

Machine Learning for Quantum Mechanical Properties of Atoms in Molecules

Matthias Rupp,^{1,*} Raghunathan Ramakrishnan,¹ and O. Anatole von Lilienfeld^{1,†}

¹*Institute of Physical Chemistry and National Center for Computational Design and Discovery of Novel Materials (MARVEL), Department of Chemistry, University of Basel, Klingelbergstr. 80, CH-4056 Basel, Switzerland*

(Dated: August 26, 2015)

We introduce machine learning models of quantum mechanical observables of atoms in molecules. Instant out-of-sample predictions for proton and carbon nuclear chemical shifts, atomic core level excitations, and forces on atoms reach accuracies on par with density functional theory reference. Locality is exploited within non-linear regression via local atom-centered coordinate systems. The approach is validated on a diverse set of 9 k small organic molecules. Linear scaling of computational cost in system size is demonstrated for saturated polymers with up to sub-mesoscale lengths.

PACS numbers: 03.65.-w,31.15.A-,31.15.E-,02.60.Ed

This work has subsequently been published in the *Journal of Physical Chemistry Letters* 6(16): 3309–3313, American Chemical Society, 2015. To access the final edited and published work see DOI 10.1021/acs.jpcclett.5b01456.

Accurate solutions to the many-electron problem in molecules have become possible due to progress in hardware and methods. [1–4] Their prohibitive computational cost, however, prevents both routine atomistic modeling of large systems and high-throughput screening. [5] Machine learning (ML) models can be used to infer quantum mechanical (QM) expectation values of molecules, based on reference calculations across chemical space. [6, 7] Such models can speed up predictions by several orders of magnitude, demonstrated for relevant molecular properties such as enthalpies, entropies, polarizabilities, electron correlation, and, electronic excitations. [8–10]

A major drawback is their lack of transferability, e.g., ML models trained on bond dissociation energies of small molecules will not be predictive for larger molecules. In this work, we introduce ML models for properties of atoms in molecules. These models exploit locality to achieve transferability to larger systems *and* across chemical space, for systems that are locally similar to the ones trained on (Fig. 1). These aspects have only been treated in isolation before. [6, 11]

We model spectroscopically relevant observables, namely ¹³C and ¹H nuclear magnetic resonance (NMR) chemical shifts [12] and 1s core level ionization energies (CIE), as well as atomic forces, crucial for structural relaxation and molecular dynamics. Nuclear shifts and ionization energies are dominated by inherently local core electron-nucleus interactions. Atomic forces are expectation values of the differential operator applied to an atom’s position in the Hamiltonian [13], and scale quadratically with inverse distance.

Inductive modeling of QM properties of atoms in molecules constitutes a high-dimensional interpolation problem with spatial and compositional degrees of freedom. QM reference calculations provide training exam-

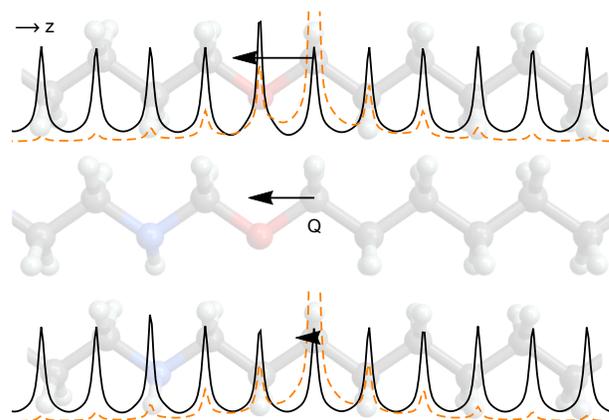


FIG. 1. Sketch illustrating local nature of atomic properties for the example of force $\langle \Psi | \partial_{\mathbf{R}_Q} \hat{H} | \Psi \rangle$ acting on a query atom in a molecule (mid), inferred from similar atoms in training molecules (top, bottom). Shown are force vectors (arrows), integrated electron density $\int dx dy n(\mathbf{r})$ (solid) and integrated electronic term of Hellmann-Feynman force along z (dashed).

ples $\{(\mathbf{x}_i, y_i)\}_{i=1}^n$, where the \mathbf{x}_i encode atoms in their molecular environment and y_i are atomic property values. ML interpolation between training examples then provides predicted property values for new atoms.

The electronic Hamiltonian is determined by number of electrons, nuclear charges $\{Z_I\}$ and positions $\{\mathbf{R}_I\}$, which can be challenging for direct interpolation. [14] Proposed requirements for representations include uniqueness, continuity, as well as invariance to translation, rotation, and nuclear permutations. [15] For scalar properties (NMR, CIE), we use the sorted Coulomb matrix [6] to represent a query atom Q and its environment: $M_{II} = 0.5Z_I^2$ and $M_{IJ} = Z_I Z_J / |\mathbf{R}_I - \mathbf{R}_J|$, where atom indices I, J run over Q and all atoms in its environment, sorted by distance to Q . Note that all molecules in this study are neutral, and no explicit encoding of charge is necessary.

Atomic forces are vector quantities requiring a basis, which should depend only on the local environment; in particular, it should be independent of the global frame

of reference used to construct the Hamiltonian in the QM calculation. We project force vectors into a local coordinate system centered on atom Q , and predict each component separately. Later, the predicted force vector is reconstructed from these component-wise predictions.

We use principal component analysis (PCA) to obtain an atom-centered orthogonal three-dimensional local coordinate system. In analogy to the electronic term in Hellmann-Feynman forces, $\int d\mathbf{r} (\mathbf{r} - \mathbf{R}_Q) Z_Q n(\mathbf{r}) / \|\mathbf{r} - \mathbf{R}_Q\|^3$ [13], we weight atoms by $Z_I / \|\mathbf{R}_I - \mathbf{R}_Q\|^3$, increasing influence of heavy atoms and decreasing influence of distant atoms. Non-degenerate PCA axes are unique only up to sign; we address this by defining the center of charge to be in the positive quadrant. A matching matrix representation is obtained via $\mathbf{M}_I = (Z_I, X'_I, Y'_I, Z'_I)$, where X', Y', Z' are projected atom coordinates, and rows are ordered by distance to central atom Q , yielding an $m \times 4$ matrix, where m is number of atoms. In both representations, we impose locality by constraining Q 's environment to neighboring atoms within a sphere of radius τ .

For interpolation between atomic environments we use kernel ridge regression (KRR) [16], a non-linear regularized regression method effectively carried out implicitly in a high-dimensional Hilbert space ("kernel trick"). [17] Predictions are linear combinations over all training examples in the basis of a symmetric positive definite kernel k : $f(\mathbf{z}) = \sum_{i=1}^n \alpha_i k(\mathbf{x}_i, \mathbf{z})$, where α are regression weights for each example, obtained from a closed-form expression minimizing the regularized error on the training data. See Refs. [6, 18–20] for details. As kernel k , we use the Laplacian kernel $k(\mathbf{x}, \mathbf{z}) = \exp(-\|\mathbf{x} - \mathbf{z}\|_1 / \sigma d)$, where $\|\cdot\|_1$ is the L^1 -norm, σ is a length scale, and $d = \dim(\mathbf{x})$. This kernel has shown best performance for prediction of molecular properties. [18]

Our models contain three free parameters: cut-off radius τ , regularization strength λ , and kernel length scale σ . Regularization strength λ , controlling the smoothness of the model, was set to a small constant (10^{-10}), forcing the model to fit the QM values closely. As for length scales σ , note that for the Laplacian kernel, non-trivial behavior requires $\|\cdot, \cdot\|_1 \approx \sigma$. We set σ to four times the median nearest neighbor L^1 -norm distance in the training set. [21] Cut-off radii τ were then chosen to minimize RMSE in initial experiments (Fig. 2). For the comparatively insensitive F_C , other statistics (maxAE, R^2) yielded an unambiguous choice.

We used three datasets for validation: For NMR chemical shifts and CIEs, both scalar properties, we employed a dataset of 9k synthetically accessible organic molecules containing 7–9 C, N, or O atoms, with open valencies saturated by H, a subset of a larger dataset. [23, 24] Relaxation and property calculations were done at the DFT/PBE0/def2TZVP level of theory [25–30] using Gaussian [31]. For forces, we distorted molecular equilibrium geometries using normal mode analysis [32–34]

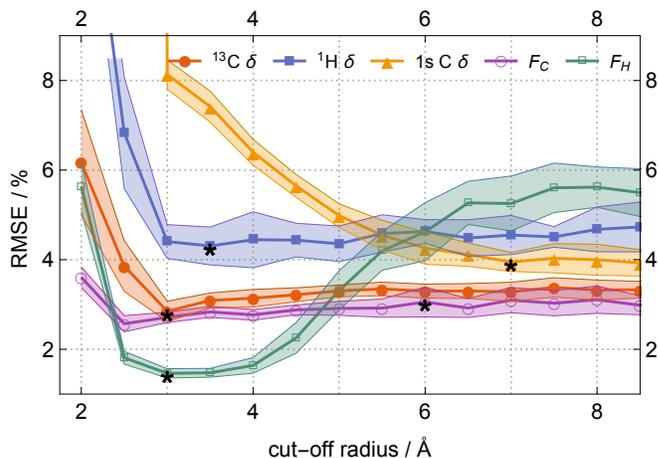


FIG. 2. Locality of properties, measured by model performance as a function of cut-off radius τ . Root mean square error (RMSE) shown as fraction of corresponding property's range [22] for nuclear shifts ($^{13}\text{C } \delta$, $^1\text{H } \delta$), core level ionization energy ($1s \text{ C } \delta$), and atomic forces (F_C , F_H). Asterisks * mark chosen values. Shaded areas indicate 1.6 standard deviations over 15 repetitions.

by adding random perturbations in the range $[-0.2, 0.2]$ to each normal mode, sampling homogeneously within an harmonic approximation. Adding spatial degrees of freedom considerably increases the intrinsic dimensionality of the learning problem. To accommodate this, we reduced dataset variability to a subset of 168 constitutional isomers of $\text{C}_7\text{H}_{10}\text{O}_2$, with 100 perturbed geometries for each isomer. Computationally inexpensive semi-empirical quantum chemistry approximations are readily available for forces. We exploit this to improve accuracy by modeling the difference between baseline PM7 [35] and DFT reference forces (Δ -learning [10]). To demonstrate linear scaling of computational cost with system size, we used a third dataset of organic saturated polymers, namely linear polyethylene, the most common plastic, with random substitutions of some CH units with NH or O for chemical variety. All prediction errors were measured on out-of-sample hold-out sets never used during training.

Table I presents performance estimates for models trained on 10k randomly chosen atoms, measured on a hold-out set of 1k other atoms. Comparison with literature estimates of typical errors of the employed DFT reference method suggests in all cases that the ML models achieve similar accuracy—at negligible computational cost after training. Statistical learning theory shows that under certain assumptions the accuracy of a ML model asymptotically improves with increasing training set size as $O(1/\sqrt{n})$. [36] Fig. 3 presents corresponding learning curves for all properties. Errors are shown as percentage of property ranges [22], enabling comparison of properties with different units. All errors start off in the single digit

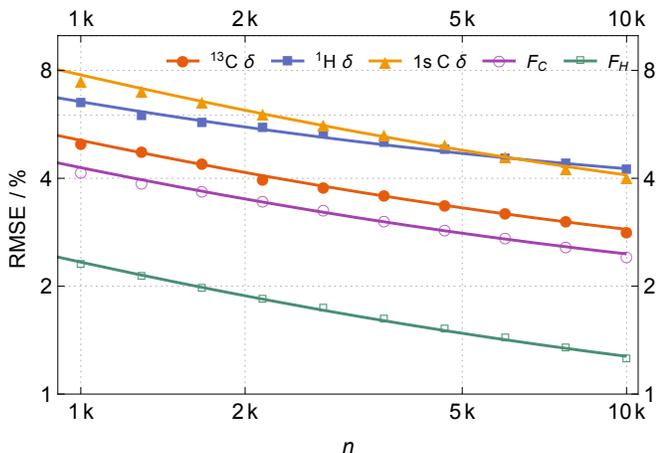


FIG. 3. Systematic improvement in accuracy of atomic property predictions with increasing training set size n . Root mean square error (RMSE) shown as fraction of corresponding property’s range [22] for nuclear shifts ($^{13}\text{C } \delta$, $^1\text{H } \delta$), core level ionization energy ($1s \text{ C } \delta$), and atomic forces (F_{C} , F_{H}). Values from 15 repetitions; see Table I for ranges and standard deviations. Solid lines are fits to theoretical asymptotic performance of $O(1/\sqrt{n})$.

percent range at 1 k training atoms, and decay systematically to roughly half their initial value at 10 k training atoms.

ML predictions and DFT values for chemical shifts of all 50 k carbon atoms in the dataset are featured in Fig. 4. The shielding of the nuclear spin from the magnetic field is strongly dependent on the atom’s local chemical environment. In accordance with the diversity of the dataset, we find a broad distribution with four pronounced peaks, characteristic of up- or downshifts of the resonant frequencies of nuclear carbon spin. The peaks at 30, 70, 150, and 210 ppm typically correspond to saturated sp^3 -hybridized carbon atoms, strained sp^3 -hybridized carbons, conjugated or sp^2 -hybridized carbon atoms, and carbon atoms in carbonyl groups, respectively. A ML model trained on only 500 atoms already reproduces all major peaks; larger training sets yield systematically improved distributions. For 10 k training examples predictions are hardly distinguishable from the DFT reference, except for a small deviation at 140 ppm. Using the same model, we predicted shifts for 847 k carbon atoms in all 134 k molecules published in Ref. [24]. The resulting distribution is roughly similar, reflecting similar chemical composition of molecules in this much larger dataset, which is beyond the current limits of DFT reference calculations employed here.

The presented approach to model atomic properties scales linearly: Since only a finite volume around an atom is considered, its numerical representation is of constant size; [45] in particular, it does not scale with the system’s overall size. Comparing atoms, and thus kernel evaluations, therefore requires constant computational effort,

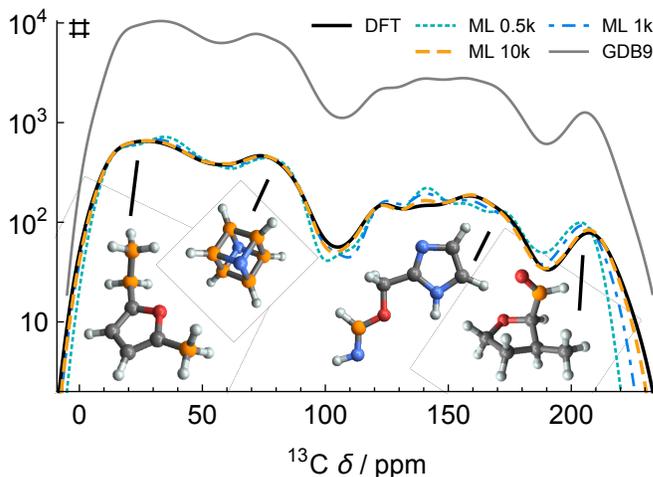


FIG. 4. Distribution of 50 k ^{13}C chemical shifts in 9 k organic molecules. ML predictions for increasing training set sizes approach DFT reference values. Molecular structures highlight chemical diversity and effect of molecular environment on chemical shift of query atom (orange; see main text). GDB9 corresponds to ML predictions for 847 k carbon atoms in 134 k similar molecules published in Ref. [24].

rendering the overall computational cost of predictions linear in system size, with small prefactor. Furthermore, a form of chemical extrapolation can be achieved despite the fact that ML models are interpolation models. As long as local chemical environments of atoms are similar to those in the training set, the model can interpolate. Consequently, using similar local “building blocks”, large molecules can be constructed that are very different from the ones used in the training set, but amenable to prediction.

To verify this, we trained a ML model on atoms drawn from the short polymers in the third dataset, then applied the same model to predict properties of atoms in polymers of increasing length. Training set polymers had a backbone length of 29 C,N,O atoms; for validation, we used up to ten times longer backbones, reaching lengths of 355 Å and 696 atoms in total. Fig. 5 presents numerical evidence for excellent near-constant accuracy of model predictions, independent of system size, validated by DFT. Although trained only on the smallest instances, the model’s accuracy varies negligibly with system size, confirming both transferability and chemically extrapolative predictive power of the ML model.

Individual ML predictions are 4–5 orders of magnitude faster than reference DFT calculations. Overall speed-up depends on dataset and reference method, and is dominated by training set generation, i.e., the ratio between number of predictions and training set size. DFT and ML calculations were done on a high-performance compute cluster and a laptop, respectively.

In conclusion, we have introduced ML models for QM properties of atoms in molecules. Performance and ap-

TABLE I. Prediction errors for ML models trained on 10k atoms and predicting properties of 1k other out-of-sample atoms. Calculated properties are NMR chemical shifts (^{13}C δ , ^1H δ), core level ionization energy (1s C δ), and forces (F_C , F_H).^a

Property	Ref.	Range	MAE	RMSE	maxAE	R^2	σ	$\tau/\text{\AA}$
^{13}C δ /ppm	2.4 [30, 37, 38]	6 – 211	3.9 ± 0.28	5.8 ± 0.30	36 ± 8.0	0.988 ± 0.001	20 ± 3.4	3
^1H δ /ppm	0.11 [38–40]	0 – 10	0.28 ± 0.01	0.42 ± 0.02	3.2 ± 1.1	0.954 ± 0.005	0.53 ± 1.2	3.5
1s C δ /mE _h	7.5 [41–43]	-165 – -2	4.9 ± 0.12	6.5 ± 0.27	34 ± 17	0.971 ± 0.002	181 ± 0.0	7
F_C /mE _h /a ₀	1 [44]	-99 – 96	3.6 ± 0.10	4.7 ± 0.15	29 ± 5.5	0.983 ± 0.002	0.69 ± 0.1	6
F_H /mE _h /a ₀	1 [44]	-43 – 43	0.8 ± 0.02	1.1 ± 0.03	7.4 ± 2.6	0.996 ± 0.003	0.35 ± 0.0	3

^a Shown are MAE of DFT reference from literature (Ref.), property ranges [22], mean absolute error (MAE), root mean squared error (RMSE), maximum absolute error (maxAE), squared correlation (R^2) and hyperparameters (kernel length scale σ , cut-off radius τ). Averages \pm standard deviations over 15 randomly drawn training sets.

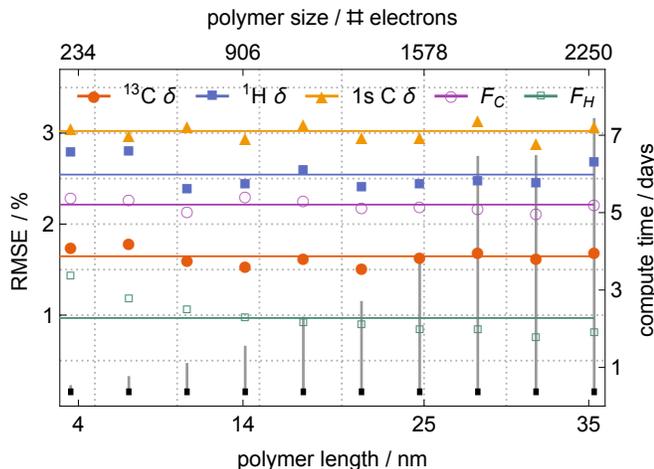


FIG. 5. Linear scaling and chemical extrapolation for ML predictions of saturated polymers of increasing length. Shown are root mean square error (RMSE), given as fraction of corresponding property’s range [22], as well as indicative compute times of cubically scaling DFT calculations (gray bars) and ML predictions (black bars, enlarged for visibility), which scale linearly with low prefactor. See Table I for property ranges.

plicability have been demonstrated for chemical shifts, core level ionization energies, and atomic forces of 9k chemically diverse organic molecules and 168 isomers of $\text{C}_7\text{H}_{10}\text{O}_2$, respectively. Accuracy of predictions is on par with the QM reference method. We have used the ML model to predict chemical shifts of all 847k carbon atoms in the 134k molecules published in Ref. [24]. Locality of modeled atomic properties is exploited through use of atomic environments as building blocks. Consequently, the model scales linearly in system size, which we have demonstrated for saturated linear polymers over 30 nm in length. Results suggest that the model could be useful in mesoscale studies.

For the investigated molecules and properties the locality assumption, implemented as a finite cut-off radius in the representation, has proven sufficient. This might not necessarily be true in general. The Hellmann-Feynman force, for example, depends directly on the electron den-

sity, which can be altered substantially due to long-range substituent effects such as those in conjugated π -bond systems. For other systems and properties, larger cut-offs or additional measures might be necessary.

The presented ML models could also be used for nuclear shift assignment in NMR structure determination, for molecular dynamics of macro-molecules, or condensed molecular phases. We consider efficient sampling, i.e., improving the ratio of performance to training set size (“sample efficiency”), and improving representations to be primary challenges in further development of these models.

We thank Tristan Bereau, Zhenwei Li, and Kuang-Yu Samuel Chang for helpful discussions. OAvL acknowledges the Swiss National Science Foundation for support (SNF grant PP00P2_138932). Calculations were performed at sciCORE (scicore.unibas.ch) scientific computing core facility at University of Basel. This research used resources of the Argonne Leadership Computing Facility at Argonne National Laboratory, which is supported by the Office of Science of the U.S. DOE under contract DE-AC02-06CH11357.

* mrupp@mrupp.info; Current address: Fritz Haber Institute of the Max Planck Society, Faradayweg 4–6, 14195 Berlin, Germany.

† anatole.vonlilienfeld@unibas.ch

- [1] C. Bekas and A. Curioni, *Comput. Phys. Comm.* **181**, 1057 (2010).
- [2] I. Duchemin and F. Gygi, *Comput. Phys. Comm.* **181**, 855 (2010).
- [3] G. H. Booth, A. J. Thom, and A. Alavi, *J. Chem. Phys.* **131**, 054106 (2009).
- [4] A. Benali, L. Shulenburger, N. A. Romero, J. Kim, and O. A. von Lilienfeld, *J. Chem. Theor. Comput.* **10**, 3417 (2014).
- [5] T. Helgaker, P. Jørgensen, and J. Olsen, *Molecular Electronic-Structure Theory* (Wiley, Chichester, England, 2000).
- [6] M. Rupp, A. Tkatchenko, K.-R. Müller, and O. A. von Lilienfeld, *Phys. Rev. Lett.* **108**, 058301 (2012).
- [7] O. A. von Lilienfeld, *Int. J. Quant. Chem.* **113**, 1676 (2013).

- [8] G. Montavon, M. Rupp, V. Gobre, A. Vazquez-Mayagoitia, K. Hansen, A. Tkatchenko, K.-R. Müller, and O. A. von Lilienfeld, *New J. Phys.* **15**, 095003 (2013).
- [9] R. Ramakrishnan, M. Hartmann, E. Tapavicza, and O. A. von Lilienfeld, arXiv , 1504.01966 (2015).
- [10] R. Ramakrishnan, P. O. Dral, M. Rupp, and O. A. von Lilienfeld, *J. Chem. Theor. Comput.* **11**, 2087 (2015).
- [11] Z. Li, J. R. Kermode, and A. D. Vita, *Phys. Rev. Lett.* **114**, 096405 (2015).
- [12] Calculated as $\delta = (\sigma_{\text{ref}} - \sigma_{\text{sample}})/(1 - \sigma_{\text{ref}})$.
- [13] R. P. Feynman, *Phys. Rev.* **56**, 340 (1939).
- [14] L. M. Ghiringhelli, J. Vybiral, S. V. Levchenko, C. Draxl, and M. Scheffler, *Phys. Rev. Lett.* **114**, 105503 (2015).
- [15] O. A. von Lilienfeld, R. Ramakrishnan, M. Rupp, and A. Knoll, *Int. J. Quant. Chem.* **115**, 1084 (2015).
- [16] T. Hastie, R. Tibshirani, and J. Friedman, *The Elements of Statistical Learning*, 2nd ed. (Springer, New York, 2009).
- [17] B. Schölkopf and A. Smola, *Learning with Kernels* (MIT Press, Cambridge, 2002).
- [18] K. Hansen, G. Montavon, F. Biegler, S. Fazli, M. Rupp, M. Scheffler, O. A. von Lilienfeld, A. Tkatchenko, and K.-R. Müller, *J. Chem. Theor. Comput.* **9**, 3543 (2013).
- [19] K. Vu, J. Snyder, L. Li, M. Rupp, B. F. Chen, T. Khelif, K.-R. Müller, and K. Burke, *Int. J. Quant. Chem.* **115**, 1115 (2015).
- [20] L. Li, J. C. Snyder, I. M. Pelaschier, J. Huang, U.-N. Niranjan, P. Duncan, M. Rupp, K.-R. Müller, and K. Burke, arXiv , 1404.1333 (2014).
- [21] J. Moody and C. J. Darken, *Neural. Comput.* **1**, 281 (1989).
- [22] Instead of $\max(p) - \min(p)$, which are extrema that strongly fluctuate across subsets of the data, we use the more robust 99% and 1% quantiles.
- [23] L. Ruedigkeit, R. van Deursen, L. C. Blum, and J.-L. Reymond, *J. Chem. Inf. Model.* **52**, 2864 (2012).
- [24] R. Ramakrishnan, P. O. Dral, M. Rupp, and O. A. von Lilienfeld, *Scientific Data* **1**, 140022 (2014).
- [25] K. Burke, *J. Chem. Phys.* **136**, 150901 (2012).
- [26] A. D. Becke, *J. Chem. Phys.* **140**, 18A301 (2014).
- [27] J. P. Perdew, K. Burke, and M. Ernzerhof, *Phys. Rev. Lett.* **77**, 3865 (1996).
- [28] F. Weigend and R. Ahlrichs, *Phys. Chem. Chem. Phys.* **7**, 3297 (2005).
- [29] F. Weigend, *Phys. Chem. Chem. Phys.* **8**, 1057 (2006).
- [30] C. Adamo and V. Barone, *Chem. Phys. Lett.* **298**, 113 (1998).
- [31] M. J. Frisch, G. W. Trucks, H. B. Schlegel, G. E. Scuseria, M. A. Robb, J. R. Cheeseman, G. Scalmani, V. Barone, B. Mennucci, G. A. Petersson, H. Nakatsuji, M. Caricato, X. Li, H. P. Hratchian, A. F. Izmaylov, J. Bloino, G. Zheng, J. L. Sonnenberg, M. Hada, M. Ehara, K. Toyota, R. Fukuda, J.-Y. Hasegawa, M. Ishida, T. Nakajima, Y. Honda, O. Kitao, H. Nakai, T. Vreven, J. Montgomery, John A., J. E. Peralta, F. Ogliaro, M. Bearpark, J. J. Heyd, E. Brothers, K. N. Kudin, V. N. Staroverov, R. Kobayashi, J. Normand, K. Raghavachari, A. Rendell, J. C. Burant, S. S. Iyengar, J. Tomasi, M. Cossi, N. Rega, J. M. Millam, M. Klene, J. E. Knox, J. B. Cross, V. Bakken, C. Adamo, J. Jaramillo, R. Gomperts, R. E. Stratmann, O. Yazyev, A. J. Austin, R. Cammi, C. Pomelli, J. W. Ochterski, R. L. Martin, K. Morokuma, V. G. Zakrzewski, G. A. Voth, P. Salvador, J. J. Dannenberg, S. Dapprich, A. D. Daniels, Ö. Farkas, J. B. Foresman, J. V. Ortiz, J. Cioslowski, and D. J. Fox, “Gaussian 09 revision D.01,” Gaussian Inc., Wallingford CT 2009.
- [32] J. W. Ochterski, *Vibrational Analysis in Gaussian*, Gaussian Inc., Wallingford, Connecticut, USA (1999).
- [33] W. Schneider and W. Thiel, *Chem. Phys. Lett.* **157**, 367 (1989).
- [34] M. J. Mills and P. L. Popelier, *Theor. Chim. Acta* **131**, 1137 (2012).
- [35] J. J. P. Stewart, *J. Comput. Chem.* **10**, 209 (1989).
- [36] S. Amari, N. Fujita, and S. Shinomoto, *Neural. Comput.* **4**, 605 (1992).
- [37] K. Dybiec and A. Gryff-Keller, *Magn. Reson. Chem.* **47**, 63 (2009).
- [38] D. Flaig, M. Maurer, M. Hanni, K. Braunger, L. Kick, M. Thubauville, and C. Ochsenfeld, *J. Chem. Theor. Comput.* **10**, 572 (2014).
- [39] P. d’Antuono, E. Botek, B. Champagne, M. Spassova, and P. Denkova, *J. Chem. Phys.* **125**, 144309 (2006).
- [40] D. E. Hill, N. Vasdev, and J. P. Holland, *Comput. Theor. Chem.* **1051**, 161 (2015).
- [41] V. Myrseth, J. D. Bozek, E. Kukk, L. J. Sæthre, and T. D. Thomas, *J. Electron Spectros. Relat. Phenom.* **122**, 57 (2002).
- [42] T. Karlsen, K. J. Børve, L. J. Sæthre, K. Wiesner, M. Bässler, and S. Svensson, *J. Am. Chem. Soc.* **125**, 7866 (2002).
- [43] A. Holme, K. J. Børve, L. J. Sæthre, and T. D. Thomas, *J. Chem. Theor. Comput.* **7**, 4104 (2011).
- [44] Gaussian’s default convergence threshold of $1\text{ mE}_h/a_0$ was used as target accuracy.
- [45] Although the size of the representation may vary, it is bounded from above by a constant.