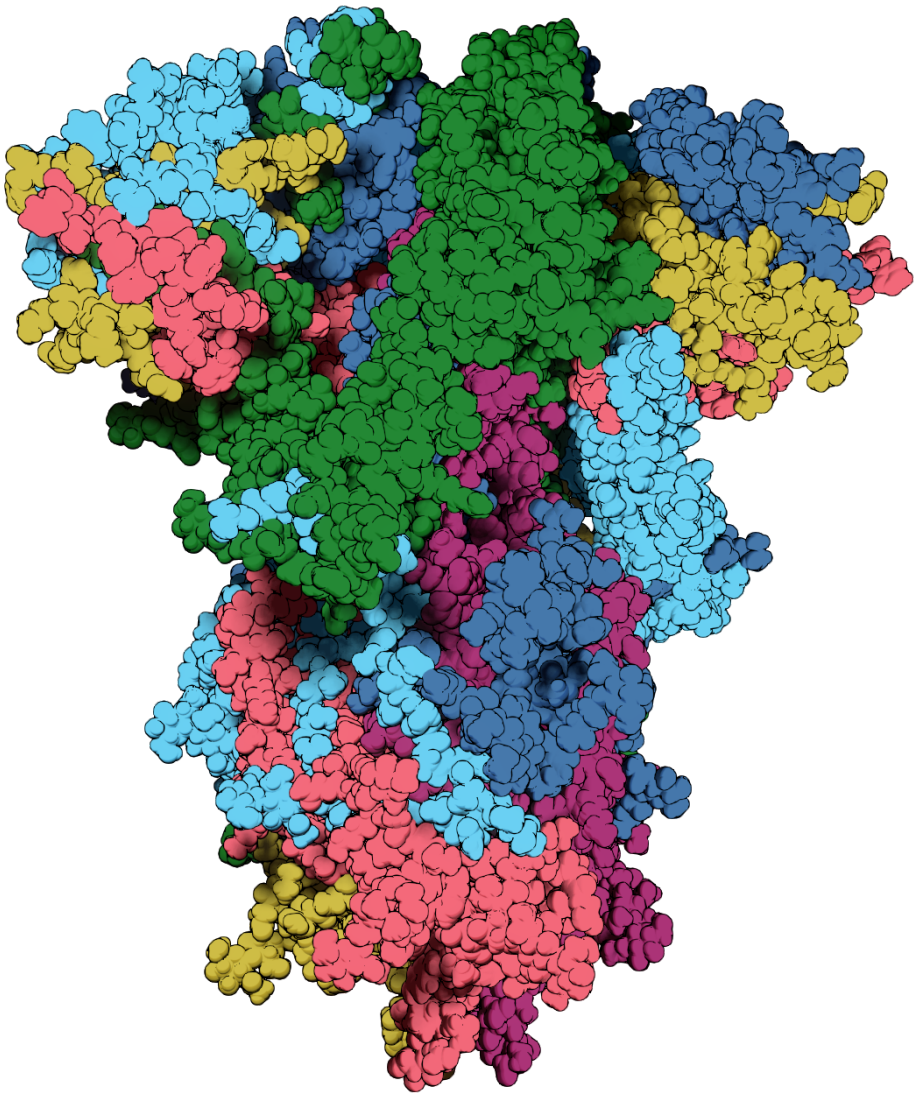

Figure 6.13.: Space-filling visualisation of the SARS-CoV-2 spike glycoprotein [Wal+20b; Wal+20a, PDB key: 6VXX]. Each atom is represented by a sphere of species-appropriate van der Waals radius, according to [Man+09]. Atoms are coloured by membership of the same covalently-bonded subunit, corresponding to a connected component of the covalent bond graph provided implicitly by the OpenBabel toolkit [OBo+11] after explicit hydrogenisation. Colours are reused for multiple subunits, but are assigned so that no two spatially-adjacent subunits share the same colour. For further details, see Section A.9.

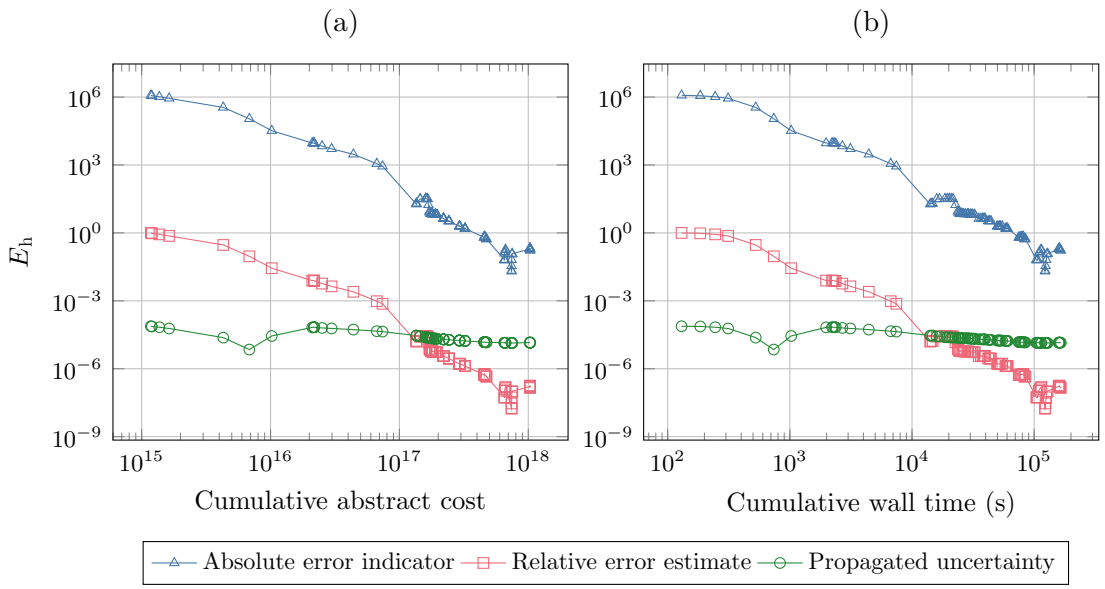

Figure 6.14.: Error metrics for an adaptive convex SUPANOVA calculation of the HF/cc-pCVTZ total energy of the SARS-CoV-2 spike glycoprotein (PDB: 6VXX). Each plot shows the per-iteration absolute value of the adaptive error indicators and propagated uncertainities, as well as an estimate of the relative error, calculated as the absolute value of the ratio of the error indicator to the corresponding approximation. The left-hand plot shows these metrics as a function of the total abstract cost of all elements in the adaptively-obtained index set at each iteration; the right-hand plot, as a function of the cumulative wall time required by the calculation up to and including the relevant iteration.

the single-axis convex SUPANOVA grid $\Pi = \mathcal{M}_g[G'/F]$, similarly to previous examples.

This calculation used the parallel implementation of the adaptive index-set calculation algorithm outlined in Section A.8. The required single-point calculations at each iteration were distributed across 146 nodes of an HPC cluster, for a total of 4864 concurrently-utilised cores. Additionally, one node was used to execute the adaptive index-set algorithm itself, and another to broker task distribution. Although this is a non-trivial set of computational resources, it still represents only a few percent of those used in [Aki+21].

We performed a single self-contained calculation run, using the Threshold refinement strategy with threshold $\alpha = 0.1$. All calculations were performed using PySCF [Sun15; Sun+17; Sun+20], with equivalent calculation settings to those used for previously-described results in this chapter. No reference value was available to measure accuracy against. We were also unable to calculate a reference abstract cost for the complete full system calculation, due to the computational effort required to assess the number of non-negligible ERIs. Thus, we simply allowed the calculation to run for a period of some

46 h, and assess the quality of the resulting approximation via only the error indicator and the propagated calculation uncertainty.

All involved subproblem potentials were explicitly evaluated from scratch during the execution of this calculation; thus, we are able here to validly consider real-world wall times as well as abstract costs. Plots of the error indicator and propagated uncertainty are given in Figure 6.14, measured both against the total abstract cost of the adaptively obtained index set at each iteration, and also against the total cumulative wall time expended at each iteration. The two sets of plots are very similar in shape, which offers some confirmation that the abstract cost model of Section 2.5 is and has been a reliable method for assessing the expense of the various calculations described throughout this thesis, at least in the Hartree-Fock case. Some small differences can be attributed to particular iterations of the adaptive algorithm in which only a few elements were added to the index set, at most a low constant multiple of the number of available calculator processes and with costs varying by several orders of magnitude. As these iterations introduce parallel inefficiency, their marginal cost in wall time terms is much greater than according to the abstract cost model.

Regardless of whether it is plotted against abstract cost or wall time, the absolute value of the error indicator exhibits an overall smooth pattern of decay. It is interesting to note that the propagated uncertainty, which assumes a per-calculation uncertainty of $\varepsilon = 10^{-8}$, does not change greatly over the course of the calculation. This is important; the approximated total energy at the final iteration is calculated as a combination sum over 67 478 distinct subproblem potentials, and so it is reassuring to know that this value seems to be reasonably free of numerical issues such as those discussed in Section 5.3 in the context of a standard MBE. It seems reasonable to hope that, were the calculation continued to a point where the error indicator fell below the chemical accuracy threshold, the propagated uncertainty would itself still be below this threshold.

With due regard of the various caveats discussed in this chapter, and particularly under the reasonable but unproven assumption that the constructed radial interaction graph is sufficiently descriptive of the system — as it was for chignolin and 1KDF — we can then tentatively suggest that the HF/cc-pVTZ total energy of 6VXX as we have considered it is approximately $E^{\mathrm{HF}}_{\text{cc-pVTZ}} = -1\,183\,006.07(17)\,E_{\mathrm{h}}$. This result does not achieve chemical accuracy. It is perhaps worth noting, however, that chemical accuracy would require here a relative error around 10^{-10}, which is two orders of magnitude greater again than that required for chemical accuracy over 1KDF.

It must be stressed that this calculation is thoroughly naïve, and ignores many important considerations that should more properly be taken into account. In particular, we have implicitly assumed here that the system under study possesses a net-zero total charge, as do all fragments, but this is likely not valid; see for instance comments about charged residues in [Aki+21]. The work in that source also deemed necessary a comprehensive pre-calculation manipulation of the glycoprotein; see [Aki+21, Supp. info.]. Even were such issues to be addressed, significant further benchmarking and validation would be

required before the result could be viewed as reliable. An adaptation to the calculation of the interaction energies described in [Aki+21] and a detailed comparison with that work would be an obvious first step.

For such a comparison to be meaningful, a treatment of the correlation energy to at least the MP2 level of theory would be necessary. We were prevented from doing so here only by a technical limitation of the particular pairing of software and hardware that we used. Specifically, the MP2 implementation available in PySCF relies on the use of disk storage to hold partially-transformed molecular integrals once the size of the ERI tensors become too large to hold entirely in memory. Most of the compute nodes used here were equipped with only relatively small amounts of local storage (approximately 80 GB), and thus could not successfully compute MP2 correlation energies for subproblems involving many more than approximately 1000 contracted basis functions. Some of the compute nodes, however, were equipped with substantially more storage (up to approximately 3 TB); calculations on these nodes were able to compute MP2 energies in reasonable time for subproblems up to and exceeding 2000 basis functions. As such, we believe that an equivalent calculation at the MP2/cc-pVTZ or MP2/cc-pCVTZ level would be quite feasible, given either a reasonable increase in the amount of per-node local storage, or alternatively, the use of a suitable, fully integral-direct MP2 implementation, similar to, e.g., that described in [WHR96].

In conclusion, although we are cautious not to read too much into the result from the perspective of accuracy, this calculation demonstrates that an application of the convex SUPANOVA approach to extremely large molecules is technically feasible, given access to a reasonable set of computational resources. Moreover, the resulting combination sum seems to be numerically stable.

7. Multilevel SUPANOVA decompositions

In the three preceeding chapters, we have considered several different applications of the order-theoretic combination technique to the calculation of energetic properties of various molecular systems. We briefly recapitulate the ideas there covered.

In Chapter 4, basing on previous work in [Zas+18], we outlined a generalised composite method (GCM) for approximating the FCI/CBS total and atomisation energies of molecules via combination sums taken over a four-axis poset grid. In this grid, in particular, one axis indexes increasingly detailed basis sets, and another indexes higher- and higher-quality treatments of electron correlation. All involved axes are simple chain posets, and the resulting poset grid Π was no more complex than that used implicitly by the standard combination technique in its original form.

In Chapters 5 and 6, we investigated MBE-style combination sums taken over a poset grid composed of a single poset axis. This axis was initially taken to be the boolean algebra B_M of all subsets of the nuclear indices, or, similarly, the poset B_K of all subsets of a particular fragmentation of those indices, corresponding to conventional forms of the MBE. We then considered what we refer to as the SUPANOVA class of decompositions, cf. [GHH14; Heb14; CGH18], which involve subposets of the poset of induced subgraphs of some full-system interaction graph G. An arbitrary SUPANOVA truncation after an order ideal of $\mathcal{M}_g(G)$, the poset (axis) corresponding to the geodesically convex subgraphs of G, is by construction combination-consistent with a truncation of the underlying nuclear MBE. Adaptively-refined combination sums over a convex SUPANOVA axis appear numerically stable and, in some limited circumstances, seem to offer a performance benefit relative to a conventional full-system calculation.

Both the GCM and SUPANOVA approaches are, however, restricted in their applicability. Like other high-accuracy composite methods [RS15; Kar16], the GCM can only be usefully brought to bear on very small molecular systems. This is due to the prohibitive scaling behaviour along each poset axis, particularly in terms of the costs of higher-quality approximations of the correlation energy. But the discussion and results in Chapter 4, as well as the literature there reviewed, make clear that such calculations are necessary if the GCM or indeed any composite method is to approximate either the true total energy or the true atomisation energy of a system to a level that approaches chemical accuracy, let alone surpasses it.

By contrast, much like other fragmentation methods [Gor+11; CB15; RS15; Her19], the various SUPANOVA-type decompositions that we have discussed can be applied to very large systems. However, it seems that relatively higher-order SUPANOVA contribution

terms are necessary in order to approximate the total energy of such systems with reasonable accuracy relative to a full-system calculation at an equivalent level of theory. The scope of the terms required depends on the choice of fragmentation and interaction graph. This has implications for the accurate calculation of correlation energies of large systems using SUPANOVA decompositions. The same high-order computational scaling of truncated coupled cluster approximations that limits the GCM will also impact the calculation and feasibility of the many higher-order subproblem potentials required for a SUPANOVA calculation, especially if a better-quality basis set is to be used for the subproblem potential calculations.

In this final chapter, we investigate a joining of the GCM and SUPANOVA techniques. Guided by the example of the ML-BOSSANOVA method [CGH18], the FCI/CBS total energy of a molecular system is subjected to a *multilevel SUPANOVA* (ML-SUPANOVA) decomposition, which contains terms representing the additional contribution of subsystem calculations not only relative to their individual subsubsystems, as in the standard SUPANOVA decomposition, but also in terms of calculations over the same subsystems using lower-quality treaments of electron correlation and/or basis sets. Informally, we hope that each term in a formal SUPANOVA decomposition of the FCI/CBS total energy can itself be efficiently approximated by a GCM combination sum; and moreover, that since SUPANOVA terms $\tilde{V}_{\mathbf{u}}$ are generally expected to be smaller in magnitude for larger subsystems $\mathbf{u}$, then overall accuracy to some particular threshold may require only a "cheaper" GCM combination sum for larger subsystems than for smaller.

The ML-SUPANOVA decomposition requires no additional theory to construct. It is obtained by simply assembling a three-axis poset grid as the direct product of two GCM poset axes with an appropriate SUPANOVA poset axis. For simplicity, and as discussed in Chapter 4, we omit the two "fine-tuning" axes that were considered in that chapter. The adaptive algorithm described in Chapter 3 is applicable here almost without change; one minor adjustment is required for technical reasons, which we shall discuss below.

As in previous chapters, the basic idea we explore here — that of a multilevel MBE-style expansion — is not in and of itself novel, and part of our aim is to demonstrate that the order-theoretic combination technique can be used to easily rederive and extend on some existing approaches. Thus, we begin with a brief summary of some multilevel approaches which have already been applied in the setting of subsystem-based techniques.

7.1. Multilevel energy-based fragmentation methods

For the purposes of this chapter, we consider a *multilevel* technique to be one that involves multiple quantum chemical calculations, not all of which are performed at the same level of computational theory. This is consistent with usage throughout the existing literature, both with respect to subsystem techniques [ŘS09; MR11; CGH18], and also more generally [Zas+18]. We note in particular that the ONIOM-style approaches

mentioned in Section 5.1.1 are multilevel techniques according to this definition, although they are sometimes referred to as *multilayer(ed)* [Sve+96; Gor+11; RH12; LH16]; see again [Chu+15] for a comprehensive summary of the ONIOM family.

Multilevel-style extensions have been proposed for a number of the energy-based fragmentation methods that were outlined in Chapters 5 and 6. What follows here is not intended to be exhaustive, and again, we are mostly seeking to collect energy expressions for comparison. We refer again to and are influenced by the more complete reviews listed at the beginning of Section 5.1, particularly [Gor+11; Chu+15; CB15; RS15; Her19], and also to [RH12].

Multilevel fragment-based techniques are often grounded in the ONIOM methodology, and we are also particularly influenced in the following by the summary in Section 2.3.2 of [Chu+15]. The *multicentered QM/QM* (MC QM/QM) approach was initially presented by Hopkins and Tschumper [HT03] as an extension of an ONIOM-style framework. Here, some number m of disjoint *centers*[1] are selected for additional theoretical attention. The resulting MC QM/QM energy expression becomes, in the ONIOM-based notation of the source [HT03, (3)],

$$E^{\mathrm{MC}}_{\mathrm{QM/QM}} = E_{\mathrm{Low}}(\mathrm{Real}) + \sum_{i=1}^{m} E_{\mathrm{High}}(\mathrm{Model}_i) - E_{\mathrm{Low}}(\mathrm{Model}_i), \tag{7.1}$$

that is, just the full-system low-level energy and a correction for the high-level energy of each distinct center. The MC QM/QM has since been generalised from this one-body form to two-body and higher-order forms [HT05; Tsc06; Bat+11], and in particular has been adjusted to account for m potentially overlapping centers via an equivalent application of the cardinality form of the principle of inclusion/exclusion to those discussed in Chapter 5. In this setup, from [Bat+11, (1)] but using a more compact notation,

$$E^{\mathrm{MC}}_{\mathrm{QM/QM}} = E_{\mathrm{Low}}(F_1 \cup \cdots \cup F_m) + \sum_{\emptyset \subset \mathbf{u} \subseteq [m]} (-1)^{|\mathbf{u}|-1} \left[E_{\mathrm{High}}\left(\bigcap_{i \in \mathbf{u}} F_i \right) - E_{\mathrm{Low}}\left(\bigcap_{i \in \mathbf{u}} F_i \right) \right]. \tag{7.2}$$

Here, we use F_i to indicate the ith center, which would correspond to Model_i in (7.1) in the disjoint case. Clearly, also $E_{\mathrm{Low}}(F_1 \cup \cdots \cup F_m)$ would be $E_{\mathrm{Low}}(\mathrm{Real})$.

The *molecules-in-molecules* (MIM) approach of Mayhall and Raghavachari [MR11] is also derived from the ONIOM formulation. The MIM method provides for the construction of a hierarchy of increasingly-detailed MIMn energy expressions, which are defined as a series of additive corrections in terms of energies obtained by a fragmentation method with increasingly high orders of theory applied to the per-fragment calculations. For

[1] In context, we spell "center" as in the source.

example [MR11, (5), (6), (7)],

$$E^{\mathrm{MIM1}} = E^{r}_{\mathrm{High}} \tag{7.3}$$

$$E^{\mathrm{MIM2}} = E^{r}_{\mathrm{High}} - (E^{r}_{\mathrm{Low}} - E^{\infty}_{\mathrm{Low}}), \quad \text{and} \tag{7.4}$$

$$E^{\mathrm{MIM3}} = E^{r}_{\mathrm{High}} - (E^{r}_{\mathrm{Med}} - E^{r'}_{\mathrm{Med}}) - (E^{r'}_{\mathrm{Low}} - E^{\infty}_{\mathrm{Low}}). \tag{7.5}$$

In these expressions, the superscripts $r < r' < \infty$ represent levels of detail used in the fragmentation procedure, such as cutoff distances used to construct families of potentially overlapping fragments.[2] $E^{\infty}_{\mathrm{Low}}$ is the total energy of the full system, calculated using a standard method. In [MR11], the energy expressions according to the fragmentations, e.g., E^{r}_{High}, are also based generally on the cardinality PIE and specifically on the approach taken by the GEBF [LLJ07]. Later consideration of these values provided one motivation for the development of the MOBE by the same authors in [MR12]; see again discussion in Chapter 5. An equivalence between the MIM and MC QM/QM methods has also been recognised [LH19].

Relatively recently, Iyengar and co-workers have developed a multilevel graph-theoretical energy-based fragmentation method also based upon ONIOM [RHI18; RI18; KI19; RI20; RKI20; Zha+21; KDI21; ZI22], which they have applied particularly in the context of molecular dynamics. We will rephrase their scheme using our terminology in order to simplify the presentation and later discussion. The approach requires first the construction of a fragment interaction graph, G. Several protocols for construction of this graph have been suggested, based variously on connectivity in the covalent bond graph [RHI18; RKI20], on techniques drawn from computational geometry [RHI18], and upon thresholded spatial distances between atoms [RI20; RKI20; KDI21; Zha+21]. The method is then defined in terms of the set of complete induced subgraphs $G[\mathbf{u}]$ of G, each of which is called in the original context a *rank-r simplex* for $r = |\mathbf{u}| - 1$. We introduce here the notation $\mathrm{comp}_R[G] := \{\mathbf{u} \subseteq [M] \mid |\mathbf{u}| \leq R + 1 \text{ and } G[\mathbf{u}] \text{ is complete}\}$ to be the set of all vertex subsets that induce complete subgraphs of G up to and including some size $R + 1$, for $R \in \mathbb{N}$; clearly, these subsets also induce all rank-r simplexes for $0 \leq r \leq R$.

The total energy of the full molecular system is approximated in terms of some or all of these simplexes as, with adaptation from [Zha+21, (3) and (4)],

$$E^{\text{simplex-ONIOM}}_{R} = E^{\mathrm{Low}}_{[M]} + \sum_{r=0}^{R} (-1)^{r} \sum_{\substack{\mathbf{u} \in \mathrm{comp}_R[G] \\ |\mathbf{u}| = r+1}} \left(E^{\mathrm{High}}_{\mathbf{u}} - E^{\mathrm{Low}}_{\mathbf{u}} \right) \left(\sum_{m=r}^{R} (-1)^{m} p^{m}_{\mathbf{u}} \right). \tag{7.6}$$

Here, we write $E^{\mathrm{Low}}_{\mathbf{u}}$ and $E^{\mathrm{High}}_{\mathbf{u}}$ to mean total energies calculated for the subsystem formed as the union of fragments indexed by $\mathbf{u}$, calculated using low and high levels of theory respectively, as usual. For each $\mathbf{u} \in \mathrm{comp}_R[G]$, the term $p^{m}_{\mathbf{u}}$ counts the

[2] It is stated in [MR11] that $r > r'$, but in context, we believe this to be a typographic error.

number of simplexes $G[\mathbf{v}]$ of rank $m \leq R$ that contain $G[\mathbf{u}]$ as a subgraph, that is, $p_{\mathbf{u}}^{m} := |\{\mathbf{v} \in \mathrm{comp}_R[G] \mid \mathbf{v} \supseteq \mathbf{u},\ |\mathbf{v}| = m+1\}|$. The maximum simplex rank R to be considered is an adjustable parameter.

Equation (7.6) first appeared, in slightly different formulation, in [RHI18, (A1)], there motivated by a counting argument. Alternative formulations in, e.g., [KI19; RI20; RKI20] are constructed by interesting and unusual but rather informal appeals to concepts drawn from topology, particularly the Euler characteristic of a simplicial complex. It has also been recognised that the sum on the right-hand side of (7.6) is the difference of two equivalent truncations of standard fragment MBEs, one each in terms of the calculations performed with low and high levels of theory [RI20; RKI20; Zha+21].

In earlier work from the same group [LI15; LHI16; HLI17], a standard cardinality-PIE overlapping-fragment energy approximation was used in a similar ONIOM-style expression [LI15, (2); LHI16, (1); HLI17, (1)]; this expression is equivalent to (7.2). It is claimed in [RI18] that such an expression in the simplex case [RI18, (1)] is "isomorphic" [RI18, p. 5548] to (7.6), although a rigorous argument is not provided. The form (7.6) is preferred, since it is "more efficient" [RI18, p. 5548], in the sense that a calculation of the involved coefficients does not require brute-force enumeration of all possible k-fold intersections of the simplex-inducing vertex sets [RHI18].

Multilevel fragment-based techniques may also be obtained by direct manipulation of the standard MBE form (5.4). For instance, Beran [Ber09] motivates his *hybrid many-body interaction* (HMBI) model by selectively merging two distinct MBEs, with terms calculated using two different qualities of model theory. Using his notation, then and for example [Ber09, (2.4) and (2.5)],[3]

$$E_{\mathrm{tot}}^{\mathrm{high}} \approx \sum_i E_i^{\mathrm{high}} + \sum_{ij} \Delta^2 E_{ij}^{\mathrm{high}} + \left(\sum_{ijk} \Delta^3 E_{ijk}^{\mathrm{low}} + \cdots + \sum_{ijk\cdots} \Delta^N E_{ijk\cdots}^{\mathrm{low}} \right) \tag{7.7}$$

$$= \sum_i E_i^{\mathrm{high}} + \sum_{ij} \Delta^2 E^{\mathrm{high}} + \left(E_{\mathrm{tot}}^{\mathrm{low}} - \sum_i E_i^{\mathrm{low}} + \sum_{ij} \Delta^2 E_{ij}^{\mathrm{low}} \right). \tag{7.8}$$

Here, a complete full-system calculation is required to deliver $E_{\mathrm{tot}}^{\mathrm{low}}$. Beran explicitly notes in [Ber09] that the working equations of his model reduce to those of the MC QM/QM approach.

The *multilevel fragment-based approach* (MFBA) of Řezáč and Salahub [ŘS09] also involves a fragment MBE with contribution terms calculated with disparate levels of theory. This MBE is initially and explicitly truncated after two-body terms. We use an informal variant of our usual notation and terminology here, observing that there is some room for definitional ambiguity in the original paper [ŘS09]. The empty-set subproblem potential

[3] And also correcting for an apparent typographical error in (2.5) of [Ber09], namely an extraneous leading i in front of a summation.

is taken to be always zero and so we do not write it anywhere explicitly. A high level of model theory is used to calculate the singleton contribution potentials $\tilde{V}^{\text{high}}_{\{i\}} = V^{\text{high}}_{\{i\}}$, as well as a subset of the pair contribution potentials $\tilde{V}^{\text{high}}_{\{i,j\}} = V^{\text{high}}_{\{i,j\}} - V^{\text{high}}_{\{i\}} - V^{\text{high}}_{\{j\}}$. Specifically, given a covalent bond interaction graph G and fixing some cutoff threshold $r_{\text{cut}} > 0$, if we write X to be the set of all two-element subsets $\{i,j\} \subseteq [M]$ such that either atoms i and j interact directly in G or $\|R_i - R_j\| \leq r_{\text{cut}}$, then the full MBFA energy is given by

$$E^{\text{MFBA}} = \sum_{i=1}^{M} \tilde{V}^{\text{high}}_{\{i\}} + \sum_{\substack{i<j \\ \{i,j\}\in X}} \tilde{V}^{\text{high}}_{\{i,j\}} + \sum_{\substack{i<j \\ \{i,j\}\notin X}} \tilde{V}^{\text{low}}_{\{i,j\}}. \tag{7.9}$$

When considering the application of their EE-MB scheme to the calculation of MP2 total energies, Dahlke and Truhlar explicitly separate the treatment of the Hartree-Fock total energy and the correlation energy [DT07a]. They decompose each according to an MBE [DT07a, (8), (9)],

$$E = E_{\text{HF}} + E_{\text{corr}} \tag{7.10}$$

$$= (V^{(1)}_{\text{HF}} + V^{(2)}_{\text{HF}} + \cdots + V^{(M)}_{\text{HF}}) + (V^{(1)}_{\text{corr}} + V^{(2)}_{\text{corr}} + \cdots + V^{(M)}_{\text{corr}}), \tag{7.11}$$

where we use a slight modification to the notation of the source and write $V^{(k)}_{\text{HF}}$ for the summed k-body contributions to the Hartree-Fock energy, and similarly $V^{(k)}_{\text{corr}}$ for the summed k-body contribution to the correlation energy. Dahlke and Truhlar explicitly truncate only the correlation-energy MBE, and replace the Hartree-Fock MBE just with the Hartree-Fock total energy.

The FCR method has also recently been extended to a multilevel setting [HK21]. Given two distinct downward-closed sets of fragments, which we will write $\{\text{FCR}_{\text{LL}}\}$ and $\{\text{FCR}_{\text{HL}}\}$ for low-level and high-level respectively, the relevant energy equation is, from [HK21, (15)] and adjusting notation for consistency with (5.10) above,

$$E^{\text{ML-FCR}} = \sum_{\mathbf{f}_l \in \{\text{FCR}_{\text{HL}}\}} p^{\text{HL}}_{\mathbf{f}_l} E^{\text{HL}}_{\mathbf{f}_l}(\{z\}_{\mathbf{f}_l}) + \sum_{\mathbf{f}_l \in \{\text{FCR}_{\text{HL}}\}} p^{\text{HL}}_{\mathbf{f}_l} E^{\text{LL}}_{\mathbf{f}_l}(\{z\}_{\mathbf{f}_l}). \tag{7.12}$$

The terms $E^{\text{HL}}_{\mathbf{f}_l}$ and $E^{\text{LL}}_{\mathbf{f}_l}$ should by this point be self-explanatory. It is reasoned that, for consistency and to avoid overcounting, it should hold that [HK21, (16)]

$$p^{\text{HL}}_{\mathbf{f}_l} + p^{\text{LL}}_{\mathbf{f}_l} = p^{\{\text{FCR}_{\text{HL}}\}\cup\{\text{FCR}_{\text{LL}}\}}_{\mathbf{f}_l}, \tag{7.13}$$

where the coefficients $p^{\{\text{FCR}_{\text{HL}}\}\cup\{\text{FCR}_{\text{LL}}\}}_{\mathbf{f}_l}$ on the right-hand side are standard FCR coefficients from (5.11) in terms of the union of the low-level and high-level sets of fragments. The high-level coefficients $p^{\text{HL}}_{\mathbf{f}_l}$ are directly chosen to be those provided by (5.11) in terms

of $\{\mathrm{FCR}_{\mathrm{HL}}\}$, and the low-level coefficients $p^{\mathrm{LL}}_{\mathbf{f}_l}$ are then delivered by (7.13). It is stated in [HK21] that this approach can be generalised to multiple layers, rather than only two, although an explicit construction is not given.

Finally, the ML-BOSSANOVA scheme of Chinnamsetty et al. [CGH18] is a direct extension of the original BOSSANOVA method as outlined in Section 6.2. Again, we will make some minor notational adjustments to the original description of ML-BOSSANOVA, for consistency with the remainder of this thesis. Rather than considering a single Born-Oppenheimer potential V^{BO}, as in the standard BOSSANOVA case, the ML-BOSSANOVA formulation considers a family $\{V^{\mathrm{BO}}_p\}_{p\in\mathbb{N}}$ of such potentials, with entries indexed by a level parameter p indicating somehow increasing basis set thoroughness. Each distinct potential V^{BO}_p is decomposed as in (6.8), that is,

$$V_{p,[M]} = \tilde{V}_{\emptyset} + \sum_{\substack{\mathbf{u}\in\mathrm{conn}\,[G]\\ |\mathbf{u}|=1}} \tilde{V}_{p,\mathbf{u}} + \sum_{\substack{\mathbf{u}\in\mathrm{conn}\,[G]\\ |\mathbf{u}|=2}} \tilde{V}_{p,\mathbf{u}} + \cdots + \sum_{\substack{\mathbf{u}\in\mathrm{conn}[G]\\ |\mathbf{u}|=N}} \tilde{V}_{p,\mathbf{u}}. \tag{7.14}$$

Then, each contribution potential $\tilde{V}_{p,\mathbf{u}}$ is itself decomposed as

$$\tilde{V}_{p,\mathbf{u}} = \sum_{q=0}^{p} \tilde{\omega}_{q,\mathbf{u}}, \tag{7.15}$$

where each [CGH18, (14)]

$$\tilde{\omega}_{q,\mathbf{u}} = \tilde{V}_{q,\mathbf{u}} - \tilde{V}_{q-1,\mathbf{u}}, \tag{7.16}$$

with specifically $\tilde{\omega}_{0,\mathbf{u}} = \tilde{V}_{0,\mathbf{u}}$. Then, under some conditions on $\{V^{\mathrm{BO}}_p\}_{p\in\mathbb{N}}$,

$$V^{\mathrm{BO}}_{\infty} = \sum_{p\in\mathbb{N}} \sum_{\mathbf{u}\in\mathrm{conn}\,[G]} \tilde{\omega}_{p,\mathbf{u}}, \tag{7.17}$$

with V^{BO}_{∞} written to indicate the Born-Oppenheimer potential for a notional CBS-limit solution to the Schrödinger equation.

The ML-BOSSANOVA construction, particularly (7.16), is explicitly motivated by reference to the standard combination technique [CGH18]. Brief mention is made in [CGH18] to the fact that the complete set of terms can be partially ordered, and it bears repeating that the ML-BOSSANOVA technique was a primary inspiration for the development of both the general order-theoretic combination technique construction in Chapter 3, as well as the accompanying adaptive algorithm.

7.2. Multilevel extensions to ANOVA-like decompositions

Extending the order-theoretic formulation of MBE-style energy-based fragmentation methods to an arbitrarily multilevel setting requires only an adjustment of the underlying

poset grid Π. Rather than choosing Π to be just a single boolean algebra axis, we take it to be the direct product of exactly one such axis with at least one additional axis corresponding to a discretisation treatment or other computational adjustment for a numerical solution to the electronic problem. This latter axis, or axes, can be either finite or infinite, depending on the desired application. More precisely, a full multilevel MBE-style poset grid can be equivalently obtained as a direct product $\Pi = B_M \times P_1 \times P_2 \times \cdots$, where B_M is a boolean algebra for the nuclear fragmentation, and P_1, P_2, $\cdots$ are chain posets indexing refinements to the treatment of the electronic problem in the subproblem potentials $V_{(\mathbf{u},p,q,\cdots)}$. If the boolean algebra is replaced with an arbitrary subposet of induced subgraphs of some full-system interaction graph, we obtain a very general definition of an ML-SUPANOVA decomposition.

The addition of only a single chain axis to a boolean algebra is sufficient to reproduce all of the existing multilevel schemes that we mentioned in the previous section, with the partial exception of ML-BOSSANOVA, to which we will come shortly. We will demonstrate this explicitly by showing precisely which order ideals I suffice to reproduce energy expressions from the previous section as combination sums. Let us be clear that, as noted above, pairwise similarities and connections between some such schemes, and also connections to the standard MBE, are either very deliberately explicit in their original derivations or have already been observed in the literature. A careful catalogue of all such previous observations is beyond the scope of this work, but suffice it to say that we have not uncovered some previously elusive truth. Rather, we simply showcase here the generality of the order-theoretic combination technique formalism as a means of alternative derivation. Even here, we mention that the ML-FCR method [HK21] is in this particular setting an equally general and basically formally equivalent tool; we shall return to this point very shortly.

Consider in general $\Pi = B_M \times [n]$ for some $n \geq 1$. If I is an order ideal of Π, it is not hard to see that I can be written as $I = \bigcup_{i=1}^{n} I_i \times \{i\}$, where each $I_i \subseteq B_M$ and $I_1 \supseteq I_2 \supseteq \cdots \supseteq I_n$. If $i = n$, then

$$D^{(I)}_{(\mathbf{u},i)} = \sum_{\substack{\mathbf{v} \supseteq \mathbf{u} \\ \mathbf{v} \in I_i}} \mu_{\Pi}((\mathbf{u},i),(\mathbf{v},i)) = \sum_{\substack{\mathbf{v} \supseteq \mathbf{u} \\ \mathbf{v} \in I_i}} \mu_{B_M}(\mathbf{u},\mathbf{v}) = D^{(I_i)}_{\mathbf{u}}, \tag{7.18}$$

by (3.25) and (3.27), and where we use $D^{(I_i)}_{\mathbf{u}}$ on the right-hand side to mean the combination coefficient of $\mathbf{u}$ in an I_i-truncation with respect to the standard B_M, rather than

the multilevel Π. If $i < n$, then similarly,

$$D^{(I)}_{(\mathbf{u},i)} = \sum_{j=i}^{n} \sum_{\substack{\mathbf{v}\supseteq\mathbf{u} \\ \mathbf{v}\in I_j}} \mu_I((\mathbf{u},i),(\mathbf{v},j)) \tag{7.19}$$

$$= \sum_{j=i}^{n} \sum_{\substack{\mathbf{v}\supseteq\mathbf{u} \\ \mathbf{v}\in I_j}} \mu_{B_M}(\mathbf{u},\mathbf{v})\mu_{[n]}(i,j) \tag{7.20}$$

$$= \sum_{\substack{\mathbf{v}\supseteq\mathbf{u} \\ \mathbf{v}\in I_i}} \mu_{B_M}(\mathbf{u},\mathbf{v}) - \sum_{\substack{\mathbf{v}\supseteq\mathbf{u} \\ \mathbf{v}\in I_{i+1}}} \mu_{B_M}(\mathbf{u},\mathbf{v}) \tag{7.21}$$

$$= D^{(I_i)}_{\mathbf{u}} - D^{(I_{i+1})}_{\mathbf{u}}. \tag{7.22}$$

If we consider $\Pi = B_M \times [2]$, and identify two downward-closed subsets $I_1 = \{\mathrm{FCR_{LL}}\} \subseteq B_M$ and $I_2 = \{\mathrm{FCR_{HL}}\} \subseteq B_M$ for use with the ML-FCR method, taken such that $I_1 \supseteq I_2$,[4] we see immediately that (7.18) and (7.22) are exactly the definitions for the two-layer ML-FCR coefficients. This is unsurprising, since as mentioned in Chapter 5, the FCR method provides for an I-truncation over an arbitrary order ideal of B_M, and thus delivers combination coefficients on B_M that are fully equivalent to those from the order-theoretic combination technique. The full (7.18) and (7.22) for arbitrary $\Pi = B_M \times [n]$ provide presumably just the multiple-layer generalisation envisaged by the authors of [HK21]. The arguments that follow therefore support claims to generality made in [HK21], in that other multilevel methods beyond those there explicitly considered can be obtained from the ML-FCR by an appropriate choice of order ideal. In the broader context, however, we observe again, much as in Chapter 5, that the ML-FCR approach is potentially unwieldy in comparison to ML-SUPANOVA-style approaches over restricted subposets of subgraphs, since it requires full enumeration of an order ideal so as to identify non-zero coefficients in the general case. This also makes it less amenable to the possible application of adaptivity.

We mention here in passing that a recent implementation of a generalised multicentre ONIOM scheme [See+22] can make use of the FCR approach for individual model calculations. Although we have not investigated in any detail, it seems likely that the final summations produced by this setup would be consistent with a generalisation of the ML-FCR as just discussed. We leave this for the moment as conjecture.

Returning to $\Pi = B_M \times [2]$, where now the entries of the chain poset correspond specifically to the exclusion or inclusion of the MP2 correlation energy in the subproblem potentials $V_{(\mathbf{u},1)}$ and $V_{(\mathbf{u},2)}$ respectively, the resulting decomposition of the full-system

[4] In fact, the ML-FCR definition also rather neatly handles the case where $I_2 \supset I_1$, by simply nulling out every low-level subproblem term $E^{\mathrm{LL}}_{\mathbf{f}_l}$ in (7.12) and thus reducing to a standard, non-multilevel I_2-truncation.

MP2 Born-Oppenheimer potential $V^{\mathrm{BO}} = V_{[M],2}$ is just and exactly that used by Dahlke and Truhlar [DT07a] in (7.10). Assuming the use of an n-body expansion of the correlation energy E_{corr}, the particular truncation of the full decomposition that they studied clearly corresponds to the order ideal I as above, with $I_1 = B_M$ and $I_2 = \{\mathbf{u} \in B_M \mid |\mathbf{u}| \leq n\}$. This should be intuitively obvious, but to make it fully explicit, note that $D^{(I)}_{(\mathbf{u},2)} = D^{(I_2)}_{\mathbf{u}}$ is just the standard n-body coefficient for $\mathbf{u} \in I_2$. Then, since $D^{(I_1)}_{\mathbf{u}}$ is unity for $\mathbf{u} = [M]$ and zero otherwise, we have from (7.18) that

$$D^{(I)}_{(\mathbf{u},1)} = \begin{cases} 1 - D^{(I)}_{(\mathbf{u},2)} & \text{if } \mathbf{u} = [M], \\ -D^{(I)}_{(\mathbf{u},2)} & \text{otherwise.} \end{cases} \tag{7.23}$$

Assuming everywhere-zero empty-set subproblem potentials[5] $V_{(\emptyset,1)} = V_{(\emptyset,2)} = 0$, it is then clear that the full truncation S_I is given by

$$S_I = V_{([M],1)} + \sum_{\mathbf{u} \in I_2} D^{(I_2)}_{\mathbf{u}} (V_{(\mathbf{u},2)} - V_{(\mathbf{u},1)}) \tag{7.24}$$

when $n < M$, so a full-system HF calculation and an n-body truncation of the MP2 correlation energy, and just by $S_I = V_{([M],2)}$ when $n = M$. The HMBI energy (7.8) of [Ber09] emerges equivalently for the specific choice $n = 2$, if *exclusion or inclusion of the MP2 correlation energy* is replaced with *use of a low or a high level model* in the above.

To obtain the MFBA energy expression (7.9) suggested by [ŘS09], let instead $I_1 = \{\mathbf{u} \in B_M \mid |\mathbf{u}| \leq 2\}$, and then take $I_2 = \{\mathbf{u} \in B_M \mid |\mathbf{u}| \leq 1\} \cup X$, where X is defined as above (7.9). Clearly $I_2 \subseteq I_1$. The first two summations of the right-hand side of (7.9) are just a standard non-multilevel MBE truncation in terms of I_2, again assuming a zero empty-set subproblem potential. In this truncation, when $|\mathbf{u}| = 2$, it is easy to see that $D^{(I_2)}_{\mathbf{u}} = 1$ when $\mathbf{u} \in X$, and $D^{(I_2)}_{\mathbf{u}} = 0$ when $\mathbf{u} \notin X$. Each atom i is involved in at most $M - 1$ pairs in X; write the number of such pairs as m_i. For $|\mathbf{u}| = 1$, then, it follows that $D^{(I_2)}_{\mathbf{u}} = 1 - m_i$. Similarly, we see that for I_1, i.e., a standard two-body expansion, $D^{(I_1)}_{\mathbf{u}} = 1$ when $|\mathbf{u}| = 2$, and $D^{(I_1)}_{\mathbf{u}} = 2 - M$ when $|\mathbf{u}| = 1$. Now, in the full multilevel I-truncation in terms of Π, we have from (7.18) that $D^{(I)}_{(\mathbf{u},2)} = D^{(I_2)}_{\mathbf{u}}$, and from (7.22), $D^{(I)}_{(\mathbf{u},1)} = D^{(I_1)}_{\mathbf{u}} - D^{(I_2)}_{\mathbf{u}}$. Thus, to show that the truncation $S_I = E^{\mathrm{MFBA}}$, we need only demonstrate that the values $D^{(I)}_{(\mathbf{u},1)}$ correspond to the summed coefficients of all low-level terms $V^{\mathrm{low}}_{\{i\}}$ and $V^{\mathrm{low}}_{\{i,j\}}$ introduced implicitly by the terms $\tilde{V}^{\mathrm{low}}_{\{i,j\}}$ in (7.9). Each $V^{\mathrm{low}}_{\{i,j\}}$ appears at most once, depending on whether or not $\{i,j\} \in X$. Since $D^{(I)}_{(\{i,j\},1)}$ is then either 0 or 1 respectively, these coefficients agree. Since there are $M - 1$ pairs for each i,

[5] There seems to be no mention of an overcounting correction for point-charge interactions in [DT07a].

the number of such pairs *not* in X is $M-1-m_i$, and each $V^{\text{low}}_{\{i\}}$ appears that many times, each carrying a leading minus. This is consistent with $D^{(I)}_{(\{i\},1)} = (2-M)-(1-m_i)$, so we are done.

Obtaining the MC QM/QM energy equation is also straightforward. We consider the general case (7.2) as adapted from [Bat+11], where centers are allowed to be overlapping. Here, $\Pi = B_M \times \{1,2\}$ again suffices, and we use equivalent subproblem/contribution potential choices as above. As discussed in Chapter 5, any overlapping-fragment energy equation obtained via the cardinality form of the PIE produces a truncation of an MBE in terms of some particular order ideal of B_M. So, let I_2 be this order ideal for the specific choice of overlapping fragments at hand, and take $I_1 = B_M$. Then, just as above, we have

$$S_I = V_{([M],1)} + \sum_{\mathbf{u}\in I_2} D^{(I_2)}_{\mathbf{u}} (V_{(\mathbf{u},2)} - V_{(\mathbf{u},1)}) \tag{7.25}$$

$$= V_{([M],1)} + S^{(2)}_{I_2} - S^{(1)}_{I_2}, \tag{7.26}$$

where we again abuse our notation and write $S^{(2)}_{I_2}$ to mean the I_2-truncation of a standard MBE defined in terms of the high-level subproblem potentials only, and similarly $S^{(1)}_{I_2}$ in terms of the low-level subproblem potentials. The arguments in Section 5.2.3 suffice to show the full termwise equivalence of (7.26) with (7.2).

The level-1 MIM energy E^{MIM1} is, of course, also just an appropriate truncation of a standard MBE matching a PIE-based overlapping-fragments expression, and not a true multilevel sum; see again [MR12], and more broadly discussion in Section 5.2.3. The argument for obtaining the overlapping version of the MC QM/QM energy above also applies to the provision of E^{MIM2}, once (7.4) is slightly rearranged to be $E^{\text{MIM2}} = E^{\infty}_{\text{Low}} - (E^{r}_{\text{High}} - E^{r}_{\text{Low}})$. As mentioned above, this formal equivalence was previously noted at least in [LH19]. Now particularly informally, if we consider instead $\Pi = B_M \times [3]$ and assume that every fragment F_i constructed according to the MIM threshold parameter r is a subset of an F'_j constructed according to $r' > r$ — as would be the case for a distance-thresholding approach as used in [MR11] — then clearly the order ideal I_3 that produces the PIE-style summations E^{r}_{high} and E^{r}_{med} is a subset of that I_2 for the summations $E^{r'}_{\text{med}}$ and $E^{r'}_{\text{low}}$. Taking $I_1 = B_M$ and forming $I = \bigcup_{i=1}^{3} I_i \times \{i\}$, we can also obtain

$$S_I = V_{([M],1)} + S^{(2)}_{I_2} - S^{(1)}_{I_2} + S^{(3)}_{I_3} - S^{(2)}_{I_3}, \tag{7.27}$$

which is just the expression for E^{MIM3} up to notation and rearrangement. The extension to an arbitrary MIMn is obvious.

To obtain the multilevel simplex-based energy expression given by Iyengar and co-

workers, we restrict ourselves first to the simpler, non-multilevel form

$$E_R^{\text{simplex}} = \sum_{r=0}^{R} (-1)^r \sum_{\substack{\mathbf{u} \in \text{comp}_R[G] \\ |\mathbf{u}|=r+1}} E_{\mathbf{u}} \left(\sum_{m=r}^{R} (-1)^m p_{\mathbf{u}}^{r,m} \right), \tag{7.28}$$

also adapted from [Zha+21, (3)]. It is stated implicitly in [RI20] and explicitly in [RKI20; Zha+21] that this is equivalent to a particular truncation of a standard fragment MBE form. This is demonstrated for the cases $R = 1$ and $R = 2$ in, e.g., [RI20], but we are not aware of a full proof given for the general case. We remark here that a more general version of (7.28) in terms of any downward-closed subset of $2^{[M]}$ can be obtained by only very minor rewriting of [KC16, (14) and (15)]. This would suffice, since as observed in [RI20; Zha+21; ZI22], if $G[\mathbf{v}]$ is complete and $\mathbf{u} \subseteq \mathbf{v}$, then also $G[\mathbf{u}]$ is complete, so $\text{comp}_R[G]$ is just such a subset. But since the rederivation of (7.28) as an I-truncation of an MBE in our formalism is straightforward, we give it explicitly for completeness.

For any fixed $R \geq 0$, write for notational convenience $I'_R = \text{comp}_R[G]$. Noting as above that $\text{comp}_R[G]$ is an order ideal of B_M, consider the I'_R-truncation of a standard fragment MBE. Then as usual,

$$S_{I'_R} = \sum_{\mathbf{u} \in I'_R} D_{\mathbf{u}}^{(I'_R)} V_{\mathbf{u}} = \sum_{\mathbf{u} \in I'_R} V_{\mathbf{u}} \sum_{\substack{\mathbf{v} \supseteq \mathbf{u} \\ \mathbf{v} \in I'_R}} \mu_{B_M}(\mathbf{u}, \mathbf{v}) = \sum_{\mathbf{u} \in I'_R} V_{\mathbf{u}} \sum_{\substack{\mathbf{v} \supseteq \mathbf{u} \\ \mathbf{v} \in I'_R}} (-1)^{\mathbf{v}-\mathbf{u}}. \tag{7.29}$$

Using $p_{\mathbf{u}}^m$ as defined above, and also that $\max_{\mathbf{u} \in I'_R} |\mathbf{u}| = R + 1$, then,

$$S_{I'_R} = \sum_{k=0}^{R+1} \sum_{\substack{\mathbf{u} \in I'_R \\ |\mathbf{u}|=k}} V_{\mathbf{u}} \sum_{m=k}^{R+1} \sum_{\substack{\mathbf{u} \subseteq \mathbf{v} \in I'_R \\ |\mathbf{v}|=m}} (-1)^{m-k} \tag{7.30}$$

$$= \sum_{k=0}^{R+1} (-1)^{-k} \sum_{\substack{\mathbf{u} \in I'_R \\ |\mathbf{u}|=k}} V_{\mathbf{u}} \sum_{m=k}^{R+1} (-1)^m p_{\mathbf{u}}^m \tag{7.31}$$

$$= \sum_{k=-1}^{R} (-1)^k \sum_{\substack{\mathbf{u} \in I'_R \\ |\mathbf{u}|=k+1}} V_{\mathbf{u}} \sum_{m=k+1}^{R} (-1)^m p_{\mathbf{u}}^m. \tag{7.32}$$

This is equivalent to (7.28), up to notation and the assumption that $V_\emptyset = 0$.

In the full multilevel setting, then, a similar argument to those used for the HMBI, MC QM/QM, and MIM methods above shows that choosing $I_1 = B_M$ and $I_2 = \text{comp}_R[G]$ leads to an I-truncation in terms of Π that is just (7.6), again up to notation. As mentioned above, that the right-hand side term of (7.6) is a termwise difference of

truncated MBEs has been recognised [RI20; RKI20; Zha+21]; a connection to the HMBI is also explicitly drawn in [RI20].

If we take $\mathrm{comp}_R^+[G] := \mathrm{comp}_R[G] \cup \{[M]\}$ to be $\mathrm{comp}_R[G]$ with $[M]$ explicitly adjoined if not already present, and denote the maximum rank of any simplex in G by $\hat{R}$, then (7.28) for any R can also be considered as a truncation of a SUPANOVA expansion of V^{BO} constructed in terms of $\mathrm{comp}_{\hat{R}}^+$. As mentioned in the previous chapter, there is a direct relationship between this SUPANOVA expansion and one constructed in terms of $\mathcal{M}_g[G]$, that is, in terms of the geodesically convex subgraphs of G. For any given G, it is easy to see that every complete induced subgraph is also a geodesically convex subgraph, although the converse is not necessarily true.[6] Thus, every $\mathrm{comp}_R[G]$ is a subposet of $\mathcal{M}_g[G]$, and moreover, clearly an order ideal of $\mathcal{M}_g[G]$. From this perspective, we can view a SUPANOVA expansion in terms of $\mathcal{M}_g[G]$ as being an extended version of a SUPANOVA expansion in terms of $\mathrm{comp}_{\hat{R}}^+$, containing additional higher-order terms.

This suggests a potential benefit of the combination of our adaptive SUPANOVA algorithm and the poset $\mathcal{M}_g[G]$ relative to the approach of Iyengar and co-workers. In effect, it is the protocol by which the interaction graph G is constructed that selects the MBE terms retained in the truncation E_R^{simplex}, possibly restricted further by the maximum rank R [KI19]. It is in this sense that their approach is adaptive, since a distance-based protocol will select different terms according to different conformations of the underlying molecular system [RHI18; KI19]. Once that protocol has been applied, however, the selection of terms is fixed. By contrast, our adaptive approach is able first to explore any strongly local subsystems as provided by simplexes in G, without necessary preselection of a maximal rank R, but then also able to extend resiliently to other, larger and less "compact" but still convex subsystems as the available simplexes in G are exhausted.

Finally, if a multilevel poset grid is formed instead as the direct product of $\mathrm{conn}[G]$ and an infinite chain axis over $\mathbb{N}$ representing a notional level of basis set theory to be used in the construction of Born-Oppenheimer potentials V_p^{BO}, then the resulting decomposition of V^{BO} is just the full ML-BOSSANOVA expansion (7.17) of [CGH18],

[6] Notions of convexity are invoked in two particular discussions of this scheme [RI20; Zha+21]. The use of mathematical terminology and notation is here unfortunately imprecise. To our best understanding, however, a geometric definition of convexity is used to justify the fact that, in our terminology, the set of vertex subsets which induce simplexes in G is an order ideal of B_M (a "(truncated) power set" [Zha+21, p. 2673]), and thus that (7.28) does not overcount any k-body terms. It is also mentioned in passing in [RKI20] that "[t]he many-body contributions in Equation (3) form an ordered set" [RKI20, p. 4], there referencing what is essentially (7.6) here, but this order does not seem to be used explicitly. At any rate, it does not seem to us that an abstract definition of convexity in the purely graph-theoretical setting is ever considered, at least not one like that given in Section 6.5 above and used in our construction of a SUPANOVA decomposition in terms of convex subgraphs.

There are, however, deep connections between more abstract versions of the topological ideas invoked in [KI19; RI20; RKI20] and abstract convexity, order theory, and Möbius inversion; see, e.g., [EJ85; Vel93, Chap. 3; Sta12, Chap. 3]. Further exploration is likely to be profitable, but we leave this for future work.

effectively by direct construction. Note here that the previously-identified issues with the use of conn$[G]$ when the underlying graph G contains the forbidden subgraphs C_4' or C_n for $n \geq 4$ will persist in this construction, and in such cases the ML-BOSSANOVA grid conn$[G] \times \mathbb{N}$ will not be fully consistent with the grid $B_M \times \mathbb{N}$.

It is easy to see that the direct product of any two posets is a meet semilattice if and only if both of those posets are meet semilattices. Thus, if we consider conn$[G] \times P_1 \times P_2 \times \cdots$ as a subposet of $B_M \times P_1 \times P_2 \times \cdots$, for any choice of P_1, P_2, ... such that every P_i is a meet semilattice, it is clear from Theorem 5.2.8 that the latter grid will not be combination consistent with the former. If, however, we use $\mathcal{M}_g[G]$ — or any SUPANOVA axis which is isomorphic to a meet subsemilattice of B_M — the resulting poset will be a meet subsemilattice, resp. isomorphic to a meet subsemilattice of $B_M \times P_1 \times P_2 \times \cdots$, and so will be combination consistent with it.

Regardless of whether or not combination consistency is guaranteed, the adaptive algorithm given in Chapter 3 can also be used to explore index sets over any such ML-SUPANOVA poset grid. If a truly adaptive selection strategy is applied, however — that is, not the All strategy — there is one minor practical detail that requires attention. The issue surrounds the benefit/cost ratios of poset grid elements involving the empty set from the SUPANOVA axis.

Suppose that $\Pi = P_1 \times P_2 \times \cdots P_n$ is an ML-SUPANOVA poset grid, where P_1 is the SUPANOVA axis. Suppose also that the values $\mathcal{L}[V_{(\emptyset, p_2, \cdots, p_n)}]$ are all equal for any choice of $p_2 \in P_2, \cdots, p_n \in P_n$. This can occur, for example, in the case when the involved contribution potentials represent vacuum embedding calculations, so both $\mathcal{L}[V_{(\emptyset, p_2, \cdots, p_n)}]$ and $\mathcal{C}(\emptyset, p_2, \cdots, p_n)$ are uniformly zero whenever any value p_i is not a $\hat{0}$ of P_i. In this case, the benefit/cost ratio $\mathcal{L}[\tilde{V}_{(\emptyset, p_2, \cdots, p_n)}]/\mathcal{C}(\emptyset, p_2, \cdots, p_n) = 0/0$ is not defined. A simple workaround would be to just set the benefit/cost ratio to zero; however, this would defer the adaptive activation of $(\emptyset, p_2, \cdots, p_n)$ until after the activation of all points $(\mathbf{u}, p_2, \cdots, p_n)$. Similarly, setting the benefit/cost ratio to a notional ∞ would allow the Threshold and Best strategies to activate only points $(\emptyset, p_2, \cdots, p_n)$ at each iteration, to the complete exclusion of any other points $(\mathbf{u}, p_2, \cdots, p_n)$. Either way, the adaptive index-set algorithm will not be able to explore the full poset grid.

To avoid such situations, the algorithm can be adjusted to include the concept of a "freely-available point", which can be automatically included into the index set if and when required. Specifically, if a point in the current index set at a given iteration is expanded, and a successor to that point would be admissible except for the absence of some number of missing freely-available points, then those missing points are tested to see whether their inclusion in the index set would break the property of downward-closure. If not, then both the original successor point *and* the missing freely-available points are chosen for addition to the index set. All freely-available points required at an iteration are explicitly added to the index set, and any relevant information incorporated into all necessary tensors, directly before any newly-added successor points are considered

at line 11 of Algorithm 3.3. The relevant changes to the algorithm are mechanical but untidy, and since the results given in this chapter do not rely on them, we do not describe them explicitly here.

7.3. Case study: heptane (C_7H_{16})

We consider now the linear alkane heptane (C_7H_{16}) [CSHept]. Heptane was used as a primary test case for the initial evaluation of the ML-BOSSANOVA method in [CGH18], where it was demonstrated that that method was capable of efficiently and accurately approximating the Hartree-Fock total energy of heptane at the cc-pV6Z basis set level. This was in reinforcement of earlier results from [Heb14; GHH14], which considered standard BOSSANOVA approximations of the HF/6-311G* total energy of heptane.

We investigated the ability of various truncated ML-SUPANOVA expansions to approximate two distinct quantities. The first quantity is the total energy of heptane at the CCSD(T)/cc-pCV5Z level [PB82; Rag+89; WD95], as calculated using PySCF [Sun15; Sun+17; Sun+20]. This value is thus known to within the convergence tolerance of the used iterative solver, to whit, $10^{-9}\,E_\mathrm{h}$. The calculation of this value was nontrivial: heptane equipped with cc-pCV5Z presents a computational problem in terms of 1895 contracted GTO basis functions. Such a problem lies in the upper range of currently-feasible conventional CCSD(T) calculations [GKN20].[7] As in previous chapters, for full calculation details for single-point and composite total energy calculations, and also for detailed citations for basis sets, see Appendix A.

Ideally, we would also like to consider an approximation targeting the true FCI/CBS total energy of heptane. The exact value of this quantity is of course not practically calculable. As a reference estimate thereof, we considered using the ccCA-PS3 total energy [DeY+09], which is the only all-electron composite method that we discussed in Chapter 4 which can be feasibly applied to heptane. As there discussed, the ccCA-PS3 estimate is not likely to be reliable past the $\mathrm{kcal\,mol^{-1}}$ level. It may, however, be viewed as an estimate of the CCSD(T) total energy extrapolated towards the CBS limit [Kar16, Tab. 1].

Both total energy values are given in Table 7.1. Even given that CCSD(T) is not variational, there is already reason to be suspicious of the ccCA-PS3 value, which is greater than the CCSD(T)/cc-pCV5Z value by approximately $0.011\,E_\mathrm{h}$. Therefore, rather than attempt to approximate the FCI/CBS total energy with only the ccCA-PS3 value as a benchmark, we consider instead the atomisation energy. Here, we can also apply the G4(MP2) method [CRR07b]. Both G4(MP2) and ccCA-PS3 atomisation energies for heptane are given in Table 7.2, alongside the CCSD(T)/cc-pCV5Z atomisation energy.

[7] Indeed, the implementation described in [GKN20] involves a "density-fitting" approximation, which reduces the computational load due to the calculation and storage of the AO ERIs, at the cost of a certain amount of accuracy.

Method	E (E_h)	Abstract cost
ccCA-PS3	−276.359 040	3.832×10^{15}
CCSD(T)/cc-pCV5Z	−276.369 764	4.735×10^{17}

Table 7.1.: Total energies and accompanying abstract costs of calculation for heptane (C_7H_{16}), according to the ccCA-PS3 composite method, and to a conventional single-point CCSD(T)/cc-pCV5Z calculation.

Method	E^atom (E_h)	Abstract cost
G4(MP2)	3.480 747	1.565×10^{13}
ccCA-PS3	3.485 641	3.832×10^{15}
CCSD(T)/cc-pCV5Z	3.473 717	4.735×10^{17}

Table 7.2.: Atomisation energies and accompanying abstract costs of calculation for heptane (C_7H_{16}), according to the G4(MP2) and ccCA-PS3 composite methods, and to a conventional single-point CCSD(T)/cc-pCV5Z calculation.

These values differ by up to approximately $0.012\,E_\mathrm{h}$. Since the G4(MP2) method was consistently more accurate than the ccCA-PS3 method in the results of Chapter 4, we choose the G4(MP2) value as our reference, $E^\mathrm{atom} = 3.480\,747\,E_\mathrm{h}$. Here, however, we should expect no more than chemical accuracy, and even that is dependent on the applicability of the HLC in this situation.

For this case study, we consider a three-axis ML-SUPANOVA poset grid Π. The SUPANOVA axis is taken to be just the standard BOSSANOVA poset, $\mathrm{conn}[G]$, of those subsets of the nuclear indices which induce connected subgraphs of the dehydrogenated covalent bond graph of heptane. Since heptane is an acyclic system, $\mathrm{conn}[G]$ is both a lattice and identical to the poset $\mathcal{M}_g[G]$ of those vertex subsets inducing geodesically convex subgraphs of G.

The remaining two axes are taken to be chain posets. The first represents electron correlation treatments at four different levels of theory: Hartree-Fock, then MP2, then CCSD, then CCSD(T). As in Chapter 4, we associate these with an index $2 \leq m \leq 5$. The second axis is the choice of cc-pCVnZ basis set. When aiming to approximate the G4(MP2) atomisation energy of heptane, we index this axis $2 \leq n \leq 5$; when aiming to approximate the atomisation energy, we use all available cc-pCVnZ basis sets and index $2 \leq n \leq 8$. To simplify the discussion, we explicitly do not apply either of the two fine-tuning axes discussed in Chapter 4; thus, all subproblem potentials $V_{(\mathbf{u},m,n)}$ are backed by all-electron calculations. We consider here only calculations using vacuum embedding potentials. For the atomisation energy calculations, we use notional subproblem potentials $V^\mathrm{atom}_{(\mathbf{u},m,n)}$ which evaluate the atomisation energy rather the total energy.

As in previous chapters, we use our standard abstract cost model, and the evaluation

functional $\mathcal{L}$ is chosen to be point evaluation of a subproblem potential at a particular nuclear geometry. For details on subproblem potentials, see again Section A.7. Here, we have used a geometry of heptane that was obtained from the ChemSpider database [CSHept] and then subjected to a further geometry optimisation using DFT with the B3LYP functional and the cc-pVDZ basis set; see again Section A.6.

In order to evaluate the benefit of the three-axis grid relative to simpler constructions, we also consider the following four "subgrids" obtainable as distinct combinations of either one or two of those three axes:

- A single SUPANOVA axis, with each fragment calculation performed at the CCSD(T)/cc-pCV5Z level. This is explicitly not a multilevel setup, and instead leads, in effect, to a standard BOSSANOVA decomposition as per [Heb14; GHH14].
- A two-axis GCM grid, corresponding to refinements of the full-system calculation for different basis sets and correlation treatments. This is closest to the original CQML setup in [Zas+18].
- A two-axis grid composed of the SUPANOVA axis and the basis set axis, with all calculations performed at the CCSD(T) level. This leads, in this case, to just an implementation of the ML-BOSSANOVA method of [CGH18].
- A two-axis grid composed of the SUPANOVA axis and the correlation-treatment axis, with the basis set fixed at cc-pCV5Z.

We consider calculations in terms each of these poset grids, subject to the restrictions outlined above for the two different reference value targets. We do not consider the last-listed grid for approximation of the atomisation energy, since it would provide only an approximation to at best CCSD(T)/cc-pCV5Z quality rather than the CCSD(T)/cc-pCV8Z quality available from the complete grid. The use of an equivalent grid with the basis set fixed at cc-pCV8Z was considered, but the involved subproblem calculations became expensive too rapidly to allow the collection of meaningful data.

Here, we used only the All strategy. This was a practical choice made in order to reduce the size of the set of results which we must plot and consider. All calculations are given for as many iterations as were possible, up to practical limitations. We observe here that the cc-pCV5Z reference calculation, which was run as a single standalone calculation, was so expensive that it was practically difficult to exceed its abstract cost via the use of the order-theoretic combination technique. Per-iteration results for the calculations targeting the CCSD(T)/cc-pCV5Z reference are given in the left-hand plot of Figure 7.1, and for those targeting the FCI/CBS atomisation energy according to G4(MP2) in the right-hand plot.

Previous results in [GHH14; Heb14; CGH18] have already shown heptane and other alkanes to be highly amenable to SUPANOVA-style treatments, and the results here are no exception. With the exception of the (full-system) GCM, adaptively-refined

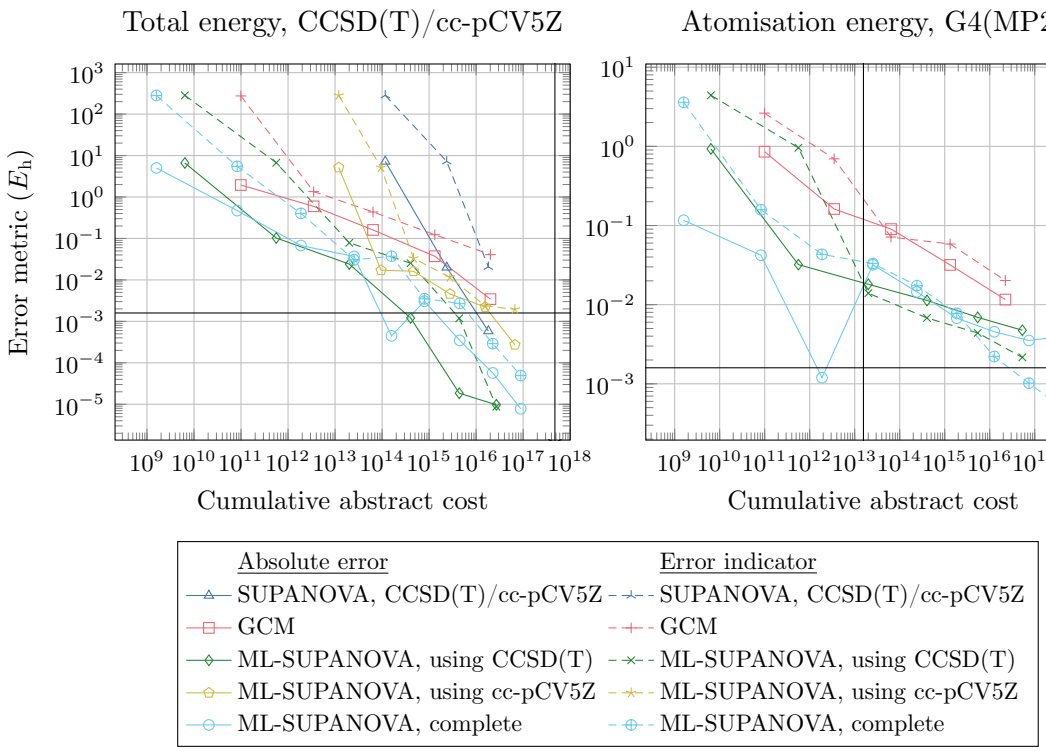

Figure 7.1.: Error metrics for ML-SUPANOVA and related calculations on heptane (C_7H_{16}). The left-hand plot shows per-iteration absolute errors measured relative to a reference CCSD(T)/cc-pCV5Z total energy calculation, as well as per-iteration error indicators. The right-hand plot shows equivalent per-iteration values for atomisation energy calculations, measured against a G4(MP2) calculation as reference. Vertical and horizontal lines indicate the abstract cost of the respective reference calculations and chemical accuracy ($\approx 0.0016\,E_h$) respectively.

combination sums in terms of each of the five poset grids considered can approximate the reference CCSD(T)/cc-pCV5Z total energy to within chemical accuracy, with significant speedups of up to more than three orders of magnitude. The cheapest calculation which achieves this is produced by the full ML-SUPANOVA calculation, but this appears to be just numerical coincidence, since the accuracy at the following iteration is again of less than chemical accuracy. The two-axis ML-SUPANOVA calculation with fixed CCSD(T) treatment of correlation, i.e., the ML-BOSSANOVA calculation, falls below chemical accuracy more consistently and at only slightly more cost. The other two-axis ML-SUPANOVA calculation, with a fixed cc-pCV5Z basis set and varying treatment of correlation, does not perform as well, and is indeed beaten to chemical accuracy by the standard non-multilevel SUPANOVA (BOSSANOVA) calculations.

Although the cost/error behaviour of basis-set ML-SUPANOVA calculations appear to be slightly more favourable than those of the complete ML-SUPANOVA calculations, the error indicator on the latter is more reliable, tracking the true error for the most part to within an order of magnitude. The error indicators for the correlation-only ML-SUPANOVA grid and the GCM grid are also reliable; those for the non-multilevel SUPANOVA axis are less so.

The results for the approximations of the atomisation energy are not as clear. Since heptane is still a relatively small system and the G4(MP2) calculation for heptane is quite affordable, it is not surprising that none of the three calculations considered here offers a significant relative performance benefit; we are more interested in the general trends of error behaviour. Once past the first few iterations, all three grids produce decays in error that go roughly at the same rate, although the non-SUPANOVA GCM calculation is notably and again unsurprisingly more expensive for equivalent accuracy. The complete ML-SUPANOVA grid calculation appears to pick up a very slightly faster rate of decay than does the two-dimensional basis-only grid calculation, although this is not clearly-indicated enough to draw any kind of conclusion. Again, the error indicator of the complete grid calculation appears to be slightly more reliable in the limit than that of the basis-only grid calculation, but the latter behaves better here than it did in the total energy case.

With the exception of an obvious numerical fluke in the results of the complete ML-SUPANOVA grid, none of the combination technique solutions approximates the G4(MP2) reference solution to within chemical accuracy. It is not clear here whether this is due to unreliability or inaccuracy of the reference itself, if the CCSD(T)/cc-pCV8Z atomisation energy which the ML-SUPANOVA solutions are actually approximating even agrees with G4(MP2) to that level of accuracy, or some combination of the two. This problem — that of obtaining reliable reference results against which to benchmark — is vexing, and it seems that the best we can say here is that all of the tested grids do allow for systematically-improvable approximations of the atomisation energy, although possibly at a slower rate of error decay compared to that of the total energy.

7.4. Case study: limonin ($C_{26}H_{30}O_8$)

For the final case study of this thesis, we return to the heterocyclic molecule limonin ($C_{26}H_{30}O_8$) [CSLimo], at the same optimised geometry as in Section 6.6. We saw there that an adaptive SUPANOVA calculation over a poset axis of convex subgraphs leads to a systematically-improvable estimation of the MP2/cc-pCVTZ total energy of limonin for both vacuum and mixed-basis embeddings. There, however, the computational benefit of the SUPANOVA approach was limited; although an adaptive vacuum-embedding SUPANOVA calculation was able to approximate the MP2/cc-pCVTZ energy to within chemical accuracy at a slight cost improvement relative to the reference, this improvement

was not especially persuasive.

We investigate now whether a multilevel treatment of limonin may be able to provide a more substantial performance improvement. Although not a particularly "large" system, we were still unable to obtain a CCSD(T) reference energy for limonin at cc-pCVTZ or higher. Instead, we consider only the MP2/cc-pCVQZ [MP34; SO89; Dun89; WD95] total energy of limonin, which required a calculation in terms of 3756 contracted basis functions. The poset grid we use to approximate this quantity is now formed as the direct product of a convex SUPANOVA axis for the same quotient graph $G = G'/F$ of the covalent bond graph G' as was used in Section 6.6, and chain axes with elements $2 \leq m \leq 3$ for the correlation treatment (i.e., either HF or MP2), and $2 \leq n \leq 4$ for the cc-pCVnZ basis set.

As for heptane, we also attempted to approximate the atomisation energy of limonin as estimated by the G4(MP2) model, $E^{\text{atom}}_{\text{G4(MP2)}} \approx 11.183\,448\,E_{\text{h}}$. An application of G4(MP2) is no longer quite so computationally trivial as has been the case for results discussed previously in this thesis, as it carries an abstract cost of 3.593×10^{16}. The ccCA-PS3 atomisation energy of limonin proved to be too expensive to calculate (abstract cost 1.811×10^{19}). When calculating atomisation energies, we extended the grids to include electron correlation treatments up to CCSD(T) and basis sets up to cc-pCV8Z, as above.

We consider now a matching set of All-strategy calculations over the same poset grids considered in the previous section. As an exception, we did not consider the GCM-only grid, since the calculations required past the third iteration became too expensive. Per-iteration error metrics are plotted in Figure 7.2, with a plot format equivalent to that in Figure 7.1.

Again, the results for the total energy approximations are inconclusive. All of the grids do produce sequences of approximations that decay in an overall smooth and consistent way. Both the standard SUPANOVA grid and the correlation treatment-only ML-SUPANOVA grid display single iterations at roughly the same cost which are more accurate than would be expected from trend, but which are not supported by the error indicator; in the end and at their most expensive iterations, both of these grids only graze chemical accuracy. The complete ML-SUPANOVA grid and the basis-only grid are here again competitive with each other. Although the basis-only grid does produce the cheapest result that falls below chemical accuracy, the decay of errors for the complete grid is truly monotonic. However, in both cases, chemical accuracy is obtained only at very limited speedup relative to the reference calculation, and unreliably and unpredictably so, since in both cases the error indicator is slightly more than an order of magnitude greater than the true error.

The lack of any significant speedup makes these results appear underwhelming. It should be considered, however, that the problem setting is only limited, and there is not much scope for the multilevel approaches to perform. There is still a reasonably clear benefit to the use of the complete and basis-only ML-SUPANOVA results. It seems uncontroversial to suggest that were we to somehow obtain a CCSD(T) total energy

for a high-quality basis set such as cc-pCV5Z or better, that a true speedup would be observed, although perhaps not as comprehensive as in the case of heptane. It is also generally encouraging to see that the ML-SUPANOVA technique produces systematically-improvable results here — with the use of convex subgraphs in a highly-cyclic molecular system — just as in the topologically much more simple case of heptane.

Considering the atomisation energy results as obtained by the two tested grids, namely the complete ML-SUPANOVA grid and the basis-only grid, we see again similar performance between the two in the later iterations. However, the utility of both as approximations of the G4(MP2) reference atomisation energy is still severely limited, since neither agrees with that value to within $0.1\,E_\mathrm{h}$. Once again, it is difficult to conclude more here, given the questions that surround the reliability of the reference value, but at the very least, the ML-SUPANOVA results are again systematically improvable. Given that the cost of the G4(MP2) method has significantly increased in this case, we can tentatively hypothesise that ML-SUPANOVA techniques may be of use in similar approximations for larger systems again, where the G4(MP2) cost will become prohibitive. Further investigation is certainly required.

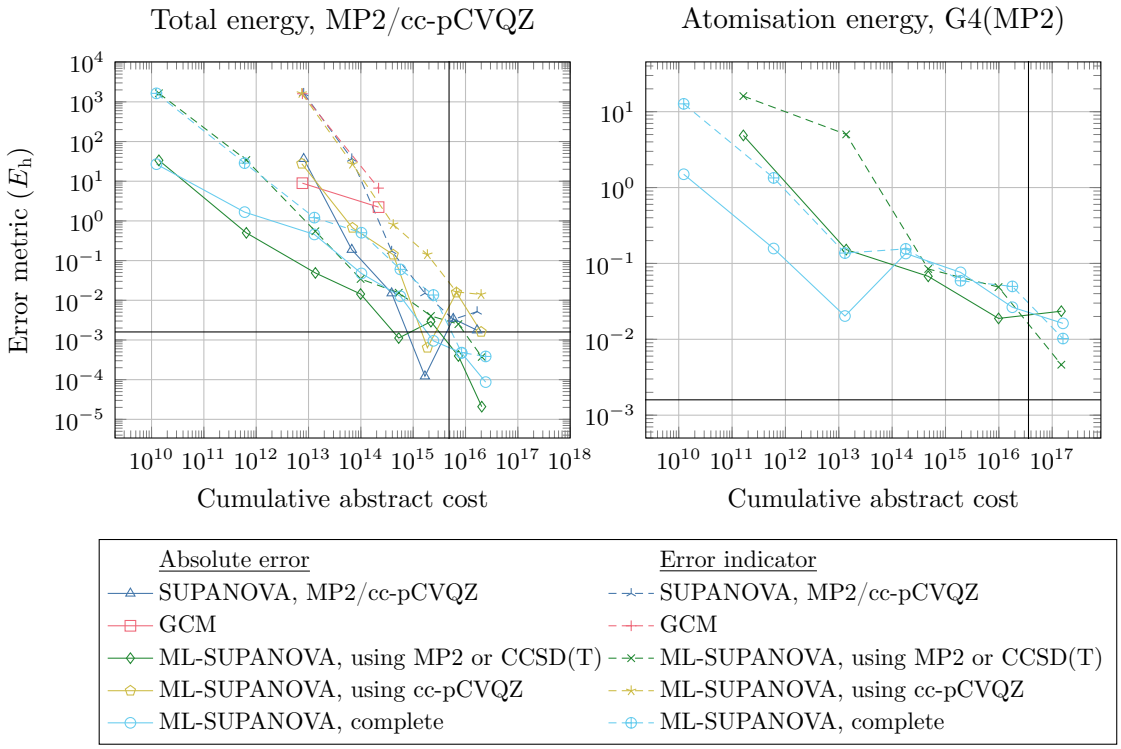

Figure 7.2.: Error metrics for ML-SUPANOVA and related calculations on limonin ($C_{26}H_{30}O_8$). The left-hand plot shows per-iteration absolute errors measured relative to a reference MP2/cc-pCVQZ total energy calculation, as well as per-iteration error indicators. The right-hand plot shows equivalent per-iteration values for atomisation energy calculations, measured against a G4(MP2) calculation as reference. The notation "using MP2 or CCSD(T)" corresponds to the left- and right-hand plots respectively. Vertical and horizontal lines indicate the abstract cost of the respective reference calculations and chemical accuracy ($\approx 0.0016\,E_h$) respectively.

8. Conclusion

In this thesis, we constructed an order-theoretic generalisation of the standard combination technique, and considered several applications of that generalisation to the approximation of energetic properties derived from solutions to the Schrödinger equation for molecular systems. We conclude by first giving a brief summary of our results, and then finally by outlining several promising avenues for potential future work.

8.1. Summary

In Chapter 2, we summarised a variety of standard *ab initio* techniques for the numerical solution of the electronic Schrödinger equation. An abstract cost model for estimating the relative cost difference between *ab initio* calculations was formulated.

We began Chapter 3 by sketching a conventional development of the standard combination technique [GSZ92; Gar12b; TW18]. Viewed from the perspective of order theory, the standard combination technique is defined in terms of a partially ordered set [Heg03; HGC07; Har16a; Won16] that can be constructed as the direct product of some number of totally ordered sets, or chains. We extended the construction of the combination technique to the setting of a poset grid, formed as the direct product of members of a very general class of partially ordered sets. The introduction of combinatorial and order-theoretic concepts, particularly the Möbius function and the technique of Möbius inversion [Sta12], was used to derive generalised analogues of the summation formulae underlying the standard combination technique in this setting. Extending on existing work surrounding and related to the standard combination technique, e.g., [Gri98; GG03; Heg03; CGH18; TW18], we outlined an adaptive algorithm for the discovery of quasi-optimal combination sums over order ideals of the underlying poset grid. We refer collectively to this extended formulation and the accompanying adaptive algorithm as the *order-theoretic combination technique*.

Extending upon observations and work by [CGH18] and particularly the CQML method of [Zas+18], we discussed in Chapter 4 the application of what is effectively just the standard combination technique to the problem of obtaining very highly accurate approximations to the true total energy of a molecular system. Since this application is inspired by consistent patterns in the energy equations of conventional composite methods, which were previously noted by [Zas+18], we refer to it as the *generalised composite method* (GCM). Case studies of the calculation of total and/or atomisation energies for three small molecules suggest that the GCM is a well-behaved approximation

technique, and can deliver results to reasonably high accuracy. Although there is no major cost-based benefit of the GCM relative to existing composite methods, and indeed the former performs not quite as well as the latter, the GCM provides a more systematic framework that may be amenable to future analysis.

Turning to larger molecular systems, Chapter 5 was devoted to a detailed discussion of the *many-body expansion* (MBE) and of a variety of related energy-based fragmentation methods. We contrasted several existing mathematical viewpoints on the MBE, and discussed the development of truncated MBEs as combination sums that are produced by the order-theoretic combination technique in terms of order ideals of either the boolean algebra B_M, or as a subposet thereof. With reference to an earlier result of Lafuente and Cuesta [LC05], we considered the conditions under which such a subposet will be *combination consistent* with the full boolean algebra, in the sense that a combination sum over any arbitrary order ideal of that subposet can be placed into termwise correspondence with an equivalent combination sum in terms of an order ideal of B_M. This idea was extended to the general order-theoretic combination technique setting.

Practical application of the order-theoretic combination technique over B_M produced an *adaptive many-body expansion*, which we tested in the calculation of the total energies of two clusters of water molecules. The results were consistent with previous work in the literature, e.g. [RLH14; Lao+16; LH17], in that both numerical and cost-scaling issues make standard n-body expansions neither particularly efficient nor accurate. The application of adaptivity proved to have only limited utility here.

In Chapter 6, we considered a general class of ANOVA-like decompositions of functions such as V^{BO}, phrased in terms of interaction graphs such as the covalent bond graph of a molecular system. This family of *SUPANOVA decompositions*, which can be viewed as a special case of a decomposition due to Klein [Kle86], contains in particular the BOSSANOVA technique of Heber and co-workers [Heb14; GHH14]. A careful consideration of the construction of the BOSSANOVA decomposition in light of the conditions for combination consistency established in Chapter 5 provided a rigorous explanation for previously-observed issues with the BOSSANOVA decomposition in cases of molecular systems with cyclic covalent bond graphs.

Addressing these issues, we considered instead a SUPANOVA decomposition based not on the poset of connected subgraphs of an interaction graph, but instead on that of geodesically convex subgraphs of that graph. This poset is guaranteed by construction to be combination consistent with the underlying boolean algebra. Although the use of convex subgraphs to construct formally similar expansions has been previously suggested [Kle86], and other existing techniques have been defined in terms of what turns out to be a simpler subset of these subgraphs [RHI18; RI18; KI19; RI20], this is the first time, to our knowledge, that they have been explicitly and fully considered in a modern energy-based fragmentation setting. Several case studies demonstrated that the resulting *convex SUPANOVA* technique is capable in some cases of providing systematically-improvable approximations of energetic quantities of potentially very large molecules, although this is

not necessarily guaranteed, and care must be taken to ensure that an interaction graph is chosen which includes sufficient information about the interaction structure of the system under study. With an eye to the latter, and reapplying well-known concepts both from fragmentation methods as well as computational chemistry more generally, we considered *radial interaction graphs* that explicitly incorporate distance-thresholded interactions between atoms as well as those implied by bond connectivity. Such graphs seem to be capable of remedying certain deficiencies in the interaction information contained in a covalent bond graph. However, although the application of the adaptive algorithm to the resulting convex SUPANOVA decomposition produces well-behaved sequences of approximations, the results in this chapter make clear that truly persuasive speedups relative to standard calculations still remain elusive, even for large molecules.

Finally, we briefly considered in Chapter 7 a straightforward combination of the GCM and SUPANOVA techniques. Again both inspired and informed by existing multilevel techniques in the literature, including particularly the ML-BOSSANOVA expansion [CGH18], this *ML-SUPANOVA* method involves the decomposition of an energetic quantity such as V^{BO} both in terms of basis set and correlation treatment, and in terms of connectivity-based subsystems. An insurmountable practical difficulty was encountered here, in that reference-quality solutions to the Schrödinger equation are practically unattainable for any system large enough that we might hope to obtain a true benefit from the ML-SUPANOVA approach. Restricting ourselves to smaller systems where reference solutions *are* available, the results are mixed. Certainly the ML-SUPANOVA technique produces approximations which increase in accuracy at a rate that improves upon those of the basic SUPANOVA formulation; however, the data are unclear as to whether a multilevel technique considering both basis set level and correlation treatment is any more efficient than one involving only basis set level, such as is provided by the ML-BOSSANOVA technique. At the very least, the use of a convex SUPANOVA decomposition rather than the original BOSSANOVA decomposition makes this technique now safely applicable to systems with cyclic interaction graphs. Also, if the adaptive error indicator provided by the order-theoretic combination technique is to be believed, a full three-dimensional ML-SUPANOVA technique may be useful when applied to larger systems.

8.2. Future work

In full conclusion, we now provide a sketched outline of interesting areas and ideas which we were not able to fully explore in this thesis, either for reasons of scope, of time, or of space. The following are given in no particular order.

Use of CBS extrapolation in the GCM

Although the GCM outlined in Chapter 4 was seen to provide a systematically-refinable approximation of the FCI/CBS energy of a molecule, its performance was no better than comparable to that of the existing composite methods that motivated it. One potential reason for the better performance of these composite methods is their aggressive use of unidirectional extrapolation techniques, particularly of HF, MP2, CCSD and CCSD(T) results towards the CBS limit [Taj+04; CRR07b; Kar+06; DeY+09]. Since any GCM index set is required to be an order ideal, it follows that if, for example, an MP2/cc-pCVQZ point is included in an index set, then so too are the MP2/cc-pCVDZ and MP2/cc-pCVTZ points required for an extrapolation from the cc-pCVQZ level towards the CBS limit. Moreover, if that index set was built adaptively, the calculated energies for these points must also have been previously calculated. It would be possible, therefore, to extend the GCM to make use of extrapolated subproblem potentials, rather than conventional ones. This may allow more accurate GCM results at a negligible increase in computational cost. Such an extension could also be applied in the more general ML-SUPANOVA setting; here, we note in connection some previous work applying CBS extrapolation to many-body terms in [MGP99].

Quantum embedding techniques

In Chapters 5 and 6, we investigated the use of mixed-basis contribution potentials. During the research which led to the production of this thesis, we originally considered these in the hope of avoiding certain practical issues associated with the use of link atoms for the treatment of dangling covalent bonds; see, e.g., [CD06; ŘS09; Le+12]. However, although the obtained results indeed seem to offer systematically-refinable approximations of energetic quantities, the still-high costs of mixed-basis embedding subproblem potentials appears to rule out their useful application in practice, particularly in light of the number of individual calculations required for SUPANOVA calculations.

In light of this and also of comments in, e.g., [HNK18], there is good reason to suspect that mixed-basis set embeddings may not be well-suited to the task. It would be of great interest to consider a similar SUPANOVA formulation that uses instead true quantum embeddings as briefly discussed in Section 5.1.1, be they either DFT-in-DFT, WFT-in-DFT, or WFT-in-WFT. Here, the previous work mentioned in Section 5.4 would be the natural place to start.

The use of quantum embeddings might be particularly worthwhile in the context of an ML-SUPANOVA formulation, where one poset axis indexes the correlation treatment of the embedding region in a WFT-in-DFT or WFT-in-WFT scheme. This idea is hinted at in [Man+12], and what is recognisably a simple multilevel scheme is used in [GW12] to incorporate dispersion contributions calculated via MP2 into two-body terms. Also, the title of [RKI20] suggests an investigation of DFT-in-DFT and WFT-in-DFT embeddings

in a multilevel setting. However, only ONIOM-style subtractive embedding approaches are there considered, rather than true quantum embeddings.

Approximations of nuclear gradients

We mentioned in Chapter 2 that the total energy of a molecular system is not a particularly useful quantity in and of its own right and that, in practice, one is usually more interested in either energy differences such as atomisation energies, or the forces on the nuclei as provided by the gradient of V^{BO} with respect to the nuclear coordinates. In this thesis, we considered the former in a limited fashion, but did not explicitly consider the latter.

It is explicit in the construction of the order-theoretic combination technique that any linear evaluation functional $\mathcal{L}$ can be used. Although we used here a functional corresponding to point evaluation of a subproblem potential, it would be theoretically trivial to replace this functional with one for point evaluation of the nuclear gradient of V^{BO}. That MBE-type sums of energies can be modified in this way to produce gradient calculations is well-known [DT07b; SDS09; MR11; LH17; Liu+19]. From a practical perspective, the combination sum would then become one of tensors, rather than of scalars. This might cause issues due to increased storage requirements for the collection of all evaluated gradients, particularly for large systems. However, these issues could in turn be treated by the use of appropriate sparse data structures, such as the sparse tensors used in Chapter 3, or indeed any sparse array implementation. There are also well-known issues regarding the impact of link atoms on the correctness of so-evaluated gradients [ŘS09] which would need to be considered.

Further verification of applicability and of results

The experimental studies provided in this thesis represent only initial investigations of the described techniques in particular cases, and the conclusions so drawn should not be extrapolated to the general case without further analysis. Indeed, while developing the algorithms and tools presented here, we also tested them informally on other molecular systems. As for those cases which we have explicitly considered, the behaviour of the adaptive algorithm and the quality of the resulting approximations seemed to be highly dependent on the structure of the involved molecule, its connectivity as modelled either by a bond graph or other interaction graph, and/or the precise approach used to generate a fragmentation of the involved atoms. In some cases, combinations of these factors appeared to present significant difficulties for the adaptive algorithm. Rigorous investigation of these cases was not feasible for reasons of time, space, and computational resources.

It is therefore important that a deeper understanding be sought of the precise conditions which molecular systems must fulfil in order that these algorithms can be expected to produce well-behaved approximations. It would also be important to investigate the

application of all of these techniques to more complicated scenarios, such as charged and/or non-equilibrium systems. Here, the natural beginning would be to revisit the existing and more rigorous analysis of the underlying nearsightedness conditions, as in, e.g., [Goe99; BBR13; Heb14].

Drawing more general conclusions about the performance and reliability of these techniques will require their application and statistical evaluation across larger datasets of molecules. Ideally, these should be chosen so as to be representative of particular subdomains of chemical compound space; see, e.g., [Zas+18; Bar+21; Kei+21] and references therein. Several such datasets have been published and widely used for evaluating the performance of quantum machine-learning techniques, see, e.g., [Ram+14; Nar+19]. Here, however, care must be taken to choose the applied subproblem potentials with reference to the methods used to calculate the reference values (particularly total/atomisation energies) in these datasets.

Theoretical analysis of benefit decay conditions

Closely related to the previous point, a key requirement for combination sums over SUPANOVA-type decompositions to be practically useful is a decaying upper bound on the magnitude of the surplus terms $\mathcal{L}[\tilde{V}_{\mathbf{u}}]$ and similar. Although such a decay is expected, and is clearly visible in, e.g., the data discussed at the end of Chapter 5, we have provided here no more than a handwaving justification for its existence.

Were rigorous bounds available here, then it might be possible to perform true error analysis on SUPANOVA-class combination sums. Existing techniques from the literature surrounding the standard combination technique and sparse grids in general may prove to be useful here. This would also be assisted by the results we have given regarding combination consistency, since if one could provide some kind of bound on the error of a combination sum over an arbitrary order ideal of a nuclear MBE — or even an MBE in terms of basis functions, rather than nuclei, like the construction underlying the BOSSANOVA method [Heb14] — then a generalisation to a SUPANOVA decomposition over another poset may not be too difficult, particularly if the Möbius function of that poset were to be known in a non-recursive form.

Alternative applications of the order-theoretic combination technique

Although the order-theoretic formulation of the combination technique that was presented in Chapter 3 is given in very general terms, we have only applied it here in a single setting. Given the broad applicability of the standard combination technique and similar methods that work in what we now recognise as simple poset grids of chain axes — see Section 3.2 and, e.g., [TW18] — it seems reasonable to suspect that there may be other areas of numerical science where the relaxed poset formalism may be useful.

A natural starting point for the discovery of further applications would be to consider the

listing of function space lattices given in [Heg03], and then investigating which (presumably high-dimensional) problem settings base naturally around those lattices. Similarly, an extension of the opticom technique given in [HGC07] to our construction strikes us as both plausible and interesting. The work of Wong [Won16] and Harding [Har16b] will surely also be relevant here.

It may also be possible to apply the order-theoretic combination technique in other areas of computational and quantum chemistry. We might begin with a deeper investigation of the chemical graph-theoretic cluster expansions listed by Klein in [Kle86], which interact with the wavefunction at a lower level than the "black-box" potentials we have considered here. Comments made in that source regarding a connection between convex subgraphs and subsets of occupied orbitals in a variant of the coupled cluster expansion are particularly intriguing, and we wonder whether the construction of some form of multilevel scheme might be possible in this setting.

Possible applications of the fast Möbius/Zeta transforms

In closing, we mention one final point related to Möbius inversion. We use here our usual notation of the order-theoretic combination technique, and assume some appropriate poset grid Π and families of model functions $\{f_{\mathbf{p}}\}_{\mathbf{p}\in\Pi}$ and hierarchical surpluses $\{\tilde{f}_{\mathbf{p}}\}_{\mathbf{p}\in\Pi}$ related as in Chapter 3 and specifically as per Theorem 3.3.4. It is natural to wonder how one might most efficiently calculate the model functions $f_{\mathbf{p}}$ (or, more precisely, evaluated model functions $\mathcal{L}[f_{\mathbf{p}}]$) given the necessary surpluses $\tilde{f}_{\mathbf{s}}$, or inversely, the hierarchical surpluses $\tilde{f}_{\mathbf{p}}$ given the necessary model functions $f_{\mathbf{s}}$. Provably optimal approaches for these calculations, called the *fast zeta transform* and *fast Möbius transform* respectively, have been developed for particular posets and classes of poset Π; see, e.g., [Ken92; Bjö+15]. One relatively recent variant is tuned for cases when many of the surpluses $\tilde{f}_{\mathbf{p}}$ are exactly zero [CDC19; CDC21; Cha21].

We became aware of this body of work only very late in the preparation of this thesis, and we have not had the opportunity to properly evaluate it in our context. However, there seems to be an immediate relevance. In particular, recall that we mentioned in Footnote 4 on page 112 that evaluating the combination sum S_I of an order ideal can be simply rephrased as the calculation of $\tilde{f}_{\hat{1}}$, where $\hat{1}$ is an additionally-adjoined maximal element of $J = I \cup \{\hat{1}\}$ and $f_{\hat{1}}$ is defined to be zero. Thus, an optimal Möbius transform for such a J would immediately provide an optimal mechanism for evaluating S_I. See, e.g., [Bjö+15; CDC19] for interesting possibilities when J is a lattice, which will occur at least whenever Π is a meet semilattice.

The adaptive approach we took in Chapter 3 to develop both I and the combination coefficients required to evaluate S_I is not directly comparable, since our focus is on the adaptive calculation of I itself. In particular, the evaluation of S_I at each loop iteration is simply linear in $|I|$, since the combination coefficients have been progressively constructed and are thus completely known. Informally, however, some $|I|$ evaluations of

MöbiusTensor($\mathbf{p}$) will have been necessarily required over the course of the algorithm up to that point, one for each $\mathbf{p} \in I$. Since each such evaluation and summation into the full combination tensor might cost at least $\mathcal{O}(\Lambda_{\mathbf{p}}^2)$ as per discussion in Section 3.5.3, the worst-case complete evaluation cost of S_I can be cubic in $|I|$. Here, it would be particularly interesting to investigate whether an alternative formulation of the adaptive algorithm could include some of kind of mechanism based upon the fast Möbius transform for more efficient updating of the involved values at each iteration. It might also be worth reconsidering the complexity of the fast Möbius transform in light of the costs to actually evaluate each $f_{\mathbf{p}}$; it seems to our initial reading of the above-cited literature regarding the fast Möbius transform that these values are assumed to be known *a priori* and thus available in constant time, which is not generally the case in our setting.

A. Calculation details

A.1. Quantum chemical software

The *ab initio* quantum chemical calculations reported in this thesis were performed with the MRCC [Kál+20; MRCC], NWChem [Apr+20], and PySCF [Sun15; Sun+17; Sun+20] software packages. We used the 2022 version of MRCC, and obtained main-branch versions of NWChem and PySCF from their respective GitHub repositories; the most recent preceeding release for NWChem was version 7.0.2, and for PySCF, version 2.0.1. Some trivial local modifications were made to both PySCF and NWChem for technical reasons. MRCC was used to perform some HF, MP2, CCSD, CCSD(T), CCSDT, CCSDT(Q), CCSDTQ, CCSDTQ(P), and CCSDTQP single-point calculations, using both RHF and UHF wavefunctions [SO89; MP34; Číž66; PB82; NB87; Rag+89; KB92; KS01; Bom+05; KG05; KG08; Rol+13; Kál14; GKN20]. NWChem was used to perform some single-point HF and MP2 calculations using RHF wavefunctions, see additionally [WH95; Fos+96; WHR96], and to perform KS-DFT/B3LYP [KS65; Ste+94] geometry optimisations for some test molecules. In all calculations, NWChem was explicitly prevented from removing linearly-dependent basis functions (`set lindep:n_dep 0`). PySCF was used to perform some HF, MP2, CCSD, and CCSD(T) single-point calculations using RHF wavefunctions; see additionally [Pul80; Pul82].

We also used the xTB package [Ban+20], via a Python API [xTBPy], to calculate the GFN2-xTB [BEG19] partial charges used in Section 5.4. The IAO partial charges [Kni13] used in the same section were calculated using PySCF, following an example script contained in the PySCF GitHub repository. The OpenBabel toolkit [OBo+11] was used to convert molecular structures between representation formats and, in some cases, to calculate bond structures and/or add hydrogen atoms.

We used functionality from the LIBCINT library [Sun15] to perform the ERI calculations necessary to implement the abstract cost model described in Section 2.5. All costs described in this thesis were calculated with LIBCINT version 5.1.1. Here, the internal cutoff parameter `PTR_EXPCUTOFF` was set to `fabs(log(1.0e-12))`. Standard single-point calculations performed with PySCF did not use manually modified LIBCINT settings.

A.2. Basis sets

The following basis sets were used for calculations described in this thesis:

- 6-31G* and 6-311G* [DHP71; HDP72; HP73; Kri+80].
- The Dunning-Hay DZ basis set [DH77].
- cc-pVnZ and aug-cc-pVnZ, for $2 \leq n \leq 8$ [Dun89; KDH92; WD93; PWD94; WMD96; FP99; MGP99; MWD99; FS00; FPD08; FP09; FPH10; FPH11; Tho+21].
- cc-pCVnZ and aug-cc-pCVnZ, for $2 \leq n \leq 8$ [WD95; Pet+97; Tho+21].
- aug-cc-pwCVTZ and aug-cc-pwCVQZ, for evaluating W4 costs in Chapter 4 [PD02].
- Three specialised basis sets required by the G4(MP2) composite method [Cur+98; CRR07b; CRR07b]: G3MP2LargeXP, and the G4(MP2) specialisations of aug-cc-pVTZ and aug-cc-pVQZ.

Most of the calculations we have described involved only basis sets for H, C, N, and O, but the cc-pVTZ calculations on the proteins 1KDF and 6VXX described in Section 6.8 also required basis sets for S.

In most cases, specifications for these basis sets were obtained from the Basis Set Exchange (BSE) [Pri+19]. The tight function values for cc-pCV6Z/aug-cc-pCV6Z were as provided by NWChem; we understand these to be the standard values originally due to Wilson and Dunning, as cited in [Pet+97]. The specifications for aug-cc-pCV7Z and aug-cc-pCV8Z were obtained from the supporting information of [Tho+21]. The cc-pCV7Z and cc-pCV8Z specifications were generated by removing diffuse functions from aug-cc-pCV7Z and aug-cc-pCV8Z. cc-pV7Z and aug-cc-pV7Z were generated by removal of core and/or diffuse functions from aug-cc-pCV7Z as appropriate. Specifications for cc-pV8Z and aug-cc-pV8Z were kindly provided by David Feller [Fel22]. The specifications of the various non-standard basis sets required for G4(MP2) calculations were obtained from the supporting information of [CRR07b].

Most basis set specifications were manipulated using the BSE Python library [Pri+19] as follows. Entries for multiple angular momenta (`SP` shells, etc.) were split into distinct entries (`uncontract_spdf(...)`). Entries were then converted to a single general contraction per angular momentum value, and these contractions were then optimised (`optimize_general(...)`) [HHT95]. Finally, all general contractions were again uncontracted and pruned (`uncontract_general(...)` and `prune_basis(...)`). These manipulations were done for reasons of computational efficiency, and to our understanding, should have no meaningful impact on results obtained.

Calculations were generally performed using spherical harmonic (i.e., pure) primitive Gaussian functions. Cartesian functions were used for calculations with 6-31G* and 6-311G**. Mixed pure/Cartesian basis sets in the sense discussed in [PLV20] were not used for any calculations.

A.3. Monoatomic total energies

This section describes the total energy calculations for the four monoatomic systems (H, C, N, and O) that were required for the calculation of the total atomisation energies described in Chapters 4 and 7.

Calculations were performed using spin-unrestricted (UHF) Hartree-Fock wavefunctions. For frozen-core calculations over C, N, and O, one spatial orbital and thus two spin orbitals were frozen. Spin multiplicities of 1, 3, 4, and 3 were set for H, C, N, and O respectively.

All monoatomic total energy calculations were performed using MRCC, with the exception of those for oxygen under the aug-cc-pCV8Z basis set; see below. The prescreening tolerance for calculation of two-electron integrals was uniformly set to 10^{-14} (`itol=14`). Energy convergence thresholds for both SCF and iterative coupled cluster calculations were set to 10^{-10} (`scftol=10`, `cctol=10`). MRCC's thresholding parameter for SCF density matrix convergence was set at 10^{-11} (`scfdtol=11`). Canonical MP2 energies were extracted from program output for the relevant CCSD calculations. Frozen-core calculations for C, N, and O were specified by the MRCC setting `core=1`.

Calculations for oxygen with the aug-cc-pCV8Z basis set were performed using PySCF. A convergence tolerance of 10^{-9} was set for both RHF and CCSD iterative solvers; PySCF's `direct_scf_tol` parameter was set to 10^{-14}. For frozen-core calculations, we set the PySCF solver attribute `frozen = 1`.

In the particular case of carbon, where frozen-core values for coupled cluster theories considering treatments higher than CCSDTQ were required, we simply used those for CCSDTQ. In this case, a CCSDTQ treatment corresponds to full coupled cluster, thus FCI.

A.4. Total energies, H_2O, O_3, and C_3H_5NO

This section contains full calculation details for the total energy calculations used in Chapter 4, both explicitly (for H_2O, in Section 4.4.1) and implicitly (for the atomisation energies of H_2O, O_3, and C_3H_5NO analysed in Sections 4.4.2 and 4.5).

All calculations at the RHF, MP2, CCSD, and CCSD(T) levels of theory were performed using PySCF. A convergence tolerance of 10^{-8} was set for both RHF and CCSD iterative solvers; PySCF's `direct_scf_tol` parameter, which we understand to control integral prescreening, was set to 10^{-14}.

All remaining correlated calculations (i.e., CCSDT and higher) were performed using MRCC. The following convergence tolerances were set: an RHF energy convergence threshold of 10^{-8} (`scftol=8`), an RHF density-matrix convergence threshold of 10^{-9} (`scfdtol=8`), a coupled cluster energy convergence threshold of 10^{-8} (`cctol=8`) and an integral prescreening threshold of 10^{-14} (`itol=14`). MRCC's implementation of the Rys

quadrature algorithm for ERI evaluation [RDK83; LRL91; Flo09] was explicitly selected (`intalg=rys`).

The use of automatically-detected symmetry was enabled for both PySCF and MRCC calculations for H_2O and O_3. The use of symmetry was not enabled for C_3H_5NO.

For frozen-core calculations over H_2O, one spatial orbital was frozen, corresponding to two spin orbitals (PySCF: solver attribute `frozen = 1`; MRCC: `core=1`). For frozen-core calculations over O_3, three spatial orbitals were frozen (PySCF: solver attribute `frozen = 3`; MRCC: `core=3`). For frozen-core calculations over C_3H_5NO, five spatial orbitals were frozen (PySCF: solver attribute `frozen = 5`; MRCC: `core=5`).

A.5. Standard composite method calculations

The G4(MP2), ccCA-PS3, and HEAT total and atomisation energies given and discussed in Chapters 4 and 7 were calculated consistent with the formulae given in Chapter 4. We calculated the G4(MP2) HLCs consistent with the description in [CRR07a]; the number of alpha/beta electrons in each system was calculated using PySCF. The implementation of G4(MP2) in the open-source `composite-thermochemistry-nwchem` package [Ern16] was a helpful reference here.

All individual single-point calculations required to compute G4(MP2) energies for non-monoatomic systems were performed using PySCF, using equivalent settings to those given in Section A.4. The full-system total energy values used to calculate ccCA-PS3 and HEAT energies in Chapter 4 were those described in Section A.4. The ccCA-PS3 total energies described in Chapter 7 were calculated directly using PySCF, again using settings as in Section A.4. For the calculation of all composite-method monoatomic energies, we used the single-point monoatomic energy values described in Section A.3. All given atomisation energies were calculated directly from corresponding molecular and monoatomic energies according to (2.27).

A.6. Preliminary geometry optimisations

The test molecules hexane, heptane, benzene, limonin, and chignolin considered in Chapters 6 and 7 were optimised to plausible equilibria after being obtained from the ChemSpider database or the PDB. Geometry optimisations were performed using NWChem, employing DFT calculations with the B3LYP functional [KS65; Ste+94] and the cc-pVDZ basis set. An energy convergence threshold of $10^{-6}\,E_\mathrm{h}$ was explicitly set; otherwise, NWChem's default thresholds were applied for both energy calculations and geometry optimisation. The ERIs used in these calculations were obtained via the SIMINT library [PC16].

A.7. Subproblem potentials

The point-evaluated vacuum and electrostatic-embedding subproblem potentials $\mathcal{L}[V_{\mathbf{u}}]$ used in the calculations discussed in Chapters 5, 6, and 7 were calculated using PySCF. Iterative convergence thresholds were set to $10^{-8}\,E_\mathrm{h}$ for both RHF and CCSD calculations, and PySCF's `direct_scf_tol` integral prescreening threshold was set to 10^{-12}. Where required, dangling single bonds were capped using hydrogen link atoms, placed as per [RH12, (9)]. Here, we used the covalent radii given in [Cor+08]. Dangling double bonds, which were only encountered in the representative calculation for benzene in Section 6.3, were treated with two hydrogen link atoms. These atoms were arrayed in a manner which we understand to be consistent with the treatment in [Heb14; Heb17], except that we also used covalent radii to decide on the distance of each hydrogen from the parent carbon atom, as above.

The mixed-basis embedding subproblem potentials were evaluated using NWChem, with iterative convergence thresholds also set to $10^{-8}\,E_\mathrm{h}$ and integral prescreening parameters (e.g., `tol2e`) set to 10^{-12}.

While performing the work described in Chapters 5, 6, and 7 of this thesis, we variously precalculated, cached, and reused the results of subproblem potential evaluations. All such results were cached at full precision, and significant care was taken to ensure that any adaptive calculation result obtained using cached results would be completely consistent with a "clean" reexecution of the same calculation, i.e., one that did not use cached results. Since costs are generally calculated and reported as per the abstract cost model described in Section 2.5, it is immaterial from this perspective whether the value of an evaluated subproblem potential is drawn from cache or calculated directly. The only case study given in this thesis in which real-world execution times were reported was the calculation of the total energy of the spike glycoprotein described in Section 6.8.2; here, all involved subproblem potentials were explicitly evaluated from scratch during the course of the calculation.

A.8. Order-theoretic combination technique implementation

We generated the results in this thesis using an implementation of the adaptive algorithm of Chapter 3 that was written in the Python programming language, with some performance-critical functionality implemented in C++. As well as various standard and general-purpose libraries and tools, we made heavy use of the Numpy [Har+20], Scipy [Vir+20], and NetworkX [HSS08] scientific libraries. A sparse tensor data structure was implemented directly, influenced by [Sparse] but tailored for our problem. Tensor reductions like, e.g., $\textsc{Reduce}(D \odot V)$ were handled in arbitrary precision using the Arb library [Joh17], using

100 bits of precision for intermediate values.[1] The required poset axis interfaces were implemented basically consistently with the high-level outlines in Appendix B, up to a variety of performance optimisations. Fallback Möbius-vector calculations for $\mathrm{conn}[G]$ and $\mathcal{M}_g[G]$ were performed using an implementation of (3.50), applying memoisation techniques in order to offset certain performance issues inherent to Python as far as possible.

Although the adaptive algorithm is given in Chapter 3 in a purely sequential form, a significant amount of parallelism is available; cf. here [CGH18, Sec. 3.4.3]. The simplest to exploit is the functional evaluation for each newly-added element in the innermost loop of the algorithm (see Algorithm 3.3, line 13). Our implementation was capable of distributing these evaluations across a standard HPC cluster according to a simple distributed-memory parallelisation scheme. The techniques used here are basic and can be found in any introductory textbook on parallel programming.

Our scheme uses the ZeroMQ message-passing protocol [PZMQ]. A *broker* process maintains connections to a collection of available *calculator* processes, which are responsible for launching evaluation calculations and collating subsequent results. At each iteration of the adaptive algorithm, a specified *batch* of calculations is dispatched from the *organiser* process (i.e., the process executing the adaptive algorithm) to the broker, which is then responsible for distributing the batched work items amongst the calculator processes, and collecting and returning the eventual results to the organiser process. We remark that similar approachs have been suggested in the fragmentation-method setting by, e.g., [Gan+06], but this is a very standard and natural pattern in high-performance and distributed computing in general.

The calculations involved in any given batch may vary substantially in terms of required runtime, potentially by several orders of magnitude. To minimise situations where a single long-running job delayed the completion of an entire batch, we applied a simple greedy-scheduling approach to load-balancing. Here, the abstract costs $\mathcal{C}(\mathbf{p})$ for all calculations in a batch are calculated *a priori*. These cost calculations are themselves dispatched to and performed in parallel by the calculator processes, since the walltime required for each cost-model evaluation itself may be relatively non-trivial: consider the abstract cost model outlined in Chapter 2, which requires the evaluation of a set of ERIs and thus requires a computation time on the order of milliseconds, rather than the nanoseconds that would be expected for the evaluation of a simple algebraic expression. Once this is done, batch tasks are assigned to available calculators in decreasing order of abstract cost. Under the assumption that true calculation costs are modelled accurately by the cost function, the resulting work distribution is close to optimal, at least when the computing capacities available to all calculator processes are identical; see, e.g., [Gra69, Thm. 2]. As above, we note that similar schemes have been previously implemented

[1] The choice of this value was essentially arbitrary, but we mention here that we found success using an almost identical precision to perform the otherwise unrelated calculations reported in [BTZ22].

in the fragmentation method setting [Gan+06; ŘS09]. More sophisticated scheduling approaches would certainly be possible, particularly if some deeper knowledge of the underlying calculations were to be applied, but anecdotally and informally, we observed our approach to work acceptably well in the general case.

A.9. Plots and visualisations

With the exception of Figures 4.2, 5.4, and 5.5, the plots and visualisations in this thesis use colours drawn from the "bright" palette of Tol [Tol21]. Three-dimensional molecule visualisations were produced using the Blender software package [Blender], shaded according to the *direct shadow overlay* (DSO) method described by Hansen [Han20].

In the visualisation of the SARS-CoV-2 spike glycoprotein in Figure 6.13, atoms are coloured by their membership in individual connected components of the underlying covalent bond graph of the full system, as implicitly calculated by the OpenBabel toolkit [OBo+11]. Colours were assigned according to an approximately minimal graph colouring of a derived graph. This graph contained one vertex for each connected component in the original bond graph, and an edge joining two vertices whenever any two atoms in the respective connected components of the original graph lay within a cutoff radius of 5 Å of each other. The graph colouring was calculated using the NetworkX function `networkx.coloring.greedy_color(...)` according to the default `largest_first` strategy; the NetworkX documentation for this function cites [KM04].

B. Poset axis interfaces

We give here a brief and high-level algorithmic summary of the required interfaces for the various poset axes used in this thesis, as per Section 3.5.2. For basic computer science results and terminology, we refer most strongly to [Cor+22]. We rely heavily on operations on sets. We assume rather naïvely here and throughout that a set $\mathbf{u}$ can be constructed and its elements iterated over with cost $\mathcal{O}(|\mathbf{u}|)$, and can have a single element added or removed at cost $\mathcal{O}(1)$; this is achieved in the average case by the built-in `set` data structure provided by Python [PWTC], but cf. more detailed discussion in [Cor+22, Chap. 11].

B.1. Chain poset

An implementation of the poset axis interface for an arbitrary chain poset P with a $\hat{0}$, finite or infinite, is given in Algorithm B.1. Here, we take the indexing bijection $\phi = \rho$, where ρ is the standard rank function on P; thus, ϕ is order-preserving in both directions. The details of the AxisIndex, Predecessors, and Successors functions follow immediately. The form of MöbiusVector follows directly from (3.27) in Example 3.3.12, restricted to the case $d = 1$ and possibly up to the finiteness of P; see alternatively [Sta12, Example 3.8.1].

B.2. Boolean algebra of rank n, B_n

An outline of the poset axis interface for the boolean algebra B_N is given in Algorithm B.2. The implementation of AxisIndex again reduces to one of an indexing bijection, $\phi : B_N \to \{0, \dots, 2^N - 1\}$. Construction of such a bijection is a standard problem in applied combinatorics [Leh64; NW78, Chap. 1; SW86; KS98, Chap. 2], where it is referred to as *ranking* the subsets of $[N]$. Probably the simplest approach is to identify each subset $\mathbf{u} \in 2^{[N]}$ with a binary string $a = a_1 a_2 \cdots a_N$, such that $a_i = 1$ if $i \in \mathbf{u}$ and $a_i = 0$ otherwise. A suitable bijection is then obtained by interpreting these strings as binary numbers, as in, e.g., the SubsetLexRank algorithm in [KS98, Alg. 2.1]:

$$\phi(\mathbf{u}) = \sum_{1 \le i \le N} a_i \cdot 2^{N-i} = \sum_{i \in \mathbf{u}} 2^{N-i}. \tag{B.1}$$

Evaluating the last sum requires $\mathcal{O}(|\mathbf{u}|)$ operations, assuming constant-cost arithmetic and exponentiation.

Algorithm B.1 Poset-axis interface implementation for a chain poset P.

```
1:  function AxisIndex(t ∈ P)
2:      return φ(t)

3:  function Predecessors(t ∈ P)
4:      if φ(t) > 0 then
5:          return {φ^{-1}(φ(t) − 1)}
6:      else
7:          return ∅

8:  function Successors(t ∈ P)
9:      if P is infinite or φ(t) < |P| − 1 then
10:         return {φ^{-1}(φ(t) + 1)}
11:     else
12:         return ∅

13: function MöbiusVector(t ∈ P)
14:     ▷ Initialise a zero-filled 1D sparse tensor with shape (|P|), which might be (∞). ◁
15:     M ← an empty sparse tensor
16:     i ← φ(t)
17:     M_i ← 1
18:     if i ≥ 1 then
19:         M_{i−1} ← −1
20:     return M
```

Although ϕ is order-preserving, it maps even some very small sets to very large indices; for example, $\phi(\{1\}) = 2^{N-1}$, which exceeds the range of a standard 64-bit unsigned integer for any $N > 64$. As a result, indices must be represented using arbitrary-precision integers. The impact of this can be ameliorated to an extent by the use of a different bijection. That just given ranks subsets as per the well-known *lexicographic ordering*; see, e.g., [KS98, Chap. 2]. We can build instead an alternative indexing bijection that orders, for every $1 \leq k \leq N$, all ranked subsets of size less than k before any of size k, and all those latter lexicographically. This ordering is also well-known; see and cf., e.g., [Eps+92, Sec. 2.5; KS98, Exercise 2.10; Fil13]. Given an algorithm for lexicographically ranking only the size-k subsets of $[N]$ into $\{0, \ldots, \binom{N}{k} - 1\}$ like, e.g., kSubsetLexRank [KS98, Alg. 2.7], then, clearly,

$$\phi(\mathbf{u}) = \sum_{k=0}^{|\mathbf{u}|-1} \binom{N}{k} + \textsc{kSubsetLexRank}(\mathbf{u}, |\mathbf{u}|, N). \tag{B.2}$$

Algorithm B.2 Poset-axis interface implementation for the boolean algebra B_n.

```
1: function AxisIndex(u ∈ B_N)
2:     return φ(u)

3: function Predecessors(u ∈ B_N)
4:     return {u − {i} | i ∈ u}

5: function Successors(u ∈ B_N)
6:     return {u ∪ {i} | i ∈ [N] − u}

7: function MöbiusVector(u ∈ B_N)
8:     ▷ Initialise a zero-filled 1D sparse tensor with shape (2^N). ◁
9:     M ← an empty sparse tensor
10:    for all v ⊆ u do
11:        i ← AxisIndex(v)
12:        M_i ← (−1)^(|u|−|v|)
13:    return M
```

An expression for kSubsetLexRank in terms of $\mathcal{O}(\mathbf{u})$ binomial coefficients is given by Lehmer [Leh64]; see also [SW86, Exercise 10; KS98, Alg. 2.9 and Thm. 2.4; use13]. Note, however, that if N is large, then arbitrary-precision integers may still be required to represent the indices of relatively small subsets; for example, if $N = 7524$, which is the number of fragments of the spike glycoprotein considered in Section 6.8.2, then $\phi(\mathbf{u}) > 2^{64} - 1$ for all subsets $\mathbf{u} \in B_N$ with $|\mathbf{u}| \geq 7$ and for some with $|\mathbf{u}| = 6$.

Given $\mathbf{u}, \mathbf{v} \subseteq [N]$, it is immediately clear that $\mathbf{u} \prec \mathbf{v}$ exactly when $\mathbf{u} = \mathbf{v} - \{i\}$ for some $i \in \mathbf{v}$. The resulting algorithms for Predecessors($\mathbf{u}$) and Successors($\mathbf{u}$) are obvious, and carry complexities of $\mathcal{O}(|\mathbf{u}| \cdot (|\mathbf{u}| - 1)) = \mathcal{O}(|\mathbf{u}|^2)$ and $\mathcal{O}(|\mathbf{u}| \cdot (N - |\mathbf{u}|)) = \mathcal{O}(N|\mathbf{u}|)$ respectively, due to the repeated replication of $\mathbf{u}$.

The implementation of MöbiusVector($\mathbf{u}$) follows immediately from the standard expression (3.32) for the Möbius function of B_N as given in Example 3.3.13. Since $\mu_{B_N}(\mathbf{v}, \mathbf{u}) \neq 0$ for any $\mathbf{v} \subseteq \mathbf{u}$, the function MöbiusVector($\mathbf{u}$) must return a sparse tensor with exactly $2^{|\mathbf{u}|}$ nonzero entries, one for each possible $\mathbf{v}$. Here, an algorithm for enumerating all subsets $\mathbf{v} \subseteq \mathbf{u}$ is required. We do not describe such an algorithm here — see, e.g., [NW78, Chap. 1; KS98, Chap. 2] — but one will require at best $\mathcal{O}(2^{|\mathbf{u}|} \cdot |\mathbf{u}|)$ operations, from the sum of the sizes of all possible subsets of $\mathbf{u}$, and this complexity is readily achievable in practice. The single $\mathcal{O}(|\mathbf{u}|)$ call to AxisIndex for each subset required to generate the indices for the sparse tensor does not affect the overall complexity of MöbiusVector, which is then $\mathcal{O}(2^{|\mathbf{u}|} \cdot |\mathbf{u}|)$ under the assumption of constant-cost writes to elements of the sparse tensor.[1]

[1]We note in passing that careful implementations that iterate directly over indices rather than explicit

B.3. Poset of subsets inducing connected induced subgraphs, conn[G]

For an arbitrary $G = (V, E)$, we are not aware of a suitable ranking function for members of conn[G] similar to those mentioned for the boolean algebra case above. One based on a lookup table could in principle be implemented by simply preenumerating all of the connected induced subgraphs, using an approach like that given in [Wer05], but this will generally not be feasible for medium to large and/or non-sparse graphs. Since conn[G] is just a subposet of $B_{|V|}$, however, it is possible to just reuse AxisIndex as defined in the previous section. Although not technically a bijection onto $\{0, \ldots, |\text{conn}[G]| - 1\}$, this functions perfectly well for the purposes of the order-theoretic combination technique; one simply uses sparse tensors with notionally longer axes.

We are neither aware of nor can we give a non-recursive expression for the Möbius function $\mu_{\text{conn}[G]}$ in the general case.[2] Instead, a fallback implementation of MöbiusVector as described in Section 3.5.3 can be used.

Pseudocode for the remaining two required poset-axis functions is given in Algorithm B.3. Their derivation follows straightforwardly from results as well as insight in [KS96; Wer05]. Specifically, Lemma 3.2 of [KS96] provides, up to notational differences, that $\mathbf{u} \prec \mathbf{v}$ in conn[G] if and only if $\mathbf{u} \subseteq \mathbf{v}$ and $|\mathbf{v}| - |\mathbf{u}| = 1$. It is also easy to see, as is implicitly relied on in [Wer05], that $\mathbf{u} \prec \mathbf{v}$ implies that, if we write $\mathbf{v} - \mathbf{u} = \{w\}$, then the additional vertex w must be adjacent in G to some vertex $u \in \mathbf{u}$; otherwise, $G[\mathbf{u}]$ and $G[\{w\}]$ would both be connected components of $G[\mathbf{v}]$ and $\mathbf{u} \cup \{w\}$ would not induce a connected subgraph of G.

Given a fixed graph G, we write the *neighbourhood* of $u \in V$ as $N(u) = \{v \in V \mid \{u, v\} \in E\}$, as usual [Cor+22]. The Successors function needs only iterate over all vertices $u \in \mathbf{u}$, and return copies of $\mathbf{u}$ extended by any neighbour $w \in N(u)$ which is not itself in $\mathbf{u}$. If G is stored using an adjacency-list representation or similar [Cor+22, Chap. 20], then $N(u)$ can be evaluated with linear complexity in the size of the output, which is at most $|V| - 1$ in the case of a complete graph. Thus, the worst-case complexity of Successors is given by $\mathcal{O}(|\mathbf{u}|^2|V|)$. If the size of any $N(u)$ is bounded above by some small constant independent of G, as is basically the case for the relatively sparse graphs we consider in this thesis, this becomes $\mathcal{O}(|\mathbf{u}|^2)$.

Calculating the Predecessors of some vertex set $\mathbf{u}$ that induces a connected subgraph of G is only slightly more complicated. In connection with [KS96, Lem. 3.2], we note that the problem is just that of enumerating the connected induced subgraphs of $G[\mathbf{u}]$ of size $|\mathbf{u}| - 1$. More general versions of this problem are well-studied [AF96; Wer05; KS21], but we use here only a naïve brute-force style search. Clearly, $\mathbf{v} \prec \mathbf{u}$ requires

subsets can bring this down to $\mathcal{O}(2^{|\mathbf{u}|})$, for both kinds of indexing bijection described above.

[2] Although this would be readily possible in the specific case where G is a tree, since conn[G] would then be a convex geometry. See Section 6.5.

Algorithm B.3 Partial poset-axis interface implementation for $\mathrm{conn}[G] \subseteq B_M$.

```
1: function Predecessors(u ∈ conn[G])
2:     if |u| = 1 then
3:         return {∅}
4:     predecessors ← ∅
5:     for all u ∈ u do
6:         if G[u − {u}] is connected then
7:             predecessors ← predecessors ∪ {u − {u}}
8:     return predecessors

9: function Successors(u ∈ conn[G])
10:    if u = ∅ then
11:        return {{u} | u ∈ V}
12:    successors ← ∅
13:    for all u ∈ u do
14:        successors ← successors ∪ {u ∪ {v} | v ∈ N(u), v ∉ u}
15:    return successors
```

that $\mathbf{v} = \mathbf{u} - \{u\}$ for some $u \in \mathbf{u}$, so each possible $\mathbf{u} - \{u\}$ is a candidate predecessor. However, not every such set is guaranteed to induce a connected subgraph of G.

Thus, we explicitly test the graph induced by each such candidate for connectedness. A well-known and standard way to achieve this is by performing a breadth-first search (BFS) [Cor+22, Sec. 20.2] of $G[\mathbf{u} - \{u\}]$, starting from an arbitrary vertex. It follows immediately from [Cor+22, Thm. 20.5] that $G[\mathbf{u} - \{u\}]$ is connected if and only if the BFS terminates without encountering every vertex in $\mathbf{u} - \{u\}$. Again assuming adjacency-list representation, forming $G[\mathbf{u} - \{u\}]$ costs at most $\mathcal{O}(|\mathbf{u}| \cdot |V|)$, again for a complete graph. Since BFS has complexity linear in the number of vertices plus the number of edges in a graph G, i.e., $\mathcal{O}(|V| + |E|)$ [Cor+22], the worst-case cost for the BFS of $G[\mathbf{u} - \{u\}]$ is $\mathcal{O}(|\mathbf{u}|^2)$, again when $G[\mathbf{u} - \{u\}]$ is complete. Thus, the worst-case cost of Predecessors($\mathbf{u}$) scales as $\mathcal{O}(|\mathbf{u}|^3 + |\mathbf{u}| \cdot |V|)$. Again, this cost is usually not encountered in practice here.

Note also that the implementations of Predecessors and Successors consider the special cases of the predecessors of a singleton subset, of which the empty set is the only one, and the successors of the empty set, which are all possible singleton sets.

B.4. Geodesic convexity, $\mathcal{M}_g[G]$

The geodesic convexity $\mathcal{M}_g[G]$ of those sets which induce geodesically-convex subgraphs of some connected graph $G = ([M], E)$ is also isomorphic to a subposet of B_M. Just as

for conn[G] above, we can reuse any valid implementation of the AxisIndex($\mathbf{u}$) function for B_M for $\mathcal{M}_g[G]$. We are not aware of a specialised expression for the Möbius function of $\mathcal{M}_g[G]$ in the general case, and must again rely on one of the fallback implementations of MöbiusVector($\mathbf{u}$) outlined in Section 3.5.3.

The implementation of Successors($\mathbf{u}$) for $\mathcal{M}_g[G]$ is also related to that of conn[G]. An if-and-only-if version of the following result is shown and used in the proof of [AK16, Thm. 20], there restricted to a particular class of graph G and without explicit mention of conn[G]. The proof for the only-if direction functions without modification in the case of a general G; the following reuses the same idea of proof, but is phrased in specific terms of conn[G].

Lemma B.4.1 (Extended/adapted from [AK16])**.** *Let $\mathbf{u} \in \mathcal{M}_g[G]$ be a convex set of vertices of some connected non-empty graph $G = (V, E)$, such that $\mathbf{u} \subset V$ induces a convex subgraph of G. Let $\mathbf{v} \in \mathcal{M}_g[G]$ be such that $\mathbf{u} \prec_{\mathcal{M}_g[G]} \mathbf{v}$, and say in context that $\mathbf{v}$ is a* convex cover *of $\mathbf{u}$. Then there exists some $\mathbf{v}' \in$ conn[G] such that $\mathbf{u} \prec_{\text{conn}[G]} \mathbf{v}'$ and $\mathrm{CH}_g[\mathbf{v}'] = \mathbf{v}$.*

Proof. We observe first that every convex set $\mathbf{u} \in \mathcal{M}_g[G]$ is also a member of conn[G] by definition, so $\mathcal{M}_g[G]$ is a subposet of conn[G]. Thus, if $\mathbf{u}$ is covered by some $\mathbf{v}$ in $\mathcal{M}_g[G]$, then $\mathbf{u}$ must have at least one cover in conn[G], so we have a non-empty set of candidates for $\mathbf{v}'$. It remains only to show that there exists some such candidate so that $\mathrm{CH}_g[\mathbf{v}'] = \mathbf{v}$.

Suppose first that $\mathbf{u} = \emptyset$. Since every singleton vertex subset is convex, the covers of $\emptyset$ in $\mathcal{M}_g[G]$ are exactly the singleton vertex subsets. Since this is also true for the covers of $\emptyset$ in conn[G], we are done.

Otherwise, following [AK16], since $\mathbf{u} \neq \emptyset$, we may pick an arbitrary vertex $u \in \mathbf{u}$. Since $\mathbf{u} \prec_{\mathcal{M}_g[G]} \mathbf{v}$, then $u \in \mathbf{v}$, and there must be at least one vertex $v \in \mathbf{v}$ such that $v \notin \mathbf{u}$. Since $\mathbf{v}$ is a convex set, all vertices in the geodesic interval $I_g(u, v)$ are by definition also in $\mathbf{v}$, and in particular, there must exist some $u', v' \in I_g(u, v)$ such that $u' \in \mathbf{u}$, $v' \notin \mathbf{u}$, and $\{u', v'\} \in E$. (We may choose u' to be the last vertex along some shortest path between u and v which is still in $\mathbf{u}$, and v' to be the vertex directly following u' on that path.)

Fixing now $\mathbf{v}' = \mathbf{u} \cup \{v'\}$, it is immediate from Section B.3 that $\mathbf{u} \prec_{\text{conn}[G]} \mathbf{v}'$. To see that $\mathrm{CH}_g[\mathbf{v}'] = \mathbf{v}$, note that $\mathbf{v}' \subseteq \mathbf{v}$, and so $\mathrm{CH}_g[\mathbf{v}'] \subseteq \mathrm{CH}_g[\mathbf{v}] = \mathbf{v}$, where we use the fact that CH_g is a closure operator on B_M; see, e.g., [Pel13, Thm. 1.3]. Were it the case that $\mathrm{CH}_g[\mathbf{v}'] \subset \mathrm{CH}_g[\mathbf{v}]$, then $\mathbf{u} \prec_{\mathcal{M}_g[G]} \mathbf{v}$ would not be true, which would be a contradiction. □

The converse is, however, not generally true; that is, it is not the case that if $\mathbf{u} \in \mathcal{M}_g[G]$, then $\mathbf{u} \prec_{\text{conn}[G]} \mathbf{v}$ implies $\mathbf{u} \prec_{\mathcal{M}_g[G]} \mathrm{CH}_g[\mathbf{v}]$. An easy counterexample is provided by the 3-fan substructure considered in the statement of Theorem 6.5.7 in Chapter 6 when regarded as a graph G in its own right. Consider there $\{v_1, v_2\}$, which is covered in conn[G] by

$\{v_1, v_2, v_3\}$, $\{v_1, v_2, v_4\}$, and $\{v_1, v_2, v_5\}$. It is readily verified that $\mathrm{CH}_g[\{v_1, v_2, v_3\}] = \{v_1, v_2, v_3\}$, and also that $\mathrm{CH}_g[\{v_1, v_2, v_4\}] = \{v_1, v_2, v_3, v_4\}$. Since the latter is a strict superset of the former, they cannot both cover $\{v_1, v_2\}$ in $\mathcal{M}_g[G]$.

It is well-known that $\mathrm{CH}_g[\mathbf{u}]$ can be calculated by repeated iteration of $\mathbf{u} \mapsto I_g[\mathbf{u}]$, proceeding until a fixed point is found; see, e.g., [Dou+09; Pel13]. The geodesic interval $I_g[\mathbf{u}]$ can itself be calculated with cost proportional to $\mathcal{O}(|\mathbf{u}|\cdot|E|)$, as discussed in [Dou+09]. The algorithm provided in that source involves launching a BFS of the entire graph from each vertex in $\mathbf{u}$. If such an algorithm is used, the overall calculation of $\mathrm{CH}_g[\mathbf{v}]$ scales in cost as $\mathcal{O}(|\mathrm{CH}_g[\mathbf{u}]| \cdot |E|)$ [Dou+09].

Using the above, we could calculate the set of convex covers of $\mathbf{u}$ — so, Successors($\mathbf{u}$) — by first calculating the set of connected covers of $\mathbf{u}$, then calculating the set of their convex hulls, and finally screening out any such convex hull which is a strict superset of another. We used here instead a slightly refined algorithm, which aims to avoid some of the possibly-expensive calculations of the geodesic interval that would be incurred by naïvely evaluating $\mathrm{CH}_g[\mathbf{v}]$ for each $\mathbf{v} \succ_{\mathrm{conn}[G]} \mathbf{u}$ in turn.

The idea is to perform the calculations of the convex hulls concurrently, and use the observation that, if some such $\mathrm{CH}_g[\mathbf{v}]$ is already known to be a convex cover of $\mathbf{u}$, then the convex hull $\mathrm{CH}_g[\mathbf{v}']$ of any other vertex subset $\mathbf{v}' \supset \mathrm{CH}_g[\mathbf{v}]$ cannot also be a convex cover of $\mathbf{u}$. Pseudocode for the ConvexCovers function is given in Algorithm B.4; although we give it a distinct name, note that this algorithm is completely functional as a definition of Successors for $\mathcal{M}_g[G]$. The core approach, that of somehow progressively expanding and winnowing out members of a set of candidate subgraphs, is a basic one in graph-theoretical algorithms; see, e.g., [Wer05; Dia+13]. We draw some inspiration here from a related iterative subset/subgraph expansion procedure described in [MŠ18], which stochastically samples maximal saturated chains from $\mathcal{M}_g[G]$, effectively using the idea of Lemma B.4.1. There is also a body of existing work focused on enumerating convex subgraphs of certain directed acyclic graphs; see, e.g., [Xia+21]. Although interesting, this does not seem to be directly applicable in our problem setting.

The algorithm begins by calculating the set of connected covers in conn[G] of some convex $\mathbf{u} \in \mathcal{M}_g[G]$, each of which has the form $\mathbf{u} \cup \{w\}$ for some $w \notin \mathbf{u}$ which is still a neighbour of some $u \in \mathbf{u}$. Each connected cover is considered to be an untested candidate for membership in the set of convex covers of $\mathbf{u}$. Each candidate is split into two disjoint sets of vertices, named `expanded` and `unexpanded`; the algorithm is constructed to enforce the invariant that $u \in$ `expanded` exactly when $I_g(u, v)$ has been previously and explicitly calculated for every other $v \in$ `expanded`. Initially, the set of `expanded` vertices is just $\mathbf{u}$, which is already known to be convex, so $I_g[\mathbf{u}] = \mathbf{u}$ and the invariant holds, and the set of `unexpanded` vertices is $\{w\}$.

The main loop here represents, very loosely, a similar approach to that underlying Dijkstra's well-known shortest-path algorithm [Cor+22, Sec. 22.3]. The candidates are placed in a "queue" in the form of a MinHeap, and ordered in that queue by size; we use here and in Algorithm B.4 essentially the notation of [Cor+22]. A collection of all convex

Algorithm B.4 Calculating the convex covers of some $\mathbf{u} \in \mathcal{M}_g[G]$.

```
1:  function ConvexCovers(u ∈ M_g[G = (V, E)])
2:      if u = ∅ then
3:          ▷ The empty set is covered by all singleton vertex subsets. ◁
4:          return {{u} | u ∈ V}
5:      ▷ Initialise the queue of candidate covers. Each candidate is stored as a pair of
          disjoint sets (expanded, unexpanded). The MinHeap is ordered by the sums
          of the sizes of the sets in each pair. ◁
6:      Q ← an empty MinHeap
7:      for all (u ∪ {w}) = v ≻_conn[G] u do
8:          MinHeap-Insert(Q, (u, {w}))
9:      ▷ Expand candidate covers in increasing order of size. ◁
10:     covers ← ∅
11:     while Q is not empty do
12:         (expanded, unexpanded) ← MinHeap-Extract-Min(Q)
13:         if ∃c ∈ covers such that c ⊆ (unexpanded ∪ expanded) then
14:             ▷ The convex hull of the candidate cannot be a (new) convex cover, so
                  discard it and proceed. ◁
15:         else if unexpanded = ∅ then
16:             ▷ Candidate is convex, and must be a convex cover. ◁
17:             covers ← covers ∪ {expanded}
18:         else
19:             ▷ Expand some vertex in unexpanded. ◁
20:             u ← an arbitrary element of unexpanded
21:             I ← I_g[expanded ∪ {u}]
22:             expanded ← expanded ∪ {u}
23:             unexpanded ← unexpanded − {u}
24:             unexpanded ← unexpanded ∪ (I − expanded)
25:             ▷ Requeue the updated candidate. ◁
26:             MinHeap-Insert(Q, (expanded, unexpanded))
27:     return covers
```

`covers` known so far is maintained. At each iteration of the algorithm, the smallest available candidate is selected from the queue. The candidate is checked against every entry of `covers`; if it is found to be a superset of any existing entry, then it is discarded. If not, and if there are no more `unexpanded` vertices, then it is added to `covers` for eventual return.

If the candidate is neither discarded nor kept aside, then an arbitrary `unexpanded` vertex u is chosen, and the geodesic interval $I_g[\texttt{expanded} \cup \{u\}]$ is calculated. This can be done using a single BFS with u as the source vertex, since the invariant guarantees that $I_g[\texttt{expanded}] = \texttt{expanded}$. The chosen vertex u is moved from `unexpanded` to `expanded`, and any vertices in the calculated geodesic interval which are not already in either `expanded` or `unexpanded` are added to `unexpanded`, to maintain the invariant on candidate pairs (`expanded`, `unexpanded`). The candidate is then requeued, and the algorithm repeats. Once the queue is empty, the algorithm returns `covers` as the set of convex covers of $\mathbf{u}$.

It is easy to see that progressively expanding a single candidate in this way, starting from $\mathbf{u} \cup \{w\}$ and stopping when $\texttt{unexpanded} = \emptyset$, produces a sequence $S^{(w)} = (\mathbf{s}_i^{(w)})_{i=1}^{K_w}$ of length K_w, and that $\mathbf{s}_{K_w}^{(w)} = \mathrm{CH}_g[\mathbf{u} \cup \{w\}]$.[3] This is, after all, just a very slight reorganisation of the algorithm for the construction of the geodesic convex hull outlined in [Dou+09] and mentioned above. It should also be clear that the terms of the sequence $(\mathbf{s}_i^{(w)})_{i=1}^{K_w}$ have the property that $\mathbf{s}_i^{(w)} \subseteq \mathbf{s}_j^{(w)}$ whenever $i < j$, and so $|\mathbf{s}_i^{(w)}| \leq |\mathbf{s}_j^{(w)}|$.

Suppose that the main loop in Algorithm B.4 were restricted such that candidates removed from the queue were never discarded, but only either selected for return, or expanded and requeued. The ordering property of the underlying MinHeap provides that the candidates removed at each iteration would provide a sequence $S' = (\mathbf{s}'_i)_{i=1}^{K'}$, such that each $S^{(w)}$ is a subsequence of S', that $K' = \sum_w K_w$, and that the elements of S' are also in nondecreasing order of size, although the subset ordering no longer applies. When candidates are allowed to be discarded, the equivalent sequence of elements $S = (\mathbf{s}'_i)_{i=1}^{K}$ is also in nondecreasing order of size, and is a subsequence of S' in turn.

Suppose that there exists some $\mathbf{s}'_i \in S'$ that is not an element of S. We can identify $\mathbf{s}'_i$ with $\mathbf{s}_j^{(w)}$ for some particular w and j, and the absence of $\mathbf{s}'_i = \mathbf{s}_j^{(w)}$ from S implies that some element $\mathbf{s}_k^{(w)}$ with $k < j$ was discarded upon removal from the queue.

Clearly, ConvexCovers terminates, since we are only considering finite graphs G and so K must be finite. We claim that it is also correct, in the sense that the sets $\mathbf{v}$ that are selected for return at line 17 of ConvexCovers are exactly and only those such that $\mathbf{u} \succ_{\mathcal{M}_g[G]} \mathbf{v}$.

To see this, suppose first that there exists some such $\mathbf{v}$ that is not selected. This is equivalent to saying that it is never removed from the queue, so is not an element of S,

[3] Technically, for the terms of this sequence to be unambiguously defined, we would need to make the selection of u at each step deterministic, but this is easily done in practice.

and so there exists some $\mathbf{s}_k^{(w)}$ for some w and k such that $\mathbf{s}_k^{(w)} \subseteq \mathbf{v}$ was removed from the queue and then discarded. If the subset relationship is proper, then $\mathbf{v}$ cannot be a convex cover of $\mathbf{u}$; if it is an equality, then $\mathbf{s}_k^{(w)} = \mathbf{v}$ would still have been selected for return. Both are contradictions.

Now suppose that some $\mathbf{v}$ which is not a convex cover of $\mathbf{u}$ is selected for return. Then there must exist some $\mathbf{w}$ such that $\mathbf{u} \prec_{\mathcal{M}_g[G]} \mathbf{w} \subset \mathbf{v}$. But we just established that all convex covers of $\mathbf{u}$ are selected. Since S is in nondecreasing order of size, $\mathbf{w}$ precedes $\mathbf{v}$ in S. Then $\mathbf{w}$ would have been removed from the queue, and selected for return prior to $\mathbf{v}$ being removed from the queue, which would have caused $\mathbf{v}$ to be discarded at line 14 of Algorithm B.4 and thus not selected for return.

Although we have observed an implementation of ConvexCovers to perform well in practice, we will not attempt to provide any explicit complexity results for the function here. For completeness, we note that the actual implementation in code that we used to obtain the results in this thesis was a slightly optimised variant of ConvexCovers as given here. The only significant difference in this version is that candidate pairs are not requeued at line 26 if their union is a superset of the union of any candidate pair already in the queue; the correctness of this optimisation also follows from the fact that $\mathbf{u} \subseteq \mathbf{v}$ implies $\mathrm{CH}_g[\mathbf{u}] \subseteq \mathrm{CH}_g[\mathbf{v}]$ since CH_g is a closure operator.

We were unable to derive a similar algorithm for calculating the predecessors of an arbitrary element of $\mathcal{M}_g[G]$. Informally, the main difficulty is that the geodesic convex hull is in general a many-to-one mapping; again, we mention the body of existing literature on problems related to the hull number of a connected graph, which includes the earlier-cited [Dou+09]. Thus, for the implementation of Predecessors($\mathbf{v}$) which we used to obtain the results reported in this thesis, we were forced to resort to a deeply inelegant solution which we will only describe informally.

Maintaining a set S which is initialised as $S \leftarrow \{\emptyset\}$, we repeatedly pick a member $\mathbf{v}$ of that set and evaluate ConvexCovers($\mathbf{v}$). If one of those convex covers is $\mathbf{u}$, then $\mathbf{v}$ is a predecessor of $\mathbf{u}$. All other convex covers $\mathbf{w} \succ_{\mathcal{M}_g[G]} \mathbf{v}$ are tested for the property $\mathbf{w} \subset \mathbf{u}$; those for which this property holds are reinserted into the set, and the process is repeated. Once no more elements remain in the set, the collection of elements $\mathbf{v} \prec_{\mathcal{M}_g[G]} \mathbf{u}$ so discovered is returned.

When care is taken not to consider the same $\mathbf{v}$ multiple times, this can be implemented as just a standard BFS (or depth-first search) over a directed graph with the elements of P as vertices, and such that the graph contains a directed edge from $\mathbf{u}$ to $\mathbf{v}$ if and only if $\mathbf{u} \prec_{\mathcal{M}_g[G]} \mathbf{v}$. This graph can be directly related to Hasse diagrams like that in Figure 3.1; indeed, Hasse diagrams are sometimes explicitly defined as directed graphs, as in [HGC07; AK16].

We found this approach to be perhaps surprisingly performant when combined with the use of memoisation so that each ConvexCovers($\mathbf{v}$) is only calculated once during a calculation. Despite the inelegance, it also still meets our informally-stated requirement

that we should be able to explore the successors and predecessors of $\mathbf{u}$ somehow locally to that point, since the amount of work required for the BFS is (up to the explicit calculations of ConvexCovers) linear in the sum of the size of the principal order ideal $\Lambda_{\mathbf{u}}$ and of the number of distinct cover relationships $\mathbf{v} \prec_{\Lambda_{\mathbf{u}}} \mathbf{w}$, and does not depend on the size of the complete poset $\mathcal{M}_g[G]$.

List of figures

List of tables

List of algorithms

List of symbols

General

$\mathbb{N}$	set of natural numbers, $\{0, 1, 2, \ldots\}$
$\mathbb{N}^+$	set of strictly positive natural numbers, $\{1, 2, \ldots\}$
$\mathbb{C}$	set of complex numbers
$\mathbb{R}$	set of real numbers
$\mathbb{R}^+$	set of strictly positive real numbers
$\mathbb{Z}$	set of integers
$[n]$	set $\{1, 2, \ldots, n\}$ for some $n \in \mathbb{N}$
$\emptyset$	empty set
$\subset$	strict set inclusion, such that $X \subset Y$ implies $X \neq Y$
$\subseteq$	set inclusion, with equality allowed
K	a field
V	a vector space
Y	a Banach space

Order theory

(Note: much of this notation is drawn from [Sta12])

$\leq$	order relation on a poset P
$\prec$	cover relation on a poset P
P, Q	posets with order relations $\leq_P$, $\leq_Q$
$\hat{0}$	zero (unique minimal element) of some P
$\hat{1}$	one (unique maximal element) of some P
$[s, t]$	interval between $s \leq t$ in some P
Λ_t	principal order ideal of some $t \in P$
I	an order ideal of some P
A	an antichain
$\langle A \rangle$, $\langle a_1, \ldots, a_n \rangle$	order ideal of some P generated by $A = \{a_1, \ldots, a_n\} \subseteq P$
μ	Möbius function of some P
$s \wedge t$	meet (greatest lower bound) of $s, t \in P$
$s \vee t$	join (least upper bound) of $s, t \in P$
B_n	boolean algebra of rank n

Combination techniques

$P,\ P_i$	poset axis
Π	poset grid, $\Pi = P_1 \times \cdots \times P_d$ for $d \geq 1$
$f_{\mathbf{p}}$	model function indexed by $\mathbf{p} \in \Pi$
$\mathcal{F}_\Pi$	model hierarchy, $\mathcal{F}_\Pi = \{f_{\mathbf{p}}\}_{\mathbf{p}\in\Pi}$
$\tilde{f}_{\mathbf{p}}$	hierarchical surplus of $f_{\mathbf{p}}$
I	downward-closed index set $I \subseteq \Pi$ (also an *order ideal*)
S_I	I-truncation of combination sum with respect to I
$D_{\mathbf{p}}^{(I)}$	combination coefficient of $f_{\mathbf{p}}$ in the I-truncation S_I
$\mathcal{L}$	evaluation functional, $\mathcal{L} : \mathcal{F}_\Pi \to Y$

Many-body expansions and SUPANOVA decompositions

$[M]$	in context, the set of nuclear indices for a molecular system
F_i	fragment $F_i \subseteq [M]$
F	fragmentation $F = \{F_i\}_{i=1}^K$ (usually a partition of $[M]$)
$F_{\mathbf{u}}$	combination of fragments, $F_{\mathbf{u}} = \bigcup_{i\in\mathbf{u}} F_i$
V^{BO}	Born-Oppenheimer potential energy function, $V^{\mathrm{BO}} : (\mathbb{R}^3 \times \mathbb{N})^M \to \mathbb{R}$
$V_{\mathbf{u}}$	nuclear subproblem potential for $\mathbf{u} \subseteq [M]$
$\tilde{V}_{\mathbf{u}}$	nuclear contribution potential for $\mathbf{u} \subseteq [M]$
$V_{F_{\mathbf{u}}}$	fragment subproblem potential for $\mathbf{u} \subseteq [M]$
$\tilde{V}_{F_{\mathbf{u}}}$	fragment contribution potential for $F_{\mathbf{u}} \subseteq [M]$

Graph theory

G	undirected graph, $G = (V, E)$
V	vertex set of a graph
E	undirected edge set of a graph, $E \subset V \times V$
u, v	general vertices $u, v \in V$
i, j	integer-valued vertices $i, j \in V = [N]$
u, v	vertex subsets, usually $\mathbf{u}, \mathbf{v} \subseteq V = [N]$
$G[\mathbf{u}]$	subgraph induced in G by some vertex subset $\mathbf{u} \subseteq V$
$\mathrm{conn}[G]$	set of all vertex subsets $\mathbf{u} \subseteq V$ that induce connected subgraphs $G[\mathbf{u}]$
$I_g(u, v)$	geodesic interval between two vertices $u, v \in V$
$I_g[S]$	geodesic closure of a set of vertices $S \subseteq V$
$\mathrm{CH}_g[S]$	geodesic convex hull of a set of vertices $S \subseteq V$
$\mathcal{M}_g[G]$	geodesic convexity of some connected graph G

Bibliography

[AF96] D. Avis and K. Fukuda. 'Reverse search for enumeration'. *Discrete Appl. Math.* 65(1-3), Mar. 1996, p. 21.

[AFK82] J. Almlöf, K. Faegri Jr. and K. Korsell. 'Principles for a Direct SCF Approach to LCAO-MO *Ab-Initio* Calculations'. *J. Comput. Chem.* 3(3), Sept. 1982, p. 385.

[Aig97] M. Aigner. *Combinatorial Theory*. Classics in Mathematics. Reprint of the 1979 ed. Springer, 1997.

[AK16] M. Albenque and K. Knauer. 'Convexity in partial cubes: The hull number'. *Discrete Math.* 339(2), Feb. 2016, p. 866.

[Aki+21] K. Akisawa, R. Hatada, K. Okuwaki, Y. Mochizuki, K. Fukuzawa, Y. Komeiji and S. Tanaka. 'Interaction analyses of SARS-CoV-2 spike protein based on fragment molecular orbital calculations'. *RSC Adv.* 11(6), 2021, p. 3272.

[An88] G. An. 'A note on the cluster variation method'. *J. Stat. Phys* 52(3-4), Aug. 1988, p. 727.

[Apr+20] E. Aprà, E. J. Bylaska, W. A. de Jong, N. Govind, K. Kowalski, T. P. Straatsma, M. Valiev, H. J. J. van Dam, Y. Alexeev, J. Anchell, V. Anisimov, F. W. Aquino, R. Atta-Fynn, J. Autschbach, N. P. Bauman, J. C. Becca, D. E. Bernholdt, K. Bhaskaran-Nair, S. Bogatko, P. Borowski, J. Boschen, J. Brabec, A. Bruner, E. Cauët, Y. Chen, G. N. Chuev, C. J. Cramer, J. Daily, M. J. O. Deegan, T. H. Dunning Jr., M. Dupuis, K. G. Dyall, G. I. Fann, S. A. Fischer, A. Fonari, H. Früchtl, L. Gagliardi, J. Garza, N. Gawande, S. Ghosh, K. Glaesemann, A. W. Götz, J. Hammond, V. Helms, E. D. Hermes, K. Hirao, S. Hirata, M. Jacquelin, L. Jensen, B. G. Johnson, H. Jónsson, R. A. Kendall, M. Klemm, R. Kobayashi, V. Konkov, S. Krishnamoorthy, M. Krishnan, Z. Lin, R. D. Lins, R. J. Littlefield, A. J. Logsdail, K. Lopata, W. Ma, A. V. Marenich, J. M. del Campo, D. Mejia-Rodriguez, J. E. Moore, J. M. Mullin, T. Nakajima, D. R. Nascimento, J. A. Nichols, P. J. Nichols, J. Nieplocha, A. O.-d. la-Roza, B. Palmer, A. Panyala, T. Pirojsirikul, B. Peng, R. Peverati, J. Pittner, L. Pollack, R. M. Richard, P. Sadayappan, G. C. Schatz, W. A. Shelton, D. W. Silverstein, D. M. A. Smith, T. A. Soares, D. Song, M. Swart, H. L. Taylor, G. S. Thomas, V. Tipparaju, D. G. Truhlar, K. Tsemekhman, T. V. Voorhis, Á. Vázquez-Mayagoitia, P. Verma, O. Villa, A. Vishnu, K. D. Vogiatzis, D. Wang, J. H. Weare, M. J. Williamson, T. L. Windus, K. Woliński, A. T. Wong, Q. Wu, C. Yang, Q. Yu, M. Zacharias, Z. Zhang, Y. Zhao and R. J. Harrison. 'NWChem: Past, present, and future'. *J. Chem. Phys.* 152(18), May 2020, p. 184102.

[AR96] X. Assfeld and J.-L. Rivail. 'Quantum chemical computations on parts of large molecules: the ab initio local self consistent field method'. *Chem. Phys. Lett.* 263(1-2), Dec. 1996, p. 100.

[Art+20] D. G. Artiukhin, E. L. Klinting, C. König and O. Christiansen. 'Adaptive density-guided approach to double incremental potential energy surface construction'. *J. Chem. Phys.* 152(19), May 2020, p. 194105.

[ASS13] R. T. Araújo, R. M. Sampaio and J. L. Szwarcfiter. 'The convexity of induced paths of order three'. *Electron. Notes Discrete Math.* 44, Nov. 2013, p. 109.

[ASW07] M. Aspnäs, A. Signell and J. Westerholm. 'Efficient Assembly of Sparse Matrices Using Hashing'. In: B. Kågström, E. Elmroth, J. Dongarra and J. Waśniewski, eds. *Applied Parallel Computing. State of the Art in Scientific Computing*. Lecture Notes in Computer Science, vol. 4699. Springer Berlin Heidelberg, 2007, p. 900.

[Bak+00] K. L. Bak, P. Jørgensen, J. Olsen, T. Helgaker and W. Klopper. 'Accuracy of atomization energies and reaction enthalpies in standard and extrapolated electronic wave function/basis set calculations'. *J. Chem. Phys.* 112(21), June 2000, p. 9229.

[BAM12] P. J. Bygrave, N. L. Allan and F. R. Manby. 'The embedded many-body expansion for energetics of molecular crystals'. *J. Chem. Phys.* 137(16), Oct. 2012, p. 164102.

[Ban+20] C. Bannwarth, E. Caldeweyher, S. Ehlert, A. Hansen, P. Pracht, J. Seibert, S. Spicher and S. Grimme. 'Extended tight-binding quantum chemistry methods'. *WIREs Comput. Mol. Sci.* 11(2), Aug. 2020.

[Bar+10] A. P. Bartók, M. C. Payne, R. Kondor and G. Csányi. 'Gaussian Approximation Potentials: The Accuracy of Quantum Mechanics, without the Electrons'. *Phys. Rev. Lett.* 104(13), Apr. 2010.

[Bar+17] J. Barker, J. Bulin, J. Hamaekers and S. Mathias. 'LC-GAP: Localized Coulomb Descriptors for the Gaussian Approximation Potential'. In: *Scientific Computing and Algorithms in Industrial Simulations: Projects and Products of Fraunhofer SCAI*. Ed. by M. Griebel, A. Schüller and M. A. Schweitzer. Springer International Publishing, Cham, 2017, p. 25.

[Bar+20] G. M. J. Barca, D. L. Poole, J. L. G. Vallejo, M. Alkan, C. Bertoni, A. P. Rendell and M. S. Gordon. 'Scaling the Hartree-Fock Matrix Build on Summit'. In: *SC20: International Conference for High Performance Computing, Networking, Storage and Analysis*. IEEE, Nov. 2020.

[Bar+21] J. Barker, L.-S. Berg, J. Hamaekers and A. Maass. 'Rapid Prescreening of Organic Compounds for Redox Flow Batteries: A Graph Convolutional Network for Predicting Reaction Enthalpies from SMILES'. *Batter. Supercaps* 4(9), June 2021, p. 1482.

[Bar09] J. Barker. 'Unlocking the PRISM-Cell: Accelerating Quantum Chemical Calculations using the Cell Broadband Engine'. Honours subthesis. Australian National University, Oct. 2009.

[Bar81] R. J. Bartlett. 'Many-Body Perturbation Theory and Coupled Cluster Theory for Electron Correlation in Molecules'. *Annu. Rev. Phys. Chem.* 32(1), Oct. 1981, p. 359.

[Bat+11] D. M. Bates, J. R. Smith, T. Janowski and G. S. Tschumper. 'Development of a 3-body:many-body integrated fragmentation method for weakly bound clusters and application to water clusters $(H_2O)_{n=3-10,16,17}$'. *J. Chem. Phys.* 135(4), July 2011, p. 044123.

[BBR13] M. Benzi, P. Boito and N. Razouk. 'Decay Properties of Spectral Projectors with Applications to Electronic Structure'. *SIAM Rev.* 55(1), Jan. 2013, p. 3.

[BC08] H.-J. Bandelt and V. Chepoi. 'Metric graph theory and geometry: a survey'. In: J. E. Goodman, J. Pach and R. Pollack, eds. *Surveys on Discrete and Computational Geometry: Twenty Years Later*. Vol. 453. Contemporary Mathematics. American Mathematical Society, 2008, p. 49.

[BCK16] N. Bock, M. Challacombe and L. V. Kalé. 'Solvers for $\mathcal{O}$(N) Electronic Structure in the Strong Scaling Limit'. *SIAM J. Sci. Comput.* 38(1), Jan. 2016, p. C1.

[BEG19] C. Bannwarth, S. Ehlert and S. Grimme. 'GFN2-xTB — An Accurate and Broadly Parametrized Self-Consistent Tight-Binding Quantum Chemical Method with Multipole Electrostatics and Density-Dependent Dispersion Contributions'. *J. Chem. Theory Comput.* 15(3), Feb. 2019, p. 1652.

[Ben72] C. F. Bender. 'Integral transformations. A bottleneck in molecular quantum mechanical calculations'. *J. Comput. Phys* 9(3), June 1972, p. 547.

[Ber09] G. J. O. Beran. 'Approximating quantum many-body intermolecular interactions in molecular clusters using classical polarizable force fields'. *J. Chem. Phys.* 130(16), Apr. 2009, p. 164115.

[BG04] H.-J. Bungartz and M. Griebel. 'Sparse grids'. *Acta Numer.* 13, May 2004, p. 147.

[BG70] E. Brändas and O. Goscinski. 'Variation-Perturbation Expansions and Padé Approximants to the Energy'. *Phys. Rev. A* 1(3), Mar. 1970, p. 552.

[BG75] E. A. Bender and J. R. Goldman. 'On the Applications of Mobius Inversion in Combinatorial Analysis'. *Am. Math. Mon.* 82(8), Oct. 1975, p. 789.

[BGR94] H. Bungartz, M. Griebel and U. Rüde. 'Extrapolation, combination, and sparse grid techniques for elliptic boundary value problems'. *Comput. Methods Appl. Mech. Eng.* 116(1-4), Jan. 1994, p. 243.

[BH71] D. F. Brailsford and G. G. Hall. 'Symmetry properties of one- and two-electron molecular integrals'. *Int. J. Quant. Chem.* 5(6), Nov. 1971, p. 657.

[Bjö+15] A. Björklund, T. Husfeldt, P. Kaski, M. Koivisto, J. Nederlof and P. Parviainen. 'Fast Zeta Transforms for Lattices with Few Irreducibles'. *ACM Trans. Algorithms* 12(1), Dec. 2015, p. 1.

[BK08] B. W. Bader and T. G. Kolda. 'Efficient MATLAB Computations with Sparse and Factored Tensors'. *SIAM J. Sci. Comput.* 30(1), Jan. 2008, p. 205.

[Blender] *Blender*. Version 3.0.0. URL: `https://www.blender.org`.

[BM07] R. J. Bartlett and M. Musiał. 'Coupled-cluster theory in quantum chemistry'. *Rev. Mod. Phys.* 79(1), Feb. 2007, p. 291.

[BM12] D. R. Bowler and T. Miyazaki. '$\mathcal{O}(N)$ methods in electronic structure calculations'. *Rep. Prog. Phys.* 75(3), Feb. 2012, p. 036503.

[BO13] J. Brown and O. R. Oellermann. 'Graphs with a Minimal Number of Convex Sets'. *Graphs Comb.* 30(6), Sept. 2013, p. 1383.

[BO15] J. I. Brown and O. R. Oellermann. 'On the spectrum and number of convex sets in graphs'. *Discrete Math.* 338(7), July 2015, p. 1144.

[BO27] M. Born and J. R. Oppenheimer. 'Zur Quantentheorie der Molekeln'. *Ann. Phys.* 84(20), 1927, p. 457.

[Boe+04] A. D. Boese, M. Oren, O. Atasoylu, J. M. L. Martin, M. Kállay and J. Gauss. 'W3 theory: Robust computational thermochemistry in the kJ/mol accuracy range'. *J. Chem. Phys.* 120(9), Mar. 2004, p. 4129.

[Bom+05] Y. J. Bomble, J. F. Stanton, M. Kállay and J. Gauss. 'Coupled-cluster methods including noniterative corrections for quadruple excitations'. *J. Chem. Phys.* 123(5), Aug. 2005, p. 054101.

[Bom+06] Y. J. Bomble, J. Vázquez, M. Kállay, C. Michauk, P. G. Szalay, A. G. Császár, J. Gauss and J. F. Stanton. 'High-accuracy extrapolated *ab initio* thermochemistry. II. Minor improvements to the protocol and a vital simplification'. *J. Chem. Phys.* 125(6), Aug. 2006, p. 064108.

[Boy50] S. F. Boys. 'Electronic wave functions - I. A general method of calculation for the stationary states of any molecular system'. *Proc. R. Soc. London Ser. A* 200(1063), Feb. 1950, p. 542.

[BP02] J. Baker and P. Pulay. 'An efficient parallel algorithm for the calculation of canonical MP2 energies'. *J. Comput. Chem.* 23(12), June 2002, p. 1150.

[BP06] N. B. Balabanov and K. A. Peterson. 'Basis set limit electronic excitation energies, ionization potentials, and electron affinities for the 3d transition metal atoms: Coupled cluster and multireference methods'. *J. Chem. Phys.* 125(7), Aug. 2006, p. 074110.

[BPH80] J. S. Binkley, J. A. Pople and W. J. Hehre. 'Self-consistent molecular orbital methods. 21. Small split-valence basis sets for first-row elements'. *J. Am. Chem. Soc.* 102(3), Jan. 1980, p. 939.

[BR04a] L. Bytautas and K. Ruedenberg. 'Correlation energy extrapolation by intrinsic scaling. I. Method and application to the neon atom'. *J. Chem. Phys.* 121(22), 2004, p. 10905.

[BR04b] L. Bytautas and K. Ruedenberg. 'Correlation energy extrapolation by intrinsic scaling. II. The water and the nitrogen molecule'. *J. Chem. Phys.* 121(22), 2004, p. 10919.

[BR05] L. Bytautas and K. Ruedenberg. 'Correlation energy extrapolation by intrinsic scaling. IV. Accurate binding energies of the homonuclear diatomic molecules carbon, nitrogen, oxygen, and fluorine'. *J. Chem. Phys.* 122(15), Apr. 2005, p. 154110.

[BR06] L. Bytautas and K. Ruedenberg. 'Correlation energy extrapolation by intrinsic scaling. V. Electronic energy, atomization energy, and enthalpy of formation of water'. *J. Chem. Phys.* 124(17), May 2006, p. 174304.

[BS17] N. Benedikter and J. Sok. *Advanced Mathematical Physics — The Hartree-Fock model*. Lecture notes, University of Copenhagen (`http://nielsbenedikter.de/teaching.html`). 2017. URL: `http://www.nielsbenedikter.de/advmaphys2/hartree-fock.pdf`.

[BS77] R. J. Bartlett and I. Shavitt. 'Comparison of high-order many-body perturbation theory and configuration interaction for H_2O'. *Chem. Phys. Lett.* 50(2), Sept. 1977, p. 190.

[BT96] D. Bakowies and W. Thiel. 'Hybrid Models for Combined Quantum Mechanical and Molecular Mechanical Approaches'. *J. Phys. Chem.* 100(25), Jan. 1996, p. 10580.

[BTZ22] J. Barker, C. Thiele and P. Zorin-Kranich. 'Band-Limited Maximizers for a Fourier Extension Inequality on the Circle, II'. *Exp. Math.* 32(2), Nov. 2022, p. 280.

[Bun+94] H.-J. Bungartz, M. Griebel, D. Röschke and C. Zenger. 'Pointwise convergence of the combination technique for the Laplace equation'. *East-West J. Numer. Math.* 2, 1994. Also as SFB-Bericht 342/16/93A, Institut für Informatik, TU München, 1993, p. 21.

[Can+03] E. Cancès, M. Defranceschi, W. Kutzelnigg, C. Le Bris and Y. Maday. 'Computational quantum chemistry: A primer'. In: *Special Volume, Computational Chemistry*. Vol. 10. Handbook of Numerical Analysis. Elsevier, 2003, p. 3.

[CB15] M. A. Collins and R. P. A. Bettens. 'Energy-Based Molecular Fragmentation Methods'. *Chem. Rev.* 115(12), Apr. 2015, p. 5607.

[CCB14] M. A. Collins, M. W. Cvitkovic and R. P. A. Bettens. 'The Combined Fragmentation and Systematic Molecular Fragmentation Methods'. *Acc. Chem. Res.* 47(9), June 2014, p. 2776.

[CD06] M. A. Collins and V. A. Deev. 'Accuracy and efficiency of electronic energies from systematic molecular fragmentation'. *J. Chem. Phys.* 125(10), Sept. 2006, p. 104104.

[CDC19] M. Chaveroche, F. Davoine and V. Cherfaoui. 'Efficient Möbius Transformations and Their Applications to D-S Theory'. In: N. Ben Amor, B. Quost and M. Theobald, eds. *Scalable Uncertainty Management. SUM 2019*. Vol. 11940. Lecture Notes in Computer Science. Springer International Publishing, 2019, p. 390.

[CDC21] M. Chaveroche, F. Davoine and V. Cherfaoui. 'Focal points and their implications for Möbius transforms and Dempster-Shafer Theory'. *Inf. Sci.* 555, May 2021, p. 215.

[CGH18] S. R. Chinnamsetty, M. Griebel and J. Hamaekers. 'An Adaptive Multiscale Approach for Electronic Structure Methods'. *Multiscale Model. Sim.* 16(2), Jan. 2018, p. 752.

[Cha21] M. Chaveroche. 'Efficient decentralized collaborative perception for autonomous vehicles'. PhD thesis. Université de Technologie Compiègne, 21 Sept. 2021.

[Che+17] G. D. Chen, J. Weng, G. Song and Z. H. Li. 'Generalized Switch Functions in the Multilevel Many-Body Expansion Method and Its Application to Water Clusters'. *J. Chem. Theory Comput.* 13(5), Apr. 2017, p. 2010.

[Chu+15] L. W. Chung, W. M. C. Sameera, R. Ramozzi, A. J. Page, M. Hatanaka, G. P. Petrova, T. V. Harris, X. Li, Z. Ke, F. Liu, H.-B. Li, L. Ding and K. Morokuma. 'The ONIOM Method and Its Applications'. *Chem. Rev.* 115(12), Apr. 2015, p. 5678.

[Cis+16] G. A. Cisneros, K. T. Wikfeldt, L. Ojamäe, J. Lu, Y. Xu, H. Torabifard, A. P. Bartók, G. Csányi, V. Molinero and F. Paesani. 'Modeling Molecular Interactions in Water: From Pairwise to Many-Body Potential Energy Functions'. *Chem. Rev.* 116(13), May 2016, p. 7501.

[Číž66] J. Čížek. 'On the Correlation Problem in Atomic and Molecular Systems. Calculation of Wavefunction Components in Ursell-Type Expansion Using Quantum-Field Theoretical Methods'. *J. Chem. Phys.* 45(11), Dec. 1966, p. 4256.

[CL00] E. Cancès and C. Le Bris. 'On the convergence of SCF algorithms for the Hartree-Fock equations'. *ESAIM: Math. Model. Numer. Anal.* 34(4), July 2000, p. 749.

[CLJ06] J. Cui, H. Liu and K. D. Jordan. 'Theoretical Characterization of the $(H_2O)_{21}$ Cluster: Application of an n-body Decomposition Procedure'. *J. Phys. Chem. B* 110(38), Mar. 2006, p. 18872.

[CMO97] R. Caflisch, W. Morokoff and A. Owen. 'Valuation of mortgage-backed securities using Brownian bridges to reduce effective dimension'. *J. Comput. Finance* 1(1), 1997, p. 27.

[CMS05] M. Changat, H. M. Mulder and G. Sierksma. 'Convexities related to path properties on graphs'. *Discrete Math.* 290(2-3), Feb. 2005, p. 117.

[Col12] M. A. Collins. 'Systematic fragmentation of large molecules by annihilation'. *Phys. Chem. Chem. Phys.* 14(21), 2012, p. 7744.

[Cor+08] B. Cordero, V. Gómez, A. E. Platero-Prats, M. Revés, J. Echeverría, E. Cremades, F. Barragán and S. Alvarez. 'Covalent radii revisited'. *Dalton Trans.*, 21 2008, p. 2832.

[Cor+22] T. H. Cormen, C. E. Leiserson, R. L. Rivest and C. Stein. *Introduction to Algorithms*. 4th ed. The MIT Press, 5 Apr. 2022.

[Cre11] D. Cremer. 'Møller-Plesset perturbation theory: from small molecule methods to methods for thousands of atoms'. *WIREs Comput. Mol. Sci.* 1(4), May 2011, p. 509.

[CRR05] L. A. Curtiss, P. C. Redfern and K. Raghavachari. 'Assessment of Gaussian-3 and density-functional theories on the G3/05 test set of experimental energies'. *J. Chem. Phys.* 123(12), Sept. 2005, p. 124107.

[CRR07a] L. A. Curtiss, P. C. Redfern and K. Raghavachari. 'Gaussian-4 theory'. *J. Chem. Phys.* 126(8), Feb. 2007, p. 084108.

[CRR07b] L. A. Curtiss, P. C. Redfern and K. Raghavachari. 'Gaussian-4 theory using reduced order perturbation theory'. *J. Chem. Phys.* 127(12), Nov. 2007, p. 124105.

[CS00] T. D. Crawford and H. F. Schaefer III. 'An Introduction to Coupled Cluster Theory for Computational Chemists'. In: K. B. Lipkowitz and D. B. Boyd, eds. *Reviews in Computational Chemistry*. Vol. 14. John Wiley & Sons, Inc., Jan. 2000, p. 33.

[CS97] M. Challacombe and E. Schwegler. 'Linear scaling computation of the Fock matrix'. *J. Chem. Phys.* 106(13), Apr. 1997, p. 5526.

[CSBenz] ChemSpider. *Benzene (CSID: 236)*. Royal Society of Chemistry. URL: `http://www.chemspider.com/Chemical-Structure.236.html`. Precise date of original access unknown. Referenced data current as of 29/8/2023.

[CSHept] ChemSpider. *Heptane (CSID: 8560)*. Royal Society of Chemistry. URL: `https://www.chemspider.com/Chemical-Structure.8560.html`. Precise date of original access unknown. Referenced data current as of 4/11/2022.

[CSHex] ChemSpider. *Hexane (CSID: 7767)*. Royal Society of Chemistry. URL: `http://www.chemspider.com/Chemical-Structure.7767.html`. Precise date of original access unknown. Referenced data current as of 29/8/2023.

[CSLimo] ChemSpider. *Limonin (CSID: 156367)*. Royal Society of Chemistry. URL: `http://www.chemspider.com/Chemical-Structure.156367.html`. Accessed on or approximately around 5/9/2021.

[CSPhen] ChemSpider. *Phenalene (CSID:8795)*. Royal Society of Chemistry. URL: `https://www.chemspider.com/Chemical-Structure.8795.html`. Precise date of original access unknown. Referenced data current as of 9/12/2022.

[Cur+01] L. A. Curtiss, P. C. Redfern, K. Raghavachari and J. A. Pople. 'Gaussian-3X (G3X) theory: Use of improved geometries, zero-point energies, and Hartree-Fock basis sets'. *J. Chem. Phys.* 114(1), 2001, p. 108.

[Cur+90] L. A. Curtiss, C. Jones, G. W. Trucks, K. Raghavachari and J. A. Pople. 'Gaussian-1 theory of molecular energies for second-row compounds'. *J. Chem. Phys.* 93(4), Aug. 1990, p. 2537.

[Cur+91] L. A. Curtiss, K. Raghavachari, G. W. Trucks and J. A. Pople. 'Gaussian-2 theory for molecular energies of first- and second-row compounds'. *J. Chem. Phys.* 94(11), June 1991, p. 7221.

[Cur+97] L. A. Curtiss, K. Raghavachari, P. C. Redfern and J. A. Pople. 'Assessment of Gaussian-2 and density functional theories for the computation of enthalpies of formation'. *J. Chem. Phys.* 106(3), Jan. 1997, p. 1063.

[Cur+98] L. A. Curtiss, K. Raghavachari, P. C. Redfern, V. Rassolov and J. A. Pople. 'Gaussian-3 (G3) theory for molecules containing first and second-row atoms'. *J. Chem. Phys.* 109(18), Nov. 1998, p. 7764.

[Cur+99a] L. A. Curtiss, K. Raghavachari, P. C. Redfern, A. G. Baboul and J. A. Pople. 'Gaussian-3 theory using coupled cluster energies'. *Chem. Phys. Lett.* 314(1-2), Nov. 1999, p. 101.

[Cur+99b] L. A. Curtiss, P. C. Redfern, K. Raghavachari, V. Rassolov and J. A. Pople. 'Gaussian-3 theory using reduced Møller-Plesset order'. *J. Chem. Phys.* 110(10), Mar. 1999, p. 4703.

[CZ99] G. Chartrand and P. Zhang. 'The forcing geodetic number of a graph'. *Discuss. Math. Graph Theory* 19(1), 1999, p. 45.

[Dap+99] S. Dapprich, I. Komáromi, K. S. Byun, K. Morokuma and M. J. Frisch. 'A new ONIOM implementation in Gaussian98. Part I. The calculation of energies, gradients, vibrational frequencies and electric field derivatives'. *J. Mol. Struct. THEOCHEM* 461-462, Apr. 1999, p. 1.

[Das+02] D. Das, K. P. Eurenius, E. M. Billings, P. Sherwood, D. C. Chatfield, M. Hodošček and B. R. Brooks. 'Optimization of quantum mechanical molecular mechanical partitioning schemes: Gaussian delocalization of molecular mechanical charges and the double link atom method'. *J. Chem. Phys.* 117(23), Dec. 2002, p. 10534.

[Dav75] E. R. Davidson. 'The iterative calculation of a few of the lowest eigenvalues and corresponding eigenvectors of large real-symmetric matrices'. *J. Comput. Phys* 17(1), Jan. 1975, p. 87.

[DC05] V. Deev and M. A. Collins. 'Approximate *ab initio* energies by systematic molecular fragmentation'. *J. Chem. Phys.* 122(15), Apr. 2005, p. 154102.

[DCR21] S. K. Das, S. Chakraborty and R. Ramakrishnan. 'Critical benchmarking of popular composite thermochemistry models and density functional approximations on a probabilistically pruned benchmark dataset of formation enthalpies'. *J. Chem. Phys.* 154(4), Jan. 2021, p. 044113.

[DCW06] N. J. DeYonker, T. R. Cundari and A. K. Wilson. 'The correlation consistent composite approach (ccCA): An alternative to the Gaussian-n methods'. *J. Chem. Phys.* 124(11), Mar. 2006, p. 114104.

[Del82] F.-J. Delvos. 'd-Variate Boolean interpolation'. *J. Approximation Theory* 34(2), Feb. 1982, p. 99.

[DeY+09] N. J. DeYonker, B. R. Wilson, A. W. Pierpont, T. R. Cundari and A. K. Wilson. 'Towards the intrinsic error of the correlation consistent Composite Approach (ccCA)'. *Mol. Phys.* 107(8-12), Apr. 2009, p. 1107.

[DFS04] R. Drautz, M. Fähnle and J. M. Sanchez. 'General relations between many-body potentials and cluster expansions in multicomponent systems'. *J. Phys.: Condens. Matter* 16(23), May 2004, p. 3843.

[DH77] T. H. Dunning Jr. and P. J. Hay. 'Gaussian Basis Sets for Molecular Calculations'. In: *Methods of Electronic Structure Theory*. Springer US, 1977, p. 1.

[DHP71] R. Ditchfield, W. J. Hehre and J. A. Pople. 'Self-Consistent Molecular-Orbital Methods. IX. An Extended Gaussian-Type Basis for Molecular-Orbital Studies of Organic Molecules'. *J. Chem. Phys.* 54(2), Jan. 1971, p. 724.

[Dia+13] E. S. Dias, D. Castonguay, H. Longo and W. A. R. Jradi. 'Efficient Enumeration of Chordless Cycles', 4 Sept. 2013. arXiv: `1309.1051 [cs.DS]`.

[Die17] R. Diestel. *Graph Theory*. Springer Berlin, Heidelberg, 2017.

[Dom74] C. Domb. 'Graph Theory and Embeddings'. In: *Series Expansions for Lattice Models*. Ed. by C. Domb and M. S. Green. Vol. 3. Phase Transitions and Critical Phenomena. Academic Press, London, New York, 1974, p. 1.

[Dou+09] M. C. Dourado, J. G. Gimbel, J. Kratochvíl, F. Protti and J. L. Szwarcfiter. 'On the computation of the hull number of a graph'. *Discrete Math.* 309(18), Sept. 2009, p. 5668.

[DS16] M. C. Dourado and R. M. Sampaio. 'Complexity aspects of the triangle path convexity'. *Discrete Appl. Math.* 206, June 2016, p. 39.

[DT06] E. E. Dahlke and D. G. Truhlar. 'Electrostatically Embedded Many-Body Expansion for Large Systems, with Applications to Water Clusters'. *J. Chem. Theory Comput.* 3(1), Nov. 2006, p. 46.

[DT07a] E. E. Dahlke and D. G. Truhlar. 'Electrostatically Embedded Many-Body Correlation Energy, with Applications to the Calculation of Accurate Second-Order Møller-Plesset Perturbation Theory Energies for Large Water Clusters'. *J. Chem. Theory Comput.* 3(4), June 2007, p. 1342.

[DT07b] E. E. Dahlke and D. G. Truhlar. 'Electrostatically Embedded Many-Body Expansion for Simulations'. *J. Chem. Theory Comput.* 4(1), Dec. 2007, p. 1.

[Duc88] P. Duchet. 'Convex sets in graphs, II. Minimal path convexity'. *J. Comb. Theory Ser. B* 44(3), June 1988, p. 307.

[Dun89] T. H. Dunning Jr. 'Gaussian basis sets for use in correlated molecular calculations. I. The atoms boron through neon and hydrogen'. *J. Chem. Phys.* 90(2), Jan. 1989, p. 1007.

[EA07] P. Echenique and J. L. Alonso. 'A mathematical and computational review of Hartree-Fock SCF methods in quantum chemistry'. *Mol. Phys.* 105(23-24), Dec. 2007, p. 3057.

[EH11] A. Engels-Putzka and M. Hanrath. 'A fully simultaneously optimizing genetic approach to the highly excited coupled-cluster factorization problem'. *J. Chem. Phys.* 134(12), Mar. 2011, p. 124106.

[EJ85] P. Edelman and R. Jamison. 'The theory of convex geometries'. *Geom. Dedicata* 19(3), Dec. 1985.

[Eps+92] D. B. A. Epstein, J. W. Cannon, D. F. Holt, S. V. F. Levy, M. S. Paterson and W. P. Thurston. *Word Processing in Groups*. Jones and Bartlett Publishers, 1992.

[Ern16] M. B. Ernst. *composite-thermochemistry-nwchem*. GitHub. 3 June 2016. URL: `https://github.com/mattbernst/composite-thermochemistry-nwchem`.

[ES85] M. G. Everett and S. B. Seidman. 'The hull number of a graph'. *Discrete Math.* 57(3), Dec. 1985, p. 217.

[Ess+77] J. W. Essam, J. W. Kennedy, M. Gordon and P. Whittle. 'The graph-like state of matter. Part 8.—LCGI schemes and the statistical analysis of experimental data'. *J. Chem. Soc., Faraday Trans. 2* 73(7), 1977, p. 1289.

[FBK90] M. J. Field, P. A. Bash and M. Karplus. 'A combined quantum mechanical and molecular mechanical potential for molecular dynamics simulations'. *J. Comput. Chem.* 11(6), July 1990, p. 700.

[FD03] D. Feller and D. A. Dixon. 'Coupled Cluster Theory and Multireference Configuration Interaction Study of FO, F_2O, FO_2, and FOOF'. *J. Phys. Chem. A* 107(45), Oct. 2003, p. 9641.

[FD18] D. Feller and E. R. Davidson. 'A theoretical study of the adiabatic and vertical ionization potentials of water'. *J. Chem. Phys.* 148(23), June 2018, p. 234308.

[Fed+14] D. G. Fedorov, N. Asada, I. Nakanishi and K. Kitaura. 'The Use of Many-Body Expansions and Geometry Optimizations in Fragment-Based Methods'. *Acc. Chem. Res.* 47(9), Aug. 2014, p. 2846.

[Fel+03] D. Feller, K. A. Peterson, W. A. de Jong and D. A. Dixon. 'Performance of coupled cluster theory in thermochemical calculations of small halogenated compounds'. *J. Chem. Phys.* 118(8), Feb. 2003, p. 3510.

[Fel13] D. Feller. 'Benchmarks of improved complete basis set extrapolation schemes designed for standard CCSD(T) atomization energies'. *J. Chem. Phys.* 138(7), Feb. 2013, p. 074103.

[Fel22] D. Feller. Personal communication. 2022.

[Fel92] D. Feller. 'Application of systematic sequences of wave functions to the water dimer'. *J. Chem. Phys.* 96(8), Apr. 1992, p. 6104.

[Fel93] D. Feller. 'The use of systematic sequences of wave functions for estimating the complete basis set, full configuration interaction limit in water'. *J. Chem. Phys.* 98(9), May 1993, p. 7059.

[Feu10] C. Feuersänger. 'Sparse Grid Methods for Higher Dimensional Approximation'. Dissertation. Rheinische Friedrich-Wilhelms-Universität Bonn, Sept. 2010.

[Fil13] Y. Filmus. *Algorithm for ranking members of a regular language?* Theoretical Computer Science Stack Exchange. Version of 2013-08-09. The author's profile can be found at `https://cstheory.stackexchange.com/users/40/yuval-filmus`. 9 Aug. 2013. URL: `https://cstheory.stackexchange.com/q/18590`.

[Fis18] K. Fischhuber. 'An adaptive Sparse Grid Approach for Many-Body Systems'. Master's thesis. Rheinische Friedrich-Wilhelms-Universität Bonn, 2018.

[Fis64] M. E. Fisher. 'The free energy of a macroscopic system'. *Arch. Ration. Mech. Anal.* 17(5), Jan. 1964, p. 377.

[FJ86] M. Farber and R. E. Jamison. 'Convexity in Graphs and Hypergraphs'. *SIAM J. Algebraic Discrete Methods* 7(3), July 1986, p. 433.

[FJ87] M. Farber and R. E. Jamison. 'On local convexity in graphs'. *Discrete Math.* 66(3), Sept. 1987, p. 231.

[Flo09] N. Flocke. 'On the use of shifted Jacobi polynomials in accurate evaluation of roots and weights of Rys polynomials'. *J. Chem. Phys.* 131(6), 2009, p. 064107.

[FNK12] D. G. Fedorov, T. Nagata and K. Kitaura. 'Exploring chemistry with the fragment molecular orbital method'. *Phys. Chem. Chem. Phys.* 14(21), 2012, p. 7562.

[Fog+12] U. R. Fogueri, S. Kozuch, A. Karton and J. M. L. Martin. 'A simple DFT-based diagnostic for nondynamical correlation'. *Theor. Chim. Acta.* 132(1), Dec. 2012.

[For+15] M. E. Fornace, J. Lee, K. Miyamoto, F. R. Manby and T. F. Miller. 'Embedded Mean-Field Theory'. *J. Chem. Theory Comput.* 11(2), Jan. 2015, p. 568.

[Fos+96] I. T. Foster, J. L. Tilson, A. F. Wagner, R. L. Shepard, R. J. Harrison, R. A. Kendall and R. J. Littlefield. 'Toward high-performance computational chemistry: I. Scalable Fock matrix construction algorithms'. *J. Comput. Chem.* 17(1), Jan. 1996, p. 109.

[FP07] D. Feller and K. A. Peterson. 'Probing the limits of accuracy in electronic structure calculations: Is theory capable of results uniformly better than "chemical accuracy"?' *J. Chem. Phys.* 126(11), Mar. 2007.

[FP09] D. Feller and K. A. Peterson. 'High level coupled cluster determination of the structure, frequencies, and heat of formation of water'. *J. Chem. Phys.* 131(15), Oct. 2009, p. 154306.

[FP99] D. Feller and K. A. Peterson. 'Re-examination of atomization energies for the Gaussian-2 set of molecules'. *J. Chem. Phys.* 110(17), May 1999, p. 8384.

[FPC06] D. Feller, K. A. Peterson and T. D. Crawford. 'Sources of error in electronic structure calculations on small chemical systems'. *J. Chem. Phys.* 124(5), Feb. 2006, p. 054107.

[FPD08] D. Feller, K. A. Peterson and D. A. Dixon. 'A survey of factors contributing to accurate theoretical predictions of atomization energies and molecular structures'. *J. Chem. Phys.* 129(20), Nov. 2008, p. 204105.

[FPH10] D. Feller, K. A. Peterson and J. G. Hill. 'Calibration study of the CCSD(T)-F12a/b methods for C_2 and small hydrocarbons'. *J. Chem. Phys.* 133(18), Nov. 2010, p. 184102.

[FPH11] D. Feller, K. A. Peterson and J. G. Hill. 'On the effectiveness of CCSD(T) complete basis set extrapolations for atomization energies'. *J. Chem. Phys.* 135(4), July 2011, p. 044102.

[Fri03] G. Friesecke. 'The Multiconfiguration Equations for Atoms and Molecules: Charge Quantization and Existence of Solutions'. *Arch. Ration. Mech. Anal.* 169(1), Aug. 2003, p. 35.

[FS00] D. Feller and J. A. Sordo. 'Performance of CCSDT for diatomic dissociation energies'. *J. Chem. Phys.* 113(2), July 2000, p. 485.

[FT13] D. Fourches and A. Tropsha. 'Using Graph Indices for the Analysis and Comparison of Chemical Datasets'. *Mol. Inf.* 32(9-10), Sept. 2013, p. 827.

[Gan+06] V. Ganesh, R. K. Dongare, P. Balanarayan and S. R. Gadre. 'Molecular tailoring approach for geometry optimization of large molecules: Energy evaluation and parallelization strategies'. *J. Chem. Phys.* 125(10), Sept. 2006, p. 104109.

[Gao+98] J. Gao, P. Amara, C. Alhambra and M. J. Field. 'A Generalized Hybrid Orbital (GHO) Method for the Treatment of Boundary Atoms in Combined QM/MM Calculations'. *J. Phys. Chem. A* 102(24), May 1998, p. 4714.

[Gar06] J. Garcke. 'Regression with the Optimised Combination Technique'. In: *Proceedings of the 23rd International Conference on Machine Learning*. ICML '06. Association for Computing Machinery, Pittsburgh, Pennsylvania, USA, 2006, p. 321.

[Gar07a] J. Garcke. 'A dimension adaptive sparse grid combination technique for machine learning'. *ANZIAM J.* 48, Dec. 2007, p. 725.

[Gar07b] J. Garcke. 'An optimised sparse grid combination technique for eigenproblems'. *PAMM* 7(1), Dec. 2007, p. 1022301.

[Gar12a] J. Garcke. 'A Dimension Adaptive Combination Technique Using Localised Adaptation Criteria'. In: H. G. Bock, X. P. Hoang, R. Rannacher and J. P. Schlöder, eds. *Modeling, Simulation and Optimization of Complex Processes*. Springer Berlin Heidelberg, Berlin, Heidelberg, 2012, p. 115.

[Gar12b] J. Garcke. 'Sparse grids in a Nutshell'. In: J. Garcke and M. Griebel, eds. *Sparse Grids and Applications*. Vol. 88. Lecture Notes in Computational Science and Engineering. Springer Berlin Heidelberg, Berlin, Heidelberg, 29 Aug. 2012, p. 57.

[Gar98] J. Garcke. 'Berechnung von Eigenwerten der stationären Schrödingergleichung mit der Kombinationstechnik'. Diplomarbeit. Rheinische Friedrich-Wilhelms-Universität Bonn, 1998.

[GG00] J. Garcke and M. Griebel. 'On the Computation of the Eigenproblems of Hydrogen and Helium in Strong Magnetic and Electric Fields with the Sparse Grid Combination Technique'. *J. Comput. Phys* 165(2), Dec. 2000, p. 694.

[GG03] T. Gerstner and M. Griebel. 'Dimension-adaptive tensor-product quadrature'. *Computing* 71(1), Aug. 2003, p. 65.

[GG98] T. Gerstner and M. Griebel. 'Numerical integration using sparse grids'. *Numerical Algorithms* 18(3/4), 1998, p. 209.

[GGG20] J. Gasteiger, J. Gross and S. Günnemann. 'Directional Message Passing for Molecular Graphs', 6 Mar. 2020. arXiv: 2003.03123 [cs.LG].

[GH07] M. Griebel and J. Hamaekers. 'Sparse grids for the Schrödinger equation'. *ESAIM: Math. Model. Numer. Anal.* 41(2), Mar. 2007, p. 215.

[GH14] M. Griebel and H. Harbrecht. 'On the Convergence of the Combination Technique'. In: J. Garcke and D. Pflüger, eds. *Sparse Grids and Applications — Munich 2012*. Vol. 97. Lecture Notes in Computational Science and Engineering. Springer International Publishing, Cham, 2014, p. 55.

[GHH08] M. Griebel, J. Hamaekers and F. Heber. *BOSSANOVA: A bond order dissection approach for efficient electronic structure calculations*. INS Preprint 0704. Institut für Numerische Simulation, Universität Bonn, 2008. URL: http://wissrech.ins.uni-bonn.de/research/pub/hamaekers/INSPreprint0704GriebelHamaekersHeber.pdf.

[GHH14] M. Griebel, J. Hamaekers and F. Heber. 'A Bond Order Dissection ANOVA Approach for Efficient Electronic Structure Calculations'. In: S. Dahlke, W. Dahmen, M. Griebel, W. Hackbusch, K. Ritter, R. Schneider, C. Schwab and H. Yserentant, eds. *Extraction of Quantifiable Information from Complex Systems*. Vol. 102. Lecture Notes in Computational Science and Engineering. Springer International Publishing, Cham, 2014, p. 211.

[Gil94] P. M. W. Gill. 'Molecular integrals Over Gaussian Basis Functions'. In: J. R. Sabin and M. C. Zerner, eds. *Advances in Quantum Chemistry*. Academic Press, 1994, p. 141.

[GK73] M. Gordon and J. W. Kennedy. 'The graph-like state of matter. Part 2. — LCGI schemes for the thermodynamics of alkanes and the theory of inductive inference'. *J. Chem. Soc., Faraday Trans. 2* 69, 1973, p. 484.

[GKC17] Á. Ganyecz, M. Kállay and J. Csontos. 'Moderate-Cost *Ab Initio* Thermochemistry with Chemical Accuracy'. *J. Chem. Theory Comput.* 13(9), Aug. 2017, p. 4193.

[GKN20] L. Gyevi-Nagy, M. Kállay and P. R. Nagy. 'Integral-Direct and Parallel Implementation of the CCSD(T) Method: Algorithmic Developments and Large-Scale Applications'. *J. Chem. Theory Comput.* 16(1), 2020, p. 366.

[GKZ07] M. Griebel, S. Knapek and G. Zumbusch. *Numerical Simulation in Molecular Dynamics*. Texts in Computational Science and Engineering. Springer Berlin Heidelberg, 2007.

[GL78] A. George and J. W. H. Liu. 'Algorithms for Matrix Partitioning and the Numerical Solution of Finite Element Systems'. *SIAM J. Numer. Anal.* 15(2), Apr. 1978, p. 297.

[GO10] W. Goddard and O. R. Oellermann. 'Distance in Graphs'. In: *Structural Analysis of Complex Networks*. Birkhäuser, Boston, Sept. 2010, p. 49.

[God18] C. Godsil. 'An Introduction to the Moebius Function', 18 Mar. 2018. arXiv: `1803.06664 [math.CO]`.

[Goe+17] L. Goerigk, A. Hansen, C. Bauer, S. Ehrlich, A. Najibi and S. Grimme. 'A look at the density functional theory zoo with the advanced GMTKN55 database for general main group thermochemistry, kinetics and noncovalent interactions'. *Phys. Chem. Chem. Phys.* 19, 48 2017, p. 32184.

[Goe99] S. Goedecker. 'Linear scaling electronic structure methods'. *Rev. Mod. Phys.* 71(4), July 1999, p. 1085.

[Goo02] D. Z. Goodson. 'Extrapolating the coupled-cluster sequence toward the full configuration-interaction limit'. *J. Chem. Phys.* 116(16), Apr. 2002, p. 6948.

[Gor+11] M. S. Gordon, D. G. Fedorov, S. R. Pruitt and L. V. Slipchenko. 'Fragmentation Methods: A Route to Accurate Calculations on Large Systems'. *Chem. Rev.* 112(1), Aug. 2011, p. 632.

[Gra69] R. L. Graham. 'Bounds on Multiprocessing Timing Anomalies'. *SIAM J. Appl. Math.* 17(2), Mar. 1969, p. 416.

[Gre82] C. Greene. 'The Möbius Function of a Partially Ordered Set'. In: I. Rival, ed. *Ordered Sets*. Vol. 83. NATO Advanced Study Institutes Series. Springer Netherlands, Dordrecht, 1982, p. 555.

[Gri06] M. Griebel. 'Sparse grids and related approximation schemes for higher dimensional problems'. In: L. Pardo, A. Pinkus, E. Suli and M. J. Todd, eds. *Foundations of Computational Mathematics (FoCM05), Santander*. Cambridge University Press, 2006, p. 106.

[Gri19] M. Griebel. Personal communication. 2019.

[Gri98] M. Griebel. 'Adaptive Sparse Grid Multilevel Methods for elliptic PDEs based on finite differences'. *Computing* 61(2), 1998, p. 151.

[GSZ92] M. Griebel, M. Schneider and C. Zenger. 'A combination technique for the solution of sparse grid problems'. In: P. de Groen and R. Beauwens, eds. *Iterative Methods in Linear Algebra*. Also as SFB Bericht, 342/19/90 A, Institut für Informatik, TU München, 1990. IMACS, Elsevier, North Holland, 1992, p. 263.

[GT95] M. Griebel and V. Thurner. 'The efficient solution of fluid dynamics problems by the combination technique'. *Int. J. Numer. Methods Heat Fluid Flow* 5(3), Mar. 1995, p. 251.

[Gue+00] C. F. Guerra, F. M. Bickelhaupt, J. G. Snijders and E. J. Baerends. 'Hydrogen Bonding in DNA Base Pairs: Reconciliation of Theory and Experiment'. *J. Am. Chem. Soc.* 122(17), Apr. 2000, p. 4117.

[GW12] J. Gao and Y. Wang. 'Communication: Variational many-body expansion: Accounting for exchange repulsion, charge delocalization, and dispersion in the fragment-based explicit polarization method'. *J. Chem. Phys.* 136(7), Feb. 2012.

[GWC99] N. Govind, Y. A. Wang and E. A. Carter. 'Electronic-structure calculations by first-principles density-based embedding of explicitly correlated systems'. *J. Chem. Phys.* 110(16), Apr. 1999, p. 7677.

[HA89] M. Häser and R. Ahlrichs. 'Improvements on the direct SCF method'. *J. Comput. Chem.* 10(1), Jan. 1989, p. 104.

[Hal+03] K. Hald, A. Halkier, P. Jørgensen, S. Coriani, C. Hättig and T. Helgaker. 'A Lagrangian, integral-density direct formulation and implementation of the analytic CCSD and CCSD(T) gradients'. *J. Chem. Phys.* 118(7), Feb. 2003, p. 2985.

[Hal+99] A. Halkier, T. Helgaker, P. Jørgensen, W. Klopper and J. Olsen. 'Basis-set convergence of the energy in molecular Hartree-Fock calculations'. *Chem. Phys. Lett.* 302(5-6), Mar. 1999, p. 437.

[Hal34] P. Hall. 'A Contribution to the Theory of Groups of Prime-Power Order'. *Proc. London Math. Soc.* s2-36(1), 1934, p. 29.

[Hal36] P. Hall. 'The Eulerian functions of a group'. *The Quarterly Journal of Mathematics* os-7(1), 1936, p. 134.

[Hal51] G. G. Hall. 'The molecular orbital theory of chemical valency VIII. A method of calculating ionization potentials'. *Proc. R. Soc. London Ser. A* 205(1083), Mar. 1951, p. 541.

[Ham09] J. Hamaekers. 'Tensor Product Multiscale Many-Particle Spaces with Finite-Order Weights for the Electronic Schrödinger Equation'. Dissertation. Rheinische Friedrich-Wilhelms-Universität Bonn, July 2009.

[Han20] J. Hansen. *Four Different Methods for Making Cel-Shaders in Blender Eevee (2.8, 2.9+)*. 4 Oct. 2020. URL: https://medium.com/@josephclaytonhansen/four-different-methods-for-making-cel-shaders-in-blender-eevee-2-8-2-9-6d976ce2555d (visited on 22/10/2022).

[Har+08] M. E. Harding, J. Vázquez, B. Ruscic, A. K. Wilson, J. Gauss and J. F. Stanton. 'High-accuracy extrapolated *ab initio* thermochemistry. III. Additional improvements and overview'. *J. Chem. Phys.* 128(11), Mar. 2008, p. 114111.

[Har+20] C. R. Harris, K. J. Millman, S. J. van der Walt, R. Gommers, P. Virtanen, D. Cournapeau, E. Wieser, J. Taylor, S. Berg, N. J. Smith, R. Kern, M. Picus, S. Hoyer, M. H. van Kerkwijk, M. Brett, A. Haldane, J. F. del Río, M. Wiebe, P. Peterson, P. Gérard-Marchant, K. Sheppard, T. Reddy, W. Weckesser, H. Abbasi, C. Gohlke and T. E. Oliphant. 'Array programming with NumPy'. *Nature* 585(7825), Sept. 2020, p. 357.

[Har16a] B. Harding. 'Adaptive Sparse Grids and Extrapolation Techniques'. In: J. Garcke and D. Pflüger, eds. *Sparse Grids and Applications — Stuttgart 2014*. Vol. 109. Lecture Notes in Computational Science and Engineering. Springer International Publishing, 2016, p. 79.

[Har16b] B. Harding. 'Fault Tolerant Computation of Hyperbolic Partial Differential Equations with the Sparse Grid Combination Technique'. PhD thesis. Australian National University, Apr. 2016.

[Har72] F. Harary. *Graph Theory*. Addison-Wesley Publishing Company, Inc., Reading, Massachusetts, Oct. 1972.

[HC96] Z. He and D. Cremer. 'Sixth-order many-body perturbation theory. IV. Improvement of the Møller-Plesset correlation energy series by using Padé, Feenberg, and other approximations up to sixth order'. *Int. J. Quant. Chem.* 59(1), 1996, p. 71.

[HDP72] W. J. Hehre, R. Ditchfield and J. A. Pople. 'Self-Consistent Molecular Orbital Methods. XII. Further Extensions of Gaussian-Type Basis Sets for Use in Molecular Orbital Studies of Organic Molecules'. *J. Chem. Phys.* 56(5), Mar. 1972, p. 2257.

[Hea96] M. Head-Gordon. 'Quantum Chemistry and Molecular Processes'. *J. Phys. Chem.* 100(31), Jan. 1996, p. 13213.

[Heb14] F. Heber. 'Ein systematischer, linear skalierender Fragmentansatz für das Elektronenstrukturproblem'. Dissertation. Rheinische Friedrich-Wilhelms-Universität Bonn, Mar. 2014.

[Heb17] F. Heber. *MoleCuilder – a molecular builder*. Version 1.6.0. 19 Mar. 2017. URL: https://www.molecuilder.de.

[Heg+16] M. Hegland, B. Harding, C. Kowitz, D. Pflüger and P. Strazdins. 'Recent Developments in the Theory and Application of the Sparse Grid Combination Technique'. In: H.-J. Bungartz, P. Neumann and W. E. Nagel, eds. *Software for Exascale Computing - SPPEXA 2013-2015*. Vol. 113. Lecture Notes in Computational Science and Engineering. Springer International Publishing, Cham, 2016, p. 143.

[Hég+16] B. Hégely, P. R. Nagy, G. G. Ferenczy and M. Kállay. 'Exact density functional and wave function embedding schemes based on orbital localization'. *J. Chem. Phys.* 145(6), Aug. 2016, p. 064107.

[Heg03] M. Hegland. 'Adaptive sparse grids'. *ANZIAM J.* 44, Apr. 2003, p. 335.

[Heh+70] W. J. Hehre, R. Ditchfield, R. F. Stewart and J. A. Pople. 'Self-Consistent Molecular Orbital Methods. IV. Use of Gaussian Expansions of Slater-Type Orbitals. Extension to Second-Row Molecules'. *J. Chem. Phys.* 52(5), Mar. 1970, p. 2769.

[Hel+97] T. Helgaker, W. Klopper, H. Koch and J. Noga. 'Basis-set convergence of correlated calculations on water'. *J. Chem. Phys.* 106(23), June 1997, p. 9639.

[Her19] J. M. Herbert. 'Fantasy versus reality in fragment-based quantum chemistry'. *J. Chem. Phys.* 151(17), Nov. 2019, p. 170901.

[HG11] J. W. Hollett and P. M. W. Gill. 'The two faces of static correlation'. *J. Chem. Phys.* 134(11), Mar. 2011, p. 114111.

[HGC07] M. Hegland, J. Garcke and V. Challis. 'The combination technique and some generalisations'. *Linear Algebra Appl.* 420(2-3), Jan. 2007, p. 249.

[HH13] B. Harding and M. Hegland. 'A robust combination technique'. *ANZIAM J.* 54, Aug. 2013, p. 394.

[HHL10] S. Hua, W. Hua and S. Li. 'An Efficient Implementation of the Generalized Energy-Based Fragmentation Approach for General Large Molecules'. *J. Phys. Chem. A* 114(31), July 2010, p. 8126.

[HHT95] T. Hashimoto, K. Hirao and H. Tatewaki. 'Comment on Dunning's correlation-consistent basis sets'. *Chem. Phys. Lett.* 243(1-2), Sept. 1995, p. 190.

[Hil12] J. G. Hill. 'Gaussian basis sets for molecular applications'. *Int. J. Quant. Chem.* 113(1), Oct. 2012, p. 21.

[Hir03] S. Hirata. 'Tensor Contraction Engine: Abstraction and Automated Parallel Implementation of Configuration-Interaction, Coupled-Cluster, and Many-Body Perturbation Theories'. *J. Phys. Chem. A* 107(46), Oct. 2003, p. 9887.

[HK21] J. Hellmers and C. König. 'A unified and flexible formulation of molecular fragmentation schemes'. *J. Chem. Phys.* 155(16), Oct. 2021, p. 164105.

[HKT08] T. Helgaker, W. Klopper and D. P. Tew. 'Quantitative quantum chemistry'. *Mol. Phys.* 106(16-18), Aug. 2008, p. 2107.

[HLI17] C. Haycraft, J. Li and S. S. Iyengar. 'Efficient, "On-the-Fly", Born-Oppenheimer and Car-Parrinello-type Dynamics with Coupled Cluster Accuracy through Fragment Based Electronic Structure'. *J. Chem. Theory Comput.* 13(5), Apr. 2017, p. 1887.

[HLR05] M. Y. Hayes, B. Li and H. Rabitz. 'Estimation of Molecular Properties by High-Dimensional Model Representation'. *J. Phys. Chem. A* 110(1), Dec. 2005, p. 264.

[HLT93] F. Harary, E. Loukakis and C. Tsouros. 'The geodetic number of a graph'. *Math. Comput. Model.* 17(11), June 1993, p. 89.

[HM99] T. K. Harris and A. S. Mildvan. 'High-Precision Measurement of Hydrogen Bond Lengths in Proteins by Nuclear Magnetic Resonance Methods'. *Proteins Struct. Funct. Genet.* 35(3), May 1999, p. 275.

[HMK05] L. Huang, L. Massa and J. Karle. 'Kernel energy method illustrated with peptides'. *Int. J. Quant. Chem.* 103(6), 2005, p. 808.

[HMS70] D. Hankins, J. W. Moskowitz and F. H. Stillinger. 'Water Molecule Interactions'. *J. Chem. Phys.* 53(12), Dec. 1970, p. 4544.

[HN81] F. Harary and J. Nieminen. 'Convexity in graphs'. *J. Differ. Geom.* 16(2), Jan. 1981.

[HNK18] B. Hégely, P. R. Nagy and M. Kállay. 'Dual Basis Set Approach for Density Functional and Wave Function Embedding Schemes'. *J. Chem. Theory Comput.* 14(9), July 2018, p. 4600.

[HOJ13] T. Helgaker, J. Olsen and P. Jørgensen. *Molecular Electronic-Structure Theory.* Wiley-Blackwell, 1 Feb. 2013.

[Hol08] M. Holtz. 'Sparse Grid Quadrature in High Dimensions with Applications in Finance and Insurance'. Dissertation. Rheinische Friedrich-Wilhelms-Universität Bonn, 2008.

[Hon+04] S. Honda, K. Yamasaki, Y. Sawada and H. Morii. '10 Residue Folded Peptide Designed by Segment Statistics'. *Structure* 12(8), Aug. 2004, p. 1507.

[HP73] P. C. Hariharan and J. A. Pople. 'The influence of polarization functions on molecular orbital hydrogenation energies'. *Theor. Chim. Acta.* 28(3), 1973, p. 213.

[HP97] P. W. Hemker and C. Pflaum. 'Approximation on partially ordered sets of regular grids'. *Appl. Numer. Math.* 25(1), Oct. 1997, p. 55.

[HPF88] M. Head-Gordon, J. A. Pople and M. J. Frisch. 'MP2 energy evaluation by direct methods'. *Chem. Phys. Lett.* 153(6), Dec. 1988, p. 503.

[HSP69] W. J. Hehre, R. F. Stewart and J. A. Pople. 'Self-Consistent Molecular-Orbital Methods. I. Use of Gaussian Expansions of Slater-Type Atomic Orbitals'. *J. Chem. Phys.* 51(6), Sept. 1969, p. 2657.

[HSS08] A. A. Hagberg, D. A. Schult and P. J. Swart. 'Exploring Network Structure, Dynamics, and Function using NetworkX'. In: G. Varoquaux, T. Vaught and J. Millman, eds. *Proceedings of the 7th Python in Science Conference.* Pasadena, CA USA, 2008, p. 11.

[HT03] B. W. Hopkins and G. S. Tschumper. 'A multicentered approach to integrated QM/QM calculations. Applications to multiply hydrogen bonded systems'. *J. Comput. Chem.* 24(13), Aug. 2003, p. 1563.

[HT05] B. W. Hopkins and G. S. Tschumper. 'Integrated quantum mechanical approaches for extended π systems: Multicentered QM/QM studies of the cyanogen and diacetylene trimers'. *Chem. Phys. Lett.* 407(4-6), May 2005, p. 362.

[Hul14] A. Hullmann. 'The ANOVA decomposition and generalized sparse grid methods for the high-dimensional backward Kolmogorov equation'. Dissertation. Rheinische Friedrich-Wilhelms-Universität Bonn, 2014.

[Huz85] S. Huzinaga. 'Basis sets for molecular calculations'. *Comput. Phys. Rep.* 2(6), May 1985, p. 281.

[HX20] J. P. Heindel and S. S. Xantheas. 'The Many-Body Expansion for Aqueous Systems Revisited: I. Water–Water Interactions'. *J. Chem. Theory Comput.* 16(11), Oct. 2020, p. 6843.

[HY04] S. Honda and K. Yamasaki. *NMR Structure of designed protein, Chignolin, consisting of only ten amino acids (Ensembles).* Apr. 2004.

[IWT13] M. Isegawa, B. Wang and D. G. Truhlar. 'Electrostatically Embedded Molecular Tailoring Approach and Validation for Peptides'. *J. Chem. Theory Comput.* 9(3), Feb. 2013, p. 1381.

[Jac+13] L. D. Jacobson, R. M. Richard, K. U. Lao and J. M. Herbert. 'Efficient Monomer-Based Quantum Chemistry Methods for Molecular and Ionic Clusters'. In: R. A. Wheeler, ed. *Annual Reports in Computational Chemistry.* Vol. 9. Elsevier, 2013. Chap. 2, p. 25.

[Jen17] F. Jensen. *Introduction to Computational Chemistry.* Wiley, 6 Feb. 2017.

[JG12] J. John and V. M. Gleeta. 'The connected hull number of a graph'. *South Asian J. Math.* 2(5), 2012, p. 508.

[JG91] J. H. Jensen and M. S. Gordon. 'Splicing I: Using mixed basis sets in *ab initio* calculations'. *J. Comput. Chem.* 12(4), May 1991, p. 421.

[JHS11] C. A. Jiménez-Hoyos, T. M. Henderson and G. E. Scuseria. 'Generalized Hartree-Fock Description of Molecular Dissociation'. *J. Chem. Theory Comput.* 7(9), July 2011, p. 2667.

[JKS95] M. S. Jacobson, A. E. Kézdy and S. Seif. 'The poset on connected induced subgraphs of a graph need not be Sperner'. *Order* 12(3), 1995, p. 315.

[Joh17] F. Johansson. 'Arb: Efficient Arbitrary-Precision Midpoint-Radius Interval Arithmetic'. *IEEE Trans. Comput.* 66(8), Aug. 2017, p. 1281.

[Jon+20] L. O. Jones, M. A. Mosquera, G. C. Schatz and M. A. Ratner. 'Embedding Methods for Quantum Chemistry: Applications from Materials to Life Sciences'. *J. Am. Chem. Soc.* 142(7), Jan. 2020, p. 3281.

[Kál+20] M. Kállay, P. R. Nagy, D. Mester, Z. Rolik, G. Samu, J. Csontos, J. Csóka, P. B. Szabó, L. Gyevi-Nagy, B. Hégely, I. Ladjánszki, L. Szegedy, B. Ladóczki, K. Petrov, M. Farkas, P. D. Mezei and Á. Ganyecz. 'The MRCC program system: Accurate quantum chemistry from water to proteins'. *J. Chem. Phys.* 152(7), Feb. 2020, p. 074107.

[Kál14] M. Kállay. 'A systematic way for the cost reduction of density fitting methods'. *J. Chem. Phys.* 141(24), Dec. 2014, p. 244113.

[Kar+06] A. Karton, E. Rabinovich, J. M. L. Martin and B. Ruscic. 'W4 theory for computational thermochemistry: In pursuit of confident sub-kJ/mol predictions'. *J. Chem. Phys.* 125(14), Oct. 2006, p. 144108.

[Kar16] A. Karton. 'A computational chemist's guide to accurate thermochemistry for organic molecules'. *WIREs Comput. Mol. Sci.* 6(3), Feb. 2016, p. 292.

[Kar22] A. Karton. 'Quantum mechanical thermochemical predictions 100 years after the Schrödinger equation'. In: *Annual Reports in Computational Chemistry.* Elsevier, 2022, p. 123.

[Kar90] M. Karplus. 'Three-dimensional "Pople diagram"'. *J. Phys. Chem.* 94(14), July 1990, p. 5435.

[KB77a] B. Klahn and W. A. Bingel. 'Completeness and linear independence of basis sets used in quantum chemistry'. *Int. J. Quant. Chem.* 11(6), June 1977, p. 943.

[KB77b] B. Klahn and W. A. Bingel. 'The convergence of the Rayleigh-Ritz Method in quantum chemistry I. The Criteria of Convergence'. *Theor. Chim. Acta.* 44(1), Mar. 1977, p. 9.

[KB77c] B. Klahn and W. A. Bingel. 'The convergence of the Rayleigh-Ritz Method in quantum chemistry II. Investigation of the Convergence for Special Systems of Slater, Gauss and Two-Electron Functions'. *Theor. Chim. Acta.* 44(1), Mar. 1977, p. 27.

[KB92] S. A. Kucharski and R. J. Bartlett. 'The coupled-cluster single, double, triple, and quadruple excitation method'. *J. Chem. Phys.* 97(6), Sept. 1992, p. 4282.

[KB97] D. J. Klein and D. Babić. 'Partial Orderings in Chemistry'. *J. Chem. Inf. Comput. Sci.* 37(4), July 1997, p. 656.

[KC06] J. Kongsted and O. Christiansen. 'Automatic generation of force fields and property surfaces for use in variational vibrational calculations of anharmonic vibrational energies and zero-point vibrational averaged properties'. *J. Chem. Phys.* 125(12), Sept. 2006, p. 124108.

[KC16] C. König and O. Christiansen. 'Linear-scaling generation of potential energy surfaces using a double incremental expansion'. *J. Chem. Phys.* 145(6), Aug. 2016, p. 064105.

[KDH92] R. A. Kendall, T. H. Dunning Jr. and R. J. Harrison. 'Electron affinities of the first-row atoms revisited. Systematic basis sets and wave functions'. *J. Chem. Phys.* 96(9), May 1992, p. 6796.

[KDI21] A. Kumar, N. DeGregorio and S. S. Iyengar. 'Graph-Theory-Based Molecular Fragmentation for Efficient and Accurate Potential Surface Calculations in Multiple Dimensions'. *J. Chem. Theory Comput.* 17(11), Oct. 2021, p. 6671.

[KDM11] A. Karton, S. Daon and J. M. L. Martin. 'W4-11: A high-confidence benchmark dataset for computational thermochemistry derived from first-principles W4 data'. *Chem. Phys. Lett.* 510(4-6), July 2011, p. 165.

[Kei+21] J. A. Keith, V. Vassilev-Galindo, B. Cheng, S. Chmiela, M. Gastegger, K.-R. Müller and A. Tkatchenko. 'Combining Machine Learning and Computational Chemistry for Predictive Insights Into Chemical Systems', 12 Feb. 2021. arXiv: 2102.06321 [physics.chem-ph].

[Ken92] R. Kennes. 'Computational aspects of the Mobius transformation of graphs'. *IEEE Trans. Syst. Man Cybern.* 22(2), 1992, p. 201.

[KG05] M. Kállay and J. Gauss. 'Approximate treatment of higher excitations in coupled-cluster theory'. *J. Chem. Phys.* 123(21), Dec. 2005, p. 214105.

[KG08] M. Kállay and J. Gauss. 'Approximate treatment of higher excitations in coupled-cluster theory. II. Extension to general single-determinant reference functions and improved approaches for the canonical Hartree-Fock case'. *J. Chem. Phys.* 129(14), Oct. 2008, p. 144101.

[KI19] A. Kumar and S. S. Iyengar. 'Fragment-Based Electronic Structure for Potential Energy Surfaces Using a Superposition of Fragmentation Topologies'. *J. Chem. Theory Comput.* 15(11), Sept. 2019, p. 5769.

[Kir+21] T. Kirsch, J. M. H. Olsen, V. Bolnykh, S. Meloni, E. Ippoliti, U. Rothlisberger, M. Cascella and J. Gauss. 'Wavefunction-Based Electrostatic-Embedding QM/MM Using CFOUR through MiMiC'. *J. Chem. Theory Comput.* 18(1), Dec. 2021, p. 13.

[Kit+99] K. Kitaura, E. Ikeo, T. Asada, T. Nakano and M. Uebayasi. 'Fragment molecular orbital method: an approximate computational method for large molecules'. *Chem. Phys. Lett.* 313(3-4), Nov. 1999, p. 701.

[Kle86] D. J. Klein. 'Chemical graph-theoretic cluster expansions'. *Int. J. Quant. Chem.* 30(S20), Mar. 1986, p. 153.

[KM04] A. Kosowski and K. Manuszewski. 'Classical coloring of graphs'. In: *Graph Colorings*. Ed. by M. Kubale. Vol. 352. Contemporary Mathematics. American Mathematical Society, Providence, RI, 2004, p. 1.

[KM06] A. Karton and J. M. L. Martin. 'Comment on: "Estimating the Hartree-Fock limit from finite basis set calculations" [Jensen F (2005) Theor Chem Acc 113:267]'. 115, 2006, p. 330.

[Kni13] G. Knizia. 'Intrinsic Atomic Orbitals: An Unbiased Bridge between Quantum Theory and Chemical Concepts'. *J. Chem. Theory Comput.* 9(11), Oct. 2013, p. 4834.

[Koh96] W. Kohn. 'Density Functional and Density Matrix Method Scaling Linearly with the Number of Atoms'. *Phys. Rev. Lett.* 76(17), Apr. 1996, p. 3168.

[Kor22] R. Kormos. *What is the origin of the many-body expansion?* Physics Stack Exchange. Version of 2022-05-14. The author's profile can be found at https://physics.stackexchange.com/users/332105/rian-kormos. 14 May 2022. URL: https://physics.stackexchange.com/q/702036.

[KR97] R. Kobayashi and A. P. Rendell. 'A direct coupled cluster algorithm for massively parallel computers'. *Chem. Phys. Lett.* 265(1-2), Jan. 1997, p. 1.

[Kra07] J. Kraus. 'Option Pricing using the Sparse Grid Combination Technique'. Master's thesis. University of Waterloo, Technical University of Munich, 2007.

[Kri+80] R. Krishnan, J. S. Binkley, R. Seeger and J. A. Pople. 'Self-consistent molecular orbital methods. XX. A basis set for correlated wave functions'. *J. Chem. Phys.* 72(1), Jan. 1980, p. 650.

[KS01] M. Kállay and P. R. Surján. 'Higher excitations in coupled-cluster theory'. *J. Chem. Phys.* 115(7), Aug. 2001, p. 2945.

[KS21] C. Komusiewicz and F. Sommer. 'Enumerating connected induced subgraphs: Improved delay and experimental comparison'. *Discrete Appl. Math.* 303, Nov. 2021, p. 262.

[KS65] W. Kohn and L. J. Sham. 'Self-Consistent Equations Including Exchange and Correlation Effects'. *Phys. Rev.* 140(4A), Nov. 1965, p. A1133.

[KS96] A. Kézdy and S. Seif. 'When is a Poset Isomorphic to the Poset of Connected Induced Subgraphs of a Graph?' *Southwest J. Pure Appl. Math.* 1, 28 May 1996, p. 42.

[KS98] D. L. Kreher and D. R. Stinson. *Combinatorial Algorithms: Generation, Enumeration, and Search*. CRC Press, Boca Raton, 1998.

[KSM17] A. Karton, N. Sylvetsky and J. M. L. Martin. 'W4-17: A diverse and high-confidence dataset of atomization energies for benchmarking high-level electronic structure methods'. *J. Comput. Chem.* 38(24), July 2017, p. 2063.

[KT13] S. Kazachenko and A. J. Thakkar. 'Water nanodroplets: Predictions of five model potentials'. *J. Chem. Phys.* 138(19), May 2013, p. 194302.

[Kuo+08] F. Y. Kuo, I. H. Sloan, G. W. Wasilkowski and H. Woźniakowski. *On decompositions of multivariate functions*. Extended preprint. May 2008. URL: `https://www.maths.unsw.edu.au/sites/default/files/amr08_5_0.pdf`.

[Kuo+09] F. Y. Kuo, I. H. Sloan, G. W. Wasilkowski and H. Woźniakowski. 'On decompositions of multivariate functions'. *Math. Comput.* 79(270), Nov. 2009, p. 953.

[Lag+20] R. Lago, M. Obersteiner, T. Pollinger, J. Rentrop, H.-J. Bungartz, T. Dannert, M. Griebel, F. Jenko and D. Pflüger. 'EXAHD: A Massively Parallel Fault Tolerant Sparse Grid Approach for High-Dimensional Turbulent Plasma Simulations'. In: H.-J. Bungartz, S. Reiz, B. Uekermann, P. Neumann and W. E. Nagel, eds. *Software for Exascale Computing — SPPEXA 2016–2019*. Vol. 136. Lecture Notes in Computational Science and Engineering. Springer International Publishing, Cham, 2020, p. 301.

[Lao+16] K. U. Lao, K.-Y. Liu, R. M. Richard and J. M. Herbert. 'Understanding the many-body expansion for large systems. II. Accuracy considerations'. *J. Chem. Phys.* 144(16), Apr. 2016, p. 164105.

[LC04] L. Lafuente and J. A. Cuesta. 'Density Functional Theory for General Hard-Core Lattice Gases'. *Phys. Rev. Lett.* 93(13), Sept. 2004.

[LC05] L. Lafuente and J. A. Cuesta. 'Cluster density functional theory for lattice models based on the theory of Möbius functions'. *J. Phys. A: Math. Gen.* 38(34), Aug. 2005, p. 7461.

[Le+12] H.-A. Le, H.-J. Tan, J. F. Ouyang and R. P. A. Bettens. 'Combined Fragmentation Method: A Simple Method for Fragmentation of Large Molecules'. *J. Chem. Theory Comput.* 8(2), Jan. 2012, p. 469.

[LeB05] C. Le Bris. 'Computational chemistry from the perspective of numerical analysis'. *Acta Numer.* 14, Apr. 2005, p. 363.

[Lec76] B. Leclerc. 'Arbres et dimension des ordres'. *Discrete Math.* 14(1), 1976, p. 69.

[Lee+17] S. J. R. Lee, K. Miyamoto, F. Ding, F. R. Manby and T. F. Miller. 'Density-based errors in mixed-basis mean-field electronic structure, with implications for embedding and QM/MM methods'. *Chem. Phys. Lett.* 683, Sept. 2017, p. 375.

[Leh64] D. H. Lehmer. 'The Machine Tools of Combinatorics'. In: *Applied Combinatorial Mathematics*. Ed. by E. F. Beckenbach. John Wiley & Sons, Inc., New York, London, Sydney, 1964. Chap. 1, p. 5.

[Lei+00] M. L. Leininger, W. D. Allen, H. F. Schaefer III and C. D. Sherrill. 'Is Møller-Plesset perturbation theory a convergent *ab initio* method?' *J. Chem. Phys.* 112(21), June 2000, p. 9213.

[Lev+12] H. R. Leverentz, K. A. Maerzke, S. J. Keasler, J. I. Siepmann and D. G. Truhlar. 'Electrostatically embedded many-body method for dipole moments, partial atomic charges, and charge transfer'. *Phys. Chem. Chem. Phys.* 14(21), 2012, p. 7669.

[Lev12] A. Levitt. 'Convergence of gradient-based algorithms for the Hartree-Fock equations'. *ESAIM: Math. Model. Numer. Anal.* 46(6), Mar. 2012, p. 1321.

[LH16] J. Liu and J. M. Herbert. 'Pair-Pair Approximation to the Generalized Many-Body Expansion: An Alternative to the Four-Body Expansion for ab Initio Prediction of Protein Energetics via Molecular Fragmentation'. *J. Chem. Theory Comput.* 12(2), Jan. 2016, p. 572.

[LH17] K.-Y. Liu and J. M. Herbert. 'Understanding the many-body expansion for large systems. III. Critical role of four-body terms, counterpoise corrections, and cutoffs'. *J. Chem. Phys.* 147(16), Oct. 2017, p. 161729.

[LH19] K.-Y. Liu and J. M. Herbert. 'Energy-Screened Many-Body Expansion: A Practical Yet Accurate Fragmentation Method for Quantum Chemistry'. *J. Chem. Theory Comput.* 16(1), Nov. 2019, p. 475.

[LHI16] J. Li, C. Haycraft and S. S. Iyengar. 'Hybrid Extended Lagrangian, Post-Hartree-Fock Born-Oppenheimer ab Initio Molecular Dynamics Using Fragment-Based Electronic Structure'. *J. Chem. Theory Comput.* 12(6), June 2016, p. 2493.

[LI15] J. Li and S. S. Iyengar. '*Ab Initio* Molecular Dynamics Using Recursive, Spatially Separated, Overlapping Model Subsystems Mixed within an ONIOM-Based Fragmentation Energy Extrapolation Technique'. *J. Chem. Theory Comput.* 11(9), Aug. 2015, p. 3978.

[Liu+19] J. Liu, B. Rana, K.-Y. Liu and J. M. Herbert. 'Variational Formulation of the Generalized Many-Body Expansion with Self-Consistent Charge Embedding: Simple and Correct Analytic Energy Gradient for Fragment-Based *ab Initio* Molecular Dynamics'. *J. Phys. Chem. Lett.* 10(14), June 2019, p. 3877.

[LL05] C. Le Bris and P.-L. Lions. 'From atoms to crystals: a mathematical journey'. *Bull. Am. Math. Soc.* 42(03), Apr. 2005, p. 291.

[LLJ07] W. Li, S. Li and Y. Jiang. 'Generalized Energy-Based Fragmentation Approach for Computing the Ground-State Energies and Properties of Large Molecules'. *J. Phys. Chem. A* 111(11), Feb. 2007, p. 2193.

[Löw55] P.-O. Löwdin. 'Quantum Theory of Many-Particle Systems. III. Extension of the Hartree-Fock Scheme to Include Degenerate Systems and Correlation Effects'. *Phys. Rev.* 97(6), Mar. 1955, p. 1509.

[LRL91] R. Lindh, U. Ryu and B. Liu. 'The reduced multiplication scheme of the Rys quadrature and new recurrence relations for auxiliary function based two-electron integral evaluation'. *J. Chem. Phys.* 95(8), Oct. 1991, p. 5889.

[LS74] E. H. Lieb and B. Simon. 'On solutions to the Hartree-Fock problem for atoms and molecules'. *J. Chem. Phys.* 61(2), July 1974, p. 735.

[LS77] E. H. Lieb and B. Simon. 'The Hartree-Fock theory for Coulomb systems'. *Commun. Math. Phys.* 53(3), Feb. 1977, p. 185.

[LT06] H. Lin and D. G. Truhlar. 'QM/MM: what have we learned, where are we, and where do we go from here?' *Theor. Chim. Acta.* 117(2), July 2006.

[LYY96] T.-S. Lee, D. M. York and W. Yang. 'Linear-scaling semiempirical quantum calculations for macromolecules'. *J. Chem. Phys.* 105(7), Aug. 1996, p. 2744.

[Man+09] M. Mantina, A. C. Chamberlin, R. Valero, C. J. Cramer and D. G. Truhlar. 'Consistent van der Waals Radii for the Whole Main Group'. *J. Phys. Chem. A* 113(19), Apr. 2009, p. 5806.

[Man+12] F. R. Manby, M. Stella, J. D. Goodpaster and T. F. Miller. 'A Simple, Exact Density-Functional-Theory Embedding Scheme'. *J. Chem. Theory Comput.* 8(8), July 2012, p. 2564.

[Mar21] J. M. L. Martin. 'Basis set convergence and extrapolation of connected triple excitation contributions (T) in computational thermochemistry: the W4-17 benchmark with up to k functions', 26 Mar. 2021. arXiv: `2103.14370 [physics.chem-ph]`.

[Mar75] J. W. Martin. 'Many-body forces in metals and the Brugger elastic constants'. *J. Phys. C: Solid State Phys.* 8(18), Sept. 1975, p. 2837.

[Mar96] J. M. L. Martin. 'Ab initio total atomization energies of small molecules — towards the basis set limit'. *Chem. Phys. Lett.* 259(5-6), Sept. 1996, p. 669.

[Mas+98] P. E. Maslen, C. Ochsenfeld, C. A. White, M. S. Lee and M. Head-Gordon. 'Locality and Sparsity of Ab Initio One-Particle Density Matrices and Localized Orbitals'. *J. Phys. Chem. A* 102(12), Feb. 1998, p. 2215.

[Mat14] K. Matuschke. 'Trigonometrische Interpolation auf verallgemeinerten dünnen Gittern mit beliebiger Levelstruktur'. Diplomarbeit. Rheinische Friedrich-Wilhelms-Universität Bonn, 2 Sept. 2014.

[MD78] L. E. McMurchie and E. R. Davidson. 'One- and Two-Electron Integrals over Cartesian Gaussian Functions'. *J. Comput. Phys* 26(2), Feb. 1978, p. 218.

[MGP99] S. L. Mielke, B. C. Garrett and K. A. Peterson. 'The utility of many-body decompositions for the accurate basis set extrapolation of *ab initio* data'. *J. Chem. Phys.* 111(9), Sept. 1999, p. 3806.

[MM95] F. Maseras and K. Morokuma. 'IMOMM: A new integrated *ab initio* + molecular mechanics geometry optimization scheme of equilibrium structures and transition states'. *J. Comput. Chem.* 16(9), Sept. 1995, p. 1170.

[MMM16] K. Miyamoto, T. F. Miller and F. R. Manby. 'Fock-Matrix Corrections in Density Functional Theory and Use in Embedded Mean-Field Theory'. *J. Chem. Theory Comput.* 12(12), Nov. 2016, p. 5811.

[MO99] J. M. L. Martin and G. de Oliveira. 'Towards standard methods for benchmark quality *ab initio* thermochemistry — W1 and W2 theory'. *J. Chem. Phys.* 111(5), Aug. 1999, p. 1843.

[Mon85] B. Monjardet. 'A use for frequently rediscovering a concept'. *Order* 1(4), 1985, p. 415.

[Mor94] T. Morita. 'Formal Structure of the Cluster Variation Method'. *Prog. Theor. Phys. Suppl.* 115, 1994, p. 27.

[MP34] C. Møller and M. S. Plesset. 'Note on an Approximation Treatment for Many-Electron Systems'. *Phys. Rev.* 46(7), Oct. 1934, p. 618.

[MR11] N. J. Mayhall and K. Raghavachari. 'Molecules-in-Molecules: An Extrapolated Fragment-Based Approach for Accurate Calculations on Large Molecules and Materials'. *J. Chem. Theory Comput.* 7(5), Apr. 2011, p. 1336.

[MR12] N. J. Mayhall and K. Raghavachari. 'Many-Overlapping-Body (MOB) Expansion: A Generalized Many Body Expansion for Nondisjoint Monomers in Molecular Fragmentation Calculations of Covalent Molecules'. *J. Chem. Theory Comput.* 8(8), July 2012, p. 2669.

[MRCC] Mrcc, a quantum chemical program suite written by M. Kállay, P. R. Nagy, D. Mester, L. Gyevi-Nagy, J. Csóka, P. B. Szabó, Z. Rolik, G. Samu, J. Csontos, B. Hégely, Á. Ganyecz, I. Ladjánszki, L. Szegedy, B. Ladóczki, K. Petrov, M. Farkas, P. D. Mezei, and R. A. Horváth. See `www.mrcc.hu`.

[MŠ18] T. Marc and L. Šubelj. 'Convexity in complex networks'. *Netw. Sci.* 6(2), Feb. 2018, p. 176.

[Mul+18] J.-M. Muller, N. Brunie, F. de Dinechin, C.-P. Jeannerod, M. Joldes, V. Lefèvre, G. Melquiond, N. Revol and S. Torres. *Handbook of Floating-Point Arithmetic*. Springer International Publishing, 2018.

[Mul55] R. S. Mulliken. 'Electronic Population Analysis on LCAO-MO Molecular Wave Functions. I'. *J. Chem. Phys.* 23(10), Oct. 1955, p. 1833.

[Mur+21] E. N. Muratov, R. Amaro, C. H. Andrade, N. Brown, S. Ekins, D. Fourches, O. Isayev, D. Kozakov, J. L. Medina-Franco, K. M. Merz, T. I. Oprea, V. Poroikov, G. Schneider, M. H. Todd, A. Varnek, D. A. Winkler, A. V. Zakharov, A. Cherkasov and A. Tropsha. 'A critical overview of computational approaches employed for COVID-19 drug discovery'. *Chem. Soc. Rev.* 50(16), 2021, p. 9121.

[MWD99] T. van Mourik, A. K. Wilson and T. H. Dunning Jr. 'Benchmark calculations with correlated molecular wavefunctions. XIII. Potential energy curves for He_2, Ne_2 and Ar_2 using correlation consistent basis sets through augmented sextuple zeta'. *Mol. Phys.* 96(4), Feb. 1999, p. 529.

[Nar+19] B. Narayanan, P. C. Redfern, R. S. Assary and L. A. Curtiss. 'Accurate quantum chemical energies for 133 000 organic molecules'. *Chem. Sci.* 10(31), 2019, p. 7449.

[NB87] J. Noga and R. J. Bartlett. 'The full CCSDT model for molecular electronic structure'. *J. Chem. Phys.* 86(12), June 1987, p. 7041.

[NC07] H. M. Netzloff and M. A. Collins. '*Ab initio* energies of nonconducting crystals by systematic fragmentation'. *J. Chem. Phys.* 127(13), Oct. 2007, p. 134113.

[Nie80] J. Nieminen. 'The lattice of connected subgraphs of a connected graph'. *Comment. Math.* 21(1), 1980, p. 187.

[NIS19] NIST. *CODATA internationally recommended 2018 values of the fundamental physical constants*. Accessed 4/1/2022. May 2019. URL: `https://physics.nist.gov/cuu/Constants/index.html`.

[Nob+16] F. Nobile, L. Tamellini, F. Tesei and R. Tempone. 'An Adaptive Sparse Grid Algorithm for Elliptic PDEs with Lognormal Diffusion Coefficient'. In: J. Garcke and D. Pflüger, eds. *Sparse Grids and Applications — Stuttgart 2014*. Vol. 109. Lecture Notes in Computational Science and Engineering. Springer International Publishing, Cham, 2016, p. 191.

[NTT15] F. Nobile, L. Tamellini and R. Tempone. 'Convergence of quasi-optimal sparse-grid approximation of Hilbert-space-valued functions: application to random elliptic PDEs'. *Numer. Math.* 134(2), Oct. 2015, p. 343.

[NW78] A. Nijenhuis and H. S. Wilf. *Combinatorial algorithms for computers and calculators*. 2nd ed. Computer science and applied mathematics. Academic Press, New York, USA, 1978.

[NWCDoc] *NWChem: Open Source High-Performance Computational Chemistry*. GitHub. User documentation. Referenced information current as of 20/10/2023. URL: `https://nwchemgit.github.io/`.

[OB16] J. F. Ouyang and R. P. A. Bettens. 'When are Many-Body Effects Significant?' *J. Chem. Theory Comput.* 12(12), Nov. 2016, p. 5860.

[OB21] M. Obersteiner and H.-J. Bungartz. 'A Generalized Spatially Adaptive Sparse Grid Combination Technique with Dimension-wise Refinement'. *SIAM J. Sci. Comput.* 43(4), Jan. 2021, p. A2381.

[Obe21] M. Obersteiner. 'A Spatially Adaptive and Massively Parallel Implementation of the Fault-Tolerant Combination Technique'. Dissertation. Technische Universität München, 2021.

[OBo+11] N. M. O'Boyle, M. Banck, C. A. James, C. Morley, T. Vandermeersch and G. R. Hutchison. 'Open Babel: An open chemical toolbox'. *J. Cheminf.* 3(1), Oct. 2011.

[OCB14] J. F. Ouyang, M. W. Cvitkovic and R. P. A. Bettens. 'Trouble with the Many-Body Expansion'. *J. Chem. Theory Comput.* 10(9), June 2014, p. 3699.

[Ols+96] J. Olsen, O. Christiansen, H. Koch and P. Jørgensen. 'Surprising cases of divergent behavior in Møller-Plesset perturbation theory'. *J. Chem. Phys.* 105(12), Sept. 1996, p. 5082.

[OS86] S. Obara and A. Saika. 'Efficient recursive computation of molecular integrals over Cartesian Gaussian functions'. *J. Chem. Phys.* 84(7), Apr. 1986, p. 3963.

[Pas14] N. Pastuszka. 'A Multilevel Method for the Global Optimization of Potential Energy Functions'. Master's thesis. Rheinische Friedrich-Wilhelms-Universität Bonn, 2014.

[PB82] G. D. Purvis III and R. J. Bartlett. 'A full coupled-cluster singles and doubles model: The inclusion of disconnected triples'. *J. Chem. Phys.* 76(4), Feb. 1982, p. 1910.

[PC16] B. P. Pritchard and E. Chow. 'Horizontal vectorization of electron repulsion integrals'. *J. Comput. Chem.* 37(28), Sept. 2016, p. 2537.

[PD02] K. A. Peterson and T. H. Dunning Jr. 'Accurate correlation consistent basis sets for molecular core-valence correlation effects: The second row atoms Al-Ar, and the first row atoms B-Ne revisited'. *J. Chem. Phys.* 117(23), Dec. 2002, p. 10548.

[Pel13] I. M. Pelayo. *Geodesic Convexity in Graphs*. Springer New York, 2013.

[Pet+06] K. A. Peterson, B. C. Shepler, D. Figgen and H. Stoll. 'On the Spectroscopic and Thermochemical Properties of ClO, BrO, IO, and Their Anions'. *J. Phys. Chem. A* 110(51), Dec. 2006, p. 13877.

[Pet+97] K. A. Peterson, A. K. Wilson, D. E. Woon and T. H. Dunning Jr. 'Benchmark calculations with correlated molecular wave functions XII. Core correlation effects on the homonuclear diatomic molecules B_2-F_2'. *Theor. Chim. Acta.* 97(1-4), Oct. 1997, p. 251.

[Pet98] S. Petrie. 'Pitfalls for the Frozen-Core Approximation: Gaussian-2 Calculations on the Sodium Cation Affinities of Diatomic Fluorides'. *J. Phys. Chem. A* 102(30), July 1998, p. 6138.

[PFD12] K. A. Peterson, D. Feller and D. A. Dixon. 'Chemical accuracy in ab initio thermochemistry and spectroscopy: current strategies and future challenges'. *Theor. Chim. Acta.* 131(1), Jan. 2012.

[PK05] E. Prodan and W. Kohn. 'Nearsightedness of electronic matter'. *Proc. Natl. Acad. Sci.* 102(33), Aug. 2005, p. 11635.

[PL15] S. Pezeshki and H. Lin. 'Adaptive-Partitioning QM/MM for Molecular Dynamics Simulations: 4. Proton Hopping in Bulk Water'. *J. Chem. Theory Comput.* 11(6), May 2015, p. 2398.

[PLV20] L. Pouchard, Y. Lin and H. Van Dam. 'Replicating Machine Learning Experiments in Materials Science'. *Adv. Parallel Comput.* 36, 2020, p. 743.

[PN54] J. A. Pople and R. K. Nesbet. 'Self-Consistent Orbitals for Radicals'. *J. Chem. Phys.* 22(3), Mar. 1954, p. 571.

[Pop+83] J. A. Pople, M. J. Frisch, B. T. Luke and J. S. Binkley. 'A Moller-Plesset study of the energies of AH_n molecules (A = Li to F)'. *Int. J. Quant. Chem.* 24(S17), July 1983, p. 307.

[Pop+85] J. A. Pople, B. T. Luke, M. J. Frisch and J. S. Binkley. 'Theoretical thermochemistry. 1. Heats of formation of neutral AH_n molecules (A = Li to Cl)'. *J. Phys. Chem.* 89(11), May 1985, p. 2198.

[Pop+89] J. A. Pople, M. Head-Gordon, D. J. Fox, K. Raghavachari and L. A. Curtiss. 'Gaussian-1 theory: A general procedure for prediction of molecular energies'. *J. Chem. Phys.* 90(10), May 1989, p. 5622.

[Pop65] J. A. Pople. 'Two-Dimensional Chart of Quantum Chemistry'. 43(10), Nov. 1965, p. S229.

[Pop99] J. A. Pople. 'Quantum Chemical Models (Nobel Lecture)'. *Angew. Chem. Int. Ed* 38(13-14), July 1999, p. 1894.

[Pri+19] B. P. Pritchard, D. Altarawy, B. Didier, T. D. Gibson and T. L. Windus. 'New Basis Set Exchange: An Open, Up-to-Date Resource for the Molecular Sciences Community'. *J. Chem. Inf. Model.* 59(11), Oct. 2019, p. 4814.

[Pru+12] S. R. Pruitt, M. A. Addicoat, M. A. Collins and M. S. Gordon. 'The fragment molecular orbital and systematic molecular fragmentation methods applied to water clusters'. *Phys. Chem. Chem. Phys.* 14(21), 2012, p. 7752.

[PSK89] R. D. Poshusta, T. G. Schmalz and D. J. Klein. 'Heisenberg-model cluster expansion for half-filled Hubbard and PPP models'. *Mol. Phys.* 66(2), Feb. 1989, p. 317.

[PTM91] G. A. Petersson, T. G. Tensfeldt and J. A. Montgomery. 'A complete basis set model chemistry. III. The complete basis set-quadratic configuration interaction family of methods'. *J. Chem. Phys.* 94(9), May 1991, p. 6091.

[Pul80] P. Pulay. 'Convergence acceleration of iterative sequences. The case of SCF iteration'. *Chem. Phys. Lett.* 73(2), July 1980, p. 393.

[Pul82] P. Pulay. 'Improved SCF convergence acceleration'. *J. Comput. Chem.* 3(4), 1982, p. 556.

[PWD94] K. A. Peterson, D. E. Woon and T. H. Dunning Jr. 'Benchmark calculations with correlated molecular wave functions. IV. The classical barrier height of the $H + H_2 \longrightarrow H_2 + H$ reaction'. *J. Chem. Phys.* 100(10), May 1994, p. 7410.

[PWTC] *The Python Wiki (`TimeComplexity`)*. URL: `https://wiki.python.org/moin/TimeComplexity` (visited on 23/06/2023).

[PY94] R. G. Parr and W. Yang. *Density-Functional Theory of Atoms and Molecules*. International Series of Monographs on Chemistry 16. Oxford University Press, USA, 1994, p. 352.

[PZ99] C. Pflaum and A. Zhou. 'Error analysis of the combination technique'. *Numer. Math.* 84(2), Dec. 1999, p. 327.

[PZMQ] *PyZMQ: Python bindings for ØMQ*. GitHub/PyPI. URL: `https://github.com/zeromq/pyzmq`.

[QG15] J. Quaintance and H. W. Gould. *Combinatorial Identities for Stirling Numbers. The Unpublished Notes of H. W. Gould*. World Scientific Publishing Co Pte. Ltd., Singapore, 2015.

[QLT13] H. W. Qi, H. R. Leverentz and D. G. Truhlar. 'Water 16-mers and Hexamers: Assessment of the Three-Body and Electrostatically Embedded Many-Body Approximations of the Correlation Energy or the Nonlocal Energy As Ways to Include Cooperative Effects'. *J. Phys. Chem. A* 117(21), May 2013, p. 4486.

[RA99] H. Rabitz and Ö. F. Aliş. 'General foundations of high-dimensional model representations'. *J. Math. Chem.* 25(2/3), 1999, p. 197.

[Raf73] R. C. Raffenetti. 'General contraction of Gaussian atomic orbitals: Core, valence, polarization, and diffuse basis sets; Molecular integral evaluation'. *J. Chem. Phys.* 58(10), May 1973, p. 4452.

[Rag+89] K. Raghavachari, G. W. Trucks, J. A. Pople and M. Head-Gordon. 'A fifth-order perturbation comparison of electron correlation theories'. *Chem. Phys. Lett.* 157(6), May 1989, p. 479.

[Raj+17] S. Rajbhandari, F. Rastello, K. Kowalski, S. Krishnamoorthy and P. Sadayappan. 'Optimizing the Four-Index Integral Transform Using Data Movement Lower Bounds Analysis'. In: *Proceedings of the 22nd ACM SIGPLAN Symposium on Principles and Practice of Parallel Programming*. ACM, Jan. 2017.

[Ram+14] R. Ramakrishnan, P. O. Dral, M. Rupp and O. A. von Lilienfeld. 'Quantum chemistry structures and properties of 134 kilo molecules'. *Sci. Data* 1(1), Aug. 2014.

[RDK83] J. Rys, M. Dupuis and H. F. King. 'Computation of Electron Repulsion Integrals Using the Rys Quadrature Method'. *J. Comput. Chem.* 4(2), 1983, p. 154.

[Rei04] C. Reisinger. 'Numerische Methoden für hochdimensionale parabolische Gleichungen am Beispiel von Optionspreisaufgaben'. Dissertation. Ruprecht-Karls-Universität Heidelberg, 2004.

[Rei12] C. Reisinger. 'Analysis of linear difference schemes in the sparse grid combination technique'. *IMA J. Numer. Anal.* 33(2), Sept. 2012, p. 544.

[RH12] R. M. Richard and J. M. Herbert. 'A generalized many-body expansion and a unified view of fragment-based methods in electronic structure theory'. *J. Chem. Phys.* 137(6), Aug. 2012, p. 064113.

[RH13] R. M. Richard and J. M. Herbert. 'Many-Body Expansion with Overlapping Fragments: Analysis of Two Approaches'. *J. Chem. Theory Comput.* 9(3), Feb. 2013, p. 1408.

[ŘH13] J. Řezáč and P. Hobza. 'Describing Noncovalent Interactions beyond the Common Approximations: How Accurate Is the "Gold Standard," CCSD(T) at the Complete Basis Set Limit?' *J. Chem. Theory Comput.* 9(5), Apr. 2013, p. 2151.

[RHI18] T. C. Ricard, C. Haycraft and S. S. Iyengar. 'Adaptive, Geometric Networks for Efficient Coarse-Grained *Ab Initio* Molecular Dynamics with Post-Hartree-Fock Accuracy'. *J. Chem. Theory Comput.* 14(6), May 2018, p. 2852.

[RI18] T. C. Ricard and S. S. Iyengar. 'Efficiently Capturing Weak Interactions in ab Initio Molecular Dynamics with on-the-Fly Basis Set Extrapolation'. *J. Chem. Theory Comput.* 14(11), Oct. 2018, p. 5535.

[RI20] T. C. Ricard and S. S. Iyengar. 'Efficient and Accurate Approach To Estimate Hybrid Functional and Large Basis-Set Contributions to Condensed-Phase Systems and Molecule–Surface Interactions'. *J. Chem. Theory Comput.* 16(8), June 2020, p. 4790.

[RKI20] T. C. Ricard, A. Kumar and S. S. Iyengar. 'Embedded, graph-theoretically defined many-body approximations for wavefunction-in-DFT and DFT-in-DFT: Applications to gas- and condensed-phase ab initio molecular dynamics, and potential surfaces for quantum nuclear effects'. *Int. J. Quant. Chem.* 120(21), May 2020.

[RLH13] R. M. Richard, K. U. Lao and J. M. Herbert. 'Achieving the CCSD(T) Basis-Set Limit in Sizable Molecular Clusters: Counterpoise Corrections for the Many-Body Expansion'. *J. Phys. Chem. Lett.* 4(16), July 2013, p. 2674.

[RLH14] R. M. Richard, K. U. Lao and J. M. Herbert. 'Understanding the many-body expansion for large systems. I. Precision considerations'. *J. Chem. Phys.* 141(1), July 2014, p. 014108.

[Roh13] T. Rohwedder. 'The continuous Coupled Cluster formulation for the electronic Schrödinger equation'. *ESAIM: Math. Model. Numer. Anal.* 47(2), Jan. 2013, p. 421.

[Rol+13] Z. Rolik, L. Szegedy, I. Ladjánszki, B. Ladóczki and M. Kállay. 'An efficient linear-scaling CCSD(T) method based on local natural orbitals'. *J. Chem. Phys.* 139(9), Sept. 2013, p. 094105.

[Roo51] C. C. J. Roothaan. 'New Developments in Molecular Orbital Theory'. *Rev. Mod. Phys.* 23(2), Apr. 1951, p. 69.

[Rot64] G.-C. Rota. 'On the foundations of combinatorial theory I. Theory of Möbius Functions'. *Z. Wahrscheinlichkeitstheorie Verwandte Geb.* 2(4), 1964, p. 340.

[ŘS09] J. Řezáč and D. R. Salahub. 'Multilevel Fragment-Based Approach (MFBA): A Novel Hybrid Computational Method for the Study of Large Molecules'. *J. Chem. Theory Comput.* 6(1), Dec. 2009, p. 91.

[RS13] T. Rohwedder and R. Schneider. 'Error estimates for the Coupled Cluster method'. *ESAIM: Math. Model. Numer. Anal.* 47(6), Aug. 2013, p. 1553.

[RS15] K. Raghavachari and A. Saha. 'Accurate Composite and Fragment-Based Quantum Chemical Models for Large Molecules'. *Chem. Rev.* 115(12), Apr. 2015, p. 5643.

[Rüt16] A. Rüttgers. 'Multiscale Simulation of Polymeric Fluids using Sparse Grids'. Dissertation. Rheinische Friedrich-Wilhelms-Universität Bonn, Dec. 2016.

[SA89] S. Sæbø and J. Almlöf. 'Avoiding the integral storage bottleneck in LCAO calculations of electron correlation'. *Chem. Phys. Lett.* 154(1), Jan. 1989, p. 83.

[SC16] Q. Sun and G. K.-L. Chan. 'Quantum Embedding Theories'. *Acc. Chem. Res.* 49(12), Nov. 2016, p. 2705.

[Sch03] H. B. Schlegel. 'Exploring potential energy surfaces for chemical reactions: An overview of some practical methods'. *J. Comput. Chem.* 24(12), July 2003, p. 1514.

[Sch09] R. Schneider. 'Analysis of the projected coupled cluster method in electronic structure calculation'. *Numer. Math.* 113(3), June 2009, p. 433.

[Sch82] H. B. Schlegel. 'Optimization of equilibrium geometries and transition structures'. *J. Comput. Chem.* 3(2), 1982, p. 214.

[SDS09] E. Suárez, N. Díaz and D. Suárez. 'Thermochemical Fragment Energy Method for Biomolecules: Application to a Collagen Model Peptide'. *J. Chem. Theory Comput.* 5(6), Apr. 2009, p. 1667.

[See+22] P. Seeber, S. Seidenath, J. Steinmetzer and S. Gräfe. 'Growing Spicy ONIOMs: Extending and generalizing concepts of ONIOM and many body expansions'. *WIREs Comput. Mol. Sci.*, Nov. 2022.

[SF95] H. B. Schlegel and M. J. Frisch. 'Transformation between Cartesian and pure spherical harmonic Gaussians'. *Int. J. Quant. Chem.* 54(2), Apr. 1995, p. 83.

[SG22] U. Seidler and M. Griebel. 'A Dimension-adaptive Combination Technique for Uncertainty Quantification', 12 Apr. 2022. arXiv: `2204.05574 [math.NA]`.

[SJ20] D. Schmitt-Monreal and C. R. Jacob. 'Frozen-density embedding-based many-body expansions'. *Int. J. Quant. Chem.* 120(21), Apr. 2020.

[SJ21] D. Schmitt-Monreal and C. R. Jacob. 'Density-Based Many-Body Expansion as an Efficient and Accurate Quantum-Chemical Fragmentation Method: Application to Water Clusters'. *J. Chem. Theory Comput.* 17(7), July 2021, p. 4144.

[SL97] M. Schütz and R. Lindh. 'An integral direct, distributed-data, parallel MP2 algorithm'. *Theor. Chim. Acta.* 95(1-2), Jan. 1997, p. 13.

[Sla29] J. C. Slater. 'The Theory of Complex Spectra'. *Phys. Rev.* 34(10), Nov. 1929, p. 1293.

[Sla30] J. C. Slater. 'Atomic Shielding Constants'. *Phys. Rev.* 36(1), July 1930, p. 57.

[Smo63] S. A. Smolyak. 'Quadrature and interpolation formulas for tensor products of certain classes of functions'. *Dokl. Akad. Nauk* 148(5), 1963, p. 1042.

[SO89] A. Szabo and N. S. Ostlund. *Modern Quantum Chemistry: Introduction to Advanced Electronic Structure Theory*. First Edition, Revised. McGraw-Hill Inc., New York, 1989.

[Sön+96] F. D. Sönnichsen, C. I. DeLuca, P. L. Davies and B. D. Sykes. 'Refined solution structure of type III antifreeze protein: hydrophobic groups may be involved in the energetics of the protein–ice interaction'. *Structure* 4(11), Nov. 1996, p. 1325.

[Sön+97] F. D. Sönnichsen, C. I. DeLuca, P. L. Davies and B. D. Sykes. *NORTH-ATLANTIC OCEAN POUT ANTIFREEZE PROTEIN TYPE III ISOFORM HPLC12 MUTANT, NMR, MINIMIZED AVERAGE STRUCTURE*. Version 1.3. Apr. 1997. URL: `https://www.rcsb.org/structure/1kdf`.

[Sparse] Sparse developers. *Sparse*. Revision 94d196c3. 2018. URL: `https://sparse.pydata.org/en/stable/`.

[ST09] H. M. Senn and W. Thiel. 'QM/MM Methods for Biomolecular Systems'. *Angew. Chem. Int. Ed* 48(7), Jan. 2009, p. 1198.

[Sta12] R. P. Stanley. *Enumerative Combinatorics*. 2nd ed. Vol. 1. Cambridge University Press, 2012.

[Ste+94] P. J. Stephens, F. J. Devlin, C. F. Chabalowski and M. J. Frisch. 'Ab Initio Calculation of Vibrational Absorption and Circular Dichroism Spectra Using Density Functional Force Fields'. *J. Phys. Chem.* 98(45), Nov. 1994, p. 11623.

[Sun+17] Q. Sun, T. C. Berkelbach, N. S. Blunt, G. H. Booth, S. Guo, Z. Li, J. Liu, J. D. McClain, E. R. Sayfutyarova, S. Sharma, S. Wouters and G. K.-L. Chan. 'PySCF: the Python-based simulations of chemistry framework'. *WIREs Comput. Mol. Sci.* 8(1), Sept. 2017.

[Sun+20] Q. Sun, X. Zhang, S. Banerjee, P. Bao, M. Barbry, N. S. Blunt, N. A. Bogdanov, G. H. Booth, J. Chen, Z.-H. Cui, J. J. Eriksen, Y. Gao, S. Guo, J. Hermann, M. R. Hermes, K. Koh, P. Koval, S. Lehtola, Z. Li, J. Liu, N. Mardirossian, J. D. McClain, M. Motta, B. Mussard, H. Q. Pham, A. Pulkin, W. Purwanto, P. J. Robinson, E. Ronca, E. R. Sayfutyarova, M. Scheurer, H. F. Schurkus, J. E. T. Smith, C. Sun, S.-N. Sun, S. Upadhyay, L. K. Wagner, X. Wang, A. White, J. D. Whitfield, M. J. Williamson, S. Wouters, J. Yang, J. M. Yu, T. Zhu, T. C. Berkelbach, S. Sharma, A. Y. Sokolov and G. K.-L. Chan. 'Recent developments in the PySCF program package'. *J. Chem. Phys.* 153(2), July 2020, p. 024109.

[Sun15] Q. Sun. 'Libcint: An efficient general integral library for Gaussian basis functions'. *J. Comput. Chem.* 36(22), June 2015, p. 1664.

[Sve+96] M. Svensson, S. Humbel, R. D. J. Froese, T. Matsubara, S. Sieber and K. Morokuma. 'ONIOM: A Multilayered Integrated MO + MM Method for Geometry Optimizations and Single Point Energy Predictions. A Test for Diels-Alder Reactions and $Pt(P(t\text{-}Bu)_3)_2 + H_2$ Oxidative Addition'. *J. Phys. Chem.* 100(50), Jan. 1996, p. 19357.

[SW86] D. Stanton and D. White. *Constructive Combinatorics*. Springer New York, 1986.

[Taj+04] A. Tajti, P. G. Szalay, A. G. Császár, M. Kállay, J. Gauss, E. F. Valeev, B. A. Flowers, J. Vázquez and J. F. Stanton. 'HEAT: High accuracy extrapolated *ab initio* thermochemistry'. *J. Chem. Phys.* 121(23), Dec. 2004, p. 11599.

[Tar+01] G. Tarczay, A. G. Császár, W. Klopper and H. M. Quiney. 'Anatomy of relativistic energy corrections in light molecular systems'. *Mol. Phys.* 99(21), Nov. 2001, p. 1769.

[Tho+19] J. H. Thorpe, C. A. Lopez, T. L. Nguyen, J. H. Baraban, D. H. Bross, B. Ruscic and J. F. Stanton. 'High-accuracy extrapolated *ab initio* thermochemistry. IV. A modified recipe for computational efficiency'. *J. Chem. Phys.* 150(22), June 2019, p. 224102.

[Tho+21] J. H. Thorpe, J. L. Kilburn, D. Feller, P. B. Changala, D. H. Bross, B. Ruscic and J. F. Stanton. 'Elaborated thermochemical treatment of HF, CO, N_2, and H_2O: Insight into HEAT and its extensions'. *J. Chem. Phys.* 155(18), Nov. 2021, p. 184109.

[TM11] E. B. Tadmor and R. E. Miller. *Modeling Materials: Continuum, Atomistic and Multiscale Techniques*. Cambridge University Press, 24 Nov. 2011.

[Tol21] P. Tol. *Qualitative colour schemes*. 2 Apr. 2021. URL: `https://personal.sron.nl/~pault/`.

[Ton09] L.-D. Tong. 'The forcing hull and forcing geodetic numbers of graphs'. *Discrete Appl. Math.* 157(5), Mar. 2009, p. 1159.

[Tou17] J. Toulouse. *Introduction to perturbation theory and coupled-cluster theory for electron correlation*. Lecture notes (`https://www.lct.jussieu.fr/pagesperso/toulouse/enseignement/`). 11 Oct. 2017. URL: `https://www.lct.jussieu.fr/pagesperso/toulouse/enseignement/introduction_pt_cc.pdf`. Date of original access unknown. As of 29/9/23, the linked content carries the date 11/10/2019, but is, to our reading, otherwise identical to that to which we have referred.

[Tsc06] G. S. Tschumper. 'Multicentered integrated QM:QM methods for weakly bound clusters: An efficient and accurate 2-body:many-body treatment of hydrogen bonding and van der Waals interactions'. *Chem. Phys. Lett.* 427(1-3), Aug. 2006, p. 185.

[TT08] A. Thompson and B. N. Taylor. *Guide for the Use of the International System of Units (SI)*. Tech. rep. 2008 Edition. National Institute of Standards and Technology, Gaithersburg, MD, Mar. 2008. NIST Special Publication 811.

[Tul00] J. C. Tully. 'Perspective on "Zur Quantentheorie der Molekeln"'. *Theor. Chim. Acta.* 103(3-4), Feb. 2000, p. 173.

[TV19] M. A. Tortorici and D. Veesler. 'Structural insights into coronavirus entry'. In: F. A. Rey, ed. *Complementary Strategies to Understand Virus Structure and Function*. Vol. 105. Advances in Virus Research. Elsevier, 2019, p. 93.

[TW18] R. Tempone and S. Wolfers. 'Smolyak's Algorithm: A Powerful Black Box for the Acceleration of Scientific Computations'. In: J. Garcke, D. Pflüger, C. G. Webster and G. Zhang, eds. *Sparse Grids and Applications — Miami 2016*. Vol. 123. Lecture Notes in Computational Science and Engineering. Springer International Publishing, Cham, 2018, p. 201.

[use13] user2357112 (pseud.) *Indexing subsets of size k [duplicate]*. Stack Overflow. Version of 2013-11-30. The author's profile can be found at `https://stackoverflow.com/users/2357112/user2357112`. 30 Nov. 2013. URL: `https://stackoverflow.com/a/20296042`.

[Var07] A. J. C. Varandas. 'Extrapolating to the one-electron basis-set limit in electronic structure calculations'. *J. Chem. Phys.* 126(24), June 2007, p. 244105.

[Var18] A. J. C. Varandas. 'Straightening the Hierarchical Staircase for Basis Set Extrapolations: A Low-Cost Approach to High-Accuracy Computational Chemistry'. *Annu. Rev. Phys. Chem.* 69(1), Apr. 2018, p. 177.

[Vel93] M. L. J. van de Vel. *Theory of Convex Structures*. Vol. 50. North-Holland Mathematical Library. Elsevier, Amsterdam, 1993.

[Vin+23] V. Vinod, S. Maity, P. Zaspel and U. Kleinekathöfer. 'Multi-Fidelity Machine Learning for Excited State Energies of Molecules', 18 May 2023. arXiv: 2305.11292 [physics.chem-ph].

[Vir+20] P. Virtanen, R. Gommers, T. E. Oliphant, M. Haberland, T. Reddy, D. Cournapeau, E. Burovski, P. Peterson, W. Weckesser, J. Bright, S. J. van der Walt, M. Brett, J. Wilson, K. J. Millman, N. Mayorov, A. R. J. Nelson, E. Jones, R. Kern, E. Larson, C. J. Carey, İ. Polat, Y. Feng, E. W. Moore, J. VanderPlas, D. Laxalde, J. Perktold, R. Cimrman, I. Henriksen, E. A. Quintero, C. R. Harris, A. M. Archibald, A. H. Ribeiro, F. Pedregosa, P. van Mulbregt and SciPy 1.0 Contributors. 'SciPy 1.0: Fundamental Algorithms for Scientific Computing in Python'. *Nat. Methods* 17, 2020, p. 261.

[VLH19] S. P. Veccham, J. Lee and M. Head-Gordon. 'Making many-body interactions nearly pairwise additive: The polarized many-body expansion approach'. *J. Chem. Phys.* 151(19), Nov. 2019.

[VM03] T. Vreven and K. Morokuma. 'Investigation of the $S_0 \rightarrow S_1$ excitation in bacteriorhodopsin with the ONIOM(MO:MM) hybrid method'. *Theor. Chim. Acta.* 109(3), Apr. 2003, p. 125.

[Vre+06] T. Vreven, K. S. Byun, I. Komáromi, S. Dapprich, J. A. Montgomery, K. Morokuma and M. J. Frisch. 'Combining Quantum Mechanics Methods with Molecular Mechanics Methods in ONIOM'. *J. Chem. Theory Comput.* 2(3), Apr. 2006, p. 815.

[W4-17] A. Karton, N. Sylvetsky and J. M. L. Martin. *W4-17 Database*. First accessed in August 2021. url: https://www.chemtheorist.com/w4-17-database.html.

[Wal+20a] A. C. Walls, Y. J. Park, M. A. Tortorici, A. Wall, A. T. McGuire and D. Veesler. *Structure of the SARS-CoV-2 spike glycoprotein (closed state)*. Version 2.1. Mar. 2020. url: https://www.rcsb.org/structure/6VXX.

[Wal+20b] A. C. Walls, Y.-J. Park, M. A. Tortorici, A. Wall, A. T. McGuire and D. Veesler. 'Structure, Function, and Antigenicity of the SARS-CoV-2 Spike Glycoprotein'. *Cell* 181(2), Apr. 2020, p. 281.

[WCC08] R. C. Walker, M. F. Crowley and D. A. Case. 'The implementation of a fast and accurate QM/MM potential method in Amber'. *J. Comput. Chem.* 29(7), 2008, p. 1019.

[WD93] D. E. Woon and T. H. Dunning Jr. 'Gaussian basis sets for use in correlated molecular calculations. III. The atoms aluminum through argon'. *J. Chem. Phys.* 98(2), Jan. 1993, p. 1358.

[WD95] D. E. Woon and T. H. Dunning Jr. 'Gaussian basis sets for use in correlated molecular calculations. V. Core-valence basis sets for boron through neon'. *J. Chem. Phys.* 103(11), Sept. 1995, p. 4572.

[Wei35] L. Weisner. 'Abstract theory of inversion of finite series'. *Trans. Am. Math. Soc.* 38(3), Mar. 1935, p. 474.

[Wer05] S. Wernicke. 'A Faster Algorithm for Detecting Network Motifs'. In: R. Casadio and G. Myers, eds. *Algorithms in Bioinformatics. WABI 2005*. Vol. 3692. Lecture Notes in Computer Science. Springer Berlin Heidelberg, Berlin, Heidelberg, 2005, p. 165.

[WH95] A. T. Wong and R. J. Harrison. 'Approaches to large-scale parallel self-consistent field calculations'. *J. Comput. Chem.* 16(10), Oct. 1995, p. 1291.

[WHM10] S. N. Weiss, L. Huang and L. Massa. 'A generalized higher order kernel energy approximation method'. *J. Comput. Chem.*, June 2010, p. 2889.

[WHR96] A. T. Wong, R. J. Harrison and A. P. Rendell. 'Parallel direct four-index transformations'. *Theor. Chim. Acta.* 93(6), June 1996, p. 317.

[Wil80] S. Wilson. 'Fourth-order invariant in Rayleigh-Schrödinger perturbation theory'. *Int. J. Quant. Chem.* 18(3), Sept. 1980, p. 905.

[WL76] A. Warshel and M. Levitt. 'Theoretical studies of enzymic reactions: Dielectric, electrostatic and steric stabilization of the carbonium ion in the reaction of lysozyme'. *J. Mol. Biol.* 103(2), May 1976, p. 227.

[WMD96] A. K. Wilson, T. van Mourik and T. H. Dunning Jr. 'Gaussian basis sets for use in correlated molecular calculations. VI. Sextuple zeta correlation consistent basis sets for boron through neon'. *J. Mol. Struct. THEOCHEM* 388, Dec. 1996, p. 339.

[Won16] M. Y. L. Wong. 'Theory of the sparse grid combination technique'. PhD thesis. Australian National University, Sept. 2016.

[WT10] B. Wang and D. G. Truhlar. 'Combined Quantum Mechanical and Molecular Mechanical Methods for Calculating Potential Energy Surfaces: Tuned and Balanced Redistributed-Charge Algorithm'. *J. Chem. Theory Comput.* 6(2), Jan. 2010, p. 359.

[WW93] T. A. Wesolowski and A. Warshel. 'Frozen Density Functional Approach for *ab Initio* Calculations of Solvated Molecules'. *J. Phys. Chem.* 97(30), July 1993, p. 8050.

[WW95] G. W. Wasilkowski and H. Woźniakowski. 'Explicit Cost Bounds of Algorithms for Multivariate Tensor Product Problems'. *J. Complexity* 11(1), Mar. 1995, p. 1.

[WX12] A. Wu and X. Xu. 'DCMB that combines divide-and-conquer and mixed-basis set methods for accurate geometry optimizations, total energies, and vibrational frequencies of large molecules'. *J. Comput. Chem.* 33(16), Apr. 2012, p. 1421.

[Xan94] S. S. Xantheas. '*Ab initio* studies of cyclic water clusters $(H_2O)_n$, n = 1–6. II. Analysis of many-body interactions.' *J. Chem. Phys.* 100(10), May 1994, p. 7523.

[Xia+21] C. Xiao, S. Wang, W. Liu, X. Wang and E. Casseau. 'An Optimal Algorithm for Enumerating Connected Convex Subgraphs in Acyclic Digraphs'. *IEEE Transactions on Circuits and Systems II: Express Briefs* 68(1), Jan. 2021, p. 261.

[xTBPy] *Python API for the extended tight binding program*. GitHub. Version 20.1. URL: `https://github.com/grimme-lab/xtb-python`.

[Yan91] W. Yang. 'Direct calculation of electron density in density-functional theory'. *Phys. Rev. Lett.* 66(11), Mar. 1991, p. 1438.

[YL95] W. Yang and T.-S. Lee. 'A density-matrix divide-and-conquer approach for electronic structure calculations of large molecules'. *J. Chem. Phys.* 103(13), Oct. 1995, p. 5674.

[Yse10] H. Yserentant. *Regularity and Approximability of Electronic Wave Functions*. Vol. 2000. Lecture Notes in Mathematics. Springer Berlin Heidelberg, Berlin, Heidelberg, 2010.

[Yu+17] K. Yu, C. M. Krauter, J. M. Dieterich and E. A. Carter. 'Density and Potential Functional Embedding: Theory and Practice'. In: *Fragmentation: Toward Accurate Calculations on Complex Molecular Systems*. Ed. by M. S. Gordon. First edition. John Wiley & Sons Ltd., 2017. Chap. 2, p. 81.

[Zas+18] P. Zaspel, B. Huang, H. Harbrecht and O. A. von Lilienfeld. 'Boosting Quantum Machine Learning Models with a Multilevel Combination Technique: Pople Diagrams Revisited'. *J. Chem. Theory Comput.* 15(3), Dec. 2018, p. 1546.

[Zen91] C. Zenger. 'Sparse grids'. In: W. Hackbusch, ed. *Parallel algorithms for partial differential equations: proceedings of the Sixth GAMM-Seminar, Kiel, January 19-21, 1990*. Vol. 31. Notes on Numerical Fluid Mechanics. Vieweg-Verlag, Braunschweig, 1991, p. 241.

[Zha+12] Q. Zhang, W. Zhang, Y. Li, J. Wang, L. Zhang and T. Hou. 'A rule-based algorithm for automatic bond type perception'. *J. Cheminf.* 4(26), 31 Oct. 2012.

[Zha+21] J. H. Zhang, T. C. Ricard, C. Haycraft and S. S. Iyengar. 'Weighted-Graph-Theoretic Methods for Many-Body Corrections within ONIOM: Smooth AIMD and the Role of High-Order Many-Body Terms'. *J. Chem. Theory Comput.* 17(5), Apr. 2021, p. 2672.

[ZI22] X. Zhu and S. S. Iyengar. 'Graph Theoretic Molecular Fragmentation for Multi-dimensional Potential Energy Surfaces Yield an Adaptive and General Transfer Machine Learning Protocol'. *J. Chem. Theory Comput.* 18(9), Aug. 2022, p. 5125.

Index