Interaction Potentials in Molecules and Non-Local Information in Chemical Space

Katja Hansen[1], Franziska Biegler[2], O. Anatole von Lilienfeld[3,4], Klaus-Robert Müller[2,5], and Alexandre Tkatchenko[1*]
[1] *Fritz-Haber-Institut der Max-Planck-Gesellschaft, Faradayweg 4-6, 14195, Berlin, Germany*
[2] *Machine Learning Group, Technical University of Berlin, Marchstr. 23, 10587 Berlin, Germany*
[3] *Institute of Physical Chemistry, Department of Chemistry, University of Basel, Klingelbergstrasse 80, CH-4056 Basel, Switzerland*
[4] *Argonne Leadership Computing Facility, Argonne National Laboratory, Argonne, Illinois 60439, USA*
[5] *Department of Brain and Cognitive Engineering, Korea University, Korea*
(Dated: February 26, 2014)

We develop and apply a systematic hierarchy of efficient empirical methods to estimate atomization and total energies of molecules. These methods range from a simple sum over "dressed" atoms, addition of bond energies, pairwise interatomic force fields, reaching to the more sophisticated machine learning approaches that are capable of describing collective interactions between many atoms or bonds. In the case of equilibrium molecular geometries, even simple pairwise force fields demonstrate prediction accuracy comparable to benchmark energies calculated using density-functional theory with hybrid exchange-correlation functionals. However, accounting for the collective many-body interactions proves to be essential for approaching the "holy grail" of chemical accuracy of 1 kcal/mol for both equilibrium and out-of-equilibrium geometries. This remarkable accuracy is achieved by the so-called *Bag-of-Bonds* model that exhibits strong non-locality in chemical space.

Chemical compound space (CCS) is the space populated by all possible energetically stable molecules varying in composition, size, and structure [1]. Chemical reactions and transformations due to external perturbations allow us to explore this astronomically large space in order to obtain molecules with desired properties (e.g., stability, mechanical properties, linear and non-linear response, ionization potential, electron affinity). The accurate prediction of these molecular properties in the CCS is a crucial ingredient toward rational compound design in chemical and pharmaceutical industries. Therefore, one of the major challenges is to enable quantitative calculations of molecular properties in CCS at moderate computational cost (milliseconds per molecule or faster). However, currently only wavefunction-based quantum-chemical calculations, which can take up to several days per molecule, consistently yield the desired "chemical accuracy" of 1 kcal/mol required for predictive *in silico* rational molecular design.

Furthermore, a unique mathematical definition of CCS is lacking because the mapping between molecular geometries and molecular properties is often not unique, meaning that there can be very different molecules exhibiting very similar properties. This complexity is reflected by the existence of hundreds of descriptors that aim to measure molecular similarity in chemoinformatics [2, 3]. In this context, one of our goals is to shed light into the structure and properties of CCS in terms of molecular atomization energies that is an essential molecular property measuring the stability of a molecule with respect to its constituent atoms. The total energy of a molecule can be trivially determined from its atomization energy by simply adding free atom energies. Under certain conditions chemical reaction barriers can also be correlated to the difference between total energies of two molecules. In this Letter, we gradually construct more reliable models that include one-, two- and finally many-body interactions between atoms or bonds. This cascade of models highlights the importance of many-body effects and also illustrates to which amount they can be incorporated as effective terms of lower complexity. Moreover, the so-called *Bag-of-Bonds* approach introduced in this Letter enables us to demonstrate the impact of non-local information in CCS that turns out to be crucial for achieving a prediction accuracy of 1.5 kcal/mol for a database of more than 7000 organic molecules. Our research is aimed towards the fundamental goal of understanding the structure and properties of CCS composed of molecules with arbitrary stoichiometry. It is thus complementary to other important efforts for constructing potential-energy surfaces of molecules and solids [4–7].

Evidently, the dimensionality of CCS grows exponentially with increasing molecular size. However, typical databases of synthetically accessible molecules are rather restricted in their composition. To avoid systematic bias yet enable complete exploration of a subset of CCS, we selected all 7165 molecules from the GDB database containing up to seven "heavy" (C, N, O, S) atoms saturated with hydrogens to satisfy valence rules [8, 9] (this database is referred to as GDB-7 throughout this work). In contrast to other widely employed databases, GDB includes *all* molecular graphs corresponding to a set of simple organic chemistry rules. The initial geometries in the GDB-7 database were generated using OpenBabel from their associated SMILES descriptors [10]. Subsequently, the geometries were optimized using density functional theory calculations with the PBE exchange-correlation functional [11] in the FHI-aims code [12]. Finally, the atomization energies were computed using the hybrid PBE0 functional [13] at the optimized PBE geometries. The PBE0 functional yields atomization energies with an overall accuracy better than 5 kcal/mol

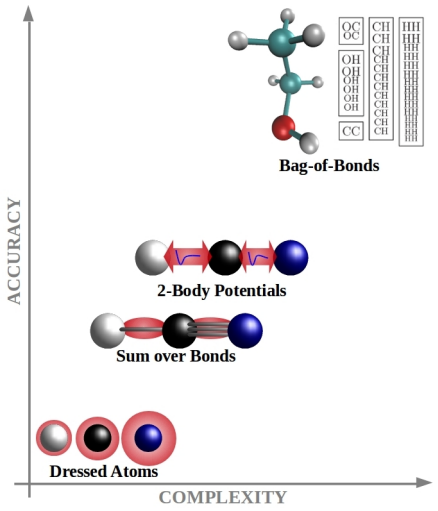

FIG. 1. Schematic overview of employed modeling approaches: The dressed atoms model incorporates only the atoms and weights them according to their type. The Sum-over-Bonds and 2-Body potentials consider pairs of atoms and the interactions between them. The most sophisticated model is the Bag-of-Bonds (BoB) model, which is illustrated for the ethanol molecule (C_2H_5OH). The BoB model implements an effective collective energy expression based on all interatomic distances within a molecule.

when compared with experiment [14], and these atomization energies will be used as a reference.

The hierarchy of models that we use to predict atomization energies is schematically illustrated in Figure 1, ordered from crudest models to the more sophisticated ones in order of complexity. The performance of the different models, evaluated using a rigorous five-fold cross-validation approach, is shown in Table I. The predictive power of various methods as a function of training set size is plotted in Figure 4, and analyzed later in this work.

In the most naive approximation, the energy of a molecule $\mathbf{M}$ can be written as a sum of effective atomic contributions

$$\hat{E}_{\mathrm{DA}}(\mathbf{M}) = \sum_{\alpha}^{\substack{\mathrm{Atom}\\ \mathrm{types}}} n_\alpha E_\alpha, \qquad (1)$$

where $\hat{E}_{\mathrm{DA}}$ refers to the approximated atomization energy, E_α is a "dressed" atomic energy for atom type α (with $\alpha \in H, C, N, O, S$), and n_α is the number of α-type atoms. This *dressed atoms* model yields an accuracy of 15.1 kcal/mol on the GDB-7 database, which is not dreadful considering the simplicity of this dramatic approximation.

Molecules form as a result of chemical bonding, hence an approach that considers bonds rather than atoms is expected to perform much better. We define a bond by the type of bonded atoms and bond order (single, double, triple). This definition leads to the so-called *sum-over-bonds* model, which improves significantly over dressed atoms model, achieving an accuracy of 9.9 kcal/mol on the GDB-7 database. However, the sum-over-bonds model is still unable to treat changes in bond distances and interatomic interactions beyond nearest neighbors.

TABLE I. Performance of different models evaluated in five-fold cross-validation on the GDB-7 database: The data was randomly split into five sets of 1433 molecules each. Each of these five sets serves once as test set while the remaining 5732 molecules are used for training. The table shows the mean performance of the models over the five runs. Performance is measured in each run as mean absolute error (MAE) on the test set. The variance of the performance over the five sets was below 3% for all models.

Model	MAE [$kcal/mol$]
Dressed atoms	15.1
Sum-over-bonds	9.9
Lennard-Jones potential	8.7
Polynomial pot. ($n = 6$)	5.6
Polynomial pot. ($n = 10$)	3.9
Polynomial pot. ($n = 18$)	3.0
Bag-of-Bonds ($p = 2$, Gaussian)	4.5
Bag-of-Bonds ($p = 1$, Laplacian)	1.5
Rupp et al. [15]	10.0
best model in Ref. [16]	3.1

Both of these effects can be included by constructing an effective pairwise interatomic potential

$$\hat{E}_{\mathrm{PP}}(\mathbf{M}) = \sum_{ij}^{\mathrm{Atom\ pairs}} \sum_{r}^{r\ \mathrm{of\ type}\ ij} \Phi_{ij}(r), \qquad (2)$$

where $\Phi(r)$ is an effective potential function for each type of atom pair ij (carbon–carbon C–C, carbon–nitrogen C–N, and so on). We note that different functional forms can be adopted for $\Phi(r)$, ranging from Lennard-Jones and Morse-type to more general polynomial potentials. Already the usage of Lennard-Jones potential yields an accuracy of 8.7 kcal/mol on the GDB-7 database. This is because the Lennard-Jones potential reproduces the basic features which a general interatomic potential should possess – repulsive wall at short distances, a well-defined minimum and the van der Waals r^{-6} decay of the interaction at large interatomic distances.

More general pairwise potentials can be constructed by a systematic expansion of $\Phi(r)$ using powers of the inverse distance r^{-n}. The performance of such polynomial models as a function of the maximum degree n is shown in Table I. Already for $n = 6$, the accuracy of 5.6 kcal/mol for polynomial potentials exceeds that of the optimal Lennard-Jones potential. The improvement in atomization energies saturates around $n = 18$, reaching an accuracy of 3.0 kcal/mol. To put this number in perspective, we recall that 3 kcal/mol is below the error of the reference PBE0 atomization energies when compared to experiment [14]. Moreover, the performance of the most sophisticated machine learning models in Ref. [16]

applied to a similar dataset was 3.1 kcal/mol. Therefore, it is remarkable that a simple and very efficient model based on pairwise potentials is able to capture the subtle energetic contributions required to predict atomization energies for equilibrium molecular geometries.

Seeking to better understand this finding, we plot the optimized C–C potential in Figure 2 for different values of n. For small values of n the potential is smooth yet more flexible than the simple Lennard-Jones potential, explaining the increased accuracy of such polynomial expansions. The increase in the degree of the polynomial leads to the appearance of shoulders and minima related to different bond orders or hybridization states. In fact, these features appear at interatomic distances well known from empirical determinations of bond orders and energies [17]. We thus conclude that the increase in the degree of the polynomial enables the potential to "learn" about chemical bonding. Similar observations as for the C–C potential are made for other atom types; we demonstrate this fact by showing similar C–N and and C–O potentials in the supplemental material [18]. The improvement in the predictive power of polynomial potentials does not only arise from their ability to distinguish between different bonding scenarios. The decay of these potentials with interatomic distance is rather slow, with energy contributions beyond nearest neighbors (> 1.5 Å) having an essential role on the scale of the obtained error (see inset in Fig. 2). We note in passing that another attractive feature of interatomic potentials is that by construction they can exactly reproduce the limit of dissociated atoms, a condition that is difficult to fulfill even in state-of-the-art *ab initio* theory. While we used polynomial potentials in this work, other choices of basis functions are certainly possible. To explore whether a different choice of pairwise potentials would be beneficial, we employed machine learning techniques to construct a general spline-based form of a pairwise potential. We found no significant accuracy gain when utilizing such a general form, thus concluding that a polynomial is sufficiently flexible.

While the performance of pairwise potentials is already quite good, they have a few notable drawbacks. For example, their performance for out-of-equilibrium molecular geometries will be strongly degraded. In order to demonstrate this, we extended the GDB-7 database by scaling all the interatomic distances in the molecules by a factor of 0.9 and 1.1, enlarging the database by a factor of three. When trying to learn the atomization energies for both equilibrium and out-of-equilibrium molecular geometries, the performance of pairwise potentials diminished by 16.7 kcal/mol compared to pure equilibrium geometries. This test demonstrates that while pairwise potentials can be successfully applied in preliminary studies of stabilities for equilibrium geometries (when these are given from some other method), more sophisticated approaches are required to optimize molecular geometries.

Going beyond pairwise potentials, we propose a more

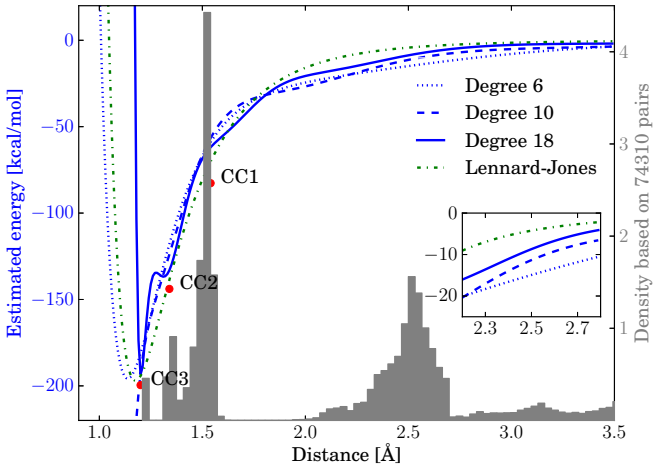

FIG. 2. Polynomial potentials for C–C interaction: The normalized gray histogramm refers to the distribution of C–C distances within the GDB-7 dataset and is associated with the right-hand axis. The red dots represent the energies of the C–C single, double and triple bond as given by fits to experimental bond energies [17]. In blue, polynomial two-body potentials (as trained in cross validation) are shown. The inset shows the difference between potentials for distances between 2.2 Å and 2.8 Å.

sophisticated machine learning approach to determine interatomic potentials which we call *Bag-of-Bonds* (BoB). The BoB concept is inspired by text mining descriptors utilized in computer science [19, 20] and is illustrated in Figure 1 using the ethanol molecule as an example. The BoB model represents molecules as concatenated vectors (bags), where each bag represents a particular bond type (i.e. C–C, C–N, and so on). Motivated by the Coulomb matrix concept of Rupp *et al.* [15], each entry in every bag is computed as $Z_i Z_j / |\mathbf{R}_i - \mathbf{R}_j|$, where Z_i and Z_j are the nuclear charges while $\mathbf{R}_i$ and $\mathbf{R}_j$ are the positions of the two atoms participating in a given bond. In order to vectorize this information, we simply concatenate all bags of bonds in a specified order, filling each bag with zeros in order to give the bags equal sizes across all molecules in the GDB-7 database and sorting the entries in each bag according to their magnitude. This representation is naturally invariant under molecular rotations and translations, whereas the permutational invariance is enforced by the sorting step. We note in passing that unlike the sorted Coulomb matrix [15] the BoB descriptor is not able to distinguish between homometric molecules [21], however our database is devoid of such cases.

Within the BoB model, the energy of a molecule $\mathbf{M}$ is written as a sum over weighted exponentials centered on every molecule I in the training set

$$\hat{E}_{\mathrm{BoB}}(\mathbf{M}) = \sum_{I=1}^{N} \alpha_I \exp\left(-d(\mathbf{M}, \mathbf{M}_I)/\sigma\right), \qquad (3)$$

where $d(\mathbf{M}, \mathbf{M}_I) = \sum_j \|M^j - M_I^j\|_p$ measures the distance between the BoB vectors corresponding to molecules $\mathbf{M}$ and $\mathbf{M}_I$ ($\|x\|_p$ refers to the l_p norm of x), α_I are the regression coefficients, the kernel width σ is optimized for each choice of p, and I runs over all molecules $\mathbf{M}_I$ in the training set of size N. The values of α_I coefficients and σ are determined by the kernel ridge regression procedure as described in detail elsewhere [16].

The flexibility in choosing the kernel function (or metric) in CCS allows us to investigate the locality properties of chemical space for the prediction model in terms of atomization energies. The high sensitivity of the BoB model on the employed kernel function is demonstrated in Table I, where a Gaussian kernel ($p = 2$) leads to an accuracy of 4.5 kcal/mol versus a much improved performance of 1.5 kcal/mol for a Laplacian kernel ($p = 1$). It is noteworthy that the Laplacian BoB model is able to come close to the elusive "chemical accuracy" of 1 kcal/mol, which is a widely accepted level of accuracy required for truly predictive *ab initio* calculations on molecular systems. To further elucidate the role of non-local information in chemical space in the prediction of atomization energies, we have systematically studied the dependence of the prediction accuracy on the metric norm p employed in Eq. 3. We find that the optimal value of p is close to unity and the predictive capability decreases significantly for $p < 0.5$ and $p > 1.5$. For larger values of p, e.g. $p = 2$, the resulting model is more local and yields worse results. For kernel-based models it is possible to calculate the contribution to the predicted value for each compound in the training set. Adding up all contributions from compounds close to the compound in question we obtain a "local estimate" of the predicted value. Figure 3 illustrates how this local estimate of the atomization energy converges towards the predicted value with growing molecular neighborhoods in the case of ethanol molecule for Gaussian and Laplacian kernels. The BoB model based on the Gaussian kernel is unable to use information from "distant" molecules and leads to a lousy prediction, while the Laplacian kernel is able to optimally utilize non-local information in CCS. Similar results are found for other molecules in the GDB-7 database.

In contrast to pairwise potentials, the good performance of the BoB approach extends also to non-equilibrium molecular geometries. For the extended GDB-7 database with stretched and compressed geometries described above, the prediction error of BoB increases only by 0.8 kcal/mol. This is a direct reflection of the ability of the BoB approach to correctly capture the intricate collective interactions between many bonds within organic molecules. To understand this aspect better, we can decompose the BoB Laplacian kernel model for a molecule $\mathbf{M}$ as $\exp(-\sum_j^n |M^j - M_I^j|/\sigma) = \prod_j^n \exp(-|M^j - M_I^j|/\sigma)$. Taylor-series expansion of the exponential as a function of internuclear Coulomb repulsion and the subsequent product will include contributions up to infinite order in terms of bond pairs between molecules $\mathbf{M}$ and $\mathbf{M}_I$. Simple sum over bonds and pairwise potential approaches can be constructed as lower-order expansions of the BoB model, given sufficient training data. In fact, a connection between the BoB model and pairwise potentials can be established by approximately rewriting the BoB kernel as $\sum_l^{n_b} \prod_{j \in b_l} \exp(-|M^j - M_I^j|/\sigma)$, where b refers to a certain type of bond (e.g., C–C) and n_b is the length of the bag corresponding to the bond type b. We found that such partial linearization of the BoB model leads to a significant loss of accuracy on the order of a few kcal/mol, clearly demonstrating the crucial role of many-body effects accounted for by the non-linear infinite-order nature of the Laplacian kernel. Exploring further connections between the BoB model and explicit many-body potentials offers a promising direction for future work.

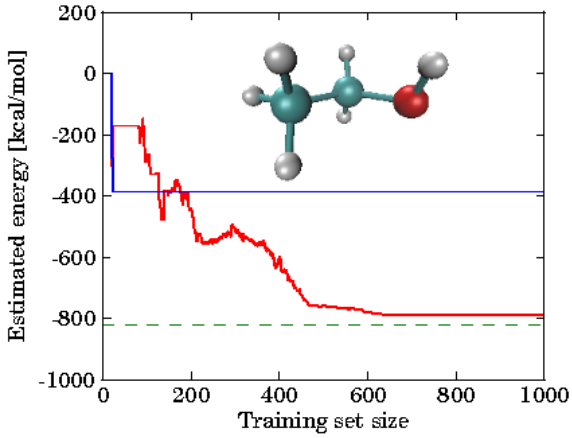

FIG. 3. (Color online) Estimated energy of the ethanol molecule (C_2H_5OH) as predicted by the BoB model using Gaussian (blue line) and Laplacian (red line) kernels. The PBE0 reference energy is indicated by the dashed green line. The training set was composed of N randomly selected molecules from the GDB-7 database, up to a maximum of 1000 molecules.

Another advantage of the BoB model over pairwise potentials is its better transferability and smooth prediction improvement with the number of training samples, as shown in Figure 4. Already when using just 1000 random molecules out of GDB-7 for training, the BoB model demonstrates prediction accuracy comparable to the best optimized polynomial potential with degree 18, which requires more than 5000 training samples to achieve the same level of accuracy. At a first glance, this is surprising considering that the polynomial potential contains less adjustable parameters. However, it is clear from Figure 4 that BoB represents a more robust machine learning model with proper regularization and that further improvement in accuracy is possible by sim-

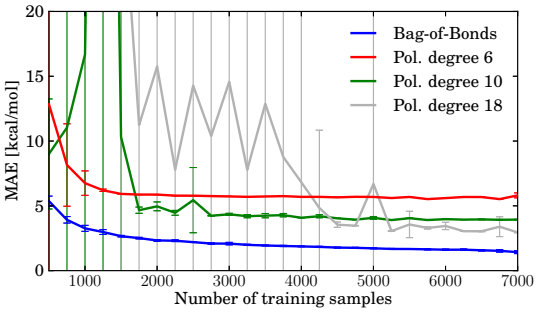

FIG. 4. Mean absolute error (MAE in kcal/mol) for BoB and polynomial models: Training sets from $N = 500$ to 7000 data points were sampled identically for the different methods (5 splits were consistently utilized for each training set size). The polynomial model of degree 10 and 18 exhibit high variances due to the random stratification, which for small N leads to non-robust fits.

ply enlarging the molecular database. The fact that the BoB model is able to predict atomization energies with an accuracy of at least 3 kcal/mol for the whole GDB-7 database with just 15% of molecules utilized for training (and 1.5 kcal/mol with 70% of training molecules) demonstrates the great promise of this approach for further exploration and understanding of CCS.

* tkatchenko@fhi-berlin.mpg.de

[1] P. Kirkpatrick and C. Ellis, Nature **432**, 823 (2004).
[2] G. Schneider, Nature Rev. **9**, 273 (2010).
[3] R. Todeschini and V. Consonni, *Handbook of Molecular Descriptors* (Wiley-VCH, Weinheim, 2009).
[4] S. Manzhos and T. Carrington, J. Chem. Phys. **125**, 194105 (2006).
[5] J. Behler and M. Parrinello, Phys. Rev. Lett. **98**, 146401 (2007).
[6] J. Behler, J. Chem. Phys. **134**, 074106 (2011).
[7] A. P. Bartók, M. C. Payne, R. Kondor, and G. Csányi, Phys. Rev. Lett. **104**, 136403 (2010).
[8] T. Fink, H. Bruggesser, and J.-L. Reymond, Angew. Chem. Int. Ed. **44**, 1504 (2005).
[9] T. Fink and J.-L. Reymond, J. Chem. Inf. Model. **47**, 342 (2007).
[10] R. Guha *et al.*, J. Chem. Inf. Model. **46**, 991 (2006).
[11] J. P. Perdew, K. Burke, and M. Ernzerhof, Phys. Rev. Lett. **77**, 3865 (1996).
[12] V. Blum *et al.*, Chem. Phys. Commun. **180**, 2175 (2009).
[13] J. P. Perdew, M. Ernzerhof, and K. Burke, J. Chem. Phys. **105**, 9982 (1996).
[14] B. J. Lynch and D. G. Truhlar, J. Phys. Chem. A **107**, 3898 (2003).
[15] M. Rupp, A. Tkatchenko, K.-R. Müller, and O. A. von Lilienfeld, Phys. Rev. Lett. **108**, 058301 (2012).
[16] K. Hansen *et al.*, J. Chem. Theory and Comput. **9**, 3404 (2013).
[17] S. W. Benson, J. Chem. Educ. **42**, 502 (1965).
[18] See supplemental material at http://xxx for additional information about pairwise potentials and plots of non-locality in chemical space.
[19] G. Forman, The Journal of Machine Learning Research **3**, 1289 (2003).
[20] T. Joachims, *Text categorization with support vector machines: Learning with many relevant features* (Springer, 1998).
[21] A. L. Patterson, Nature **143**, 939 (1939).