James Barker, Johannes Bulin, Jan Hamaekers and
Sonja Mathias

Localized Coulomb Descriptors for the Gaussian Approximation Potential

Localized Coulomb Descriptors for the Gaussian Approximation Potential

James Barker[1,2], Johannes Bulin[1], Jan Hamaekers[1], and Sonja Mathias[1]

[1]Fraunhofer-Institut für Algorithmen und Wissenschaftliches Rechnen SCAI, Schloss Birlinghoven, D-53754 Sankt Augustin

[2]Institut für Numerische Simulation, Universität Bonn, Wegelerstr. 6, D-53115 Bonn

Abstract

We introduce a novel class of localized atomic environment representations, based upon the Coulomb matrix. By combining these functions with the Gaussian approximation potential approach, we present LC-GAP, a new system for generating atomic potentials through machine learning (ML). Tests on the QM7, QM7b and GDB9 biomolecular datasets demonstrate that potentials created with LC-GAP can successfully predict atomization energies for molecules larger than those used for training to chemical accuracy, and can (in the case of QM7b) also be used to predict a range of other atomic properties with accuracy in line with the recent literature. As the best-performing representation has only linear dimensionality in the number of atoms in a local atomic environment, this represents an improvement both in prediction accuracy and computational cost when considered against similar Coulomb matrix-based methods.

1   Introduction

An important problem in computational materials science is the development of efficient and accurate *atomic potentials*. For an N-particle atomic system, an atomic potential (or simply *potential*) is a function $E^{(N)} : \mathbb{R}^{3N} \times \mathbb{Z}^N \to \mathbb{R}$ that maps from the set of three-dimensional Cartesian coordinates locating the atoms in that system, and their corresponding atomic numbers, to a potential energy.

The most accurate possible potential for any given system can be obtained from an (analytic) wavefunction solution to Schrödinger's equation [6], and all other potentials can be viewed as approximations to the same. Such solutions must be approximated numerically in all but the simplest of cases, a process which becomes computationally expensive for even relatively small systems. A variety of *first-principle* methods have been developed for this task, including most notably the density functional theory (DFT) [19]; these methods can be used to obtain potentials displaying high levels of accuracy, but are nevertheless always computationally expensive.

By contrast, empirical potentials are parametrized functions, with a fixed functional form that is usually physically motivated [12]. The parameters of an empirical potential must be fitted against the properties of a particular material, which can be a difficult task requiring a level of intuition; once fitted, however, the evaluation of an empirical potential for an arbitrary atomic configuration is comparatively cheap. Although it is difficult to accurately reproduce material properties other than those to which the potential was fitted, limiting their use in general cases, their applicability to the problem of exploring the energy hypersurface immediately surrounding an atomic configuration makes them useful for simulations.

In recent years, machine learning (ML) approaches have also been applied to the problem of potential generation. Such approaches can capture, or "learn" information from a wider range of materials than standard empirical potentials. The resulting ML potentials are almost as fast to evaluate as empirical potentials, and retain acceptable accuracy when used to predict the energy of a larger variety of materials.

Notable ML algorithms that have already been used for generating potentials include kernel ridge regression [7, 8, 16, 18] and multilayer neural networks [10]. Another promising approach is the Gaussian approximation potential (GAP) by Bartók et al. [2], which will be explained in detail in Section 2. This potential requires the construction of intermediate representations of the *atomic neighborhoods* surrounding each atom in the system; the choice of this representation is crucial for the performance of the GAP. Bartók originally proposed a modified bispectrum method for producing descriptors of mono-species crystalline environments. Although this approach performs well for semiconductors with respect to accuracy, the nature of the representation as coefficients of an expansion in spherical harmonics makes it costly to evaluate. Subsequently, Bartók et al. replaced the bispectrum representation with an approach called the *SOAP kernel*, which involves expanding a density function associated with an atomic environment in spherical harmonics and to directly define the similarity between any two such density functions [1].

Here, we present an alternative approach to the encoding of atomic environments, which we refer to as the *localized Coulomb* (LC) representation. This representation is based on the Coulomb matrix, originally proposed by Rupp et al. [18], and its derivative, the sorted Coulomb matrix [10, 17]. We combine the LC representation with the Gaussian approximation potential to obtain a new method for the creation and evaluation of families of atomic potentials; we term this method *LC-GAP*.

The remaining article is organized as follows. In Section 2 we briefly summarize the basics for machine learning for potential energy surfaces and describe our new Coulomb matrix based descriptors in particular in Section 2.2. In Section 3 we give numerical results for several datasets and additionally discuss the distribution of individual atomic contributions in Section 3.3. We conclude with some remarks in Section 4.

2 Potential Energy Prediction through Machine Learning

As stated above, our goal is to produce a function $E^{(N)} : \mathbb{R}^{3N} \times \mathbb{Z}^N \to \mathbb{R}$ that approximates the potential energy for an arbitrary unknown atomic system $X^* \in \mathbb{R}^{3N} \times \mathbb{Z}^N$. Given a collection of atomic systems of size N and their energies, $\{(X_i, E_i)\}_{i=1}^M$ with $X_i \in \mathbb{R}^{3N} \times \mathbb{Z}^N$,

an arbitrary ML algorithm could be used to produce, or *learn* the function $E^{(N)}$. However, such an approach will not scale well to larger systems; the dimension of $E^{(N)}$ grows linearly in the number of atoms, and the so-called "curse of dimensionality" implies that the computational cost of learning will grow exponentially in the same.

Additionally, the use of raw representations X_i as inputs for an ML process is problematic from a physical standpoint. The potential energy of any given molecule is necessarily invariant under translation and orthogonal transformation; it is also invariant under permutation of the order in which atoms are described [19]. A raw representation obeys none of these restrictions; therefore, a single unique system can be described by uncountably many raw representations. This offers a significant obstacle to the learning and prediction abilities of any ML algorithm.

The Gaussian approximation potential (GAP) framework of Bartók et al. addresses both these problems [2]. The first, by introducing an ansatz that can be used to limit the dimensionality of a representation, and the second, by using more sophisticated representations that maximize invariance. In this section, we briefly recap the GAP framework, and present some alternative, simpler, representations of atomic systems.

2.1 The GAP Framework and Gaussian Process Regression

The fundamental assumption of the GAP framework is the *atomic decomposition ansatz*: that the potential energy of an atomic system can be written as the sum of energies attributed to each of its atoms, and these atomic energies depend only on a neighborhood of the corresponding atom. For systems without long range electrostatic interactions, this assumption can be motivated by the *nearsightedness of electronic matter* [9, 13]. This allows us to write the potential energy of an N-particle system $X \in \mathbb{R}^{3N} \times \mathbb{Z}^N$ as a sum of atomic energy contributions:

$$E^{(N)}(X) = \sum_{i=1}^{N} E_{\text{atomic}}\left(L_i^{(N)}(X)\right). \tag{1}$$

Here, the $L_i^{(N)}(X) : \mathbb{R}^{3N} \times \mathbb{Z}^N \to \mathbb{R}^d$ are *atomic neighborhood representation functions*. Each of these maps a full atomic system to a d-dimensional representation, or *descriptor*, of the atomic neighborhood of the ith atom. $E_{\text{atomic}} : \mathbb{R}^d \to \mathbb{R}$ is then an unknown function that assigns an atomic energy contribution to an atomic neighborhood representation. Given Equation (1) above, the problem of learning the function $E^{(N)}$ reduces to the problem of learning the function E_{atomic}. This function is necessarily dependent on the form of the functions $L_i^{(N)}$ and the chosen dimensionality of their output d, both of which will be discussed in the next section.

As originally presented, the GAP framework uses the method of Gaussian process regression (GPR) to learn the function E_{atomic}, although other suitable ML algorithms could also be used. The input to the learning process is a *training set* of M systems, each containing some number of atoms $\{N_i\}_{i=1}^{M}$, and their known energies: $X = \left\{\left(X_i \in \mathbb{R}^{3N_i} \times \mathbb{Z}^{N_i}, E_i \in \mathbb{R}\right)\right\}_{i=1}^{M}$. The ability to learn from a training set containing systems with differing numbers of atoms is a direct result of learning the function E_{atomic} rather than E_{total}, the energy of the whole molecule. Additionally, a level of Gaussian

observation noise $\varepsilon_i \sim \mathcal{N}\left(0, \sigma_i^2\right)$ is associated with each training set energy E_i, such that

$$E_i = \left(\sum_{j=1}^{N} E_{\text{atomic}}\left(L_j^{(N)}\left(X\right)\right)\right) + \varepsilon_i.$$

Gaussian process regression predicts the function E_{atomic} as a linear combination of positive-definite kernel functions $\kappa : \mathbb{R}^d \times \mathbb{R}^d \to \mathbb{R}$ centered on the atomic neighborhood representations of each atom in every system in the training set. That is, for some N-particle system X^* whose energy we wish to predict,

$$E_{\text{atomic}}\left(L_k^{(N)}(X^*)\right) = \sum_{i=1}^{M} \alpha_i \sum_{j=1}^{N_i} \kappa\left(L_j^{(N_i)}\left(X_i\right), L_k^{(N)}\left(X^*\right)\right),$$

where the α_i are chosen during the learning process. Again, note that the use of E_{atomic} rather than E_{total} allows the prediction of potential energies for systems with arbitrary numbers of atoms, regardless of the contents of the training set.

2.2 Localized Coulomb Matrix Descriptors

The choice of representation functions $L_i^{(N)} : \mathbb{R}^{3N} \times \mathbb{Z}^N \to \mathbb{R}^d$ is critical to the behavior of the GAP framework, both in terms of computational performance and in accuracy. The primary contribution of this paper is the introduction of three new such functions, which are described below.

The genesis of our work is the Coulomb matrix descriptor, due to Rupp et al. [18], which allows the representation of complete molecules as square matrices. Consider an N-atom system $X \in \mathbb{R}^{3N} \times \mathbb{Z}^N$, consisting of Cartesian coordinates $\mathbf{R}_1, \ldots, \mathbf{R}_N$ describing atom locations, and the atomic numbers $Z_1, \ldots, Z_N$ associated with those atoms. The Coulomb matrix is then an $N \times N$ matrix M, whose entries are given as

$$M_{ij} = \begin{cases} 0.5 Z_i^{2.4} & i = j, \\ \frac{Z_i Z_j}{\|\mathbf{R}_i - \mathbf{R}_j\|_2} & i \neq j. \end{cases} \tag{2}$$

By construction, the Coulomb matrix is invariant under translations and orthogonal transformations of the set of Cartesian coordinates $\{\mathbf{R}_1, \ldots, \mathbf{R}_N\}$. It is, however, not invariant under permutation of the indexing order of the coordinates and their associated atomic numbers. To improve the handling of permuted systems, the rows and columns of the Coulomb matrix can be sorted according to their respective norms (which are identical, as the matrix is symmetric), resulting in the *sorted Coulomb matrix* [10].

The sorted Coulomb matrix is a *global descriptor* of an atomic system. As such, it can only represent molecules, and is inapplicable to infinitely-periodic crystal systems. However, it is possible to modify the Coulomb matrix into a *local descriptor*, encoding information only about the immediate neighborhood of an atom. Such a descriptor is then a candidate for unqualified use as an atomic representation function $L_i^{(N)}$ in the GAP framework, and can be applied to infinitely-periodic crystal systems as well as finite molecules.

4

Before we begin, it is useful to define a system of *local indices* to specify more clearly the structure of an atomic neighborhood around the ith atom in some system X. First, let p_1 indicate the ith atom itself. Then let K be the number of atoms that are located within some cutoff radius $r_{\text{cut}} > 0$ around $\mathbf{R}_i$, and let $p_2, \ldots, p_{K+1}$ specify these atoms. Finally, let m be the maximum neighborhood occupancy, which must be chosen in such a way that it is impossible to have greater than $m-1$ atoms surrounding any atom in either the training set, or an atom in a system which will be predicted. Then let all indices $p_{K+2}, \ldots, p_m$ specify *dummy atoms*, which have arbitrary location and atomic number zero. The use of such dummy atoms is an established procedure [10], and serves to ensure that all descriptors produced in either the training or prediction process will be the same size and therefore comparable.

With the aid of local indices, we now define the *localized Coulomb matrix* for the ith atomic neighborhood. First, similarly to Equation (2) above, define the $m \times m$ matrix $M^{(i)}$, with entries given by

$$M_{jk}^{(i)} = \begin{cases} 0.5 Z_{p_j}^{2.4} & j = k, \\ \dfrac{Z_{p_j} Z_{p_k}}{\|\mathbf{R}_{p_j} - \mathbf{R}_{p_k}\|_2^\alpha} & \text{otherwise.} \end{cases} \tag{3}$$

(Note that for notational simplicity, we do not explicitly indicate the dependency of the $(p_1, \ldots, p_m)$ local indices on the choice of i.) Then let P be the permutation matrix such that $P_{1,1} = 1$ and

$$\left\| \left(PM^{(i)} P^T \right)_{j,*} \right\|_2 \geq \left\| \left(PM^{(i)} P^T \right)_{j+1,*} \right\|_2 \tag{4}$$

for $j = 2, \ldots, m-1$. That is, P reorders the rows and columns of $M^{(i)}$ such that those corresponding to the central atom p_1 become the upper- and leftmost respectively, while permuting the remainder in descending order of norm. The permuted matrix $PM^{(i)} P^T$ is called $C_{\text{loc}}^{(i)}$ and is termed the *localized Coulomb matrix* for the ith neighborhood. Finally, the upper triangular entries of this matrix are packed in row-wise order into a vector $\mathbf{c}$ of size $d = \frac{m(m+1)}{2}$, and $L_i^{(N)}(X) := \mathbf{c}$ is the *localized Coulomb descriptor function*. (Because the matrix is symmetric, the lower-triangular entries contain no further information.)

One issue with this choice of representation function is that small changes to the state of the system that result in atoms entering or leaving the cutoff radius can cause large changes to the resulting descriptor. To reduce the significance of this effect, we modify the denominator to include distances from each atom to the central atom, creating the *decaying Coulomb matrix* $C_{\text{dec}}^{(i)} = P\hat{M}^{(i)} P^T$, with:

$$\hat{M}_{jk}^{(i)} = \begin{cases} 0.5 Z_{p_j}^{2.4} & j = k = 1, \\ \dfrac{Z_{p_j} Z_{p_k}}{\left(\|\mathbf{R}_{p_1} - \mathbf{R}_{p_j}\|_2 + \|\mathbf{R}_{p_1} - \mathbf{R}_{p_k}\|_2 + \|\mathbf{R}_{p_j} - \mathbf{R}_{p_k}\|_2 \right)^\alpha} & \text{otherwise.} \end{cases} \tag{5}$$

The permutation matrix is chosen in the same fashion as for the localized Coulomb matrix. Due to the introduction of the additional distance terms, the movement of atoms across the cutoff boundary cause significantly smaller changes, which may be lower than machine precision for some values of r_{cut} and α. The decaying Coulomb matrix can be packed into a vector of size $d = \frac{m(m+1)}{2}$ in the same manner as before.

Finally, we introduce a lower-dimensional descriptor, called the *reduced Coulomb matrix*, based in turn upon the decaying Coulomb matrix. Rather than use a permutation matrix,

we require that the atoms $p_2, \ldots, p_{k+1}$ are indexed in such a way that $\|\mathbf{R}_{p_j} - \mathbf{R}_{p_1}\|_2 \leq \|\mathbf{R}_{p_{j+1}} - \mathbf{R}_{p_1}\|_2$ for $j = 1, \ldots, m-1$, similar to the approach in [17]. Then the descriptor is constructed by taking the first row and the diagonal of the matrix $\hat{M}^{(i)}$ in Equation (5):

$$C_{\text{red}}^{(i)} = \left[\hat{M}_{11}^{(i)}, \hat{M}_{12}^{(i)}, \ldots, \hat{M}_{1m}^{(i)}, \hat{M}_{22}^{(i)}, \hat{M}_{33}^{(i)}, \ldots, \hat{M}_{mm}^{(i)} \right]. \tag{6}$$

This descriptor has dimensionality $2m - 1$, linear rather than quadratic in the maximum neighborhood occupancy, while still encoding all information about the pairwise interaction of the central particle with all its neighboring atoms.

3 Results

In order to evaluate the new atomic neighborhood functions presented in Section 2.2 above, we performed a series of numerical experiments over the QM7, QM7b, and GDB9 datasets. The QM7 dataset [3, 18] contains approximately 7100 biomolecules selected from the GDB13 database, each with up to seven heavy atoms[1], and their atomization energies[2], calculated at the PBE0 level of theory. The QM7b dataset [3, 11] is a slightly-expanded version of QM7, containing 7211 biomolecules with up to seven heavy atoms, including some with chlorine. As well as atomization energy, QM7b provides a total of thirteen extra properties per molecule (e.g. polarizability, HOMO and LUMO eigenvalues, excitation energies) calculated at different levels of theory (ZINDO, SCS, PBE0, GW). Finally, the GDB9 dataset [14] contains approximately $134,000$ biomolecules (also selected from GDB13) containing up to nine heavy atoms, along with their atomization energies calculated at both the PM7 and B3LYP levels of theory. For computational convenience, we selected a subset of the full GDB9 database (calculated at PM7), comprised of all molecules with up to eight heavy atoms, and approximately $18,000$ molecules with nine heavy atoms, chosen by stratification. We refer to this subset as the GDB9_18K dataset.

For each of the three datasets described above, we isolated subsets containing up to n heavy atoms for all $n \leq m$, where m is the maximum number of heavy atoms in any molecule in the dataset. We refer to these subsets as QM7_1, QM7_2, QM7_3, etc. A decomposition of each dataset by number of heavy atoms per molecule is given in Table 1.

Table 1: Decomposition of testing datasets by number of heavy atoms per molecule.

| Dataset | Number of heavy atoms | | | | | | | | | Total molecules |
	1	2	3	4	5	6	7	8	9	
QM7	1	3	12	43	158	950	5998	–	–	7165
QM7b	1	3	12	43	158	953	6041	–	–	7211
GDB9-18K	3	5	9	31	130	618	3197	18298	18599	40890

Each experiment consisted of learning a potential function from one such subset, called the *training set*, and then using the resulting potential to predict the atomization

---

[1]All non-hydrogen atoms are considered heavy.

[2]The atomization energy is the potential energy of a molecule that has been adjusted by the combined potential energy of its isolated atoms [5].

energies of another (non-overlapping) subset, called the *test set*. All testing was performed using an implementation of the Gaussian approximation potential, using one of the three localized Coulomb representation functions described in Section 2.2 above; we term this system *LC-GAP*. The underlying ML algorithm used was Gaussian process regression, as described in [15].

For these experiments, we used the Laplacian kernel

$$\kappa(x, y) = \sigma^2 \exp\left(\frac{-\|x - y\|_1}{l^2}\right),$$

which has shown good performance when applied to similar problems [7, 8]. The terms σ and l are kernel *hyperparameters*, the choice of which can have serious implications for the accuracy of the system.

To assess the quality of prediction for each test set, we used the mean absolute error (MAE), defined as

$$\mathrm{MAE}\,(X^*) = \frac{1}{M} \sum_{m=1}^{M} \left| E_m^{\mathrm{exact}} - E_m^{\mathrm{pred}} \right|,$$

where M is the number of entries in the test set, the terms E_m^{exact} are the atomization energies obtained from the dataset, and E_m^{pred} the atomization energies obtained by prediction. This error metric is well-established in the statistics community for assessing the prediction accuracy of regression models [20], and has been used along with the QM7 dataset to benchmark other ML-based potentials.

3.1 Comparison of Descriptor Functions on QM7

We began by investigating the performance of the three localized Coulomb descriptors (sorted, decaying, and reduced) described in Section 2.2 above. To establish the ability of these descriptors to predict atomization energies of molecules with similar properties to those used for training, we performed five-fold cross-validation on each of the QM7 subsets containing at least three heavy atoms. Kernel hyperparameters (l, σ) were chosen by minimization of the negative log-likelihood on the training set. For each descriptor, a grid of localization parameters α and neighborhood cutoffs r_{cut} were trialed.[3] A graph demonstrating the best results obtained on each subset can be found in Figure 1; a table describing the values of α and r_{cut} for each can be found in Table 2.

For all but the smallest subset, the three tested descriptor types perform similarly, all producing results on the same order of magnitude and (for QM7_5 and higher) within one kcal/mol of MAE of each other. Although the localized Coulomb representation performs the best for the QM7_4, QM7_5 and QM7_6 subsets, it is interesting to note that on the full QM7 dataset, the reduced Coulomb representation slightly outperforms both of the others. This is notable, given the lower dimensionality of that descriptor.

The cross-validation performance of each descriptor allows a direct comparison with previously-published results in the literature. Hansen et al. compared several different ML potentials by using five-fold cross-validation on the complete QM7 dataset [8]. Using non-localized kernel ridge regression coupled with randomly-sorted (global) Coulomb

---

[3]Here, we tested values $(\alpha, r_{\mathrm{cut}}) \in \{3, 4, \ldots, 7\} \times \{3.0, 3.5, 4.0, 4.5, \ldots, 7.0\}$.
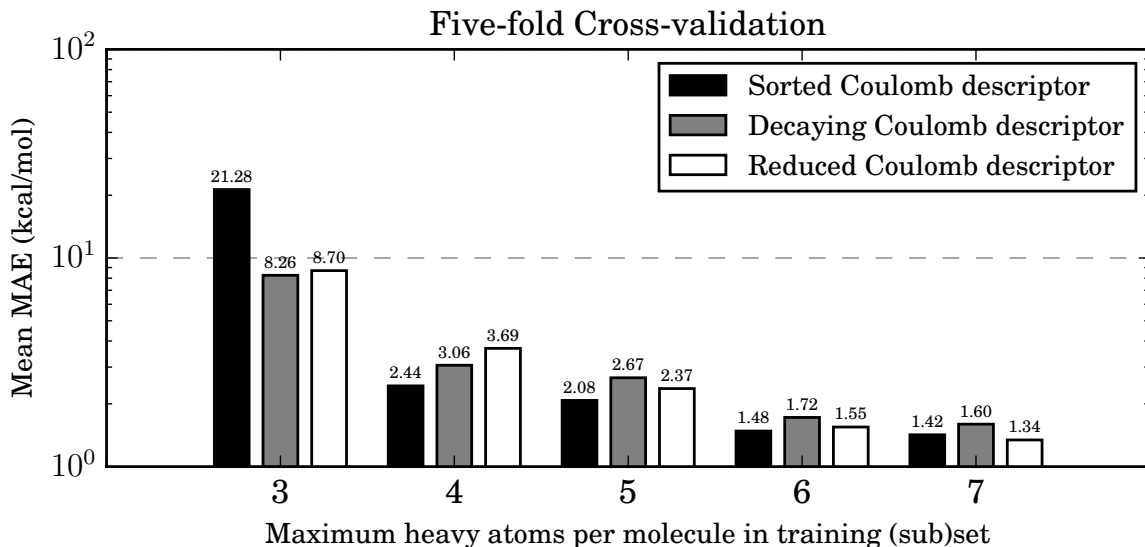
Figure 1: Mean MAEs of the predicted atomization energies on various QM7 subsets obtained during five-fold cross-validation, using LC-GAP potentials equipped with the different localizers described in Section 2.2. The precise values of α and r_{cut} used for each result can be found in Table 2.

Table 2: Best choices of descriptor parameters (α, r_{cut}) for cross-validation over QM7, by descriptor type and data subset. These choices are used to generate the results seen in Figure 1.

Descriptor	QM7_3	QM7_4	QM7_5	QM7_6	QM7_7
Localized	(4.0, 5.0)	(3.0, 3.5)	(5.0, 3.5)	(3.0, 3.5)	(3.0, 3.5)
Decaying	(3.0, 4.0)	(3.0, 3.5)	(3.0, 3.5)	(4.0, 3.5)	(3.0, 4.5)
Reduced	(3.0, 6.0)	(5.0, 6.0)	(3.0, 5.5)	(4.0, 5.5)	(5.0, 6.5)

matrices, they report an MAE of 3.07 kcal/mol. Using multilayer neural networks, they obtained an MAE of 3.51 kcal/mol. Recently, Hansen et al. also reported an MAE of 1.5 kcal/mol, obtained through kernel ridge regression combined with the "Bag of Bonds" approach [7]. We note that both the sorted and reduced Coulomb descriptors outperform this result, with MAEs of 1.42 and 1.34 kcal/mol respectively. We conclude that all three of the tested descriptors offer competitive accuracy when compared to the literature.

It is also important to consider the application of an ML-based potential to datasets with potentially different properties than those used for training. To evaluate the behavior of the three LC-GAP localizer functions in such a scenario, we performed "upwards transferability" testing over the QM7 dataset. In these tests, an LC-GAP potential is trained using one of the QM7 subsets; it is then used to predict all entries in the remainder of the complete dataset, which have (by construction) at least one heavy atom more than any molecule in the training set. The results of these tests are given in Figure 2.

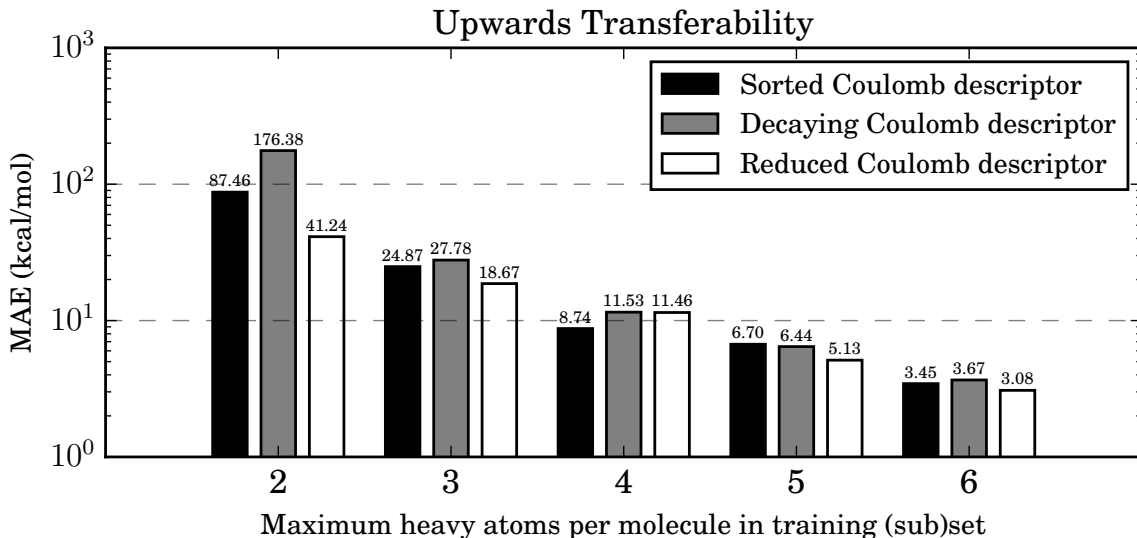The achieved MAEs are not as low as in the case of cross-validation, which is to

Figure 2: MAEs of the predicted atomization energies on the remainder of the full QM7, using LC-GAP potentials trained on QM7 subsets, equipped with the different localizers described in Section 2.2. The precise values of α and r_{cut} used for each result can be found in Table 3.

Table 3: Best choices of descriptor parameters (α, r_{cut}) for upwards transferability over QM7, by descriptor type and data subset. These choices are used to generate the results seen in Figure 2.

Descriptor	QM7_2	QM7_3	QM7_4	QM7_5	QM7_6
Localized	(4.0, 3.5)	(4.0, 3.5)	(4.0, 3.5)	(5.0, 3.5)	(4.0, 4.0)
Decaying	(4.0, 3.5)	(4.0, 3.5)	(4.0, 3.5)	(4.0, 4.0)	(5.0, 3.5)
Reduced	(4.0, 5.5)	(4.0, 5.5)	(4.0, 5.0)	(5.0, 5.0)	(5.0, 6.0)

be expected. However, potentials display consistent improvement as their training sets increase in complexity; potentials trained on the QM7_5 and QM7_6 datasets achieve MAEs below 10 kcal/mol in all cases. Unlike the cross-validation testing, the reduced Coulomb descriptor performs the best in all cases except for QM7_4. This suggests that despite its lower dimensionality, the reduced Coulomb descriptor embodies as much "useful information" as the higher-order descriptors, at a significantly lower cost with respect to both storage and computation.

3.2 Larger Datasets and Prediction of Multiple Properties

To further investigate the performance of LC-GAP, we repeated the cross-validation and upwards-transferability tests on the atomization energy figures provided by the remaining two datasets, QM7b and GDB9_18K, and their respective subsets. Due to the competitive performance of the reduced Coulomb descriptor, we did not perform

Table 4: MAE for different algorithms when training and evaluating on the whole QM7 dataset.

Algorithm	MAE [kcal/mol]
LC-GAP (localized Coulomb matrix)	1.42
LC-GAP (decaying Coulomb matrix)	1.60
LC-GAP (reduced Coulomb matrix)	1.34
Kernel ridge regression [8]	3.07
Multilayer neural network [8, 10]	3.51
Bag of Bonds [7]	1.5

Table 5: Best choices of reduced Coulomb matrix descriptor parameters (α, r_{cut}) for upwards transferability, by data set and number of heavy atoms in training set. These choices are used to generate the results seen in Figure 3.

Dataset	3	4	5	6	7	8
QM7	(3.0, 6.0)	(5.0, 6.0)	(3.0, 5.5)	(4.0, 5.5)	(5.0, 6.5)	-
QM7b	(5.0, 3.5)	(5.0, 5.5)	(5.0, 5.5)	(5.0, 5.0)	(5.0, 5.5)	-
GDB9	(5.0, 4.5)	(5.0, 4.0)	(5.0, 5.0)	(6.0, 6.0)	(5.0, 4.0)	(5.0, 4.0)

further testing with the localized and decaying Coulomb descriptors; similarly, due to computational limitations, we did not perform cross-validation testing on the entire GDB9_18K dataset. As before, kernel hyperparameters were chosen by minimization of the negative log-likelihood of the training set, and a grid of descriptor parameters (α, σ) were used. Graphs displaying the results for each set of experiments can be found in Figure 3 and Figure 4 respectively, with the previously-obtained results for QM7 provided for comparison. The kernel hyperparameters and descriptor parameters used for each are given in Table 5 and Table 6.

Table 6: Best choices of reduced Coulomb matrix descriptor parameters (α, r_{cut}) for upwards transferability, by data set and number of heavy atoms in training set. These choices are used to generate the results seen in Figure 4.

Dataset	3	4	5	6	7
QM7	(4.0, 5.5)	(4.0, 5.0)	(5.0, 5.0)	(5.0, 6.0)	-
QM7b	(4.0, 5.5)	(4.0, 5.0)	(6.0, 5.0)	(5.0, 5.0)	-
GDB9	(5.0, 6.0)	(7.0, 4.0)	(5.0, 5.5)	(5.0, 4.0)	(6.0, 4.0)

In both cross-validation and upwards-transferability tests, the QM7b results perform similarly, although generally slightly better than the QM7 results. This is unsurprising, given that the datasets are similar, and QM7b contains only slightly more molecules than
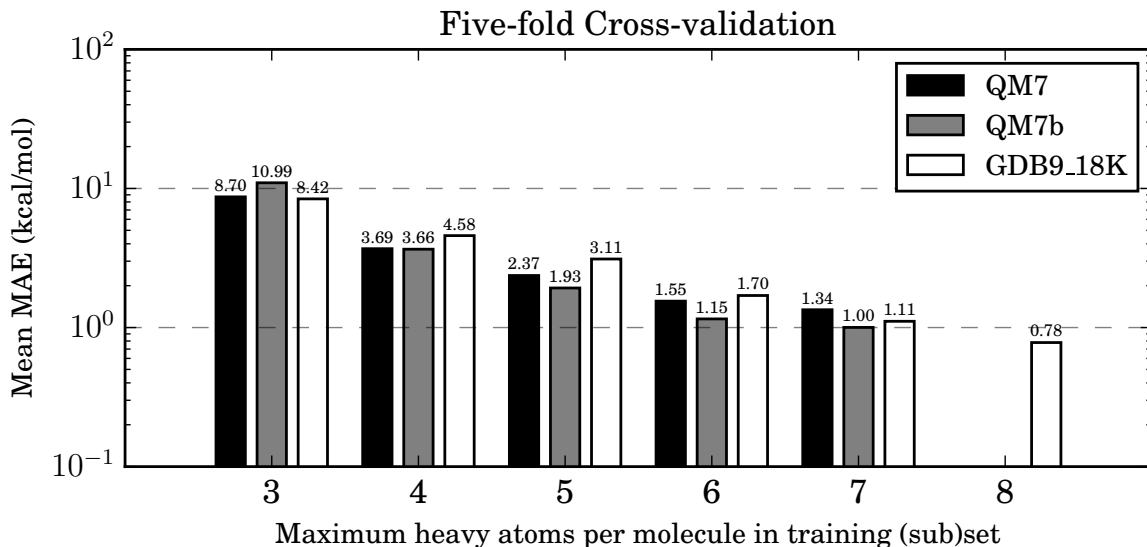
Figure 3: Mean MAEs of the predicted atomization energies on subsets of QM7, QM7b, and GDB9_18K, obtained during five-fold cross-validation, using LC-GAP potentials equipped with the reduced Coulomb descriptor (6). The precise values of α and r_{cut} used for each result can be found in Table 5.

QM7. However, we note that during cross-validation over the entirety of the dataset, LC-GAP is able to predict molecules in QM7b to a mean MAE of approximately 1.00 kcal/mol, i.e., at chemical accuracy.

In contrast, LC-GAP underperforms on GDB9 when considered against both QM7 and QM7b at the same number of heavy atoms. In the upwards-transferability case, this can be explained by the fact that the vast majority of the molecules in GDB9 have eight or nine heavy atoms, increasing the difficulty of prediction by a potential trained on molecules containing fewer heavy atoms. For cross-validation, this is potentially a result of the slightly lower number of molecules in the smaller subsets (cf. Table 1). The results for GDB9 are nevertheless encouraging: a potential trained on molecules containing eight or fewer heavy atoms can predict the remainder of the dataset (i.e., all molecules with nine heavy atoms) with an MAE of only 1.20 kcal/mol, and cross-validation over the entire dataset produces a mean MAE of 0.78 kcal/mol – well above chemical accuracy.

As well as atomization energies, QM7b contains a number of different molecular properties for each molecule in the dataset, as described above. In [4], De et al. report prediction results for these properties obtained with the SOAP kernel. To investigate the applicability of the LC-GAP system to this problem, we repeatedly performed cross-validation over the entire QM7b dataset, considering each property in turn. Importantly, we did not repeat the process of hyperparameter and descriptor parameter selection for each property; rather, we used those indicated in the atomization-energy case as above. A summary of the obtained results and a comparison with those of De et al. can be found in Table 7; additionally, scatter plots indicating the distribution of both the predicted results and their absolute errors can be found in Figure 5.

Without exception, the mean MAE results achieved by LC-GAP are on the same order
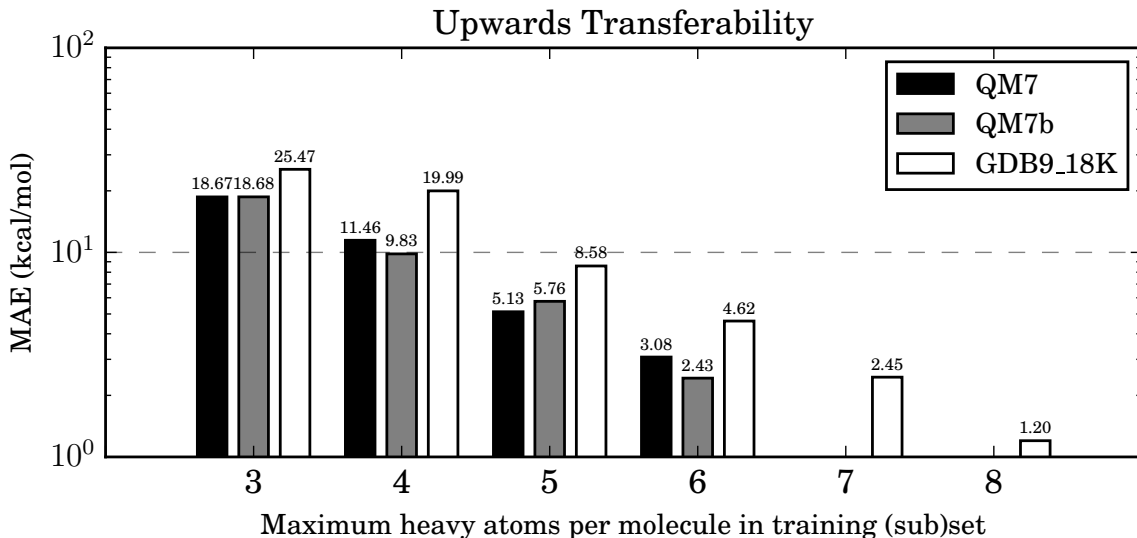
11

Figure 4: MAEs of the predicted atomization energies on the remainder of the full dataset, using LC-GAP potentials trained on subsets, equipped with the reduced Coulomb descriptor (6). The precise values of α and r_{cut} used for each result can be found in Table 6.

of magnitude as those reported in the reference data; many agree to at least one significant figure, although in no cases are the results better than those reported; considering the slight difference between the results for atomization energy, this is unsurprising. The precise distribution of the results is harder to interpret; almost all of the lower-subrow scatter plots display similar patterns of errors to one another, regardless of the difference between the reference data accuracy and that achieved by LC-GAP. Nevertheless, these results indicate that the LC-GAP potential is, in general, quite capable of predicting molecular properties using hyperparameters and descriptor parameters that were not optimized for that specific property – potentially a significant benefit, since parameter optimization is significantly more computationally intensive than simple training and application of a potential.

3.3 Distribution of Individual Atomic Contributions

Through the atomic decomposition ansatz, the GAP approach predicts the total value of any given property for a molecule as the sum of many atomic contributions, one for each atom in that molecule. Although previous work has focused only on the total value of a property, the atomic contributions are readily available, and are worthy of interest.

To inspect the distribution of the atomic contributions for atomization energy, we repeated the computation for the cross-validation of the complete QM7 dataset with the reduced Coulomb descriptor. The full set of all atomic contributions for each cross-validation split was retained, and the values for each of the five splits were compiled; the resulting set can be considered as a collection of all atomic contributions obtained during prediction of that dataset. A histogram of the distribution of all atomic contributions for

Table 7: Results obtained during cross-validation on the QM7b dataset for each of the fourteen molecular properties contained therein, with comparison figures obtained from [4]. For all results, the reduced Coulomb localizer was used; kernel hyperparameters and descriptor parameters were equivalent to those used to produce the entries in graph for the full QM7b dataset.

Property	Units	LC-GAP Mean MAE	Reference Data [4] MAE
E (PBE0)	kcal/mol	1.002 ± 0.022	0.92
$E^*_{\max}$ (ZINDO)	eV	1.717 ± 0.025	1.56
$I_{\max}$ (ZINDO)	arbitrary	0.087 ± 0.003	0.08
HOMO (ZINDO)	eV	0.410 ± 0.009	0.13
LUMO (ZINDO)	eV	0.228 ± 0.003	0.10
$E^*_{1\mathrm{st}}$ (ZINDO)	eV	0.493 ± 0.007	0.18
IP (ZINDO)	eV	0.439 ± 0.010	0.19
EA (ZINDO)	eV	0.267 ± 0.004	0.13
HOMO (PBE0)	eV	0.284 ± 0.006	0.11
LUMO (PBE0)	eV	0.091 ± 0.001	0.08
HOMO (GW)	eV	0.355 ± 0.008	0.12
LUMO (GW)	eV	0.146 ± 0.003	0.12
α (PBE0)	Å^3	0.072 ± 0.002	0.05
α (SCS)	Å^3	0.086 ± 0.002	0.02

QM7 can be found in Figure 3.3; additionally, a violin plot displaying the distribution of all atomic contributions (plotted by a kernel density estimator), as well as those of each of the five atom types contained in the dataset (hydrogen, carbon, nitrogen, oxygen, and sulfur) is given in the same figure.

The histogram displays a notable peak around approximately -70. This peak can be understood by inspection of the violin plots for the specific elements; a large majority of the contributions for hydrogen atoms cluster around this value. This suggests that many of the hydrogen atoms contribute similar amounts to the atomization energy of their respective molecules, and is interesting, because hydrogen atoms are by far the most prevalent in the QM7 dataset. By contrast, the contributions for the other atom types are relatively spread out.

Although more work needs to be done with respect to the classification of these contributions, we can draw one particular preliminary conclusion from them. The tight clustering of hydrogen atoms suggests large-scale redundancies of information in the underlying dataset; in the context of the GAP framework, this will produce covariance matrices with high condition numbers and corresponding numerical inaccuracies. This is born out by our informal observations; we have noted that covariance matrices encountered when attempting to train potentials using even larger datasets than those used here (such as the full GDB9 dataset) have prohibitively high condition, particularly for suboptimal choices of kernel and descriptor parameters. We suspect that methods for reducing the redundancies contained within the training datasets (such as low-rank approximations
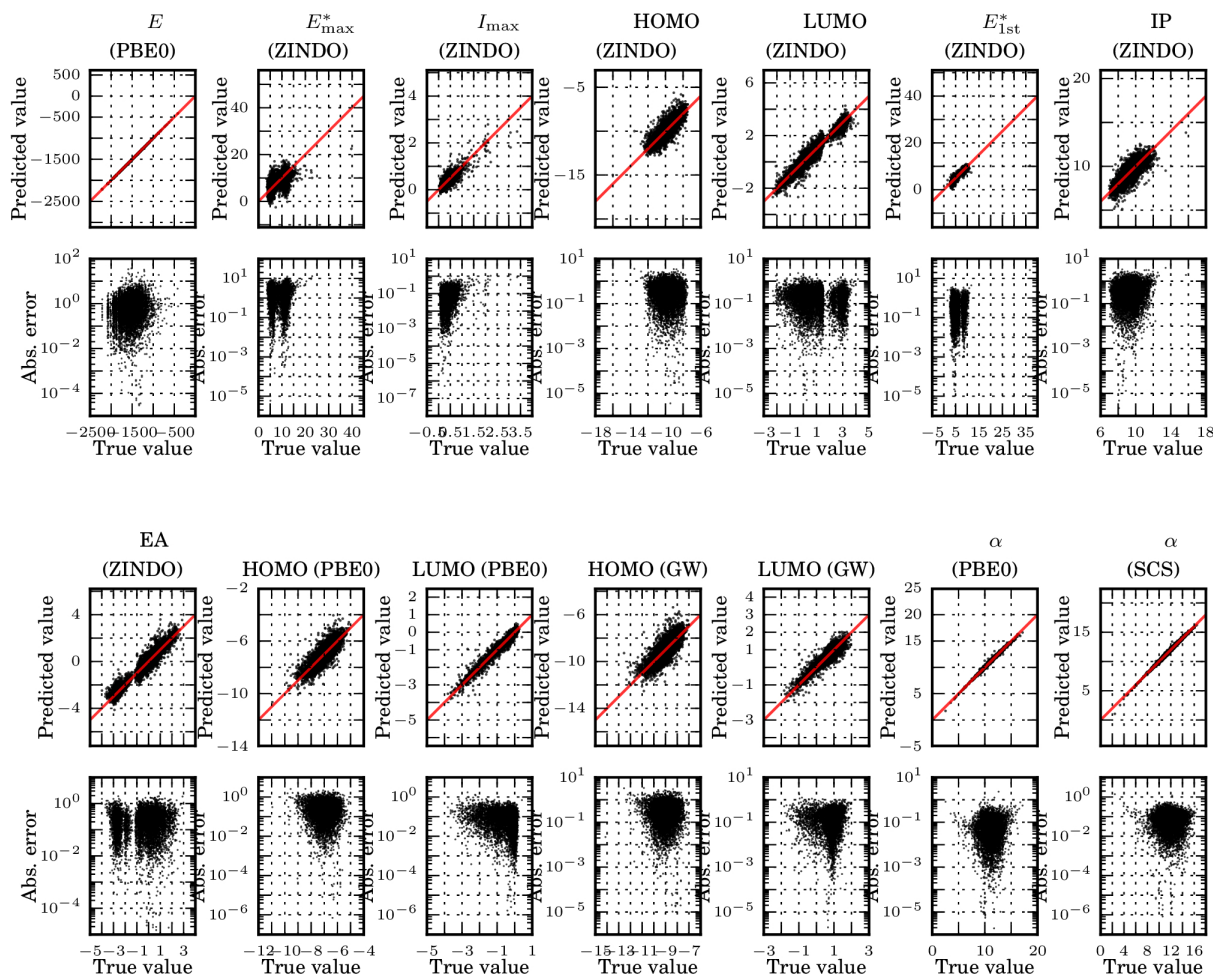
Figure 5: Scatter plots displaying the distribution of all results obtained during cross-validation over the QM7b dataset for each of the fourteen included properties, corresponding to the results in Table 7. In each main row, plots in the upper subrow display the obtained prediction values plotted against their actual values; plots in the lower subrow display the absolute error of the predictions, plotted against the actual value.

of covariance matrices and/or dataset sparsification techniques) will yield results of significantly higher quality, while reducing computational cost.

4   Conclusions and Future Work

We have introduced three new numerical representations for atomic neighborhoods, which can be applied to the Gaussian approximation potential (GAP). Numerical experiments showed that our LC-GAP implementation is capable of predicting the atomization energies of organic molecules in the QM7 dataset at levels of accuracy competitive with similar potentials already extant in the literature, for all three presented representation types. Furthermore, we have demonstrated that LC-GAP potentials trained on systems containing lower numbers of atoms can be used to predict the energy values for systems containing
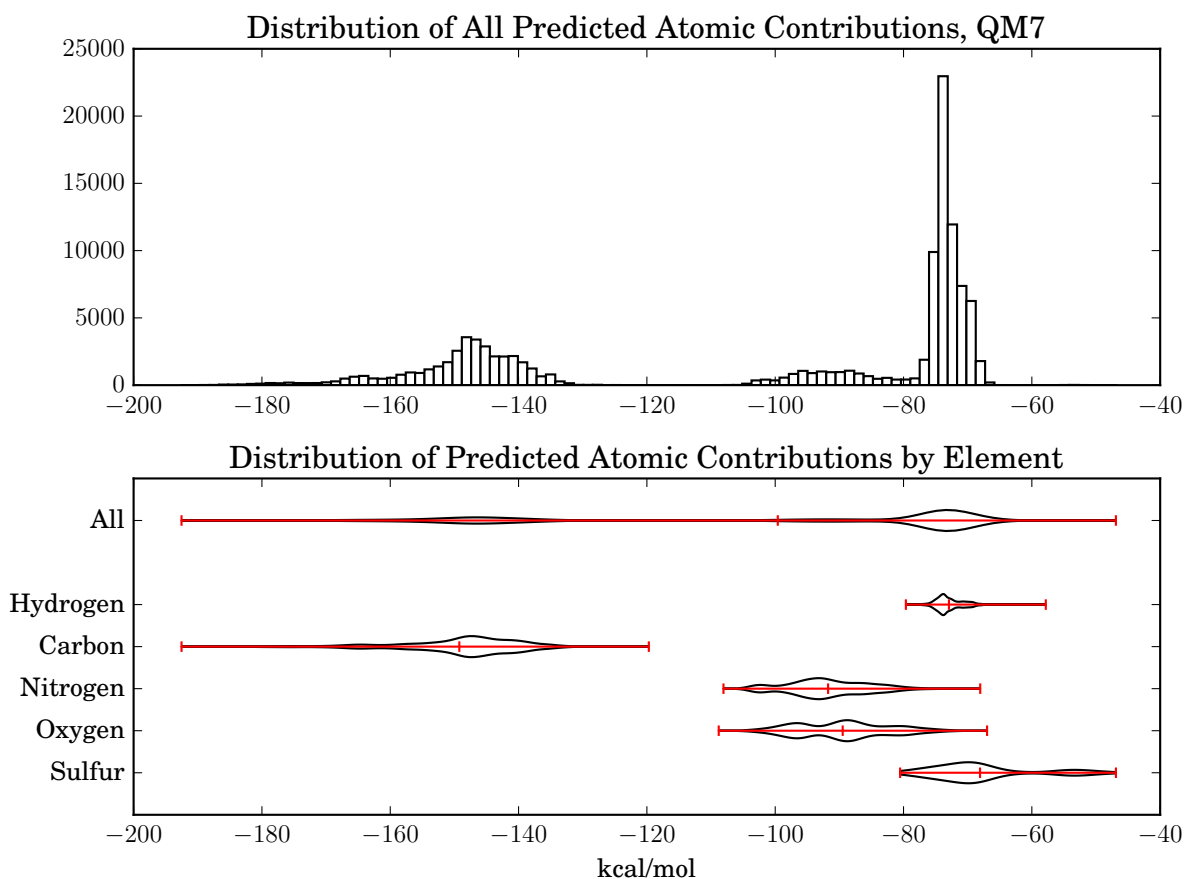
Figure 6: Plots showing the distribution of all calculated individual atomic contributions towards the atomization energies of molecules in QM7, obtained during five-fold cross-validation. The upper plot is a standard histogram. The lower plot displays an indication of the density function of the entire collection of contributions (the topmost object), as well as indications of the individual density functions of all contributions by atoms of a particular element (remaining objects). Straight lines indicate range, with mean marked; curves are kernel density estimators.

higher numbers of atoms, again with acceptable accuracy.

Of the three representations presented here, the reduced Coulomb descriptor has linear dimensionality in the number of particles in a local neighborhood, while the other two descriptor types have quadratic dimensionality. Despite this, the reduced Coulomb descriptor performs comparably well, if not better, in all cases. This descriptor also performs well when used to evaluate alternative molecular properties than the atomization energy, producing results comparable to the literature when applied to the QM7b many-properties dataset. Finally, when used for cross-validation on the QM7b and GDB9 datasets, LC-GAP equipped with the reduced Coulomb descriptor predicts atomization energies at and comfortably below chemical accuracy respectively, achieving MAEs of approximately 1.00 and 0.78 kcal/mol.

A number of questions remain about the utility of these representation functions

when training on and predicting different kinds of atomic systems. In particular, as all systems in the QM7 dataset are equilibrium-state biomolecules, it would be interesting to investigate the behavior of the LC-GAP system over non-equilibrium systems, as well as infinitely-periodic crystal systems such as semiconductors. Additionally, an analysis of per-atom contributions produced during cross-validation suggest the existence of high levels of redundancy in the QM7 dataset; we strongly suspect that the application of techniques for lessening the impact of this redundancy will produce worthwhile results.

Acknowledgement

References

[1] A. P. Bartók, R. Kondor, and G. Csányi. On representing chemical environments. *Physical Review B*, 87(18):184115, 2013.

[2] A. P. Bartók, M. C. Payne, R. Kondor, and G. Csányi. Gaussian approximation potentials: The accuracy of quantum mechanics, without the electrons. *Physical Review Letters*, 104(13):136403, 2010.

[3] L. C. Blum and J.-L. Reymond. 970 million druglike small molecules for virtual screening in the chemical universe database GDB-13. *Journal of the American Chemical Society*, 131(25):8732–8733, 2009.

[4] S. De, A. P. Bartók, G. Csányi, and M. Ceriotti. Comparing molecules and solids across structural and alchemical space. *Physical Chemistry Chemical Physics*, 18(20):13754–13769, 2016.

[5] S. Fliszár. *Atoms, Chemical Bonds, and Bond Dissociation Energies*. Lecture Notes in Chemistry. Springer-Verlag, 1994.

[6] M. Griebel, S. Knapek, and G. Zumbusch. *Numerical Simulation in Molecular Dynamics: Numerics, Algorithms, Parallelization, Applications*, volume 5 of *Texts in Computational Science and Engineering*. Springer Science & Business Media, Heidelberg, 2007.

[7] K. Hansen, F. Biegler, R. Ramakrishnan, W. Pronobis, O. A. von Lilienfeld, K.-R. Müller, and A. Tkatchenko. Machine learning predictions of molecular properties: Accurate many-body potentials and nonlocality in chemical space. *The Journal of Physical Chemistry Letters*, 6(12):2326–2331, 2015.

[8] K. Hansen, G. Montavon, F. Biegler, S. Fazli, M. Rupp, M. Scheffler, O. A. von Lilienfeld, A. Tkatchenko, and K.-R. Müller. Assessment and validation of machine learning methods for predicting molecular atomization energies. *Journal of Chemical Theory and Computation*, 9(8):3404–3419, 2013.

[9] W. Kohn. Density functional and density matrix method scaling linearly with the number of atoms. *Physical Review Letters*, 76(17):3168, 1996.

[10] G. Montavon, K. Hansen, S. Fazli, M. Rupp, F. Biegler, A. Ziehe, A. Tkatchenko, O. A. von Lilienfeld, and K.-R. Müller. Learning invariant representations of molecules for atomization energy prediction. In *Advances in Neural Information Processing Systems*, pages 440–448, 2012.

[11] G. Montavon, M. Rupp, V. Gobre, A. Vazquez-Mayagoitia, K. Hansen, A. Tkatchenko, K.-R. Müller, and O. A. von Lilienfeld. Machine learning of molecular electronic properties in chemical compound space. *New Journal of Physics*, 15(9):095003, 2013.

[12] S. J. Plimpton and A. P. Thompson. Computational aspects of many-body potentials. *MRS bulletin*, 37(05):513–521, 2012.

[13] E. Prodan and W. Kohn. Nearsightedness of electronic matter. *Proceedings of the National Academy of Sciences of the United States of America*, 102(33):11635–11638, 2005.

[14] R. Ramakrishnan, P. O. Dral, M. Rupp, and O. A. von Lilienfeld. Quantum chemistry structures and properties of 134 kilo molecules. *Scientific Data*, 1, 2014.

[15] C. E. Rasmussen and C. K. I. Williams. *Gaussian Processes for Machine Learning*. The MIT Press, 2006.

[16] M. Rupp. Machine learning for quantum mechanics in a nutshell. *International Journal of Quantum Chemistry*, 115(16):1058–1073, 2015.

[17] M. Rupp, R. Ramakrishnan, and O. A. von Lilienfeld. Machine learning for quantum mechanical properties of atoms in molecules. *The Journal of Physical Chemistry Letters*, 6(16):3309–3313, 2015.

[18] M. Rupp, A. Tkatchenko, K.-R. Müller, and O. A. von Lilienfeld. Fast and accurate modeling of molecular atomization energies with machine learning. *Physical Review Letters*, 108(5):058301, 2012.

[19] E. Tadmor and R. Miller. *Modeling Materials: Continuum, Atomistic and Multiscale Techniques*. Cambridge University Press, 2011.

[20] C. J. Willmott and K. Matsuura. Advantages of the mean absolute error (MAE) over the root mean square error (RMSE) in assessing average model performance. *Climate Research*, 30(1):79, 2005.