THE EUROPEAN
PHYSICAL JOURNAL B

Regular Article

Capturing intensive and extensive DFT/TDDFT molecular properties with machine learning[*]

Wiktor Pronobis[1], Kristof T. Schütt[1], Alexandre Tkatchenko[2,a], and Klaus-Robert Müller[1,3,4,b]

[1] Machine Learning Group, Technische Universität Berlin, 10587 Berlin, Germany
[2] Physics and Materials Science Research Unit, University of Luxembourg, Luxembourg 1511, Luxembourg
[3] Max Planck Institute for Informatics, 66123 Saarbrücken, Germany
[4] Department of Brain and Cognitive Engineering, Korea University, Seoul 136-713, South Korea

Abstract. Machine learning has been successfully applied to the prediction of chemical properties of small organic molecules such as energies or polarizabilities. Compared to these properties, the electronic excitation energies pose a much more challenging learning problem. Here, we examine the applicability of two existing machine learning methodologies to the prediction of excitation energies from time-dependent density functional theory. To this end, we systematically study the performance of various 2- and 3-body descriptors as well as the deep neural network SchNet to predict extensive as well as intensive properties such as the transition energies from the ground state to the first and second excited state. As perhaps expected current state-of-the-art machine learning techniques are more suited to predict extensive as opposed to intensive quantities. We speculate on the need to develop global descriptors that can describe both extensive and intensive properties on equal footing.

1 Introduction

Studying the valence electronic spectra of small molecules can yield insights into the properties and discovery of solar cell materials [1] and organic diodes [2]. Attractive candidates for computing such properties are time-dependent DFT or wavefunction-based methods. One popular method is to use linear response time-dependent density functional theory (LR-TDDFT) within the adiabatic approximation [3]. Although less computationally expensive than corresponding coupled-cluster approaches, computing the spectra via LR-TDDFT is still a demanding task, in particular across chemical compound space, where the properties of a diverse data set of compounds need to be obtained in a fast and reliable manner.

Recently, machine learning has been successfully applied to the fast and accurate prediction of molecular properties across chemical compound space [4–10] and molecular dynamics simulations [11–15] as well as for studying properties of quantum-mechanical densities [16,17]. An indispensable ingredient to most machine learning models are molecular descriptors, which are constructed to provide an invariant, unique and efficient representation as input to machine learning models [18–24]. A popular molecular descriptor is the bag-of-bonds (BOB) model [25], which is an extension of the Coulomb matrix (CM) approach [4] and groups the pairwise distances according to pairs of atom types.

In this work, we examine how these novel machine learning approaches can be transferred to predicting intensive properties, in particular the singlet-singlet transition energies computed with TDDFT. To trace the source of possible difficulties back to intensiveness or descriptor, we choose a set of different types of properties to be predicted with machine learning. Specifically, we select the atomization energy and the isotropic polarizability as extensive property. In addition, we choose three intensive properties: the gap between the highest occupied and lowest unoccupied molecular orbital energies (HOMO–LUMO gap), together with the transition energy of the ground state (S_0) to the lowest two vertical electronic excited states (S_1 and S_2), E_1 and E_2, respectively.

On these selected quantum mechanical properties, we perform experiments with various types of molecular descriptors. We examine 2- and 3-body translational and rotationally invariant molecular descriptors [26] which are especially suited for this study as they are invariant w.r.t. atom indexing and independent of the size of the molecule. Thus they can easily be used in combination with kernel ridge regression and artificial neural

[a] alexandre.tkatchenko@uni.lu
[b] klaus-robert.mueller@tu-berlin.de

networks. Furthermore, the used molecular representation is extensible to large molecules and solids as well as to incorporate higher-order interaction terms. Additionally, we examine and compare the performance with the neural network SchNet [9], which learns a local representation of the property under investigation.

2 Methods

2.1 Invariant 2-body interaction descriptors F_{2B}

For two atoms of the molecule with the atomic numbers and coordinates $(Z_1, \boldsymbol{r}_1)$ and $(Z_2, \boldsymbol{r}_2)$, the set of 2-body descriptors is given by

$$F_{2B, Z_1, Z_2} := \left\{ \|\boldsymbol{r}_1 - \boldsymbol{r}_2\|^{-m} \right\}_{m=1,\dots,M}, \qquad (1)$$

where we choose $M = 15$ for this study. For the whole set of 2-body descriptors F_{2B}, we concatenate the descriptors F_{2B, Z_1, Z_2} for the set of pairs of atomic numbers (Z_1, Z_2) present in the data set. The descriptors F_{2B} are invariant w.r.t. the translation and rotation of the molecule as well as the indexing of the atoms.

2.2 Invariant 3-body interaction descriptors F_{3B}

For three atoms of the molecule with the atomic numbers and coordinates $(Z_1, \boldsymbol{r}_1)$, $(Z_2, \boldsymbol{r}_2)$ and $(Z_3, \boldsymbol{r}_3)$, the set of 3-body descriptors is given by

$$F_{3B, Z_1, Z_2, Z_3} := \left\{ \frac{1}{\|\boldsymbol{r}_{12}\|^{m_1} \|\boldsymbol{r}_{13}\|^{m_2} \|\boldsymbol{r}_{23}\|^{m_3}} \right\}, \quad (2)$$

where $m_1, m_2, m_3 = 1, \dots, P$ and we choose $P = 7$ for this study. In equation 2, all combinations of three atoms of the molecule are taken into account. In this work, we use a local variant of F_{3B, Z_1, Z_2, Z_3}, where we select 3-body interactions, which are formed by two sets of bonded atoms which have a common atom. We define two atoms to be bonded, if their euclidean distance is smaller than the threshold function $B(Z_1, Z_2) := 1.1 \cdot L(Z_1, Z_2)$ and the values for the bond length function L given in the Appendix A. For the whole set of 3-body descriptors F_{3B}, we concatenate the descriptors F_{3B, Z_1, Z_2, Z_3} for the set of 3-tuples of atomic numbers (Z_1, Z_2, Z_3) present in the data set. Analogously to the 2-body descriptors, the descriptors F_{3B} are invariant w.r.t. translation and rotation and atom indexing.

2.3 SchNet

The neural network SchNet [10] is a variant of the earlier proposed deep tensor neural networks [8] is based on the principle of learning atom-wise representations directly from first-principles. Given the atoms of type $Z_1, \dots, Z_N$, initial atom embeddings $\mathbf{x}_{Z_i}^{(0)} \in \mathbb{R}^F$, where F is the dimension of the feature space, depend only on the atom type.

Table 1. Mean absolute errors of predicting the atomization energy (U0), isotropic polarizability (α), difference between the HOMO and LUMO energies (gap) and the transition energy to the first (E_1) and second (E_2) electronic excited singlet state. The properties U0, α and gap were calculated with DFT at the B3LYP/6-31G(2df,p) level of theory, the transition energies were calculated with LR-TDDFT at the PBE0/def2TZVP level of theory. The energy units are kcal/mol, the polarizability is given in units Bohr^3. Best results are marked bold.

Method	U0	α	Gap	E_1	E_2
Mean pred.	185.0	6.27	25.4	22.4	18.0
CM	4.8	0.60	7.8	12.7	10.2
BOB	2.3	0.36	4.8	11.5	9.6
F_{2B}	2.9	0.45	5.8	11.6	9.5
F_{3B}	2.9	0.45	5.3	11.3	9.4
$F_{2B} + F_{3B}$	1.1	0.33	4.6	**11.1**	**9.2**
SchNet	**1.0**	**0.22**	**3.4**	11.4	10.0

Then, a series of pairwise interaction refinements

$$\mathbf{x}_i^{t+1} = \mathbf{x}_i^t + \sum_{j \neq i} V^t(\mathbf{x}_j^t, \|\mathbf{r}_{ij}\|),$$

introduces information about the chemical environment into the embeddings. In SchNet, this is modeled using continuous-filter convolutions with filter-generating networks [9]. Through multiple of these corrections, SchNet is able to include complex many-body terms in the representation. Finally, an output neural network O predicts atom-wise property contributions, such that the final prediction is $\hat{y} = \sum_{i=1}^{N} O(\mathbf{x}_i^{(T)})$ for extensive properties and $\hat{y} = \frac{1}{N} \sum_{i=1}^{N} O(\mathbf{x}_i^{(T)})$ for intensive properties. During training, the initial embedding vectors $\mathbf{x}_{Z_i}^{(0)}$ as well as the parameters of the interaction network V and the output network O are optimized. In this work, we use $T = 6$ interaction refinements and $F = 64$ feature dimensions.

3 Experiments and discussion

We use the 21 786 molecules from the QM9 benchmark dataset with up to 8 heavy atoms of type CNOF. QM9 includes relaxed geometries and properties computed using DFT at the B3LYP/6-31G(2df,p) level of theory [27]. This data set was previously used to predict deviations from reference second-order approximate coupled-cluster (CC2) singles and doubles spectra from their TDDFT counterparts [28]. The singlet–singlet transition energies from the ground state to the first and second excited state were calculated at the LR-TDDFT [29] level employing the hybrid XC functional PBE0 [30,31] with def2TZVP basis set [32]. Instead of applying the delta learning approach [28], we attempt to learn the transition energies *directly*. We will additionally use the atomization energy U0, isotropic polarizability α and HOMO–LUMO gap from QM9 for evaluation.

For all models, we use 10k random molecules for training and the remaining unseen 11 786 molecules for computing the prediction errors. The results are listed in

Table 1. For the CM and BOB descriptors, the Laplace kernel is known to perform better as compared to the Gaussian kernel [25]. For the 2- and 3-body descriptors F_{2B} and F_{3B}, the Gaussian kernel achieves smaller prediction errors.

The mean predictor (mean pred.) is given by the average value of the property to be predicted. In general, the mean predictor yields an upper bound of the mean absolute prediction error of the machine learning models under investigation. While the Coulomb matrix (CM) uniquely encodes the structure of a given molecule, it performs worst of the evaluated descriptors. A major reason for this is that it implies a similarity measure of atom types based on the Coulomb interaction of the atomic nuclei, which does not reflect chemistry well. The bag-of-bonds (BOB) model is an extension to the Coulomb matrix where atom types are sorted into bags, thereby, avoiding an unsuited atom similarity. This significantly boosts the performance compared to the CM for the atomization energy, polarizability and gap. Still, the bags are not invariant to atom indexing, which allows for multiple possible descriptors of the same molecule.

The F_{2B} descriptors solve some of the sorting problems encountered in the CM and BOB representation. The prediction error of the F_{2B} descriptors is significantly better than the CM result, while being slightly worse compared to the BOB model. Combining the local 3-body descriptors F_{3B} with the 2-body descriptors F_{2B} significantly increases the predictive performance of the atomization energy, polarizability and the HOMO–LUMO gap (62%, 27% and 21% improvement, respectively). For the transition energies E_1 and E_2, only a minor performance gain is observed by including the local 3-body descriptors F_{3B}. SchNet slightly improves upon the descriptors $F_{2B} + F_{3B}$ for the prediction of the atomization energy, polarizability and gap, indicating that these properties can be well-represented by atom-wise contributions.

Figure 1 shows percentiles of the residuals when using various descriptors to predict the transition energy E_1. For the transition energies, the 2- and 3-body descriptors do not improve upon the performance of the baseline methods BOB and CM as much as for the extensive properties. Moreover, SchNet only achieves a performance that lies on the level of CM and BOB.

As SchNet is able to include complex many-body terms in the representation, the non-locality of the transition energies do not allow a decomposition into atomic contributions. This indicates the need for much more complex global many-body terms for predicting transition energies, possibly encoding higher order interactions with order larger than three.

As most descriptors are either size-dependent or encode a sum or average term over local many-body interactions, they are naturally better suited to predict extensive properties. Such descriptors are typically limited by the order of the explicitly included many-body interactions. This can be a problem for predicting more complex quantum mechanical properties as demonstrated by the HOMO–LUMO gap, where SchNet performs better than explicit pairwise and 3-body interaction descriptors. For the transition energies, SchNet does not improve upon the
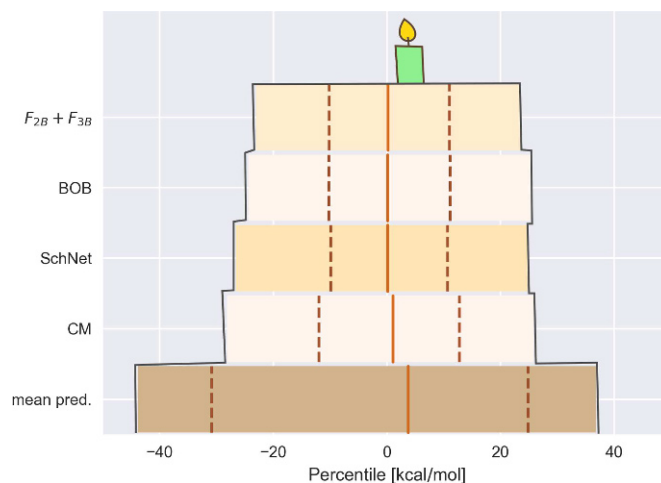

Fig. 1. Percentiles of the residuals of using kernel ridge regression in combination with various descriptors to predict the singlet-singlet transition energy TDPBE0/def2TZVP-E_1. Specifically, the left and right borders correspond to the 5% and 95% percentiles, the solid middle line is the median and the left and right dashed lines correspond to the 20% and 80% percentiles, respectively. The candle serves as a guide to the eye for the median and furthermore as a guide to the occasion of this special issue.

$F_{2B} + F_{3B}$ result. Even though the 3-body descriptors are only applied to local bond angles, they perform better than 2-body descriptors. In light of the SchNet results, this indicates that explicit many-body terms are more suitable to model transition energies using machine learning. As SchNet is designed to include high-order *local* interactions, we speculate on the need to develop *global* descriptors for intensive properties. As such properties are in general more difficult to predict than their localized counterparts, we conjecture that such kind of descriptors will describe both extensive and intensive properties on equal footing. In addition, as seen by the learning curves in Figures 2–6, more data may be exceedingly helpful for further improving the predictive performance of the intensive properties under investigation.

4 Conclusion

We have evaluated a variety of machine learning techniques for intensive and extensive properties. As expected, all of them perform better on extensive properties than on intensive quantities. For the gap, SchNet performs 25% better than the explicit combination of pairwise and 3-body descriptors. As SchNet is able to include complex many-body terms in principle, this result indicates the need for descriptors with many-body interactions with order larger than three for predicting the HOMO–LUMO gap. For the intensive properties E1 and E2, the 3-body descriptors work best, in particular combined with the 2-body terms. In contrast, the decomposition into atom-wise contributions of SchNet, while working well for extensive properties, can be considered a drawback when
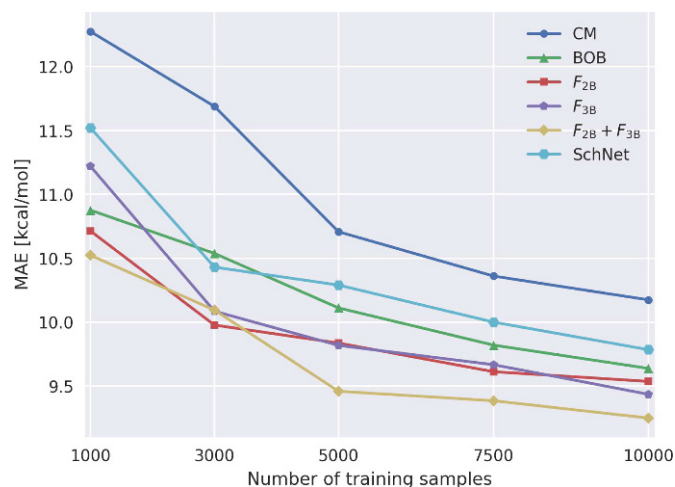
Fig. 2. Mean absolute error of predicting the singlet–singlet transition energy TDPBE0/def2TZVP-E_2 in dependence of the number of training samples. The errors are given in kcal/mol. The model hyperparameters have been determined by 10-fold cross validation.
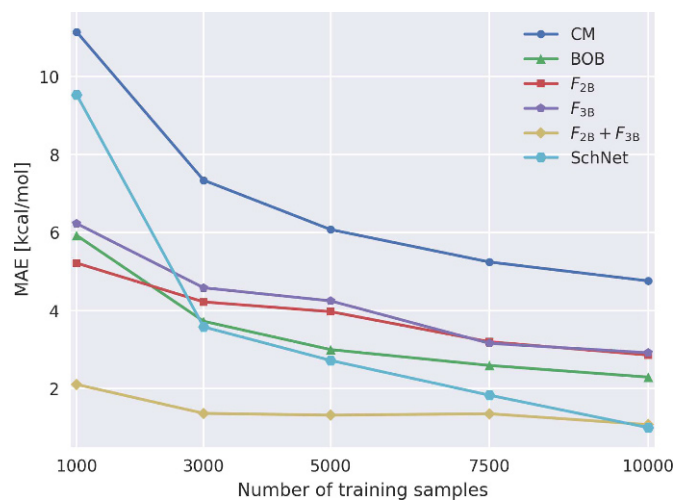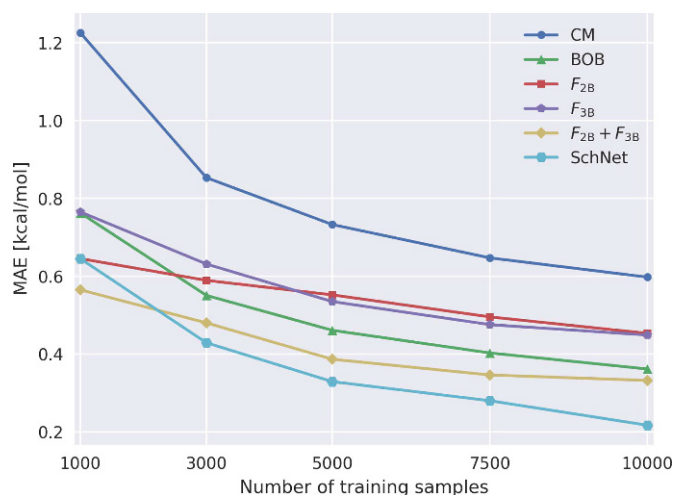

Fig. 4. Mean absolute error of predicting the B3LYP/6-31G(2df,p) polarizability in dependence of the number of training samples. The errors are given in Bohr^3. The model hyperparameters have been determined by 10-fold cross validation.

Fig. 3. Mean absolute error of predicting the B3LYP/6-31G(2df,p) atomization energy in dependence of the number of training samples. The errors are given in kcal/mol. The model hyperparameters have been determined by 10-fold cross validation.
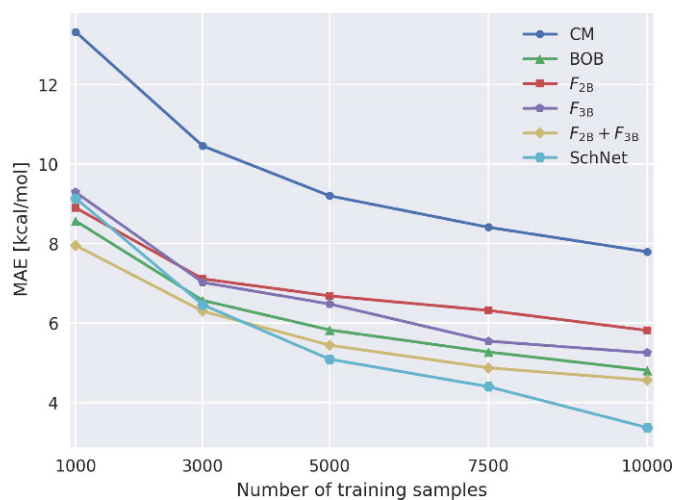

Fig. 5. Mean absolute error of predicting the B3LYP/6-31G(2df,p) HOMO–LUMO gap in dependence of the number of training samples. The errors are given in kcal/mol. The model hyperparameters have been determined by 10-fold cross validation.

attempting to predict transition energies by the averaging approach in the last output layer of SchNet.

Still, even with the best-performing descriptors the error of transition energy prediction may still be too high for any practical use. More advanced non-local descriptors will be necessary to predict transition energies more accurately, possibly encoding higher many-body terms or electronic state information. In addition, as seen by the learning curves, more data may be exceedingly helpful for further improving the predictive performance of the intensive properties under investigation.

Author contribution statement

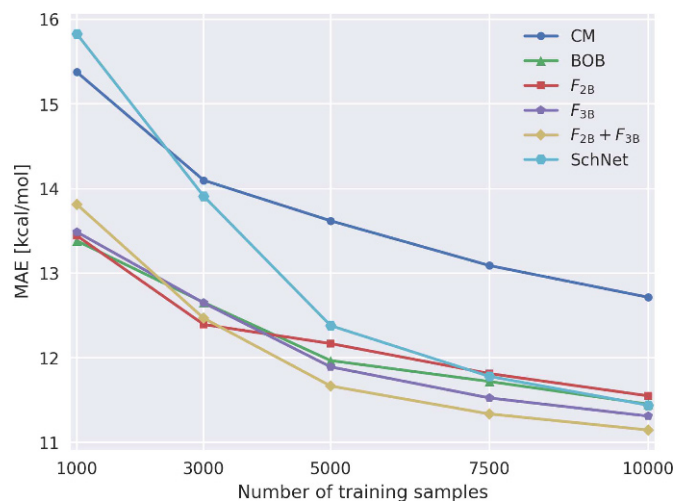WP developed the F_{2B} and F_{3B} descriptors, prepared the figures and wrote the initial manuscript. WP and KTS

Fig. 6. Mean absolute error of predicting the singlet-singlet transition energy TDPBE0/def2TZVP-E_1 in dependence of the number of training samples. The errors are given in kcal/mol. The model hyperparameters have been determined by 10-fold cross validation.

Table A.2. Bond lengths in ngström for all combinations of the elements C, H, N, O, F. Used for computing the 3-body interaction descriptors F_{3B}.

Bond-type	(Z_1, Z_2)	$L(Z_1, Z_2)$
H–H	$(1,1)$	0.74
H–C	$(1,6)$	1.08
H–O	$(1,8)$	0.96
H–N	$(1,7)$	1.01
C–C	$(6,6)$	1.51
C–O	$(6,8)$	1.43
C–N	$(6,7)$	1.47
O–O	$(8,8)$	1.48
O–N	$(8,7)$	1.40
N–N	$(7,7)$	1.45
F–H	$(9,1)$	0.92
F–C	$(9,6)$	1.35
F–O	$(9,8)$	1.42
F–N	$(9,7)$	1.36
F–F	$(9,9)$	1.42

performed the experiments. AT and KRM supervised the project. All authors discussed the results and contributed to the final manuscript.

Appendix A

A.1 Kernel ridge regression (KRR)

In kernel ridge regression (e.g. [33]), a kernel is used as similarity measure between two molecules. From this similarity measure, the prediction of a given property of a new molecule $\boldsymbol{x}$ is obtained by

$$E = \sum_{i=1}^{N} \alpha_i \cdot K(\boldsymbol{x}_i, \boldsymbol{x}), \qquad (A.1)$$

where the α_i denote the weighting coefficients obtained by training the model and N is the number of train molecules $\{\boldsymbol{x}_i\}_{i=1}^{N}$. Training the model involves a set of molecules with known labels $\{\boldsymbol{x}_i, E_i\}_{i=1}^{N}$ from which the $\boldsymbol{\alpha}$ is obtained by solving a regularized system of linear equations

$$(\lambda \cdot \mathbf{I} + \mathbf{K}) \cdot \boldsymbol{\alpha} = \boldsymbol{E}, \qquad (A.2)$$

where $\mathbf{K}_{ij} := K(\boldsymbol{x}_i, \boldsymbol{x}_j)$ and λ is the regularization parameter. Popular choices of kernels are the Gaussian kernel and the Laplace kernel, respectively. The hyper parameters of the model can be determined with cross-validation [34].

A.2 Bond lengths

Table A.2 lists the bond lengths in ngström for all combinations of the elements C, H, N, O, F used to compute the 3-body interaction descriptors F_{3B}.

References

1. M. Grätzel, Nature **414**, 338 (2001)
2. M. Gross et al., Nature **405**, 661 (2000)
3. E. Runge, E.K.U. Gross, Phys. Rev. Lett. **52**, 997 (1984)
4. M. Rupp et al., Phys. Rev. Lett. **108**, 058301 (2012)
5. K. Hansen et al., J. Chem. Theory Comput. **9**, 3404 (2013)
6. G. Montavon et al., New J. Phys. **15**, 095003 (2013)
7. F.A. Faber et al., J. Chem. Theory Comput. **13**, 5255 (2017)
8. K.T. Schütt et al., Nat. Commun. **8**, 13890 (2017)
9. K.T. Schütt et al., Adv. Neural Inf. Process. Syst. **30**, 992 (2017)
10. K.T. Schütt et al., J. Chem. Phys. **148**, 241722 (2018)
11. S. Chmiela et al., Sci. Adv. **3**, e1603015 (2017)
12. F. Noé, C. Clementi, J. Chem. Theory Comput. **11**, 5002 (2015)
13. M. Gastegger, J. Behler, P. Marquetand, Chem. Sci. **8**, 6924 (2017)
14. A.P. Bartók et al., Phys. Rev. B **88**, 054104 (2013)
15. A. Mardt et al., Nat. Commun. **9**, 5 (2018)
16. J.C. Snyder et al., Phys. Rev. Lett. **108**, 253002 (2012)
17. F. Brockherde et al., Nat. Commun. **8**, 872 (2017)

18. C.R. Collins et al., arXiv:1701.06649 [physics.chem-ph] (2017)
19. B. Huang, O. Anatole von Lilienfeld, J. Chem. Phys. **145**, 161102 (2016)
20. G. Montavon et al., in Advances in Neural Information Processing Systems, edited by F. Pereira et al. (Curran Associates, Inc., NewYork, 2012), Vol. 25, pp. 440–448
21. H. Huo, M. Rupp, arXiv:1704.06439 (2017)
22. J. Behler, M. Parrinello, Phys. Rev. Lett. **98**, 146401 (2007)
23. A.P. Bartók et al., Sci. Adv. **3**, e1701816 (2017)
24. A.P. Bartók et al., Phys. Rev. Lett. **104**, 136403 (2010)
25. K. Hansen et al., J. Phys. Chem. Lett. **6**, 2326 (2015)
26. W. Pronobis, A. Tkatchenko, K.-R. Müller, J. Chem. Theory Comput. **14**, 2991 (2018)
27. R. Ramakrishnan et al., Sci. Data **1**, 140022 EP (2014)
28. R. Ramakrishnan et al., J. Chem. Phys. **143**, 084111 (2015)
29. F. Furche, R. Ahlrichs, J. Chem. Phys. **117**, 7433 (2002)
30. J.P. Perdew, M. Ernzerhof, K. Burke, J. Chem. Phys. **105**, 9982 (1996)
31. C. Adamo, V. Barone, J. Chem. Phys. **110**, 6158 (1999)
32. F. Weigend, R. Ahlrichs, Phys. Chem. Chem. Phys. **7**, 3297 (2005)
33. C.M. Bishop, Pattern Recognition and Machine Learning (Information Science and Statistics) (Springer-Verlag, Inc., Secaucus, NJ, USA, New York, 2006)
34. K.-R. Müller et al., IEEE Trans. Neural Netw. **12**, 181 (2001)