Articles
doi.org/10.1002/syst.201900052

Chemistry Europe
European Chemical Societies Publishing

Check for updates

Special Collection

Size-Extensive Molecular Machine Learning with Global Representations**

Hyunwook Jung+,[a, b] Sina Stocker+,[a] Christian Kunkel,[a] Harald Oberhofer,[a] Byungchan Han,[b] Karsten Reuter,[a] and Johannes T. Margraf*[a]

Machine learning (ML) models are increasingly used in combination with electronic structure calculations to predict molecular properties at a much lower computational cost in high-throughput settings. Such ML models require representations that encode the molecular structure, which are generally designed to respect the symmetries and invariances of the target property. However, size-extensivity is usually not guaranteed for so-called global representations. In this contribution, we show how extensivity can be built into global ML models using, *e.g.*, the Many-Body Tensor Representation. Properties of extensive and non-extensive models for the atomization energy are systematically explored by training on small molecules and testing on small, medium and large molecules. Our results show that non-extensive models are only useful in the size-range of their training set, whereas extensive models provide reasonable predictions across large size differences. Remaining sources of error for extensive models are discussed.

1. Introduction

In recent years, machine-learning (ML) methods are increasingly applied to the prediction of molecular properties such as atomization and orbital energies, dipole moments and ionization potentials.[1–9] One of the main promises of ML in chemistry is that it allows surpassing the size and time scales accessible to accurate first-principles electronic structure calculations, *e.g.* based on density-functional theory (DFT). This is particularly relevant in a high-throughput setting, *e.g.* when a large chemical reaction network with many intermediates and transition states is to be explored, or a large chemical space is of interest.[10–13]

The wide range of ML methods that have emerged in this context raises the question which one should be used for a given application. Since the atomization energy (AE) has a long tradition as the foremost benchmark property to judge the accuracy of quantum chemical approximations,[14–16] it has also become one of the standard targets to illustrate the accuracy of novel ML methods.[1,3] The most straightforward way to construct a ML model for the AE is to use some vectorized representation v of the molecule . Constructing the ML model is then simply a regression task between v and the property of interest $y(v)$.[17] While any general linear or non-linear regression method (*e.g.* Kernel Ridge Regression, KRR or Artificial Neural Networks) can be used, the choice of the representation is critical. In particular, several physically motivated criteria such as translational, permutational, and rotational invariance and uniqueness should be fulfilled.[5,18]

The Coulomb matrix (CM) developed by Rupp et al.[4] was one of the earliest (global) molecular representations used to this end (see below for a specification of global in contrast to local representations). However, it suffers from two notable limitations, namely that the size of v depends on the number of atoms in the system and that permutational invariance can only be achieved through a canonical ordering of the vector elements.[2] This led to several subsequent improvements of the CM concept, such as the Bag-of-Bonds,[19] different histogram based methods[1] and the Many-Body Tensor Representation (MBTR).[6,18] These representations fix the main drawbacks of the CM and can thus be used to construct more accurate and data-efficient ML models of molecular properties, typically using KRR.

However, the combination of KRR with global representations still suffers from the problem that the resulting predictions are typically not size-extensive. This should in principle be a fundamental problem for predicting any extensive property like the AE. In practice, this issue can be and has been overlooked to some extent, as the databases that are hitherto typically used to test ML models (*e.g.* QM9)[20] do not contain large size differences. For example, ca. 97% of the molecules in QM9 contain 8 or 9 heavy atoms. Consequently, an approximate size-extensivity of the model can be learned by simply including all small systems in the training set explicitly.[17] However, this only obscures the fundamental problem, and such a model will fail when applied to significantly larger molecules. Similarly, the

[a] *H. Jung,+ S. Stocker,+ C. Kunkel, PD Dr. H. Oberhofer, Prof. Dr. K. Reuter, Dr. J. T. Margraf*
Chair for Theoretical Chemistry and Catalysis Research Center, Technische Universität München, Lichtenbergstraße 4, D-85747 Garching, Germany
E-mail: johannes.margraf@ch.tum.de

[b] *H. Jung,+ Prof. Dr. B. Han*
Department of Chemical and Biomolecular Engineering, Yonsei University, Seoul 03722, Republic of Korea

[+] *These authors contributed equally to the manuscript*

[**] *A previous version of this manuscript has been deposited on a preprint server (DOI: 10.26434/chemrxiv.10002020).*

Supporting information for this article is available on the WWW under https://doi.org/10.1002/syst.201900052

An invited contribution to a Special Collection on the Computational Chemistry of Complex Systems

description of chemical reactions (where a large molecule can decompose into smaller fragments) cannot be consistently achieved when the predicted energies are non-extensive.[21]

The goal of the present paper is to address the size-extensivity of ML models that use a global representation of the molecular structure, using KRR models with the MBTR of Huo and Rupp as an illustrative example.[18] We will discuss how extensive ML models can be constructed with MBTR and compare them with the conventional, non-extensive formulation. Importantly, the performance of the models is compared across different size-ranges both within the QM9 database and between databases going up to molecules with more than 80 heavy atoms.[11,22]

2. Theory

Kernel Ridge Regression: In KRR, the target property $y(v)$ (*i.e.* here the AE) of an unknown molecule with the representation v is calculated via:

$$y(v) = \sum_i w_i K(v, v_i) \tag{1}$$

where v_i are the representations of training data points and w_i are regression weights. Here, we introduced the kernel function $K(v, v')$, which provides a similarity measure between two representations v and v'. A common choice for $K(v, v')$ is the Gaussian kernel:

$$K(v, v') = \exp\left(-\frac{\|v - v'\|_2^2}{2\sigma^2}\right). \tag{2}$$

Here, σ is the kernel length scale, a hyperparameter that governs how prone the kernel is to classify systems as similar. Specifically, a large value of σ will indicate some degree of similarity between most inputs, whereas a small value will only find similarities for systems that are very close in feature space. Below, we also use the linear kernel, which simply consists of the dot-product of v and v'.

The optimal (in a least-squares sense) set of weights ω can be obtained via the expression:

$$w = (\mathbf{K} + \lambda \mathbf{I})^{-1}\mathbf{y}, \tag{3}$$

where $\mathbf{K}$ is the kernel matrix of the training set (with $K_{ij} = K(v_i, v_j)$), λ is a regularization parameter and $\mathbf{I}$ is the identity matrix. λ is another hyperparameter of the model, which represents the uncertainty of the observations.

Training a KRR model is thus a simple linear algebra operation. Obviously, the performance of the model critically depends on the choice of representation and kernel function. In analogy to the common notation of Functional/Basis-Set in DFT, this choice is designated as Representation/Kernel in the following.

Many-Body Tensor Representation: Herein, we use the MBTR of Huo and Rupp as a prototypical global representation of molecular structure.[18] Simply put, the MBTR provides a measure of how often characteristic geometric features (corresponding to different orders of a many-body expansion) occur. Canonically, these features are atom counts (1-body), inverse interatomic distances (2-body), angles (3-body), dihedrals (4-body), etc. For each body-order and element combination, a broadened distribution function of these features is constructed as a sum of Gaussians, as shown in Figure 1 for the 2-body terms in water. These Gaussians are additionally scaled by a distance-dependent weighting function, which introduces a characteristic length-scale to the representation. Beyond this length-scale atoms or molecules are effectively non-interacting.

For a given body order k and N_{species} chemical species there are in principle $N_{\max} = N_{\text{species}}^k$ such distribution functions. Although some combinations can be excluded by symmetry (*i.e.* C–H is equal to H–C), this means that the size of the MBTR vector quickly explodes with the body order. In practice, the MBTR is therefore usually limited to the lowest order terms, *i.e.* including up to 2- or 3-body contributions. The final MBTR vector is obtained by concatenating the discretized feature distribution functions $v_{k,i}$:

$$v^{MBTR} = v_{1,1} \oplus v_{1,2} \oplus \ldots \oplus v_{k_{max},N_{max}} \tag{4}$$

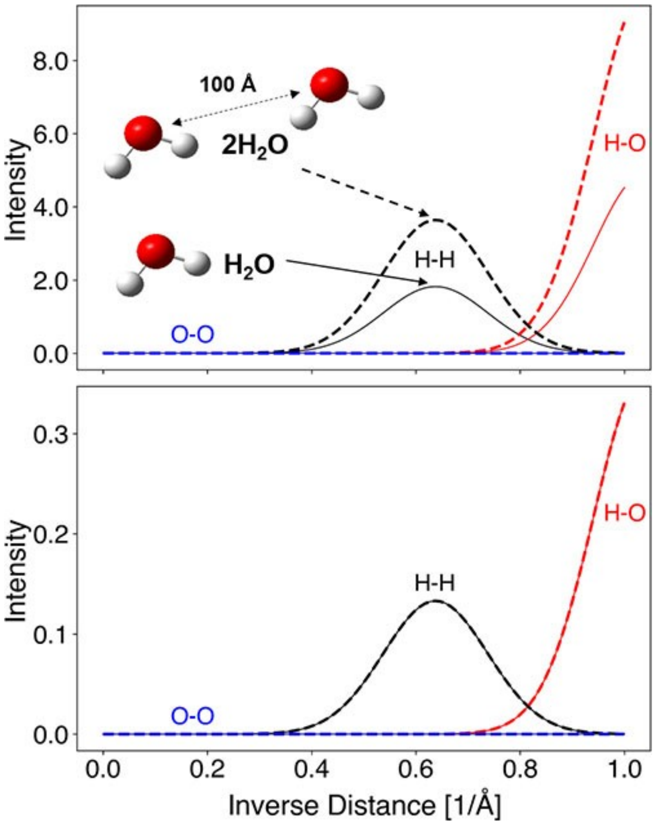

Figure 1. Sample illustration of 2-body MBTR output of a single water molecule (solid) and two distant water molecules (dashed). Interatomic interactions include: H–H (black), H–O (red), and O–O (blue). (Top) MBTR (Bottom) iMBTR

In this original formulation, the representation thus contains absolute counts of the occurrence of a given feature. In contrast, below, we also consider a *normalized* version of the MBTR, where each distribution function is normalized according to its l^2-norm:

$$v^{\mathrm{iMBTR}} = \frac{1}{\left\|v_{1,1}\right\|} \cdot v_{1,1} \oplus \frac{1}{\left\|v_{1,2}\right\|} \cdot v_{1,2} \oplus \ldots \oplus \frac{1}{\left\|v_{k_{\max},N_{\max}}\right\|} \cdot v_{k_{\max},N_{\max}} \quad (5)$$

For clarity, this normalized MBTR version is designated as iMBTR (for *intensive*).

Size-Extensivity: According to eq. 1, the target property (here the AE) is predicted as a linear combination of kernel functions. Consequently, it is advantageous if the kernel can be constructed in such a way that it adheres to conditions known to be fulfilled by the target property. For example, the AE is invariant to translations and rotations of a molecule. Consequently, MBTR-based kernels are constructed to satisfy these same invariances.

A less commonly imposed condition relates to the extensive or intensive nature of the target property. As with the invariances, the kernel should ideally reflect the extensivity or intensivity of the property of interest. Specifically, for two non-interacting molecules A,

$$K(A, 2A) = 2 \times K(A, A), \quad (6)$$

for an extensive property (such as the AE) and

$$K(A, 2A) = K(A, A), \quad (7)$$

for an intensive property (such as the ionization potential).

Unfortunately, the original MBTR/Gaussian kernel is neither intensive nor extensive. While the distribution functions that make up the representations for A and $2A$ have identical shapes, the amplitude of each peak is twice as large for $2A$ (see Figure 1, top). Since the norm of the difference between MBTR vectors enters the Gaussian kernel, it will evaluate to approximately zero (depending on the lengthscale σ). In contrast, the combination iMBTR/Gaussian leads to an *intensive* kernel. This is because the iMBTR for an arbitrary number of non-interacting molecules becomes identical to the single molecule case due to its normalization (see Figure 1, bottom). Finally, the combination MBTR/linear leads to an *extensive* kernel. This can easily be verified by considering that each element in the MBTR of 2A differs from the MBTR of A by a factor of two.

From this perspective, the MBTR/linear kernel appears to be the most appropriate choice for learning AEs. However (as the name implies) KRR with the linear kernel is simply linear regression. As the main advantage of KRR is the introduction of non-linearity (*e.g.* via the Gaussian kernel), this is not ideal.

Fortunately, we can resort to a simple trick to obtain an extensive non-linear KRR model. Specifically, an iMBTR/Gaussian model can be trained to predict the atomization energy per atom (AE/N), which is an intensive quantity. Indeed, it has already been suggested in the context of electronic structure methods that AE/N may actually be a more appropriate target for fitting and benchmarking.[21,23]

Note that this *intensive atomization energy* should not be interpreted as a *local* atomic energy (see below). Instead it can be understood as a generalization of the concept of cohesive energy for extended crystals to finite systems.[23] In Figure 2, AE/N is plotted for linear hydrocarbons (*i.e.*, alkanes, alkenes, and alkynes) of different sizes. All three curves converge to a constant value (the cohesive energy of the corresponding 1D crystal) for large systems and display a smooth dependence on the number of atoms for smaller systems. To predict the AE with the iMBTR/Gaussian model, we thus train on AE/N and subsequently simply multiply the prediction by the number of atoms. For comparison, the original MBTR/Gaussian and MBTR/linear models are trained on the AE, as usual.

Global and Local representations: So far, we have focused on the general case of a *global* representation v, which encodes the entire structure of the molecule/system with the property $y(v)$. A major advantage of global ML models is that the assumed relationship between structure and property mirrors the fact that any property can in principle be computed from the Schrödinger equation.[24,25] Meanwhile, a significant drawback is that the cost of computing global representations does not scale linearly with the size of the system. This inhibits the use of global representation as universal descriptors applicable to proteins or solids. Fortunately, this is not problematic for molecular systems with tens to hundreds of atoms. A second, more critical aspect is that global representations are not automatically size-extensive, as discussed in the previous section.

In contrast to this, a variety of *local* ML models have been developed that guarantee size-extensivity and linear scaling.[26–28] In the tradition of empirical interatomic potentials, these models approximate the total property (here the AE) as a sum of local (*e.g.* atomic) contributions:

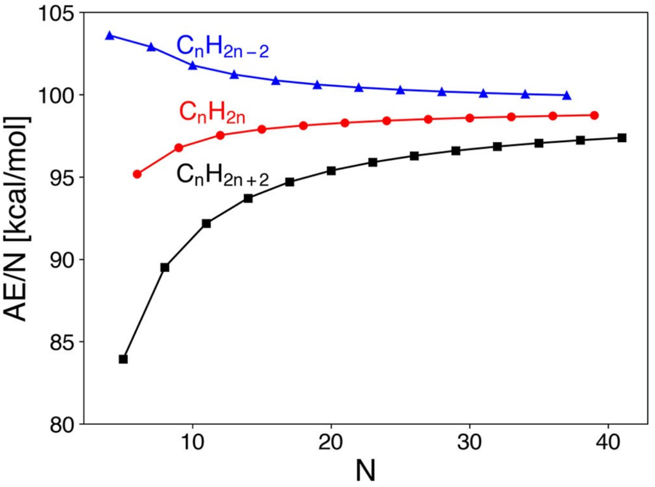

Figure 2. Atomization energy per atom for linear alkanes (C_nH_{2n+2}), alkenes (C_nH_{2n}) and alkynes (C_nH_{2n-2}) from C_1 to C_{13}.

$$y \approx \sum_{\text{atom}} y_{\text{atom}} \tag{8}$$

Here, the local properties y_{atom} (*e.g.* atomic energies in the case of AE) only depend on the immediate chemical environment of the atom. Importantly, these local energies result from an optimal decomposition of the total property and are not necessarily physically meaningful.

While the expression in Eqs. 8 is manifestly extensive, it also generally introduces an approximation to the model. For instance, in the case of total energies or AEs it effectively neglects any long-range interatomic interactions. Furthermore, the local properties (like a local energy) might not necessarily be quantum mechanical observables. In practice, the severity of this approximation is property and material dependent. For example, in many cases excellent interatomic potentials based on Eqs. 8 have been obtained.[29,30]

For kernel-based regression, there is an interesting connection between global and local representations, as there are several ways to convert local to global kernels. For example, as noted by Bártok and coworkers, a ML potential based on the local SOAP representation is equivalent to a global model using the *averaged* kernel:[3]

$$K(A,B) = \sum_{i\in A, j\in B} \frac{1}{N_A N_B} k(i,j), \tag{9}$$

where $K(A,B)$ is a global kernel comparing molecules A and B, and $k(i,j)$ is a local kernel comparing atoms i and j. Similarly, a sum of local kernels can also form a global kernel:[31]

$$K(A,B) = \sum_{i\in A, j\in B} k(i,j), \tag{10}$$

From the perspective taken in this paper, Eqs. 9 and 10 are recipes to construct global kernels from local representations, which conform to Eqs. 7 and 6, respectively. These kernels are special cases of the general case discussed herein, in the sense that local representations can be used to build extensive kernels, but not all extensive kernels must be built from local representations. Recently, Tamblyn and coworkers also suggested semi-local, extensive ML models based on deep neural networks.[32]

3. Methods

Datasets: In this paper, we use two reference databases of DFT AEs, namely the QM9 and OE62 sets.[20,22] The QM9 set includes over 134,000 drug-like organic molecules and is frequently used as a benchmark for ML studies.[1,3,6] The molecules in QM9 have a heavy atom count (HAC) of up to nine and are comprised of the elements H, C, O, N, and F. As alluded to above, most of these molecules (ca 97%) contain 8 or 9 heavy atoms. This leaves a total of 3993 molecules with a HAC = 1–7, which we will use for training.

The OE62 dataset originates from a high-throughput screening study for organic semiconductors by Schober et al. and has also been used for benchmarking different ML methods.[6,22] While somewhat smaller than QM9 (61,489 molecules) it is significantly more chemically diverse. For example, OE62 contains 16 different elements and much larger molecules, with up to 174 atoms (max. HAC = 92).

Predicting properties of the OE62 set is therefore a very hard task for ML models trained on the small molecules contained in QM9, but it should in principle be possible for a size-extensive model. However, this can only work if both datasets are consistent. We therefore focus here on a subset of 32,467 OE62 molecules that contain the same elements as QM9 (H, C, O, N, and F). Furthermore, the original QM9 data was computed at the B3LYP/6-31G(2df,p) level, whereas the OE62 database is based on the Perdew-Burke-Ernzerhof (PBE) functional with Tkatchenko-Scheffler Van-der-Waals correction (PBE-vdW), tight integration grids and a "tier2" basis set of numerical atomic orbitals.[33–35] To increase the consistency between both datasets, the atomization energies for all QM9 molecules were correspondingly recomputed with the OE62 settings (using the original QM9 geometries). This new dataset is freely available from the authors.

Hyperparameter Optimization: The hyperparameters σ and λ (from Eqs. 2 and 3) were optimized through 4-fold cross validation (CV). Specifically, the parameters that minimize the average root mean square difference (RMSD) in CV were obtained using the Nelder-Mead minimization algorithm[36,37] as implemented in the scikit-learn package.[38] MBTR vectors were obtained via the *DScribe* package, including only one- and two-body terms.[39] Unlike σ and λ, the MBTR-specific hyperparameters were not optimized, and the default values for broadening and damping functions were used (see SI).

We note that using higher order terms and optimizing all hyperparameters would certainly lead to somewhat lower errors. However, the goal of this study is not to benchmark MBTR itself but to understand the role of size-extensivity on ML models with global representations. For this purpose, we found the above choices to be adequate.

4. Results and discussion

As discussed in the theory section, we will focus on three KRR models, namely the combinations MBTR/Gaussian, iMBTR/Gaussian and MBTR/linear. In line with previous ML studies on predicting AEs, we start by checking the predictive performance of the models within a dataset.[2,6,18] Here, we focus on a subset of QM9, containing all 3,993 molecules with up to seven heavy atoms. The average RMSD from 4-fold CV on this set is shown in Table 1.

The MBTR/Gaussian kernel performs best, followed by the iMBTR/Gaussian and MBTR/linear models. This shows the benefit of the non-linear Gaussian kernel, though the results of the linear kernel are also respectable, in line with what was reported by Huo and Rupp.[18] For consistency, all errors are reported with respect to total AEs, even for the iMBTR/Gaussian

Table 1. Averaged RMSD from 4-fold cross validation KRR models trained on the 3,993 QM9 molecules with HAC = 1–7.

MBTR Normalization	Kernel	Training Target	RMSD (kcal/mol)
iMBTR	Gaussian	AE/N	3.14
MBTR	linear	AE	4.09
MBTR	Gaussian	AE	2.30

model which is trained on AE/N. It is therefore not surprising that iMBTR/Gaussian performs somewhat more poorly than MBTR/Gaussian, given that it minimizes a different loss function. Still, one might naively conclude from this analysis that the conventional MBTR/Gaussian kernel is suitable for predicting AEs, in spite of its lacking extensivity.

This picture changes radically when the models are forced to extrapolate beyond the scope of their training sets, however. To this end, we consider a separate test set of 2000 QM9 molecules with nine heavy atoms. In addition to the standard HAC = 1–7 training set, we thereby also consider training sets containing only up to four, five, and six heavy atoms, respectively, to specifically test the extrapolation capabilities of the models. The results for all models are summarized in Figure 3.

Contrary to the previous result, the original MBTR/Gaussian method now shows the worst prediction performance among the three models, which is a direct manifestation of its lacking size-extensivity. Even the (extensive) MBTR/linear model shows significantly lower RMSD compared to MBTR/Gaussian. Finally, the iMBTR/Gaussian model combines proper extensivity with the non-linearity of the Gaussian kernel and performs best. Indeed, it even provides qualitatively useful predictions (with a relative error of ca. 1–2 %) for the smallest training set, which consists of just 48 molecules with up to four heavy atoms.

An even more challenging test case is predicting the AEs of the OE62 set while still training on QM9. As mentioned above, the latter has a very narrow heavy atom distribution (peaking at 9) whereas the former has a wide distribution peaking around 20 (see Figure S1 in the SI). Furthermore, OE62 contains chemical structures that are absent from QM9, such as polycyclic aromatic compounds. As before, the models are trained on the 3,993 QM9 molecules with up to seven heavy atoms.

The correlations between predicted and reference AEs for all KRR models are shown in Figure 4. Here, the most notable feature is the abysmal performance of MBTR/Gaussian, with an RMSD of 4,327 kcal/mol. While the model actually displays reasonable accuracy up to AEs around ca. 2,500 kcal/mol (*i.e.* for molecules similar to the training set), it completely fails beyond this range. Indeed, as the kernel function vanishes for large molecules, the model predicts an AE of zero for all large molecules. This poor performance of MBTR/Gaussian vividly demonstrates its lack of size-extensivity.

In contrast, the iMBTR/Gaussian and MBTR/linear models both show good correlations with the reference across the full range of systems ($R^2 = 0.99$), with dramatically lower RMSDs of 184.4 and 138.2 kcal/mol, respectively. At first glance, this is still a large margin of error, compared to the results for QM9. It should however be noted that the error of a predicted AE should itself be size-extensive, so that larger errors are to be expected for larger systems.[21] Given that the AEs of the OE62 set range up to ca. 18,000 kcal/mol, an RMSD of ca. 100 kcal/mol is actually not that poor in relative terms. To quantify this, the RMSD can be normalized by the standard deviation of the Aes in the data set. This yields normalized RMSDs of 0.10 (iMBTR/Gaussian) and 0.08 (MBTR/linear), respectively (where 1.0 would be the performance of a random Gaussian model with appropriate mean and standard deviation).

Furthermore, this error is quite systematic, with the AEs of large systems being consistently underestimated. A linear fit of the correlation plots reveals that this is a bit more pronounced for iMBTR/Gaussian than for MBTR/linear (see Figure 4). Indeed,

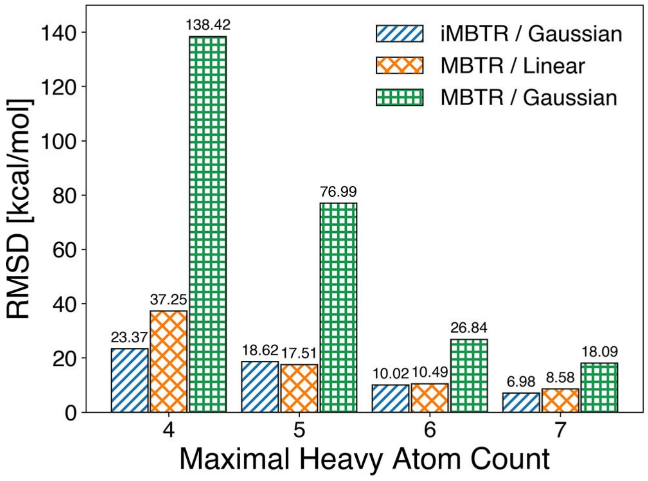

Figure 3. Accuracy of KRR models trained on small QM9 molecules (max. HAC = 4–7) when predicting larger molecules from QM9 (HAC = 9).

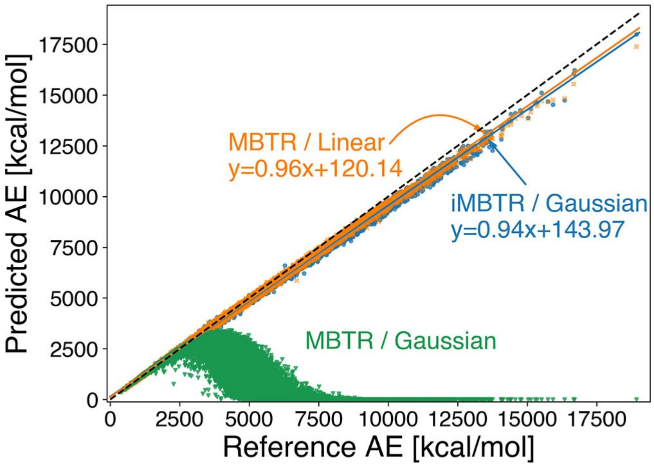

Figure 4. Correlation plots of predicted OE62 AEs for MBTR/Gaussian (▼ green), iMBTR/Gaussian (○ blue) and MBTR/linear (× orange). All models were trained on 3,993 QM9 molecules with HAC = 1–7. Prediction was performed on 32,467 OE62 molecules consisting of C, H, O, N and/or F. Linear regression lines and equations are shown for iMBTR/Gaussian (blue) and MBTR/linear (orange).

if the results of the linear regressions are subtracted from the predictions, the corresponding RMSDs are reduced to 63.85 kcal/mol (iMBTR/Gaussian) and 61.33 kcal/mol (MBTR/linear).

Of course, even in relative terms, the errors of these models are still larger than what would be expected purely based on the cross-validation RMSD of their training sets. This is because extensivity is not the only relevant size-effect. For example, long-range interactions like electrostatics and dispersion can play a significant role in stabilizing large molecules. Furthermore, electronic effects like quantum confinement may occur on the nanometer scale. These effects lead to a net stabilization of larger molecules, reflected in the systematic underestimation of the AEs mentioned above.

Consequently, AE/N is not converged for systems with seven heavy atoms, even in the fairly simple case of linear hydrocarbons (Figure 2). In Figure 5, the distribution of AE/N vs. N is shown for the full QM9 and OE62 sets. Interestingly, the basic features of this plot are remarkably similar to Figure 2. In particular, it can be seen that the mean AE/N is approximately constant for molecules with more than ca. 20 atoms. This regime corresponds to the largest molecules in QM9. The figure also provides an intuitive explanation of why the iMBTR/Gaussian method works. By choosing AE/N as the target quantity, the variability that the model must account for is decreased from ca. 18,000 kcal/mol to ca. 80 kcal/mol.

5. Conclusion

In this contribution, we have explored the size-extensivity of molecular ML models based on global representations such as the MBTR. While the conventional MBTR/Gaussian model is not ideal for either extensive or intensive properties, we showed that there are appropriate kernels for both cases, namely the MBTR/linear (extensive) and iMBTR/Gaussian (intensive). While current extensive ML models are typically built from local representations, our work shows that this is not strictly a requirement. We also showed how an intensive kernel can be used to predict an extensive property. To illustrate the significance of these results, a highly challenging ML task with large size differences between the molecules in the training and test sets was devised. We found that properly extensive models perform reasonably well in this setting, whereas the conventional MBTR/Gaussian approach fails outright.

Importantly, we stress that a non-extensive model can still be quite accurate if the size of the chemical space of interest is limited. However, in those areas of chemistry where ML is expected to have a large impact, this is not the case. In particular, for the study of large reaction networks (*e.g.* within systems chemistry or catalysis) a useful ML model must adequately describe the transition from small molecules to larger systems and even polymers (and *vice versa*). The present work represents an important stepping stone to this end.

Finally, it should be noted that the present study was purposefully designed to study the effects of size-extensivity in the limit of large size differences between training and test molecules. In practice, we expect that the systematic errors in the extensive models could be mitigated by including a limited number of larger molecules in the training set.

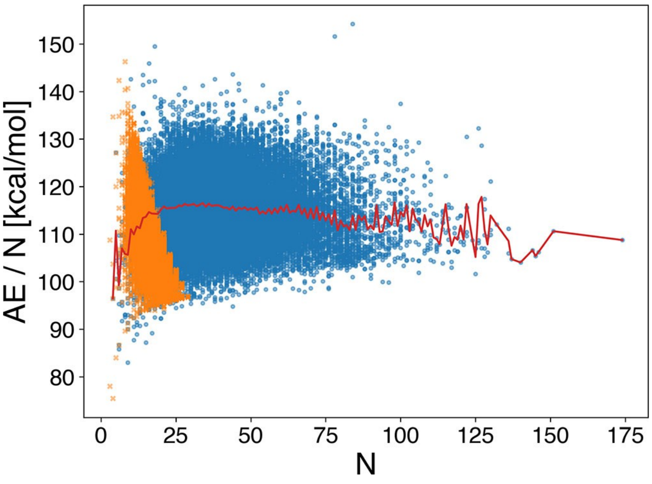

Figure 5. Plot of AE/N vs. N for molecules in QM9 (× orange) and OE62 (○ blue). The mean AE/N is shown as a red line.

Acknowledgements

This work was supported by the Korea Ministry of Environment (MOE) as "the Chemical Accident Prevention Technology Development Project" and the Global Frontier Program through the Global Frontier Hybrid Interface Materials (GFHIM) (2013-M3A6B1078882). We further acknowledge support by Deutsche Forschungsgemeinschaft (DFG) through TUM International Graduate School of Science and Engineering (IGSSE) (GSC 81). Computer time on resources of the Argonne Leadership Computing Facility was provided by the Innovative and Novel Computational Impact on Theory and Experiment (INCITE) program.

Conflict of Interest

The authors declare no conflict of interest.

Keywords: Machine learning · Kernel ridge regression · Many-body tensor representation · Size-extensivity · Atomization energy

[1] F. A. Faber, L. Hutchison, B. Huang, J. Gilmer, S. S. Schoenholz, G. E. Dahl, O. Vinyals, S. Kearnes, P. F. Riley, O. A. Von Lilienfeld, *J. Chem. Theory Comput.* **2017**, *13*, 5255–5264.

[2] K. Hansen, G. Montavon, F. Biegler, S. Fazli, M. Rupp, M. Scheffler, O. A. Von Lilienfeld, A. Tkatchenko, K.-R. Müller, *J. Chem. Theory Comput.* **2013**, *9*, 3404–3419.

[3] A. P. Bartók, S. De, C. Poelking, N. Bernstein, J. R. Kermode, G. Csányi, M. Ceriotti, *Sci. Adv.* **2017**, *3*, e1701816.

[4] M. Rupp, A. Tkatchenko, K.-R. Müller, O. A. Von Lilienfeld, *Phys. Rev. Lett.* **2012**, *108*, 58301.

[5] W. Pronobis, A. Tkatchenko, K.-R. Müller, *J. Chem. Theory Comput.* **2018**, *14*, 2991–3003.

[6] A. Stuke, M. Todorović, M. Rupp, C. Kunkel, K. Ghosh, L. Himanen, P. Rinke, *J. Chem. Phys.* **2019**, *150*, 204121.

[7] W. Pronobis, K. T. Schütt, A. Tkatchenko, K.-R. Müller, *Eur. Phys. J. B* **2018**, *91*, 178.

[8] R. Ramakrishnan, M. Hartmann, E. Tapavicza, O. A. Von Lilienfeld, *J. Chem. Phys.* **2015**, *143*, 84111.

[9] J. Kang, S. H. Noh, J. Hwang, H. Chun, H. Kim, B. Han, *Phys. Chem. Chem. Phys.* **2018**, *20*, 24539–24544.

[10] J. T. Margraf, K. Reuter, *ACS Omega* **2019**, *4*, 3370–3379.

[11] C. Kunkel, C. Schober, J. T. Margraf, K. Reuter, H. Oberhofer, *Chem. Mater.* **2019**, *31*, 969–978.

[12] C. Kunkel, C. Schober, H. Oberhofer, K. Reuter, *J. Mol. Model.* **2019**, *25*, 87.

[13] A. Bruix, J. T. Margraf, M. Andersen, K. Reuter, *Nat. Can.* **2019**, *2*, 659–670.

[14] J. A. Pople, M. Head-Gordon, D. J. Fox, K. Raghavachari, L. A. Curtiss, *J. Chem. Phys.* **1989**, *90*, 5622–5629.

[15] R. Peverati, D. G. Truhlar, *Philos. Trans. R. Soc. London* **2014**, *372*, 20120476.

[16] A. Karton, N. Sylvetsky, J. M. L. Martin, *J. Comput. Chem.* **2017**, *38*, 2063–2075.

[17] M. Rupp, *Int. J. Quantum Chem.* **2015**, *115*, 1058–1073.

[18] H. Huo, M. Rupp, *arXiv Prepr. arXiv1704.06439* **2017**.

[19] K. Hansen, F. Biegler, R. Ramakrishnan, W. Pronobis, O. A. Von Lilienfeld, K.-R. Müller, A. Tkatchenko, *J. Phys. Chem. Lett.* **2015**, *6*, 2326–2331.

[20] R. Ramakrishnan, P. O. Dral, M. Rupp, O. A. Von Lilienfeld, *Sci. Data* **2014**, *1*, 140022.

[21] J. T. Margraf, D. S. Ranasinghe, R. J. Bartlett, *Phys. Chem. Chem. Phys.* **2017**, *19*, 9798–9805.

[22] C. Schober, K. Reuter, H. Oberhofer, *J. Phys. Chem. Lett.* **2016**, *7*, 3973–3977.

[23] J. P. Perdew, J. Sun, A. J. Garza, G. E. Scuseria, *Z. Physiol. Chem.* **2016**, *230*, 737–742.

[24] R. Ramakrishnan, O. A. von Lilienfeld, *Chim. Int. J. Chem.* **2015**, *69*, 182–186.

[25] S. Chmiela, A. Tkatchenko, H. E. Sauceda, I. Poltavsky, K. T. Schütt, K.-R. Müller, *Sci. Adv.* **2017**, *3*, e1603015.

[26] J. Behler, M. Parrinello, *Phys. Rev. Lett.* **2007**, *98*, 146401.

[27] A. P. Bartók, M. C. Payne, R. Kondor, G. Csányi, *Phys. Rev. Lett.* **2010**, *104*, 136403.

[28] A. P. Bartók, R. Kondor, G. Csányi, *Phys. Rev. B* **2013**, *87*, 184115.

[29] S. Kondati Natarajan, J. Behler, *J. Phys. Chem. C* **2017**, *121*, 4368–4383.

[30] V. L. Deringer, G. Csányi, *Phys. Rev. B* **2017**, *95*, 094203.

[31] F. A. Faber, A. S. Christensen, B. Huang, O. A. von Lilienfeld, *J. Chem. Phys.* **2018**, *148*, 241717.

[32] K. Mills, K. Ryczko, I. Luchak, A. Domurad, C. Beeler, I. Tamblyn, *Chem. Sci.* **2019**, *10*, 4129–4140.

[33] V. Blum, R. Gehrke, F. Hanke, P. Havu, V. Havu, X. Ren, K. Reuter, M. Scheffl *Comput. Phys. Commun.* **2009**, *180*, 2175–2196.

[34] J. P. Perdew, K. Burke, M. Ernzerhof, *Phys. Rev. Lett.* **1996**, *77*, 3865–3868.

[35] A. Tkatchenko, M. Scheffler, *Phys. Rev. Lett.* **2009**, *102*, 073005.

[36] J. A. Nelder, R. Mead, *Comput. J.* **1965**, *7*, 308–313.

[37] M. H. Wright, *Pitman Res. Notes Math. Ser.* **1996**, 191–208.

[38] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, E. Duchesnay, *J. Mach. Learn. Res.* **2011**, *12*, 2825–2830.

[39] L. Himanen, M. O. J. Jäger, E. V. Morooka, F. Federici Canova, Y. S. Ranawat, D. Z. Gao, P. Rinke, A. S. Foster, *Comput. Phys. Commun.* **2019**, 106949.

Manuscript received: October 23, 2019
Version of record online: February 4, 2020