Institut für Numerische Simulation
Rheinische Friedrich-Wilhelms-Universität Bonn

Wegelerstraße 6 • 53115 Bonn • Germany
phone +49 228 73-3427 • fax +49 228 73-7527
www.ins.uni-bonn.de

M. Griebel, J. Hamaekers, and F. Heber

BOSSANOVA: A bond order dissection approach for efficient electronic structure calculations

INS Preprint No. 0704

January 2008

BOSSANOVA: A BOND ORDER DISSECTION APPROACH FOR EFFICIENT ELECTRONIC STRUCTURE CALCULATIONS

MICHAEL GRIEBEL, JAN HAMAEKERS, AND FREDERIK HEBER

Abstract. In this article, we present a new decomposition approach for the efficient approximate calculation of the electronic structure problem for molecules. It is based on a dimension-wise decomposition of the space the underlying Schrödinger equation lives in, i.e. $\mathbb{R}^{3(M+N)}$, where M is the number of nuclei and N is the number of electrons. This decomposition is similar to the ANOVA-approach (analysis of variance) which is well-known in statistics. It represents the energy as a finite sum of contributions which depend on the positions of single nuclei, of pairs of nuclei, of triples of nuclei, and so on. Under the assumption of locality of electronic wave functions, the higher order terms in this expansion decay rapidly and may therefore be omitted. Furthermore, additional terms are eliminated according to the bonding structure of the molecule. This way, only the calculation of the electronic structure of local parts, i.e. small subsystems of the overall system (plus some additional saturation with hydrogen) is necessary to approximate the total ground state energy. To determine the necessary subsystems, we employ molecular graph theory combined with molecular binding knowledge. In principle, the local electronic subproblems may be approximately evaluated with whatever technique is appropriate, e.g. DFT, CC or CI. From these local energies, the total energy of the overall system is then approximately put together in a telescope-like fashion. Thus, if the size of the local subproblems is independent of the size of the overall molecular system, linear scaling is directly obtained. As the size of each subproblem depends on the bond coordination of the involved atoms, we coined the method BOSSANOVA (Bond Order diSSection ANOVA). We discuss the details of our new approach and apply it – based on state-of-the-art graph algorithms – to various test systems and to C- and BN-nanotube structures.

1. Introduction. The coupling of the micro- and the mesoscale of chemical reactions is currently a field of intensive research. Where the microscale is the realm of quantum mechanical effects, the mesoscale is described by statistical mechanics and macroscopic thermodynamics. Nevertheless, there are additional strong influences onto the mesoscale by effects from the microscale. Numerically, the microscale is usually treated with Hartree-Fock (HF), Coupled Cluster (CC), Configuration Interaction (CI) or Density Functional Theory (DFT) methods that yield approximate solutions to the underlying quantum-mechanical (QM) Schrödinger equation (SE), whereas the mesoscale is the realm of classical molecular mechanics (MM) methods that use Newton's mechanics with empirically fitted potential functions.

The ultimate goal would be a seamless coupling of quantum mechanical computations where needed and classical molecular mechanics simulations where sufficient. Such approaches are generally referred to as multi-scale methods, an extensive overview is given in [1]. Any starting point must be the general Schrödinger equation for the electrons and for the nuclei of the system under consideration. The Schrödinger equation however has a dimensional complexity of $3(M + N)$, where M denotes the number of nuclei and N denotes the number of electrons. This renders a direct numerical treatment impossible due to the curse of dimension and one has to resort to model approximations. As a first step, in the Born-Oppenheimer molecular dynamics (MD) approach, the wave functions of the nuclei and electrons are separated, the subsystem of the nuclei is treated classically with Newton's mechanics and the remaining $3N$-dimensional electronic Schrödinger equation is further approximated by some of the aforementioned methods. The potential needed for Newton's mechanics is obtained from the electronic solution by the Hellmann-Feynman theorem. This way, QM and MM are globally coupled. However, a global electronic QM solution is, at least for larger molecules, still too expensive as conventional methods scale with $\mathcal{O}(M^3)$ due to the underlying problem of matrix diagonalization.

To this end, general *linear scaling* electronic structure methods are employed to overcome the dimensionality problem. As a first step, for Gaussian basis sets, the use of the fast multipole method [2] has resulted in $\mathcal{O}(M^2 \log M)$ complexity. Furthermore, a cutoff radius such as for the MP2 theory [3] was used in a Divide&Conquer approach to take advantage of the exponential decay properties of electronic wave functions. Altogether, this resulted in linear scaling [4]. Another common method is the Density Matrix Minimization technique [5, 6]. There, the density matrix is unconstrainedly minimised via a conjugate gradient scheme, using idempotency and normalization. The Fock matrix is the minimized output, after off-diagonal elements have also been truncated at a cutoff radius. Also for plane wave basis sets the electronic localization in non-metallic systems can be exploited [7]. Again, a cutoff then allows for linear scaling. Note however that there is a crossover point up to which the standard cubic scaling approaches still perform faster due to smaller prefactors in their computational complexity counts [8].

In order to reduce the constants and thus to shift this crossover point, one tries to somehow further decompose the full global electronic Hamiltonian into local parts and employs local QM there. Let us briefly summarize the most common *decomposition approaches* in the following. One of the first, the Force-Matching Method by Ercolessi [9], tries to automatically generate empirical potentials by a least-square fitting of the forces of ab-initio calculations to general many-body potential forms such as that of the Embedded Atom approach [10] or that of Abell-Tersoff [11, 12]. Due to its big parameter space, the method yields precise results only for a huge data set. Then, there is a range of methods which employ a decomposition[1] directly in $\mathbb{R}^3$: An early idea is the SIBFA (Sum of Interactions Between Fragments computed Ab initio) procedure [13], where the molecular system is additively built from constitutive molecular fragments. The electrostatic and polarization components are calculated using multipole expansions of the fragment electron density, the repulsion part is determined as a sum of bond and lone pair interactions. Furthermore, Morokuma proposed an ansatz called IMOMM [14] which combines two molecular orbital (MO) calculations to one calculation. Basically, it describes a telescopic sum over two regions, here Ω_1 and $\Omega_2 \subseteq \Omega_1$, where the energy is split as: $E^{ONIOM} = E^{MM}_{\Omega_1} + E^{QM}_{\Omega_2} - E^{MM}_{\Omega_2}$. Later variants and generalizations lead to the so-called ONIOM approach [15] where more regions were included, adding to the picture of the various layers of an onion. The method is simple, yet practical and has only little overhead. But to our knowledge no dynamic extension has been proposed yet. A generalization of SIBFA is the so-called Fragmentation Reconstruction Method (FRM) [16], where the interaction energy of a molecule is computed from hydrogenated components (closed-shell molecules). The procedure involves stringent chemical knowledge to choose the cuts as best as possible but to still keep the ground-state electronic density intact. Furthermore, a scheme for modeling the electrostatic impact of a passive MM environment on the active QM system is described in [17]. There, a short-range modified Coulomb potential is applied. However, there are difficulties choosing its functional form to obtain correct interaction properties and to avoid spill-out of electron density into the passive regime. Moreover, in [18] and [19] an interface regime between QM and MM with "link" atoms is proposed to account for the cutting of bonds. This, however, requires additional parametrization to reproduce geometrical and electronic properties on both scales and has to be adjusted from case to case. Similar techniques are used in [20, 21].

Another approach divides *time* instead of space in order to generate a coupling between QM and MM. One of these methods is learn-on-the-fly [22], which is similar to the Force-Matching method, but is run during the computation: At intermittent time steps certain clusters of the simulation domain are locally computed by QM, and the obtained local forces are used to correct the MM calculation. Note however that QM and MM involve different time step sizes. Here, especially the QM subproblems need very small time steps which causes additional complexity problems.

While all of the above methods have promising features, we feel that they generally either involve too many additional parameters, unchemically cut bonds in separating active from passive regions, or even worse, add unphysical pseudo-atoms in order to compensate for the different energy and time scales and to avoid spill-out effects of electronic density or energy. Moreover, they are plainly too simple or do not grasp the problem in its full complexity, if only a matching or interpolation with respect to energy or forces between the QM and MM parts of the overall approach is considered.

[1] Ultimately, the aim would be a decomposition of $\mathbb{R}^{3(M+N)}$, the space where the full Schrödinger equation lives in.

In this article, we propose a more sophisticated method. Our ansatz is sparked off the idea underlying Tersoff's [12] many-body potential, where the energy and the forces of an atom are assumed to depend on its local coordination. Here, for a proof-of-concept, we concentrate on covalent binding, hence on charge-neutral molecular systems and subsystems.[2] We will use this knowledge of coordination and bonds between nuclei to decompose the $R^{3(M+N)}$ of the underlying Schrödinger equation in a dimension-wise fashion. This decomposition is similar to the ANOVA-approach (analysis of variance) which is well-known in statistics. It represents the energy as a finite sum of contributions which depend on the positions of single nuclei, of pairs of nuclei, of triples of nuclei, and so on. Under the assumption of locality of electronic wave functions, the higher order terms in this expansion decay rapidly and may therefore be omitted. Furthermore, additional terms are eliminated according to the bonding structure of the molecule. This way, only the calculation of the electronic structure of local parts, i.e. small overlapping subsystems of the overall molecule system is necessary to approximate the total ground state energy. To determine the necessary subsystems, we employ molecular graph theory combined with molecular binding knowledge. This way, modern graph algorithms – that involve breadth-first and depth-first search techniques – are used to create proper local subproblems as overlapping fragments of the overall molecular system. Furthermore, hydrogenization is used to close shells and saturate bonds that have been cut. We thus also exploit locality, however not by an explicit cutoff radius as most conventional methods do, but by implicitly using it in the inherent bond structure of a molecular system. In principle, the local electronic subproblems may be approximately evaluated with whatever QM technique is appropriate, e.g. DFT, CC or CI. From these local energies, the total energy of the overall system is then approximately put together in a telescope-like fashion. Thus, if the size of the local subproblems is independent of the size of the overall molecular system, linear scaling is directly obtained. The $3(M+N)$-dimensional full global Hamiltonian is broken down within the Born-Oppenheimer Approximation to $\mathcal{O}(M)$ components, each with $M_i^{(k)}$ degrees of freedom, with an upper bound $\max_i\{M_i^{(k)}\}$ controlled by a single parameter k which we name the *bond order* of the approximation. This ansatz specifically combines the smaller prefactor of the cubic scaling methods with a general linear scaling behavior. As the size of each subproblem depends on the bond coordination of the involved atoms, we coined the method BOSSANOVA (Bond Order diSSection ANOVA).

The remainder of this article is organised as follows: In section 2 we briefly summarize the basics of the underlying Schrödinger equation. In section 3 we describe the ANOVA-like decomposition of the Schrödinger equation in the context of molecular graph theory. In section 4 we present the details of our graph-based algorithmic implementation. In section 5 we give numerical results of three model systems and of carbon and of boron-nitride nanotubes. We end with some concluding remarks in section 6.

2. Schrödinger equation in the Born-Oppenheimer approximation.

Let us consider a molecular system consisting of M nuclei and N electrons. Its time-dependent state function can be written in general as $\Psi = \Psi(\mathbf{R}_1, \ldots, \mathbf{R}_M, \mathbf{r}_1, \ldots, \mathbf{r}_N, t)$, where $\mathbf{R}_i$ and $\mathbf{r}_j$ denote positions in three-dimensional space $\mathbb{R}^3$ associated to the ith nucleus and the jth electron, respectively. The variable t denotes the time-dependency of the state function. The vector space (space of configurations), in which the coordinates of the particles are given, is therefore of dimension $3(M+N)$. In the following we will abbreviate $(\mathbf{R}_1, \ldots, \mathbf{R}_M)$ and $(\mathbf{r}_1, \ldots, \mathbf{r}_N)$ with the shorter notation $\mathbf{R}$ and $\mathbf{r}$, respectively. Also, we assume that Ψ is normalized to $\int \Psi^*(\mathbf{R}, \mathbf{r}, t)\Psi(\mathbf{R}, \mathbf{r}, t) d\mathbf{R} d\mathbf{r} = 1$.

Nuclei and electrons are charged particles. The electrostatic potential (Coulomb potential) of a point charge is $\frac{1}{r}$ in atomic units, where r is the distance from the position of the charged particle. An electron moving in this potential possesses the potential energy $V(r) = -\frac{1}{r}$. Neglecting spin and relativistic interactions and assuming that no external forces act on the system, the Hamilton operator in position

[2] Note however that our approach should work equally well also in the non-charge neutral case.

representation associated to the system of nuclei and electrons is given as the sum over the operators for the kinetic energy and the Coulomb potentials,

$$\mathcal{H}^{(N,M,Z_1,m_1,\dots,Z_M,m_M)}(\mathbf{R},\mathbf{r}) := \underbrace{-\frac{1}{2}\sum_{k=1}^{N}\Delta_{\mathbf{r}_k}+\sum_{k<j}^{N}\frac{1}{\|\,\mathbf{r}_k-\mathbf{r}_j\,\|}-\sum_{k=1}^{N}\sum_{j=1}^{M}\frac{Z_j}{\|\,\mathbf{r}_k-\mathbf{R}_j\,\|}+\sum_{k<j}^{M}\frac{Z_kZ_j}{\|\,\mathbf{R}_k-\mathbf{R}_j\,\|}}_{\mathcal{H}_e^{(N,M,Z_1,m_1,\dots,Z_M,m_M)}(\mathbf{R},\mathbf{r})}-\frac{1}{2}\sum_{k=1}^{M}\frac{1}{m_k}\Delta_{\mathbf{R}_k}, \tag{2.1}$$

where we use the number M of atoms, the number N of electrons, the nuclei mass in atomic units m_j and the atomic number Z_j as upper indices of $\mathcal{H}$ to distinguish between parameters and degrees of freedom. Here, $\|\mathbf{r}_k - \mathbf{r}_j\|$ are the distances between electrons, $\|\mathbf{r}_k - \mathbf{R}_j\|$ are distances between electrons and nuclei and $\|\mathbf{R}_k - \mathbf{R}_j\|$ are distances between nuclei. We will omit parameters from this list if they are clear from the context. This will later especially be $N, M, Z_1, m_1, \dots, Z_M, m_M$.

Now, a system of equations for the electronic and for the nuclei degrees of freedom is usually derived with the *Born-Oppenheimer approximation*. To this end, the large difference in masses between electrons and atomic nuclei is exploited to decouple the motion of the electrons from that of the nuclei.[3] Then, one assumes that the electrons adapt instantaneously to a change in the nuclear configuration and are thus always in the quantum mechanical ground state, denoted by $\phi_{(0)}^{(\mathbf{R}(t))}(\mathbf{r})$, which is associated to the actual position of the nuclei $\mathbf{R}(t)$. Note that this allows us to write $\mathcal{H}_e^{(\mathbf{R}(t))}(\mathbf{r})$ instead of $\mathcal{H}_e(\mathbf{R}(t),\mathbf{r})$ since the movement of the nuclei during the adaptation of the electron positions is negligibly small in the sense of classical dynamics. This justifies to set $\Psi(\mathbf{R},\mathbf{r},t) \approx \Psi^{BO}(\mathbf{R},\mathbf{r},t) := \sum_{j=0}^{\infty}\chi_j(\mathbf{R},t)\phi_j^{(\mathbf{R})}(\mathbf{r})$, which allows to separate the fast from the slow variables. We then obtain the following set of equations:

$$M_k\ddot{\mathbf{R}}_k(t) = -\nabla_{\mathbf{R}_k} \underbrace{\min_{\phi_{(0)}^{(\mathbf{R}(t))}}\left\{\int \phi_{(0)}^{(\mathbf{R}(t))^*}(\mathbf{r})\mathcal{H}_e^{(\mathbf{R}(t))}(\mathbf{r})\phi_{(0)}^{(\mathbf{R}(t))}(\mathbf{r})d\mathbf{r}\right\}}_{=:V_e^{BO}(\mathbf{R}(t))} \tag{2.2}$$

$$\mathcal{H}_e^{(\mathbf{R}(t))}(\mathbf{r})\phi_{(0)}^{(\mathbf{R}(t))}(\mathbf{r}) = E_0(\mathbf{R}(t))\phi_{(0)}^{(\mathbf{R}(t))}(\mathbf{r}). \tag{2.3}$$

In the end, after time discretization we have to perform in each time step the following tasks: First, we have to compute an approximate solution of the electronic Schrödinger equation in (2.3) for fixed positions $\mathbf{R}$ of the nuclei, then we have to compute from its solution the forces on the nuclei and finally the positions of the nuclei at the next time step by e.g. a Verlet time step for Newton's equations of motion of the nuclei in (2.2). To this end, we use the *Hellmann-Feynman Theorem* to obtain the electronic forces $\mathbf{F}_k(\mathbf{R}) = -\nabla_{\mathbf{R}_k}\int \phi_{(0)}^{(\mathbf{R}(t))^*}\mathcal{H}_e^{(\mathbf{R}(t))}\phi_{(0)}^{(\mathbf{R}(t))}d\mathbf{r}$ acting on the nuclei. Variants of this approach are the Ehrenfest molecular dynamics and the Car-Parrinello method. For details of the derivation, see [23] and references therein.

3. ANOVA decomposition scheme. So far, the Born-Oppenheimer molecular dynamics was employed to split the full Schrödinger problem into two parts, i.e. a classical Newton's equation of motion for the nuclei, and, in each discretized time step, the electronic problem of (2.3) which may approximately be solved by e. g. the Hartree Fock, Coupled Cluster, Configuration Interaction or Density Functional method, see [24, 25]. However, such an overall approach is only feasible for small molecules due to the high complexity of any approximate solution method for the electronic problem. To overcome this difficulty, the aforementioned coupling techniques and linear scaling methods had been developed. They

[3] The ratio of the velocity v_K of a nucleus to the velocity of an electron v_e is in general smaller than 10^{-2}.

basically all exploit locality of the electronic wave function in one way or another to reduce the complexity of the electronic problem. This excludes in general metallic systems, whose electrons may be delocalized due to a vanishing band gap.[4]

In the following, we also resort to a certain locality of the electronic wave function which is expressed in the bond structure of the molecular system and decompose the overall electronic problem into small subproblems which then may be handled efficiently. To this end, we introduce an ANOVA decomposition scheme of a molecular system into local parts by means of the bond order of the nuclei in the system. Basically, this involves a decomposition into the different many-body interactions. Here, we employ graph algorithms to derive a proper fragmentation of molecules and associated interaction energies[5] where we use bonding information and neighboring relations in the structure of molecular graphs in order to decide which terms to neglect in the ANOVA expansion of the ground state energy functional of the molecular system. As we assume the molecular system and its subsystems to be charge-neutral, the electronic locality is implicitly exploited there.

In the following subsections we describe this approach in more detail. We begin with an introduction to graph theory in subsection 3.1, where we gather necessary definitions. In subsection 3.2 we describe the total energy functional in its dependence on the nuclei coordinates and the method of ANOVA expansion of this functional. In section 3.3 we elaborate on a hydrogen saturation scheme which results in better convergence. The technical details of our algorithm used to obtain the proper molecular fragments in the ANOVA series expansion with a truncation to a certain bond order are given later in section 4.

3.1. Graph theory. Since we will use undirected graphs to represent the structure of molecular systems, we briefly give some common definitions for graphs, for further reading we refer to the literature [26, 27]. The atoms and bonds in a molecular system resemble sets of *vertices* V and *edges* K and form an *undirected graph* $G = (V, K)$. In the following, for any set B its cardinality shall be represented by $|B|$. Furthermore, the vertices of G shall be labeled, i.e. $V = \{v_1, \dots, v_{|V|}\}$, where $|V|$ is the number of atoms. Here, each v_i is uniquely associated to a nuclei with coordinate vector R_i. Then, for two vertices v_i and v_j of G there may be an edge $k = (v_i, v_j)$. These edges shall be labeled arbitrarily, i.e. $K = \{k_1, \dots, k_{|K|}\}$, where $|K|$ is the number of edges. The number $|K(v)|$ of edges $K(v) = \{(v_i, v_j) | v_i = v \vee v_j = v\}$ connected to a vertex v is called the *degree* or *valency* $d(v)$ of the vertex. Then, the *average degree* of G is the number $d(G) = \sum_{v \in \{V\}} d(v)/|V|$. Note that the mean number of covalent bonds per molecule is in general roughly equal to the number of atoms, i.e. $|K| \approx 2|V|$, and there exists an upper bound to it. A *subgraph* $G' = (V', K')$ is a subset of edges $K' \subseteq K$ and vertices $V' \subseteq V$ of a graph $G = (V, K)$. It is called an *induced subgraph* if it contains all edges $k = (v_i, v_j) \in K$ with $v_i, v_j \in V'$. In the following, we denote subgraphs of a graph G often with G' or G'' and, often, we further equip subgraphs with an index such as G'_i to emphasize their interrelation in a set of graphs $\{G'_i\}_{i \in I}$ with index set I.

For a graph G an associated *adjacency-list* or *-matrix* A is defined as

$$A_{ij} = \begin{cases} 1, \text{if } \exists k = (v_i, v_j) \in K, \\ 0, \text{if not.} \end{cases} \tag{3.1}$$

A is a symmetric matrix of dimension $|V|$.[6] A *path* is a non-empty sequence of edges $\{k_0, k_1, \dots, k_n\}$ so that the edges k_{i-1} and k_i for each $i = 1, 2, \dots$ have a common vertex v. The *length n of a path* for two given vertices v_i and v_j is defined as the number of edges of the connecting path. It is set to ∞ if there

[4] Furthermore, the notion of the locality of the wave function is important as it leads to the general chemical understanding of molecules from the general bond structure up to nucleophilic sites.

[5] Note that we neglect long-range Coulomb interaction in this article. Coulomb interaction may later be incorporated via Ewald summation or P3M techniques.

[6] In the case of a *sparse* matrix – i. e. for $|K| \sim |V|$ which is typical for molecules – it is recommended to store the adjacency matrix as a list for cost complexity reasons and as a matrix only in the case of a *dense* graph, i. e. for $|K| \sim |V|^2$.

is no such path. The *shortest path* between two vertices v_i and v_j is then the minimum of the lengths of all possible paths in G with these two vertices as endpoints, its length shall be denoted by $d_G(v_i, v_j)$. A non-empty graph is called *connected* if any two of its vertices v_i and v_j are linked by a path in G. A *cycle* is a closed path. Its vertices can be labeled $v_0, \dots, v_n$ such that the edges are $(v_{i-1}, v_i),\ \forall i \in \{1, \dots, n\}$ and $v_n = v_0$. It is also called a *circuit of length* n. A *forest* is a graph with no cycles, a connected forest is also called *tree*. If $A, B \subseteq V$ and $X \subseteq V \cup K$ are such that every path between A and B in G contains a vertex or edge from X, then X *separates* the sets A and B and X is called a *separator*. Two special kinds of separators are as follows: A *cutvertex* separates two other vertices of the given component. A *bridge* or tree edge k_i is an edge, for which there are two sets A, B that are separated by $X = \{k_i\}$, i. e. it is an edge that when removed separates the graph. Clearly, bridges are those edges that do not belong to any cycle.

Obviously, the structure of the graph which represents a given molecular system may be analyzed to a certain detail with the help of the above definitions.[7] To compute such properties of graphs, so-called depth-first search (DFS) and breadth-first search (BFS) methods are used which typically involve $\mathcal{O}(|V|)$ cost, see [27].

3.2. ANOVA expansion. We will now define the energy function for a molecular system and its ANOVA series expansion. To this end, consider a molecular system which consists of N electrons and M nuclei, each with coordinate vector $R_i \in \mathbb{R}^3$ and atomic number $Z_i \in \mathbb{N}$, $i \in \{1, \dots, M\}$. Let G be the associated graph that represents the bond structure of this molecular system. This graph may be derived from the molecular system as later described in section 4.2. Furthermore, we restrict ourselves to charge-neutral systems, i.e. the number of electrons N is equal to $\sum_i^M Z_i$ and we assume that the associated graph is connected, both for reasons of simplicity. Finally, we consider the systems only in their electronic ground state in the framework of the Born-Oppenheimer molecular dynamics. To this end, we separate the time-independent electronic Schrödinger equation as in (2.3) and define a total ground state energy function $E^{(M)} : (\mathbb{N} \times \mathbb{R}^3)^M \to \mathbb{R}$. It depends on the parameters that completely identify the system under consideration, namely the coordinates R_i and the atomic number Z_i of each nuclei with fixed and unique label $i \in \{1, \dots, M\}$, i.e.

$$E^{(M)}(\underbrace{(Z_1, R_1)}_{=:\widetilde{R}_1}, \dots, \underbrace{(Z_M, R_M)}_{=:\widetilde{R}_M}) :=$$

$$\min_{\left|\phi_{(0)}^{(\mathbf{R}(t))}\right|=1} \int \phi_{(0)}^{(\mathbf{R}(t))^*}(\mathbf{r}) \mathcal{H}_e^{(N=\sum_{i=1}^M Z_i, (Z_1,R_1),\dots,(Z_M,R_M))} \phi_{(0)}^{(\mathbf{R}(t))}(\mathbf{r}) d\mathbf{r}. \quad (3.2)$$

In the following, we further simplify the notation by defining $\widetilde{R}_i := (Z_i, R_i)$, i. e. $\widetilde{R}_i$ combines the atomic number and the coordinates of the nuclei i. Note that, due to the charge-neutrality condition $N = \sum_i^M Z_i$, the parameter N may now be eliminated from the parameter list of the Hamiltonian $\mathcal{H}$.

Now we will decompose the function $E^{(M)}$ in a multivariate telescopic sum, i.e. in a finite series expansion in the nucleic parameters, in the same way as the ANOVA decomposition (analysis of variance)[8] which is well-known in statistics [28]. This decomposition involves a splitting of the M-dimensional function into contributions which depend on the positions of single nuclei and associated charges, of

[7] Note that to each graph there is an associated dual graph, where the edges are the vertices of the primal graph and the vertices are the edges. We may thus define all of the above edge relations also for vertices and vice versa.

[8] The ANOVA decomposition of a M-dimensional function $f : [0,1]^M \to \mathbb{R}$ reads $f = \sum_{\mathbf{u} \subseteq \{1,\dots,M\}} f_{\mathbf{u}}$ with $f_{\mathbf{u}}$ depending only on the variables indicated in $\mathbf{u}$. The functions $f_{\mathbf{u}}$ satisfy the recurrence relation $f_{\{\}} = L_{\{1,\dots,M\}}(f)$, $f_{\mathbf{u}} = L_{\{1,\dots,M\}/\mathbf{u}}(f) - \sum_{\mathbf{v} \subsetneq \mathbf{u}} f_{\mathbf{v}}$ with $L_{\mathbf{w}}(f) = (\prod_{j \in \mathbf{w}} L_j)(f)$ where $L_j(f)(x_1, ..., x_M) = \int_0^1 (x_1, ..., x_M) dx_j$. Thus, f is decomposed into a constant, a sum of one-dimensional functions, a sum of two-dimensional functions, and so on. The involved functions are generated by proper partial integration and telescopic corrections according to the recurrence relation.

pairs of nuclei and associated charges, of triples of nuclei and charges, and so on. To this end, we consider the subset of the nuclei parameters $\{\widetilde{R}_i\}_{i\in I}$ described by a set of labels I with cardinality $|I| = k$ and call it the *molecular fragment* associated to I with size k. Note that we here need not to consider the electronic degrees of freedom $\mathbf{r}$, as the system is assumed to be in ground state and, hence, the electronic state functions are all fixed by the minimum condition in (3.2).

First, we define the total electronic ground state energy of lower-dimensional subsystems of the molecular system under consideration, described by the set of indices $I = \{i_1, \dots, i_k\}$,

$$E_{i_1,\dots,i_k}(\widetilde{R}_1, \dots, \widetilde{R}_k) := \min_{|\phi_{(0)}|=1} \int \phi^*_{(0)}(\mathbf{r})\mathcal{H}_e^{(\sum_{j=1}^k Z_{i_j},(Z_{i_1},\widetilde{R}_{i_1}),\dots,(Z_{i_k},\widetilde{R}_{i_k}))}\phi_{(0)}(\mathbf{r})d\mathbf{r}. \tag{3.3}$$

Note that this is in form very similar to (3.2). In the notation of the electronic ground state wave functions $\phi_{(0)}$, the dependency on $\mathbf{R}(\mathbf{t})$ was dropped as it is clear from the context. Furthermore, each $E_{i_1,\dots,i_k}$ still depends on the whole graph G as a parameter, which is not indicated explicitly here to simplify notation.

Then, the energy function $E^{(M)}$ is decomposed analogously to the ANOVA approach as

$$\begin{aligned} E^{(M)}(\widetilde{R}_1, \dots, \widetilde{R}_M) &= F_0 \\ &+ \sum_{i_1}^{M} F_{i_1}(\widetilde{R}_{i_1}) \\ &+ \sum_{i_1<i_2}^{M} F_{i_1,i_2}(\widetilde{R}_{i_1,i_2}) \\ &+ \sum_{i_1<i_2<i_3}^{M} F_{i_1,i_2,i_3}(\widetilde{R}_{i_1,i_2,i_3}) \\ &+ \dots \\ &+ F_{i_1,\dots,i_M}(\widetilde{R}_{i_1,\dots,i_M}) \\ &=: \sum_{U\subseteq\{1,\dots,M\}} F_U(\widetilde{R}_U), \end{aligned} \tag{3.4}$$

where R_U denotes the set of variables $\{R_i\}_{i\in U}$ and $U \subseteq \{1, \dots, M\}$.

Here, each term $F_{i_1,\dots,i_k}$ is defined as follows:

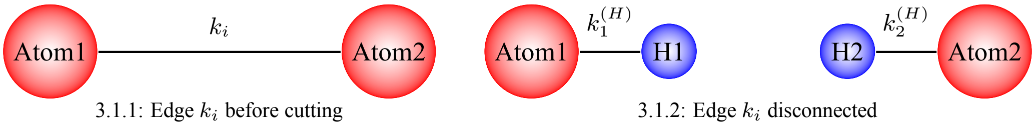
$$
\begin{aligned}
F_0 &= 0 \\
F_{i_1}(\widetilde{R}_{i_1}) &= E_{i_1}(\widetilde{R}_{i_1}) - F_0 \\
F_{i_1,i_2}(\widetilde{R}_{i_1,i_2}) &= E_{i_1,i_2}(\widetilde{R}_{i_1,i_2}) - F_{i_1}(\widetilde{R}_{i_1}) - F_{i_2}(\widetilde{R}_{i_2}) - F_0 \\
\dots \quad & \dots \\
F_{i_1,\dots,i_k}(\widetilde{R}_{i_1,\dots,i_k}) &= E_{i_1,\dots,i_k}(\widetilde{R}_{i_1,\dots,i_k}) \\
&- \sum_{U\subseteq I,|U|=k-1} F_U(\widetilde{R}_U) \\
&- \sum_{U\subseteq I,|U|=k-2} F_U(\widetilde{R}_U) \\
&\dots \\
&- \sum_{U\subseteq I,|U|=1} F_U(\widetilde{R}_U) - F_0 \\
\dots \quad & \dots,
\end{aligned}
\tag{3.5}
$$

where the constant function F_0 is set equal to zero since it corresponds to the energy of an empty molecular system.

Let us note that the decomposition is exact and contains 2^M different terms due to the power set construct. In general it might be that all terms are equally important up to the last, M-dimensional one, or in the extreme case, that the last term might be the only important one and thus nothing is gained from this decomposition. However, if the size of the terms decay fast with e.g. the order of the terms, then a proper truncation of the ANOVA series expansion results in a substantial reduction in computational complexity. We then only have to deal with a sequence of lower-dimensional subproblems which are associated to the remaining lower-dimensional energy terms of the decomposition. To this end, let us remark that the energy functions $F_{i_1,\dots,i_k}$ in (3.4) may be recognized as an expansion of many-body interaction contributions, as in [29]. This leads us to the following assumption which is central to our further approach: There is a certain decay in the contribution of each order of the ANOVA expansion and this results in a monotone convergence of the approximation error with rising order. Consequently, from a certain order onward, we may neglect the higher higher-order terms in the ANOVA decomposition. This will gain a good approximation to the true result[9] with an accuracy which is related to the order parameter at which the truncation was executed. This assumption is also strongly supported by the success of conventional two- and many-body potential functions used in classical molecular dynamics, such as short range pair-potentials like harmonic springs, the Morse potential and the Lennard-Jones potential, three- and four-body potential like angle and dihedral potential functions and more advanced many-body potential functions which involve a local coordination number (that is the local density of atoms) like Tersoff's potential [12], the embedded atom method [10] or Brenner's reactive bond order potential for hydrocarbons [30]. Here, in any case, only a small number of neighboring atoms are involved in the potential forms, for further details see [23].

Thus, our ansatz is as follows: We decompose the total energy function (3.2) in an ANOVA series expansion as in (3.4) and truncate this series to a certain order k, which we call the *bond order* of the approximation. Note that each set $I = \{i_1, \dots, i_{|I|}\}$ of nucleic parameters indices for each term $E_{i_1,\dots,i_{|I|}}(\widetilde{R}_{i_1,\dots,i_{|I|}})$ in (3.5) is directly connected to an induced subgraph $G_I = (V_I, K_I)$ of the total graph G with $V_{i_j} = \{v_{i_j}\}_{i_j \in I}$. On top of that, we neglect in a second step even further terms in the

[9] Note that, in practise, the global electronic problem is only solved approximately anyway, by e.g. DFT, CC, CI.



3.1.1: Edge k_i before cutting

3.1.2: Edge k_i disconnected

FIGURE 3.1. *Cut of an edge k_i between two vertices and replacement with two edges $k_1^{(H)}$ and $k_2^{(H)}$ to two newly introduced terminal vertices (hydrogen atoms H1 and H2).*

truncated series. These terms contain as parameters the degrees of freedom which belong to nuclei in I that are not connected by a path in the graph G_I, i. e. we additionally eliminate those terms whose induced subgraph G_I is not connected. This second elimination step is motivated by the locality of the electronic wave functions: Atoms that share a bond to a nearby atom will be much influenced by changes in the chemical vicinity of nearest or next-nearest binding partners whereas atoms that share no bond to a nearby atom will not.

The remaining terms in our decomposition will be determined from the bond order parameter and from the graph of the overall molecular system by means of modern graph algorithms whose details will be explained later in subsection 4.

3.3. Saturation with hydrogen. After the motivation of the basic principles of our decomposition scheme in the last section, we now have to face a technical difficulty: Let us consider the behavior of our approach for simple organic chain-like molecules, i.e. for various n-alkane molecules up to heptane, which are well suited to our proposed dissection scheme. In particular, due to their linear chain structure, there should be a clear decay of the magnitude of higher order contributions in the ANOVA expansion and thus a monotone convergence behavior of our approach with rising order k. For details on the algorithm, the parameters and the approximate DFT-solver, we refer to section 5. The resulting approximate energy and relative error, calculated by our BOSSANOVA series expansion approach (3.4), truncated to k-th order, is given in the upper half of table 3.1. We clearly see the anticipated convergence with rising values of k to the solution of the full electronic problem, but the results are not yet completely satisfactory with respect to the monotonicity of the convergence. One reason for that is the inept direct cutting of bonds at the end of fragments. A general concept in most fragmentation schemes is the conservation of the total electronic ground state density within each fragment as best as possible. This is also the central guideline in the choice of cuts in the Fragmentation Reconstruction Method (FRM) [16]. Clearly, if we create fragments and induced subgraphs in the ANOVA approach, we remove atoms and electrons and thus significantly change the local ground state density in the fragment with respect to the total ground state density of the molecular system.

A step to remedy this situation is a saturation of the dangling bonds of the fragments by adding hydrogen at the places where edges were cut. This correction is schematically depicted in figure 3.1 where we just show two atoms and its vertex but omitted for simplicity any further vertices and edges these atoms might be connected to. Here, let us assume that, after cutting the edge k_i, Atom1 should belong to an induced subgraph G', while Atom2 should not. Then, edge $k_i = (\text{Atom1}, \text{Atom2})$ is not present in this subgraph. Now, we insert two new terminal vertices H1 and H2 and two new edges $k_1^{(H)} = (\text{Atom1}, \text{H1})$ and $k_2^{(H)} = (\text{Atom2}, \text{H2})$ so that all dangling bonds are closed. Hence, the new vertex H1 and the edge $k_1^{(H)}$ would be added to G' next to Atom1. By this saturation procedure, we only calculate closed-shell atoms. In particular, the electronic density of the cut edges is thus conserved to a higher degree. Note that this approach is still tunable by the bond length used between new hydrogen vertices and cutvertices. In our subsequent implementation we use here the equilibrium hydrogen bond lengths of certain small molecules taken from [31].

This procedure can be understood as a re-definition of the electronic Hamiltonian $\mathcal{H}_e$ in (3.3): From the known graph G of the molecular system l additional hydrogen vertices, bonds and their graph-

TABLE 3.1

Hydrogen saturation: *Convergence of approximated n-alkane total energies in Hartree and relative error with respect to full DFT calculation with k-truncated ANOVA expansion. From top to bottom: Total energy without saturation, relative error without saturation, total energy with saturation, relative error with saturation. From left to right the order k of the truncation of the ANOVA series expansion from 1 (single-body only) to 6 (up to six-body contributions) and the full DFT calculation of the given molecule.*

Order/ Molecule	1	2	3	4	5	6	full DFT
C_7H_{16}	-48.0531	-49.7134	-49.1314	-49.2745	-49.3436	-49.3478	-49.3398
C_6H_{14}	-41.3865	-42.7421	-42.2959	-42.4153	-42.4629		-42.4570
C_5H_{12}	-34.7199	-35.7708	-35.4604	-35.5560			-35.5743
C_4H_{10}	-28.0533	-28.7996	-28.6249				-28.6915
C_3H_8	-21.3867	-21.8283					-21.8086
C_2H_6	-14.7201						-14.9251
C_7H_{16}	$2.6078 \cdot 10^{-2}$	$7.5722 \cdot 10^{-3}$	$4.2246 \cdot 10^{-3}$	$1.3235 \cdot 10^{-3}$	$7.7828 \cdot 10^{-5}$	$1.6255 \cdot 10^{-4}$	
C_6H_{14}	$2.5214 \cdot 10^{-2}$	$6.7149 \cdot 10^{-3}$	$3.7958 \cdot 10^{-3}$	$9.8382 \cdot 10^{-4}$	$1.3732 \cdot 10^{-4}$		
C_5H_{12}	$2.4016 \cdot 10^{-2}$	$5.5262 \cdot 10^{-3}$	$3.2009 \cdot 10^{-3}$	$5.1245 \cdot 10^{-4}$			
C_4H_{10}	$2.2243 \cdot 10^{-2}$	$3.7669 \cdot 10^{-3}$	$2.3209 \cdot 10^{-3}$				
C_3H_8	$1.9345 \cdot 10^{-2}$	$9.0258 \cdot 10^{-4}$					
C_2H_6	$1.3737 \cdot 10^{-2}$						
C_7H_{16}	-56.3229	-49.3201	-49.3425	-49.3402	-49.3398	-49.3398	-49.3398
C_6H_{14}	-48.2768	-42.4411	-42.4590	-42.4573	-42.4570		-42.4570
C_5H_{12}	-40.2307	-35.5621	-35.5755	-35.5744			-35.5743
C_4H_{10}	-32.1845	-28.6831	-28.6921				-28.6915
C_3H_8	-24.1384	-21.8041					-21.8086
C_2H_6	-16.0923						-14.9251
C_7H_{16}	$-1.4153 \cdot 10^{-1}$	$4.0049 \cdot 10^{-4}$	$-5.3912 \cdot 10^{-5}$	$-6.8910 \cdot 10^{-6}$	$4.0535 \cdot 10^{-7}$	0	
C_6H_{14}	$-1.3707 \cdot 10^{-1}$	$3.7605 \cdot 10^{-4}$	$-4.6400 \cdot 10^{-5}$	$-5.4172 \cdot 10^{-6}$	$2.3553 \cdot 10^{-7}$		
C_5H_{12}	$-1.3089 \cdot 10^{-1}$	$3.4216 \cdot 10^{-4}$	$-3.5981 \cdot 10^{-5}$	$-3.3732 \cdot 10^{-6}$			
C_4H_{10}	$-1.2175 \cdot 10^{-1}$	$2.9235 \cdot 10^{-4}$	$-2.0215 \cdot 10^{-5}$				
C_3H_8	$-1.0683 \cdot 10^{-1}$	$2.0561 \cdot 10^{-4}$					
C_2H_6	$-7.8200 \cdot 10^{-2}$						

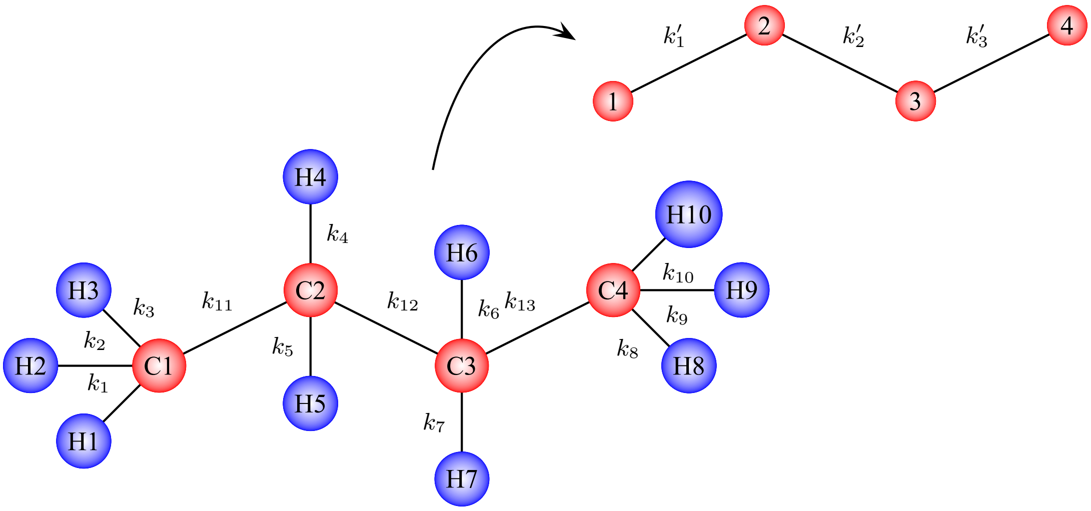


FIGURE 3.2. *Hydrogen vertices in blue are combined with their binding partners in red to new single vertices. The remaining edges and new vertices have been relabeled, denoted by single digits.*

dependent coordination $R_i^H(G)$, $1 \leq i \leq l$, are derived and the ground state energy evaluated for this system is defined as:

$$\widehat{E}_{i_1,\dots,i_k}(\widetilde{R}_1,\dots,\widetilde{R}_k) := \min_{|\phi_{(0)}|=1} \int \phi^*_{(0)}(\mathbf{r})\mathcal{H}_e^{\left(l+\sum_{j=1}^k Z_{i_j},(Z_{i_1},\widetilde{R}_{i_1}),\dots,(Z_{i_k},\widetilde{R}_{i_k}),R_1^H(G),\dots,R_l^H(G)\right)}\phi_{(0)}(\mathbf{r})d\mathbf{r}. \quad (3.6)$$

Note that this saturated energy function is denoted by $\widehat{E}$.

For the example of n-alkane molecules, we give the approximated energy and the relative error in the lower half of table 3.1, this time evaluated from saturated fragments. Clearly, we obtain a more satisfying result. The error is substantially reduced, the convergence rate for the bond order k is improved and it is now clearly decaying with rising values of k.

At this point, the following remarks on the saturation procedure are in order. Note that there has to be sufficient space for the newly introduced hydrogen atoms. If the edge k_i, where the subgraph G' is separated off, is a bridge then there is in general sufficient space. However, if k_i belongs to a cycle, there will in general not be sufficient space for the placement of additional hydrogen nuclei in the fragment molecule. Furthermore, the saturation hydrogens at either end of the remainder of the cycle may influence each other negatively. Hence, we suggest to either keep cycles always as a whole fragment if possible, or to at least create only fragments of cycles with a bond order which is particularly smaller than the cycle length. Note furthermore that the saturation procedure by means of hydrogen renders the role of hydrogen special in our approach. Thus, it is is useless to cut out a fragment at an edge involving only one hydrogen nucleus, as this will only create an additional hydrogen molecule while leaving the edge as it was before. Here, the best procedure is to remove the hydrogen nuclei degrees of freedom from the ANOVA decomposition algorithm, i. e. to drop them completely from the graph G, or to combine them with their binding partners since they are always terminal vertices anyway, see figure 3.2 for an illustration. Hence, in the following, we will not take further heed of the hydrogen atoms which are present in the molecular system.

3.4. A simple example. Let us now discuss these various aspects of our overall decomposition algorithm by means of a simple example. To this end, we consider the molecule butane, a chain molecule of four carbon atoms, which may be labeled according to their sequence 1, 2, 3 and 4, see figure 3.2. Let $I = \{1, 2, 3, 4\}$ be the set of indices. In order to simplify notation we now only note the label of each parameter and not its coordinate. We thus define $E^{(4)}[1, 2, 3, 4] := E^{(4)}(\widetilde{R}_1, \widetilde{R}_2, \widetilde{R}_3, \widetilde{R}_4)$. For the lower-dimensional terms (3.6) we simply drop the parameters, as they are clear from the set of lower indices, $\widehat{E}_{i_1,\ldots,i_l} := \widehat{E}_{i_1,\ldots,i_l}(\widetilde{R}_{i_1}, \ldots, \widetilde{R}_{i_l})$.[10] For example, the two-body term $\widehat{E}_{1,2} = \widehat{E}_{1,2}(\widetilde{R}_1, \widetilde{R}_2)$ is then the total energy of the induced subgraph $G' = (V', K')$ with the set $V' = \{v_1, v_2\} \subseteq V$ of the two vertices which are associated to the non-hydrogen nuclei coordinates R_1 and R_2, and the set $K' = \{k = (v_i, v_j) \in K \,|v_i, v_j \in V'\}$ of edges induced from $G = (V, K)$. Note here that the indices of a subset of vertices of a given graph G uniquely define the induced subgraph G'.

First, let us write down the ANOVA expansion, here explicitly only up to second order.

$$
\begin{aligned}
E^{(4)}[1, 2, 3, 4] = {} & \widehat{E}_1 + \widehat{E}_2 + \widehat{E}_3 + \widehat{E}_4 \\
& + \widehat{E}_{1,2} - (\widehat{E}_1 + \widehat{E}_2) + \widehat{E}_{1,3} - (\widehat{E}_1 + \widehat{E}_3) + \widehat{E}_{1,4} - (\widehat{E}_1 - \widehat{E}_4) + \widehat{E}_{2,3} - (\widehat{E}_2 - \widehat{E}_3) \\
& \quad + \widehat{E}_{2,4} - (\widehat{E}_2 + \widehat{E}_4) + \widehat{E}_{3,4} - (\widehat{E}_3 - \widehat{E}_4) \\
& + \ldots
\end{aligned}
$$

Now, we neglect terms which are, due to non-existent bonds, close to zero. For example, since there is no bond between atom 1 and 3 in our butane molecule we assume that the energy of the combined system of 1 and 3 is close to the sum of the two single systems and therefore $\widehat{E}_{1,3} - (\widehat{E}_1 + \widehat{E}_3)$ nearly vanishes. Analogously, also certain telescopic terms disappear which involve more than two coordinate labels. The remaining ANOVA expansion, now written explicitly out up to highest order, then reads

$$
\begin{aligned}
& E^{\ (4)}[1, 2, 3, 4] = \\
& \quad \widehat{E}_1 + \widehat{E}_2 + \widehat{E}_3 + \widehat{E}_4 \\
& \quad + \widehat{E}_{1,2} - (\widehat{E}_1 + \widehat{E}_2) + \widehat{E}_{2,3} - (\widehat{E}_2 + \widehat{E}_3) + \widehat{E}_{3,4} - (\widehat{E}_3 + \widehat{E}_4) \\
& \quad + \widehat{E}_{1,2,3} - \big(\widehat{E}_{1,2} - (\widehat{E}_1 + \widehat{E}_2) + \widehat{E}_{2,3} - (\widehat{E}_2 + \widehat{E}_3) + \widehat{E}_1 + \widehat{E}_2 + \widehat{E}_3\big) \\
& \qquad + \widehat{E}_{2,3,4} - \big(\widehat{E}_{2,3} - (\widehat{E}_2 + \widehat{E}_3) + \widehat{E}_{3,4} - (\widehat{E}_3 + \widehat{E}_4) + \widehat{E}_2 + \widehat{E}_3 + \widehat{E}_4\big) \\
& \quad + E^{(4)}[1, 2, 3, 4] - \Big(\widehat{E}_{1,2,3} - \big(\widehat{E}_{1,2} - (\widehat{E}_1 + \widehat{E}_2) + \widehat{E}_{2,3} - (\widehat{E}_2 + \widehat{E}_3) + \widehat{E}_1 + \widehat{E}_2 + \widehat{E}_3\big)\Big) \\
& \qquad\qquad - \Big(\widehat{E}_{2,3,4} - \big(\widehat{E}_{2,3} - (\widehat{E}_2 + \widehat{E}_3) + \widehat{E}_{3,4} - (\widehat{E}_3 + \widehat{E}_4) + \widehat{E}_2 + \widehat{E}_3 + \widehat{E}_4\big)\Big) \\
& \qquad\qquad - (\widehat{E}_{1,2} - (\widehat{E}_1 + \widehat{E}_2)) - (\widehat{E}_{2,3} - (\widehat{E}_2 + \widehat{E}_3)) - (\widehat{E}_{3,4} - (\widehat{E}_3 + \widehat{E}_4)) \\
& \qquad\qquad - \widehat{E}_1 - \widehat{E}_2 - \widehat{E}_3 - \widehat{E}_4.
\end{aligned} \tag{3.7}
$$

Thus, in first order we simply have all single-body terms as the four one-body fragments. In second order we get all remaining two-body fragments, e. g. $\widehat{E}_{1,2}$. However, since they still contain certain single-body energies, we have to subtract these, i. e. $-(\widehat{E}_1 + \widehat{E}_2)$, such that the remainder corresponds to the true two-body energy. The same occurs in third order, e. g. $\widehat{E}_{1,2,3}$. Here, two-body terms have to be subtracted again, i. e. $-(\widehat{E}_{1,2} + \widehat{E}_{2,3})$, now inherently with subtracted single-body terms, i. e. $+\big((\widehat{E}_1 + \widehat{E}_2) + (\widehat{E}_2 + \widehat{E}_3)\big)$. Furthermore, as the three-body term itself also contain certain single-body energies, these also have to be subtracted, i. e. $-(\widehat{E}_1 + \widehat{E}_2 + \widehat{E}_3)$, and so on. In the end, many terms cancel in this telescopic sum.

[10] Note that the full term $E^{(4)}[1, 2, 3, 4]$ is never saturated, only lower order terms $\widehat{E}_{i_1,\ldots,i_l \subseteq I}$ with $l < 4$ are.

Altogether, if the ANOVA expansion is truncated at k-th order with the bond structure of the molecule taken into consideration, the terms finally remaining for each k are for our butane example

$$\begin{aligned}
\text{1st order: } E^{(4)}[1,2,3,4] &\approx \widehat{E}_1 + \widehat{E}_2 + \widehat{E}_3 + \widehat{E}_4 \\
\text{2nd order: } E^{(4)}[1,2,3,4] &\approx \widehat{E}_{1,2} + \widehat{E}_{2,3} + \widehat{E}_{3,4} - \widehat{E}_2 - \widehat{E}_3 \\
\text{3rd order: } E^{(4)}[1,2,3,4] &\approx \widehat{E}_{1,2,3} + \widehat{E}_{2,3,4} - \widehat{E}_{2,3} \\
\text{4th order: } E^{(4)}[1,2,3,4] &= E^{(4)}[1,2,3,4]. \qquad (3.8)
\end{aligned}$$

Note that in highest order the equality is still valid due to the telescopic sum effect of the ANOVA expansion. The neglecting of certain terms due to the bond structure of the molecule merely influences the lower order approximations. Note furthermore that most of the terms in the ANOVA expansion (3.4) cancel. The combination of finally remaining terms depends strongly on the the underlying molecular graph structure and the chosen order k at which the expansion was truncated.

Now, an efficient procedure is needed which creates these remaining fragments and their proper combination for a prescribed order parameter k from general molecular structures. Its algorithmic details are described in the following section.

4. Algorithms for the decomposition. We now discuss our present implementation of the BOSS-ANOVA decomposition scheme for a molecular system. We start from the bond structure expressed in the associated graph G where hydrogen vertices have been already properly combined with their respective binding partners, compare figure 3.2. Let $I = \{1, \ldots, M\}$ be the set of nucleic indices. The goal of the algorithm is to create all necessary molecular fragments, i. e. all the possible induced subgraphs $G' = (V', K')$ from G with cardinality $|V'|$ equal to k and smaller, which appear in the ANOVA expansion truncated after the k-th term while additionally taking the bonding structure of the graph into account, compare the example (3.7).

Note that this procedure is naturally split up into two basic parts that correspond to (3.4) and (3.5): First, we have to construct the subgraph of each term $F_0, F_{i_1}, \ldots, F_{i_1,\ldots,i_k}$, with $i_1 < \ldots < i_k \subseteq I$, for the given bond order k, compare (3.4). Next, we have to construct each term $F_{i_1,\ldots,i_l}$, $1 \le l \le k$, in a recursive fashion from the terms $\widehat{E}_{i_1,\ldots,i_m}$, where $1 \le m \le l$, compare (3.5), since only these can be put forward to the approximate solver. We then perform the respective telescopic sum, eliminate this way a substantial amount of terms in the decomposition and thus determine the remaining ones plus their associated prefactors.[11] Let us remark again that the indices of the subset of vertices of a given graph G uniquely define the induced subgraph G'. Hence, only the associated labels have to be stored for each term in the ANOVA series. In the end, only the remaining terms with prefactors not equal to zero after final combination are forwarded to the approximate solver for the evaluation of the energies of the involved fragments which are finally combined to the total ground state energy.

The proposed scheme can be split into the following parts:

Graph recognition: In order to generate the graph G from a given molecular system with nuclei coordinates R_i, $i \in \{1, \ldots, M\}$, we resort to the linked-cell technique [23]. It scales linearly in the number of nuclei M. We then use use a depth-first search algorithm to recognize tree edges and cycles in the obtained graph structure which is also of $\mathcal{O}(M)$ complexity. Details will be given in subsections 4.1 and 4.2.

Order-by-order fragmentation: Now, the total energy $E^{(M)}$ has to be decomposed into terms $F_{i_1,\ldots,i_l}$, with $1 \le l \le k$, compare (3.4). To this end, for each l, every summand has to be constructed via a power set generation method. It produces all possible unordered vertex subsets $V' \subseteq V$

[11] Another possibility would be to try to determine the finally remaining terms directly. As we do not yet aim for maximum efficiency but for a proof of the linear scaling complexity of the approach, we will use the simpler approach via the explicit evaluation of the telescopic sums.

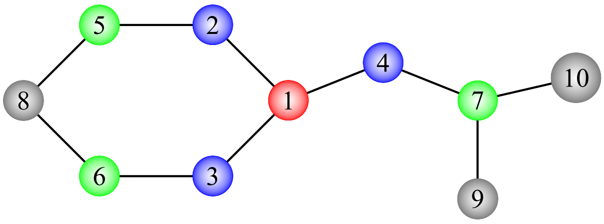


FIGURE 4.1. *Illustration of the order $k = 3$-restricted BFS-exploration of a graph structure. Gray spheres are not included in the subgraph G'. The vertices are labeled in order of their discovery starting from the root which is labeled "1". Color indicates the shortest path distance to the root in red from blue(1) and green(2). This sample graph G was derived from the non-hydrogen atoms in acetanilide.*

with cardinality $|V'| = l$ from the graph $G = (V, K)$. Thereby we drop all terms, whose induced subgraph $G'_{(l)} = (V', K')$ is not connected, where $|V'| = l$ with set K' given by the induced subgraph constraint. Then, each remaining term $F_{i_1,\dots,i_l}$, with $1 \leq l \leq k$, is constructed recursively from the terms $\widehat{E}_{i_1,\dots,i_m}$, where $1 \leq m \leq l$ and $\{i_1, \dots, i_m\} \subseteq \{i_1, \dots, i_l\}$, by calling again the power set generating method. Here, each term $\widehat{E}_{i_1,\dots,i_m}$ is assigned a prefactor of $(-1)^{(l-m)}$ for the later series summation, compare (3.5). In section 4.3 we discuss the details of a suitable memory arrangement of all these terms.

Power set generation: Now, given a specific cardinality m, there is the following task: Create all induced subgraphs $G' = (V', K')$ with $|V'| = m$ of a given graph G. That is, for any two subgraphs $G'_1 = (V'_1, K'_1)$ and $G'_2 = (V'_2, K'_2)$ it has to hold: $V'_1 \cup V'_2 \neq V'_1 \cap V'_2$. This routine shall return a list of subgraphs $\{G'_i = (V'_i, K'_i)\}_i$. Details are discussed in subsection 4.4.

Fragment list reduction: Finally, recognize all identical subgraphs G'_i by a "fingerprint" and combine to a single one. Furthermore determine the corresponding prefactor as the sum of the prefactors of all identical subgraphs. For details see subsection 4.5.

And at last, we construct the geometry of each molecular fragment from the remaining set of all unique subgraphs, call the approximate solver of the local Schrödinger problem for each fragment and sum up the returned energies, each multiplied with its associated prefactor, to obtain the k-th order BOSSANOVA approximation to the total ground state energy of the molecular system.

4.1. Graph exploration algorithms. There are two basic graph exploration algorithms, breadth-first and depth-first search. For variants and advanced implementations, we refer to [27]. Both search algorithms have a cost complexity of $\mathcal{O}(|V|)$ as they step over each vertex exactly once. They make use of a stack. Here, the order of putting and retrieving items from the stack decides on the type of algorithm: a first-in-first-out stack yields depth-first search, whereas a first-in-last-out stack yields breadth-first search. In algorithm 1, a sample implementation is given for the BFS variant. We use this procedure to construct a subgraph G' from $G = (V, K)$. To this end, we start from a root vertex $s \in V$ and an empty set G' and successively build up the desired subgraph, see figure 4.1 for an illustration. We may limit the exploration to the bond order k. Then, the maximum of the shortest path length during the exploration for any vertex v must be $d_G(s, v) < k$. The length of the stack S represents this limitation in the exploration of the graph. Furthermore, vertices are coloured during the exploration in order to distinguish between new and already visited vertices: We use here white at the beginning, gray for all visited vertices, and black for all visited vertices whose edges were all used. The array $l(v)$ labels the vertices in the order of discovery.

The implementation of DFS is analogous, now however a FIFO stack is used.[12]

[12] Now, to find cycles, tree or back edges, separation vertices and non-separable components we have to resort to the scanning variant also given in [27]. There, a *lowpoint of* v is defined as the least label $k(u)$ of a vertex u which can be reached from v

Algorithm 1: Breadth-first search

```
Data: Graph $G = (V, K)$, Root vertex $s \in V$, current vertex $v$, FILO vertex stack $S$ with push()
      and pop(), desired bond order $k$, label array $l(v)$ of size $|V|$
Result: subset $V'$ of subgraph $G'$ of $G$ with $\forall v \in V$ with $d_G(s, v) < k$ then $v \in V'$

$G' = \emptyset$;
for $\forall V \in V$ do
    mark $e$ white;
Add $s$ to $G'$;
push(s);
i=0;
$l(v) = (i++)$;
while $S \neq \emptyset$ do
    v = pop();
    for $\forall e_i = (v, t)$ with $t \in V$ do
        if $t$ is white then
            $l(v) = (i++)$;
            color $t$ gray;
            ShortestPath($t$) = ShortestPath($v$)+1;
            if ShortestPath(t) < $k$ then
                Add $t$ to $G'$;
                push(t);
    color $v$ black;
```

4.2. Graph recognition. In order to recognize the bond structure of a molecular system, a naive ansatz would compare the distance between every atom. This however would result in a cost complexity of $\mathcal{O}(M^2)$. Instead, we employ the linked-cell technique [23]. It combines the nuclei into groups by putting them into virtual, non-overlapping cells of a certain edge length. The edge length has to be greater than the largest bond distance. Then, it is guaranteed that possible bond neighbours can be found only in the very same and all directly adjacent cells. To this end, at most the neighbouring nuclei in 27 cells have to be scanned for each nuclei. Altogether, a run time complexity results which scales linearly with the number of atoms M. The procedure is described in algorithm 2. Note that the typical cutoff bond distance may be chosen freely, i. e. as twice the typical bond length, as an interelement cutoff distance. In our implementation we used the largest typical bond length in a given molecule plus an additional small safety margin.

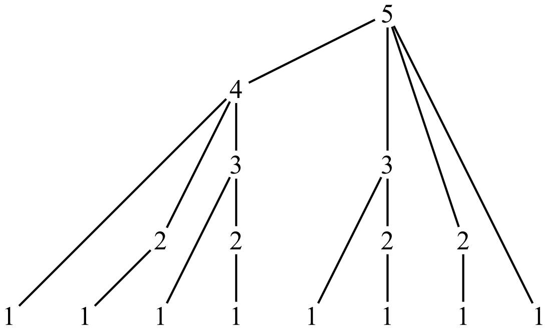
4.3. Order-by-order fragmentation. In this step, we break down the problem of creating all terms $F_{i_1,\ldots,i_l}$, $1 \leq l \leq k$, from the terms $\widehat{E}_{i_1,\ldots,i_m}$, $1 \leq m \leq l$, in their necessary multiplicities into the problem of creating all the induced subgraphs of G with given vertex cardinality equal to m. The solution to this subproblem is then described in the section 4.4.

This is generally a matter of correct accounting. There are two steps: First, create each $F_{i_1,\ldots,i_l}$, second for each $F_{i_1,\ldots,i_l}$ create all necessary $\widehat{E}_{i_1,\ldots,i_m}$. As described before, step one is executed by a loop over $1 \leq l \leq k$, calling a power set generating method for a given cardinality l working on the graph G. We obtain k sets of terms $\{F_{i_1,\ldots,i_l}\}_{\{i_1,\ldots,i_l\}\subseteq I}$ with $1 \leq l \leq k$ or k sets of subgraphs $\{G'_{l,p}\}_p$, respectively. Next, we work on each of these subgraphs $G'_{l,p}$ to recursively create all terms $\widehat{E}_{i_1,\ldots,i_m}$,

via a possible empty path consisting of tree edges followed by at most one back edge. This property can be built up during the exploration and henceforth can be exploited to find cyclic bonds, i.e. when both its vertices have the same lowpoint number.

Algorithm 2: Graph recognition

```
Data: Graph $G = (V, K = \emptyset)$ with $|V| = M$ atoms, each associated to coordinate vector $R_i$, list
      of atoms $cell[]$, cutoff bond distance $d^{(\text{bond})}$
Result: Graph $G = (V, K)$ with $|V| = M$ and each $v_i \in V$ representing atom $i$ and its bonds
        $k = (v_i, v_j) \in K$
$divisor[i]$ = floor($cell[i]/d^{(\text{bond})}$);
foreach $v_i \in V$ do
    for $0 \leq i \leq 2$ do
        $n[j] = R_i/d^{(\text{bond})} \cdot divisor[i]$;
        $index = n[2] \cdot (n[1] + n[0] \cdot divisor[1]) \cdot divisor[2]$;
        Add $v_i$ to list $cell[index]$;
foreach $0 < n[i] < divisor[i]$ and for each $0 \leq i \leq 2$ do
    $index = n[2] + (n[1] + n[0] \cdot divisor[1]) \cdot divisor[2]$;
    foreach $v_i$ in $cell[index]$ do
        foreach $index2 \in \{index,$ all indices that belong to adjacent cells$\}$ do
            foreach $v_j \in cell[index2]$ do
                $dist = |R_i - R_j|$ ;
                if $dist < d^{(bond)}$ then
                    Add bond $A_{ij} = 1$ to adjacency list;
```



Suborder	no. of lists
5	1
4	1
3	2
2	4
1	8
$\sum$	16

FIGURE 4.2. *The order-by-order fragmentation explained, here for the line associated to order $l = 5$ in a decomposition with bond order $k \geq 5$. The numbers $\{1, 2, 3, 4, 5\}$ stand for the suborder of a group of ANOVA series expansion terms, i. e. the number of bodies they contain. The top node "5" here resembles the set of subgraphs $G'_{(l=5)}$. We observe that the number of lists is always 2^{l-m-1} if l is the order and $1 \leq m < l$ is the suborder and at the same time $\sum_{m=1}^{l-1} 2^{l-m-1} + 1 = 2^{l-1}$.*

$1 \leq m \leq l$. The order of the recursion from $m = l$ down to $m = 1$ is depicted in an example dependence graph in figure 4.2: Given a molecular fragment graph $G'_{l,p}$ with vertex cardinality $l = 5$, corresponding to one such term $F_{i_1,\ldots,i_l}$, the edges indicate in what order the subgraphs with vertex cardinality $1 \leq m \leq l$ of $G'_{l,p}$ may be constructed which correspond to the terms $\widehat{E}_{i_1,\ldots,i_m}$. From the figure 4.2, we also make the following observation: For a given order l we will obtain 2^{l-1} sets of indices in the end, or each $F_{i_1,\ldots,i_l}$ in (3.5) consists of 2^{l-1} terms $E_{i_1,\ldots,i_m}$, $1 \leq m \leq l$. This gives the idea to employ an array of size 2^{l-1} as a temporal memory structure for the sets of indices which correspond to a term $E_{i_1,\ldots,i_m}$. For an illustration, compare figure 4.3. The arrows indicate the sequence of filling. They are derived directly from the tree graph structure in figure 4.2.

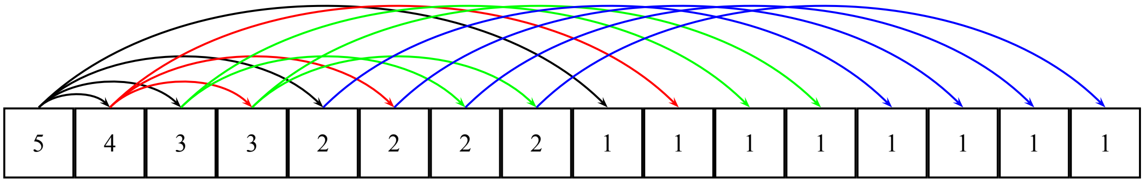


FIGURE 4.3. *Explanation of the temporary storage structure for list of subgraphs of order $l = 5$ and derived suborders m. The arrows indicate the dependence of every field on the first. Color indicates the order of creation: Black, red, green and blue. For a better understanding compare with figure 4.2.*

The overall procedure is given in algorithm 3. Note that "PowerSetGenerator" is the method described in the next section. It works on list of subgraphs, not on single subgraphs alone.

Algorithm 3: Order-by-order fragmentation

```
Data: Graph $G = (V, K)$, bond order $k$, function PowerSetGenerator()
Result: $list[i], i \in \{1, \dots, 2^k - 1\}$, of subsets $V_i'$ of subgraph $G_i'$ of $G$, with $\forall v \in V'$ with
        $d_G(s, v) < l, 1 \leq l \leq k$, and $prefactor[]$, for storage structure $list[]$ see fig. 4.3

foreach $1 \leq i \leq k$ do
    $NumberLevels = 2^{i-1}$;
    Allocate memory for $list[]$ of subgraphs and $prefactor[]$ with $NumberLevels$;
    Call PowerSetGenerator() with graph $G$, order $i$;
    Store return list in $list[0]$;
    Set $prefactor[0] = 1$;
    foreach $0 \leq sourceField < NumberLevels/2$ do
        foreach $i > j > 1$ do
            $destField = sourceField + 2^{i-j}$;
            Call PowerSetGenerator() with $list[sourceField]$, order $j$;
            Store return list in $list[destField]$;
            Set $prefactor[destField] = (-1)^{j-1}$;
```

Let us finally consider the complexity involved in algorithm 3. As the fragment generation will already scale linearly with the cardinality of the vertex set $|V|$ in the given graph G, we need to show that the number of lists only scales with the bond order k and not with the number of nuclei M. Since for each order $1 \leq l \leq k$ of the ANOVA terms we get 2^{l-1} list of subgraphs, we have a total number of lists of $\sum_{l=1}^{k} 2^{l-1} = 2^k - 1$ and, indeed, the complexity of the order-by-order fragmentation scales (exponentially) with k but not with M.

4.4. Power set generation. Now we deal with the problem to uniquely create all possible induced subgraphs $G'_{(j)} = (V'_{(j)}, K'_{(j)})$ for a given graph G and a given $|V'| = j$. Note that any induced subgraph $G'_{(j)}$ with fixed $|V'_{(j)}|$ and root vertex $s \in V'_{(j)}$ cannot contain a vertex $v \in V$ with $d_G(s, v) \geq j$. This motivates the following approach: Generate a subgraph $G' = (V', K')$ induced by all vertices which can possibly be reached from a root vertex s. Then, from this smaller subgraph, construct uniquely all possible power set combinations of vertices in V' that yield a connected subgraph and restrict it in such a way that each generated subgraph G' has to be a tree. We define the shortest path length of an edge $k = (v, w)$ with respect to a root vertex s to be $\min\{d_G(s, v), d_G(s, w)\}$ of all shortest paths whose last edge is k. We describe briefly this algorithm which is the subroutine called "PowerSetGenerator" in algorithm 3.

1. Starting once from every vertex v_i, use a breadth-first search algorithm and explore the graph up to paths with length less than or equal to k. Give each newly found vertex a label which is then kept fixed in the following.
2. At the same time, construct a list of all found edges and sort those with equal path lengths n with respect to v_i into the same list K_n.
3. Now, we have $1 \leq l \leq (k-1)$ sets of edges K_l. We start with an empty subset $V' = \emptyset$ and add here the root vertex s.
4. For level $l = 1$ to $k-1$ do the following:
   (a) Construct the reduced set of edges $K'_l = \{k = (v_i, v_j) \in K_l | v_i \text{ or } v_j \in V'\}$, i. e. all edges $k = (v_i, v_j)$ for which either v_i or v_j exists in the current set of vertices V'.
   (b) For level $l < k$, generate the power set of all possible unordered combinations $C_i \subseteq K'_l$ of this reduced set of edges K'_l. Note that $|C_i| = 2^{|K'_l|}$.
   (c) Go through the edges of each combination C_i and do the following:
      i. Add all other endpoints of the edges $e = (v_j, v_k) \in C_i$, i. e. all vertices v_j, v_k not contained in V' so far.
      ii. If $|V'| = k$, store the set of vertex indices, remove all vertices added on this level.
      iii. If not and if $|V'| < k$, go into recursion at step 4a with level $l+1$ and the current set V'.
      iv. Return from the recursion, i. e. go to lower level $l-1$ and proceed.

By means of this algorithm, we divide the neighborhood of a root vertex into levels of different shortest path lengths, see figure 4.1 for an example. Edges who connect[13] vertices $v \in V$ to the root s by the same shortest path length $d_G(v, s) = l$ are put into level l of a set of edges K_l. Note that it is advantageous to work with edges instead of vertices, as edges inherently contain a direction due to the known set V'.

We now will show that this algorithm scales at most linearly in M. First note that the BFS algorithm scales with $|V'| < (\max_{i \in \{1,\dots,M\}} d(v_i))^k$ of the created subgraph and not with $|V| = M$ of the whole graph as the exploration frontier in BFS is limited to k. Second, we have to check the number of subgraphs created from the power set. The number of possible combinations is $2^{|K'_l|}$ per level l for a reduced set of edges K'_l. Thus, we overall obtain $\sum_{l=1}^{k-1} 2^{|K'_l|}$. Let now the maximum bond degree be $c := (\max_{i \in \{1,\dots,M\}} d(v_i))$. Then, since at least one edge per vertex is a incident one from the next lower level $l-1$, $|K'_l|$ is bound by $(c-1) \cdot |I_{l-1}|$ with a given set I_{l-1} that represents all vertices added on level $l-1$. Furthermore, $|I_l|$ is bounded by $k-l$, because at least one vertex per level has to be added and there already is one root vertex. Using now partial sums of a geometric series, the cardinality of all sets of vertices for a single root can be bounded by

$$\sum_{l=1}^{k-1} 2^{(c-1)\cdot(k-l)} = \sum_{l=1}^{k-1} \left(\underbrace{2^{(c-1)}}_{=:C \geq 1} \right)^{k-l} = \sum_{l=1}^{k-1} C^{k-l} \overset{C \geq 1}{=} \frac{C^k - C}{C-1} \leq C^k \tag{4.1}$$

and thus depends only on the bond order k and the highest degree c of any vertex v in G. Since the loop runs over all possible root vertices, this algorithm scales with $\mathcal{O}(M \cdot C^k)$. Finally, as $2^k - 1$ lists of at most $\mathcal{O}(M \cdot C^k)$ fragments are generated altogether, we obtain a an overall scaling behaviour of $\mathcal{O}(M \cdot \widehat{C}^k)$ with $\widehat{C} := 2C$.

Since we restrict ourselves to the ANOVA truncation order k as the largest shortest path length, we consider from a root vertex v_i only an induced subgraph G' of G that contains all vertices whose shortest

[13] Uniquely, because the subgraph is a tree.

path length is less than or equal to k, see figure 4.1 for an illustration. It holds that this subgraph G' is a tree, if k is below the minimum of all cycle lengths in the graph G. So far, we did not guarantee the uniqueness of each constructed fragment as an induced subgraph of G'. This is however simple for a tree:[14] We demand that the root vertex v_i has always the lowest label of all vertices in a constructed subgraph. This can be realized already in the BFS step where we now exclude any bond that would lead to edges whose other endpoint's label is lower than that of our current root v_i. In order to guarantee that the constructed subgraph is indeed a tree, we must scan the molecular graph G in advance, determine the minimum of all cycle lengths n and set the ANOVA truncation order k to at most $n-1$.[15]

Note that the described algorithm creates each possible induced subgraph $G' = (V', K')$ with $|V'| = k$ and each only once. This is easy to prove if the power set generation, the unique, exploration-limited subgraph, its property of being a tree graph and the resulting edge sets K are taken properly into consideration.

4.5. Fragment list reduction. Finally, we have to identify equivalent subgraphs and combine their prefactors by summation. Only the set of indices of each subgraph is stored as a unique identification. Hence, the reduction may easily be achieved in $\mathcal{O}(M \log M)$ if efficient sorting algorithms are employed.

We briefly describe the procedure. Let $U_l := \{G'_{l,p} = (\{v_{i_1}, \dots, v_{i_l}, \}, K'_{l,i}\}_{\{i_1,\dots,i_l\}\subseteq I}$, $1 \leq l \leq k$, be the set of all subgraphs, each corresponding to an energy function term $\widehat{E}_{i_1,\dots,i_l}$ in the ANOVA series expansion. Note that I is the set of labels of all vertices $v \in V$ of the molecular graph $G = (V, K)$. Note that only subgraphs belonging to the same U_l can possibly be equal.

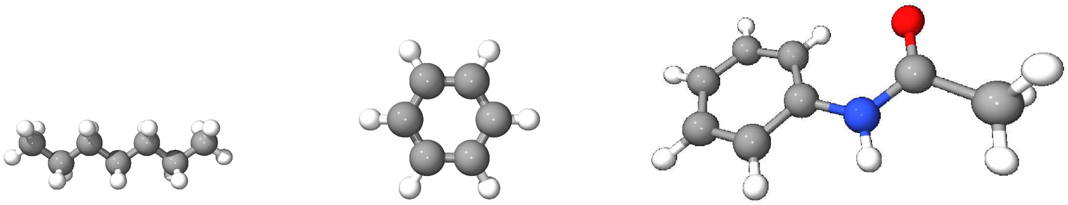
1. Heapsort [32] each set U_l with the following comparator function: For two subgraphs $G'_{l,p}$ and $G'_{l,q}$, if the vertex index of the first element of $G'_{l,p}$ is smaller than that of $G'_{l,q}$ return -1, if it is larger return 1 and if it is equal continue with comparing the next element. If all elements are equal return 0. The sorted lists are denoted by U'_l
2. In each sorted list U'_l equivalent subgraphs can only be situated next to each other. Use this property to pick out a representative of each unique subgraph and set its prefactor to the sum of all equal ones. The resulting set is denoted by $U''_l \subseteq U'_l$.
3. Go through each set U''_l and drop any subgraph with prefactor of 0. This gives the final set $U'''_l \subseteq U''_l$.

5. Numerical results. Now, we present our numerical results obtained so far. This section is divided into two parts. In the first part, we study small test molecules – first heptane, benzene and acetanilide – and then carbon and boron-nitride nanotubes to assess the the quality and correctness of our BOSSANOVA decomposition approach. In the second part, we study chain molecules, namely n-alkanes and periodic carbon and boron-nitride nanotubes, both with chirality (6,0), and discuss the scaling of the code with the bond order and with the number of molecules.

As approximate computational method for the electronic subproblems associated with the different fragments we have chosen the density functional theory with a plane wave basis set and Troullier-Martins pseudopotentials. The fragment-molecules have been evaluated employing a super cell ansatz with a minimum distance of 5 Å of every nuclei to any cell wall in order to reduce boundary effects. We have employed an energy cutoff of 128 Hartree. The nanotube fragments have been evaluated with a cutoff of 96 Hartree. We use evaluations with these cutoffs as reference results (full DFT) to compare the approximation error against. The stop condition of the conjugate gradient minimisation scheme was a relative change in the kinetic energy part of less than 10^{-5} and less than 10^{-7} in the relative change of the total energy. Hence, we expect the minimisation to be roughly converged to 10^{-5} relative error

[14] Note that an induced subgraph of a tree is again a tree.

[15] We believe that it is possible to easily create also non-tree subgraphs in a unique manner. However, common molecular structures mostly have a minimum cycle length of as large as 6, c. f. aromatic rings, and we therefore decided for the to above-described simpler tree-subgraph approach in our present implementation.



5.1.1: Heptane

5.1.2: Benzene

5.1.3: Acetanilide

FIGURE 5.1. *Illustration of our small test molecules.*

and thus aim for an approximation error of our BOSSANOVA decomposition method of up to the same accuracy.

Qualitative study. The numerical results obtained with our BOSSANOVA approach for various n-alkanes were already presented in table 3.1 (lower part). Now, beside to heptane, we applied the method also to benzene, acetanilide and periodic carbon and boron-nitride nanotube structures. Illustrations of these molecules can be found in the figure 5.1 and of the nanostructures in figure 5.2. The results are given up to order $k = 3$ in table 5.1 with relative errors in table 5.2.

We have chosen heptane here because it is a simple chain molecule and thus should be well suited to the fragmentation process. Its graph forms a tree, each vertex represents only a single bond. It is very symmetric and fragments easily. Here, we expect to see a very fast decay in the higher-order many-body contributions. Benzene on the other hand is completely different and way more complicated. Due to its aromatic ring, there should be significant 6-body contributions. Here, we are interested in how good the BOSSANOVA approximation is if the series expansion is truncated at orders less than $k = 6$. Finally, we picked acetanilide as an example of a simple organic molecule which combines both of the above features. It consists of an aromatic ring *and* a long, chain-like ligand and additionally features an oxide atom which is often hard to capture precisely in DFT calculations due to its strong electronegativity.

These small molecules were particularly used as benchmarks to assess the correctness of the implementation.[16] The results for these molecules with our BOSSANOVA approach are given in the tables 5.1 and 5.2. We see that the obtained accuracy for heptane is excellent with a relative error around 10^{-5} already at order $k = 3$, compare also table 3.1 (lower half) for analogous results with other n-alkanes. For the case of benzene the relative error is however worse by two orders of magnitude for $k < 4$. This

[16] Note that the fragments needed in the BOSSANOVA expansion may still be determined manually for these small molecules.

5.2.1: CNT (6,0)

5.2.2: BNNT (6,0)

FIGURE 5.2. *Illustration of the periodic cells of the carbon and boron-nitride nanotube structures, each with chirality (6,0).*

TABLE 5.1

Results for the k-th order BOSSANOVA approximation and the full DFT, total energy in Hartree.

Molecule/Order	1st	2nd	3rd	full DFT
Heptane	-56.3229	-49.3201	-49.3425	-49.3398
Benzene	-48.238996	-37.558394	-37.68375	-37.73463
Acetanilide	-93.13398	-76.546419	-76.825011	-76.79198
CNT (6,0)	-385.442438	-271.134204	-273.903358	-273.8944
BNNT (6,0)	-385.63609	-286.3467	-289.30785	-299.0197

TABLE 5.2

Results for the k-th order BOSSANOVA approximation and the full DFT, relative error to the full DFT result.

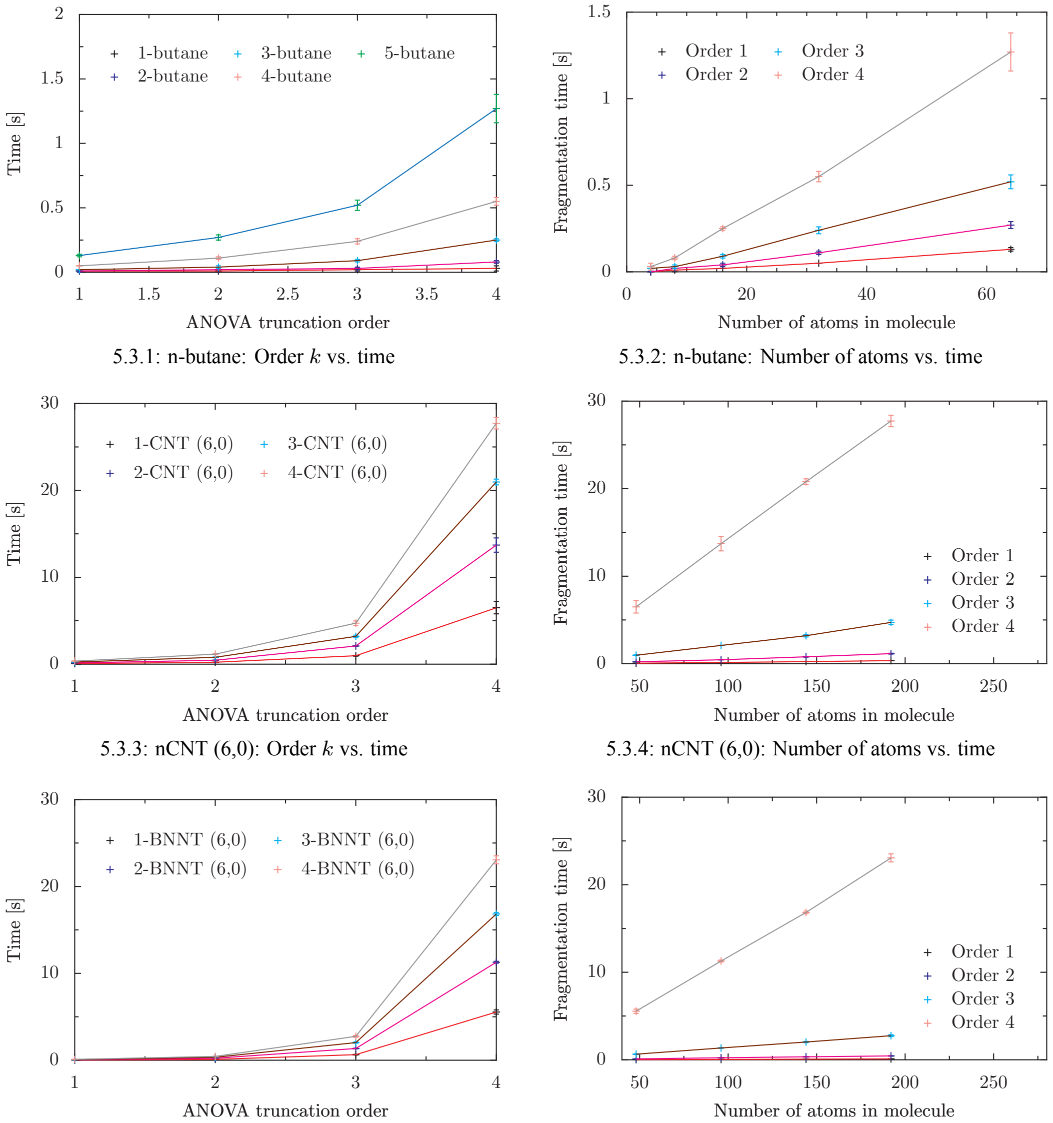
Molecule/Order	1st	2nd	3rd
Heptane	$1.4153 \cdot 10^{-1}$	$4.0049 \cdot 10^{-4}$	$5.3912 \cdot 10^{-5}$
Benzene	$2.7837 \cdot 10^{-1}$	$4.6704 \cdot 10^{-3}$	$1.3484 \cdot 10^{-3}$
Acetanilide	$2.1228 \cdot 10^{-1}$	$3.1977 \cdot 10^{-3}$	$4.3014 \cdot 10^{-4}$
CNT (6,0)	$4.0727 \cdot 10^{-1}$	$1.0078 \cdot 10^{-2}$	$3.2713 \cdot 10^{-5}$
BNNT (6,0)	$2.8967 \cdot 10^{-1}$	$4.238 \cdot 10^{-2}$	$3.2477 \cdot 10^{-2}$

indicates that the ring structure is not completely captured at the orders $k \leq 3$. Here, larger values of k are needed for good results. We furthermore see that the molecule acetanilide results in a better convergence than benzene despite its included aromatic ring. The results for carbon and boron-nitride nanotubes are also given in the tables 5.1 and 5.2. Here, we constructed both geometries using typical C-C and B-N bond lengths of such nanotubes. The periodic cell consisted of 48 atoms overall. The comparison of the results for the two geometrically identical but chemically different structures gives additional clues: A large difference in their relative error per order k indicates that the fragmentation process is sensitive to the chosen hydrogen bond length used in the saturation of dangling bonds. To this end, especially the results for the carbon nanotube are quite promising where an error of 10^{-5} is already achieved at order $k = 3$. Note however that the approximated total ground state energy is slightly lower than the value obtained from the full DFT calculation. Here, an overcompensation might have taken place in our method, which needs to be further investigated with nanotubes of varying chiralities and numbers of atoms in the unit cell. The results for the boron nitride nanotube do not show the good accuracy of the carbon tube. We believe that this is due to the imprecise B-H and N-H bond lengths and angles that, after hydrogenization of the fragments, do not maintain the total ground state density sufficiently well.[17] The C-H bond lengths and angles used in the carbon nanotube are of substantially better quality in this respect. Here, further investigations and improvements are needed in the future.

Scaling study. In the second part of this subsection we investigate the computational scaling behaviour of our BOSSANOVA implementation with respect to the number of nuclei M and with respect to the truncation order k. From the theoretical considerations of the previous section, we here expect a linear scaling complexity with M.

To this end, we studied n-butane with $n = 1, 2, 3, 4, 5$, i.e. periodically repeated and concatenated butane molecules with M from 4 to 64 that correspond to 4-alkane, 8-alkane and so on. We furthermore constructed larger nanotubes by periodically repeating the unit cell of 48 atoms along the symmetry direction where n indicates again the number of repeat cells and denote these configurations as n-CNT

[17] In order to maintain it as good as possible, one has to know the position of the center of the electronic wave function in the cut bond as the distance to either binding partner and the position of the center in the corresponding hydrogenated bonds. Then, one has to scale the hydrogen bond distance in such a way that the electronic wave function basically remains centered around the same position in the hydrogenated bond as it was in the bond before cutting. Here, we have made no investigations so far. We believe that the used C-H bond lengths reproduce such a scaling well while the used N-H and B-H bond lengths do not.



5.3.1: n-butane: Order k vs. time

5.3.2: n-butane: Number of atoms vs. time

5.3.3: nCNT (6,0): Order k vs. time

5.3.4: nCNT (6,0): Number of atoms vs. time

5.3.5: nBNNT (6,0): Order k vs. time

5.3.6: nBNNT (6,0): Number of atoms vs. time

FIGURE 5.3. *Truncation order k versus computing time (left), number of nuclei M versus time (right) for n-butane, n-CNT (6,0) and n-BNNT (6,0).*

and n-BNNT. For $n = 1, 2, 3, 4$, we thus obtained nanotubes of size one, two, three and four times that of the unit cell tube along with 48, 96, 144, 192 atoms. We then computed the BOSSANOVA fragmentation for each of these butanes and nanotubes and measured the necessary computing time.[18] The results are given in figure 5.3. Here, we can clearly see the anticipated linear scaling behaviour in the M-time-plots, especially for CNT- and BNT-systems with larger values of M. Let us remark that we naturally obtain an exponential behavior with respect to the scaling with the order k. Furthermore, we performed

[18] Here, only the time for the fragmentation process is determined. The time for the DFT solutions of the fragment problems is not included.

TABLE 5.3
Results for the least square fit with the values from figure 5.3: Exponent a in the M- and 10^k-scaling for varying values of k. .

(a) Bond order scaling

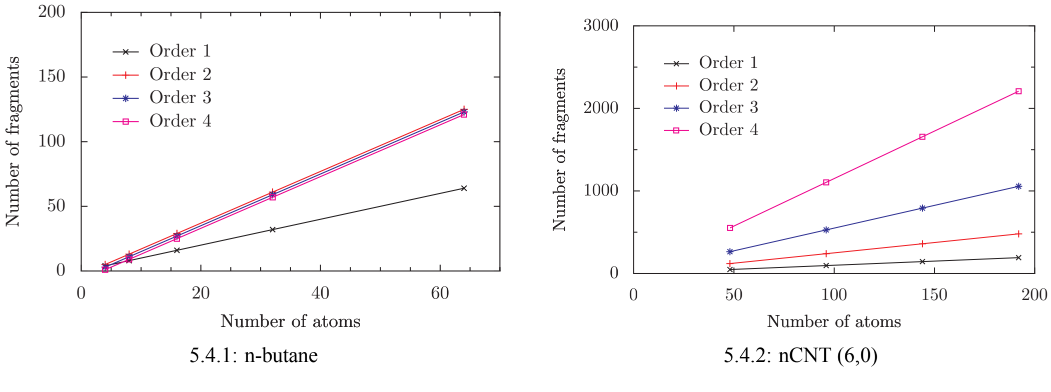
test system	a in k-scaling				
M-factor n	1	2	3	4	5
n-butane	0.40	0.66	0.84	0.80	0.75
n-CNT (6,0)	0.67	0.66	0.68	0.64	
n-BNNT (6,0)	0.89	0.88	0.87	0.86	

(b) Atom count scaling

test system	a in M-scaling			
bond order k	1	2	3	4
n-butane	1.35	1.38	1.27	1.17
n-CNT (6,0)	1.32	1.36	1.17	1.02
n-BNNT (6,0)	1.33	0.94	1.03	1.03

a least square fit to the assumed linear increase to obtain the power of the scaling with either order k, i.e. $f(k) = 10^{a \cdot k + c}$, or number of nuclei M, i. e. $\log f(M) = b \cdot M^a + c$. These fit values a are given in table 5.3. The found slopes a in $\log f(M) = b \cdot M^a + c$ underline the linear scaling. Note that the linearity is overshadowed by the principal function overhead below $M = 50$.

In the figures 5.4 and 5.5, we give the number of generated fragments for n-butane and n-CNT[19] versus the number M of atoms and versus the bond order k with our BOSSANOVA method. Note that the increase is linear with M. Furthermore, we see that there is basically no dependence on the bond order k for purely linear systems such as the n-butanes for which the average number of bonds per vertex is 2, i. e. $C \approx 1$. For more strongly connected systems such as the carbon and boron-nitride nanotubes, with an average bond per atom of $c = 3$, the number of fragments scales with $M \cdot \left((c-1)^{k-2} + c^1 + c^0\right)$ for $k > 2$ and with $M \cdot c^{k-1}$ for $k = 2$. The upper bound on the scaling behaviour, which was theoretically derived in the previous section 4.4, was $\left(2 \cdot 2^{c-1}\right)^k$ which is fairly well reproduced in our practical measurements.



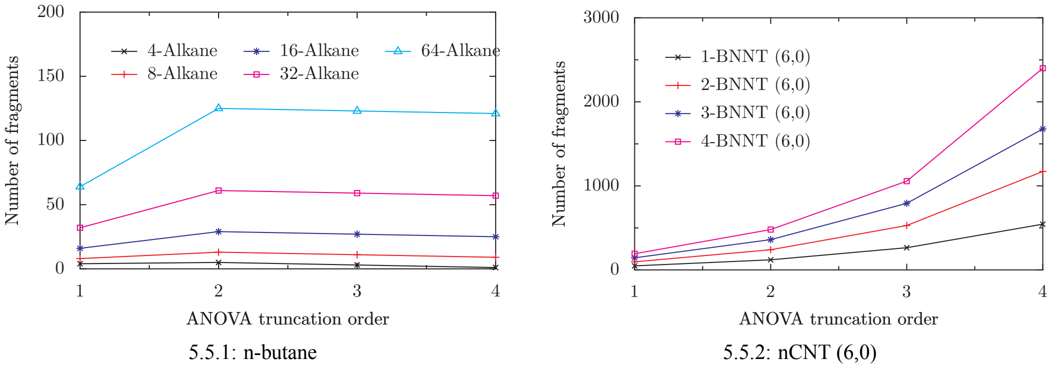
5.4.1: n-butane

5.4.2: nCNT (6,0)

FIGURE 5.4. *Final number of fragments versus the number of atoms in the ANOVA decomposition scheme.*

Note finally that we gave no measured times for the overall calculation which also would involve the DFT solution of the electronic subproblems associated to the fragments. The exact cross-over point

[19] Note that, due to the equivalent graphs, n-CNT and n-BNNT yield the same fragment counts.



5.5.1: n-butane

5.5.2: nCNT (6,0)

FIGURE 5.5. *Final number of fragments versus bond order k in the BOSSANOVA decomposition scheme.*

– that is when our linear scaling k-th order BOSSANOVA method will evaluate faster than the full QM code – depends on the specific molecular graph and on the approximate subproblem solver which can be chosen freely. We have shown that our fragmentation scheme scales linearly. It is also clear that the overall complexity is linear in the number of nuclei M, if the number of nuclei $M_i^{(k)}$ of each fragment is bounded, since this bound does not depend on M but on the truncation order k, and if the number of fragments scales linearly with M. Thus, for each approximate subproblem solver there will be a corresponding cross-over point at a certain number of nuclei M_c. For our DFT solver we found this point to be around 1000 nuclei for bond order $k = 3$.

6. Concluding Remarks. In this article we presented the BOSSANOVA decomposition approach for the approximate solution to the electronic Schrödinger equation for a given molecular system. It involves an ANOVA series expansion of an electronic energy function in the framework of the Born-Oppenheimer molecular dynamics. A truncation of this series at a certain *bond order* and the elimination of certain further terms by a locality constraint of the electronic wavefunction plus some additional hydrogen saturation results in a set of fragments of the overall molecule and associated electronic subproblems which may be solved with e.g. DFT, CI or CC methods. A proper combination of these solutions of the subproblems then leads to an approximate total ground state energy.

We gave an algorithmic description of how to derive this truncated BOSSANOVA expansion on any given graph. Furthermore we showed theoretically as well as practically that our new method indeed scales linearly with the number M of atoms in the overall problem. We gave numerical results for small organic molecules, carbon nanotubes and boron-nitride nanotubes. So far, we have seen that the BOSSANOVA approach works very well for n-alkanes and carbon based nanostructures. The obtained relative accuracy was below 10^{-5} for $k = 3$. The boron-nitride nanotube results might be further improved if the typical bond lengths to hydrogen are properly corrected. Also, more complex molecules like acetanilide, which contains various elements and bond structures at the same time, gave sufficiently accurate results. They nevertheless may be further improved by better suited hydrogen bond lengths in subsequent investigations.

Note that the impact of the neglected long-range Coulomb energy on the accuracy of the method is not yet studied. The Coulomb energies must be incorporated in a future implementation by e.g. Ewald summation or P3M techniques. Note furthermore that our BOSSANOVA approach is not rid of empirical parameters due to the necessity to saturate dangling bonds with hydrogen in the fragmentation process. Since the typical bond lengths and angles of hydrogenated systems are well assessed by measurements, we hope that a careful collection of robust values into a database may enable a broad range of application for the BOSSANOVA method.

Let us also point out that our approach is trivial to parallelize since the evaluation of each fragment by an appropriate solver can be done independently. Furthermore, since each fragment only contains a number of atoms roughly equal to the bond order k (neglecting hydrogen), the evaluation of the subproblems is possible already on very small machines with minimal memory prerequisites. Of course, also the memory cost scales only linearly. Thus, if the energy of a single fragment is calculated in seconds by e.g. a solver which is specifically tailored to the fast but precise evaluation of small and isolated systems, even a number of 10^5 or 10^6 fragments is within reach and the approximate total ground state energy evaluation of huge homogeneous molecular systems would become computationally feasible.

Finally, let us remark on how the BOSSANOVA method may be incorporated into a general coupling scheme of QM and MM. The BOSSANOVA fragmentation would be executed only in a given local domain, i.e. the active region where QM is locally needed. The resulting fragments are then forwarded to a suitable QM solver, whereas the surrounding passive environment would not be fragmented but is directly passed on to a MM solver. Our BOSSANOVA scheme is closely related to conventional many-body potentials (however in an ab-initio fashion) and to a variable many-body order. Furthermore, due to the fragmentation process, the interface region is not sharply defined. Therefore, we believe that this approach also remedies the problems of energy and electron density leaking of other local coupling methods to a certain extent.

REFERENCES

[1] W. K. Liu, E. G. Karpov, S. Zhang, and H. S. Park. An introduction to computational nanomechanics and materials. *Computer methods in applied mechanics and engineering*, 193:1529–1578, 2004.

[2] L. Greengard and V. Rokhlin. The fast multipole method for gridless particle simulation. *Computer Physics Communications*, 48:117–125, 1988.

[3] P. Y. Ayala and G. E. Scuseria. Linear scaling second-order Moeller-Plesset theory in the atomic orbital basis for large molecular systems. *Journal of Chemical Physics*, 110(8):3660–3671, 1999.

[4] C. Fonseca Guerra, J. G. Snijders, G. te Velde, and E. J. Baerends. Towards an order-N DFT method. *Theoretical Chemistry Accounts*, 99(6):391–403, 1998.

[5] X. P. Li, R. W. Nunes, and D. Vanderbilt. Density-matrix electronic-structure method with linear system-size scaling. *Physical Review B*, 47:10891–10894, 1993.

[6] M. Challacombe. A simplified density matrix minimization for linear scaling self-consistent field theory. *Journal of Chemical Physics*, 110:2332–2342, 1999.

[7] C.-K. Skylaris, P. D. Haynes, A. A. Mostof , and M. C. Payne. Introducing ONETEP: Linear-scaling density functional simulations on parallel computers. *Journal of Chemical Physics*, 122:084119–1–10, 2005.

[8] S. Goedecker. Linear scaling electronic structure methods. *Reviews of Modern Physics*, 71(4):1085–1123, 1999.

[9] F. Ercolessi and J. B. Adams. Interatomic potentials from 1st-principles calculations - the force-matching method. *Europhysics Letters*, 26(8), June 1994.

[10] M. S. Daw and M. I. Baskes. Embedded-atom method: Derivation and application to impurities, surfaces and other defects in metals. *Physical Review B*, 29(12):6443–6453, 1984.

[11] G. C. Abell. Empirical chemical pseudopotential theory of molecular and metallic bonding. *Physical Review B*, 31(10):6184–6196, 1985.

[12] J. Tersoff. Modeling solid-state chemistry: Interatomic potentials for multicomponent systems. *Physical Review B*, 39:5566–5568, 1989.

[13] N. Gresh, P. Claverie, and A. Pullman. Theoretical studies of molecular conformation. Derivation of an additive procedure for the computation of intramolecular interaction energies. Comparision with ab-inito SCF computations. *Thereotica Chimica Acta*, 66:1–20, 1984.

[14] F. Maseras and K. Morokuma. IMOMM - a new integrated ab-initio plus molecular mechanics geometry optimization scheme of equilibrium structures and transition-states. *Journal of Computational Chemistry*, 16(9):1170–1179, 1995.

[15] T. Vreven and K.Morokuma. On the application of the IMOMO (Integrated Molecular Orbital + Molecular Orbital) Method. *Journal of Computational Chemistry*, 21(16):1419–1432, 2000.

[16] A. Amovilli, I. Cacelli, S. Campanile, and G. Prampolini. Calculation of the intermolecular energy of large molecules by a fragmentation scheme: Application to the 4-n-pentyl-4'-cyanobiphenyl (5CB) dimer. *Journal of Chemical Physics*, 117:3003–3012, 2002.

[17] A. Laio, J. Van de Vondele, and U. Rothlisberger. A Hamiltonian electrostatic coupling scheme for hybrid Car-Parrinello molecular dynamics simulations. *Journal of Chemical Physics*, 116(16):6941–6947, 2002.

[18] I. Antes and W. Thiel. Adjusted connection atoms for combined quantum mechanical and molecular mechanical methods. *Journal of Physical Chemistry A*, 103(46):9290–9295, 1999.

[19] J. Sauer and M. Sierka. Combining quantum mechanics and interatomic potential functions in ab initio studies of extended systems. *Journal of Computational Chemistry*, 21(16):1470–1493, 2000.

[20] A. Van der Vaart, V. Gogonea, S. L. Dixon, and K. M. Merz jr. Linear scaling molecular orbital calculations of biological systems using the semiempirical divide and conquer method. *Journal of Computational Chemistry*, 21(16):1494–1504, 2000.

[21] G .T. Velde, F. M. Bickelhaupt, E. J. Baerends, C. F. Guerra, S. J. A. Van Gisbergen, J. G. Snijders, and T. Ziegler. Chemistry with ADF. *Journal of Computational Chemistry*, 22(9):931–967, 2001.

[22] G. Csyani, T. Albaret, M. C. Payne, and A. De Vita. Learn on the fly: A hybrid classical and quantum-mechanical molecular dynamics simulation. *Physical Review Letters*, 93(17):175503–1 – 175503–4, 2004.

[23] M. Griebel, S. Knapek, and G. Zumbusch. *Numerical Simulation in Molecular Dynamics*. Springer, 2007.

[24] A. Szabo and N. S. Ostlund. *Modern Quantum Theory - Introduction to Advanced Electronic Structure Theory*. Dover Publications, 1996.

[25] R. G. Parr and W. Yang. *Density-Functional Theory of Atoms and Molecules*. Oxford Science Publications, 1989.

[26] R. Diestel. *Graph Theory*. Number 173 in Graduate Texts in Mathematics. Springer-Verlag, Heidelberg, 2005.

[27] S. Even. *Graph Algorithms*. Computer Science Press, 1979.

[28] W. Hoeffding. A class of statistics with asympotically normal distributions. *Annals of Math. Statist.*, 19:293–325, 1948.

[29] D. Marx and J. Hutter. Ab initio molecular dynamics: Theory and implementation. In *Modern Methods and Algorithms of Quantum Chemistry*, volume 1 of *NIC Series*, pages 301–440. Forschungszentrum Juelich, Deutschland, 2000.

[30] D. W. Brenner. A second-generation reactive bond order (REBO) potential energy expression for hydrocarbons. *Journal of Physics: Condensed Matter*, 14:783–802, 2002.

[31] Russell D. Johnson III. NIST Computational Chemistry Comparison and Benchmark Database, NIST Standard Reference Database Number 101, September 2006.

[32] M. Galassi. *GNU Scientific Library Reference Manual*, revised second edition, 1992.