The Journal of Chemical Physics

Communication: Understanding molecular representations in machine learning: The role of uniqueness and target similarity

Cite as: J. Chem. Phys. **145**, 161102 (2016); https://doi.org/10.1063/1.4964627
Submitted: 31 August 2016 . Accepted: 22 September 2016 . Published Online: 24 October 2016

Bing Huang, and O. Anatole von Lilienfeld

AIP Publishing

Communication: Understanding molecular representations in machine learning: The role of uniqueness and target similarity

Bing Huang and O. Anatole von Lilienfeld[a)]

Institute of Physical Chemistry and National Center for Computational Design and Discovery of Novel Materials (MARVEL), Department of Chemistry, University of Basel, Klingelbergstrasse 80, 4056 Basel, Switzerland

(Received 31 August 2016; accepted 22 September 2016; published online 24 October 2016)

The predictive accuracy of Machine Learning (ML) models of molecular properties depends on the choice of the molecular representation. Inspired by the postulates of quantum mechanics, we introduce a hierarchy of representations which meet uniqueness and target similarity criteria. To systematically control target similarity, we simply rely on interatomic many body expansions, as implemented in universal force-fields, including Bonding, Angular (BA), and higher order terms. Addition of higher order contributions systematically increases similarity to the true potential energy *and* predictive accuracy of the resulting ML models. We report numerical evidence for the performance of BAML models trained on molecular properties pre-calculated at electron-correlated and density functional theory level of theory for thousands of small organic molecules. Properties studied include enthalpies and free energies of atomization, heat capacity, zero-point vibrational energies, dipole-moment, polarizability, HOMO/LUMO energies and gap, ionization potential, electron affinity, and electronic excitations. After training, BAML predicts energies or electronic properties of out-of-sample molecules with unprecedented accuracy and speed. *Published by AIP Publishing.* [http://dx.doi.org/10.1063/1.4964627]

Reasonable predictions of ground-state properties of molecules require computationally demanding calculations of approximated expectation values of the corresponding operators.[1] Alternatively, Kernel-Ridge-Regression (KRR) based machine learning (ML) models[2] can also *infer* the observable in terms of a linear expansion in chemical compound space.[3–6] More specifically, any observable can be estimated using $O^{\text{inf}}(\mathbf{M}) = \sum_i^N \alpha_i k(d(\mathbf{M},\mathbf{M}_i))$, where k is the kernel function (e.g., Laplacian with training set dependent width), $\mathbf{M}$ is the molecular representation (typically in matrix or vector format),[7,8] and d is a metric (often the L_1-norm). The sum runs over all reference molecules i used for training to obtain regression weights $\{\alpha_i\}$. The advantage of such ML methods consists of (i) their computational efficiency (once trained, typical speed-up is multiple orders of magnitude with respect to conventional quantum chemistry) and (ii) their accuracy can systematically be converged to complete basis set limit through addition of sufficient training instances. Their drawback is that they are incapable of extrapolation by construction, and that they require substantial training data before reaching satisfying predictive power for out-of-sample molecules. In practice, addressing the former drawback is less important since one typically knows beforehand which ranges of interatomic distances and chemical compositions are relevant to the chemical problem at hand: It is straightforward to define the appropriate domain of applicability for the application of supervised ML models in chemistry. In recent years, much work has been devoted to tackle the latter drawback, through the discovery and development of improved representations $\mathbf{M}$.[9–16]

For large N, errors of ML models have been found to decay inverse powers of N,[2] implying a linear relationship, $\log(\text{Error}) = a - b\log(N)$. Therefore, the best representation must (i) minimize the off-set a and (ii) preserve the linearity in the second term while maximizing its pre-factor b. According to the first postulate of quantum mechanics, any system is represented by its wavefunction Ψ which results from applying the variational principle to the expectation value of the Hamiltonian operator. As such, it is important to recall the one-to-one relationship between the Hamiltonian and Ψ. While some representations have been introduced which explicitly mimic the external potential or Ψ (or its corresponding electron density[20])[12,21,22] it is obvious that many observables are extremely sensitive to minute changes in Ψ. As such we prefer to focus directly on the system's Hamiltonian (and its groundstate potential energy surface) defined throughout chemical compound space.[4–6]

Within this study, we have realized that representations based on increasingly more realistic approximations to the potential energy surface afford increasingly more accurate KRR ML models. In other words, the off-set a in the linear log-log learning curve decreases as one increases similarity between representation and potential energy (target similarity). Furthermore, b appears to be a global constant, independent of the representation's target similarity, as long as the crucial[12] uniqueness criterion is met.

First, we exemplify the importance of target similarity for a mock supervised learning task: Modeling a 1D Gaussian

[a)]Electronic mail: anatole.vonlilienfeld@unibas.ch

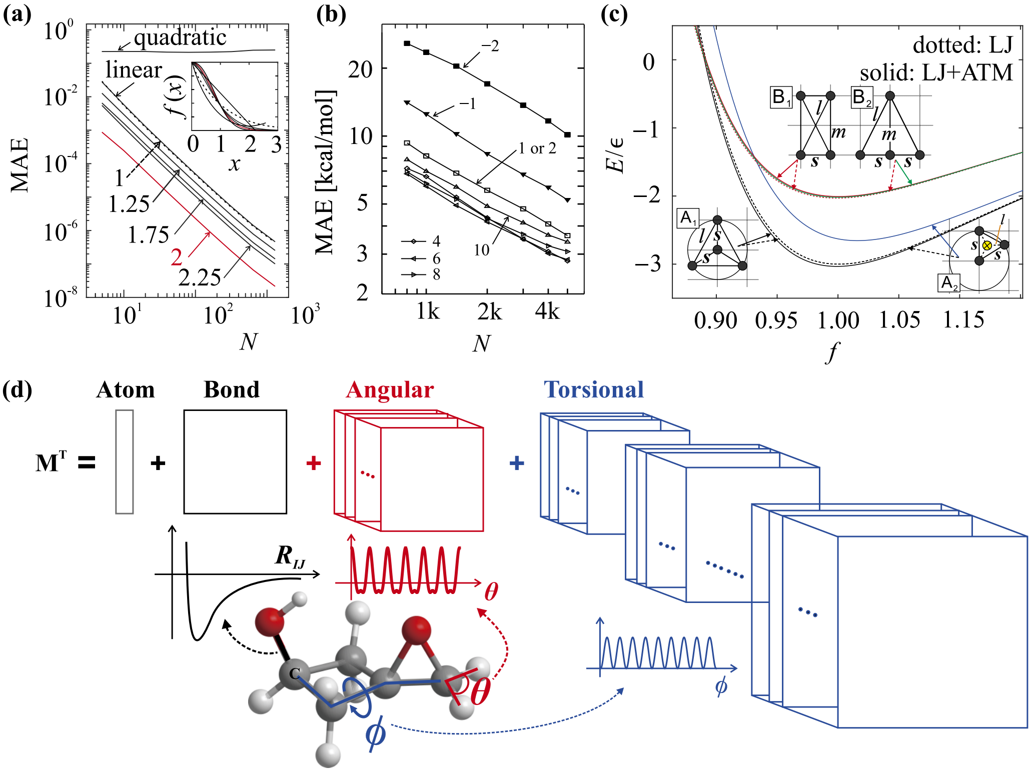

FIG. 1. Target similarity determines offset a in ML model learning curves $\log(\text{Error}) = a - b\log(N)$. Panels (a) and (b) illustrate learning curves for ML models obtained for representations of varying target similarity applied to (a) modeling a 1-D Gaussian target function or (b) enthalpy of atomization for QM7b dataset.[8] Lines in (a) correspond to models resulting from linear, quadratic, and various exponential (e^{-x^n} with $n = \{1, 1.25, 1.75, 2, \text{ and } 2.25\}$) representations. The inset shows the target function (red) as well as the representations. Learning curves in (b) correspond to models resulting from Coulomb matrices with varying definition of off-diagonal elements, $Z_I Z_J / R_{IJ}^n$, where n is specified in the figure. (c) Illustration that 3-body interactions are crucial for distinguishing two pairs of homometric Ar$_4$ clusters: pair A (A$_1$ and A$_2$, where $l = \sqrt{3}s$) and pair B (B$_1$ and B$_2$, where $m = 2s, l = \sqrt{5}s$). Horizontal axis label f scales s, where $f = 1$ corresponds to the choice $s = 3.82$ Å. LJ and ATM correspond to Lennard-Jones and Axilrod-Teller-Muto[17,18] potentials, respectively. (d) Illustration of the universal force-field[19] based construction of the BA representation.

function (inset Fig. 1(a)). As representations **M** we use linear, quadratic, and exponentially decaying functions with varying exponent of x. Learning curves of resulting ML models (Fig. 1(a)) indicate systematic improvement as the target similarity, i.e., similarity of representation to Gaussian function, increases. Note that all learning curves, with the notable exception of the quadratic one, exhibit the same slope b on the log-log plot of the learning curve: They only differ in learning curve off-set a which coincides with their target similarity. When using a Gaussian function as a representation, the smallest off-set is observed—as one would expect. The error of the ML model using the quadratic function as a representation does not decrease when adding more training data: Its minimum is at $x = 2$, and in the region $x > 2$ the function turns upward again, preventing a one-to-one map between x and representation. In other words, the quadratic function is not monotonic and therefore lacks uniqueness, introducing noise in the data which cannot converge to zero and which results in a constant error for large N. By contrast, all other representations are monotonic and conserve the one-to-one map to x. As such, they are unique representations and they all reduce the logarithm of the error in a linear fashion at the same rate as the amount of training data grows. While the rate appears to be solely determined by the uniqueness of the representation, confirming that uniqueness is a necessary condition for functional descriptors,[6] the off-set a of the resulting learning curve appears to be solely determined by target similarity.

To see if our line of reasoning also holds for real molecules, we have investigated the performance of ML models for predicting atomization energies of organic molecules using a set of unique representations with differing target similarity. More specifically, we calculated learning curves for ML models resulting from atom adjacency matrices derived from the Coulomb matrix,[7] with off-diagonal elements $M_{IJ} = Z_I Z_J / R_{IJ}^n$, where R_{IJ} is the interatomic distance between atoms I and J, and the conventional variant (giving rise to the name) is recovered for $n = 1$. For any non-zero choice of n, these matrices encode the complete polyhedron defined in the high-dimensional space spanned by all atoms in the molecules: They uniquely encode the molecule's geometry and composition, thereby ensuring a constant b. For negative n values, however, this representation becomes an unphysical model of the atomization energy: The magnitude of its off-diagonal elements *increases* with interatomic distance. Corresponding learning curves shown in Fig. 1(b) reflect this fact: As off-diagonal elements become increasingly unphysical by dialing in linear and quadratic functions in interatomic distance, respectively, the off-set a increases. Conversely, matrix representations with off-diagonal elements which follow the Coulomb and higher inverse power laws are more physical and exhibit lower off-sets. Interestingly, we note the additional improvement as we change from Coulombic $1/R$ to van der Waals $1/R^6$ like power laws. These results suggest that—after scaling—pairwise London dispersion kind of interactions are more similar to molecular atomization energies than simple Coulomb interactions. In the following, we dub the resulting representation the London Matrix (LM).

The bag-of-bond (BoB) representation, a stripped down pair-wise variant of the Coulomb matrix, has resulted in remarkably predictive ML models.[23] Based on the insights gained from the above, we use the bagging idea as a starting point for the development of our systematically improved representation. Unfortunately, when relying on bags of pair-wise interactions as a representation the uniqueness requirement is violated by arbitrarily many sets of geometries, no matter how strong the (effective) target similarity of the employed functional form. In Fig. 1(c), we illustrate this problem for two pairs of homometric molecules, each with four rare gas atoms: Once in a competition of a pyramidal/planar geometry (A) and once for a rectangular/triangular pair (B). Within both pairs the atomic clusters exhibit the exact same list of interatomic distances: 3 s/3 l for pair A and 2 s/2 m/2 l for pair B. Consequently, when using a pair-wise energy expression, the predicted curve as a function of a global scaling factor f will be indistinguishable (example shown in Fig. 1(d) using Lennard-Jones potentials with parameters for argon). This artificial degeneracy is lifted only after addition of the corresponding three-body van der Waals Axilrod-Teller-Muto[17,18,35] contribution, allowing to distinguish the homometric pairs.[24]

To construct an improved representation based on all of the above, we simply use a hierarchy consisting of bags of (1) dressed atoms ($\mathbf{M}^{\mathrm{D}}$), (2) atoms and bonds ($\mathbf{M}^{\mathrm{B}}$), (2) atoms, bonds, and angles ($\mathbf{M}^{\mathrm{A}}$), and (3) atoms, bonds, angles, and torsions ($\mathbf{M}^{\mathrm{T}}$). To indicate the many-body expansion character, we dub these feature vectors "BA-representation" (standing for bags of Bonds, Angles, Torsions, etc. pp.). The terms are illustrated in Fig. 1(d), and correspond to averaged atomic contributions to energies of molecules in training set for atoms, Morse and Lennard-Jones potentials for covalent and non-covalent intramolecular atom pair-wise bonding, respectively, as well as sinusoidal functions for angles (three body) and torsions (four body) between covalently bounded atoms. Here, we chose functional forms and parameters for BA-representations to correspond to UFF.[19] More technical details are given in the supplementary material.

We tested UFF based BAML using three previously established data sets: DFT energies and properties of ~7k organic molecules stored in the QM7b data set,[8] G4MP2 energies and DFT properties for 6k constitutional isomers of $C_7H_{10}O_2$, and DFT energies and properties for 134k organic molecules QM9 (both published in Ref. 31). Initial structures for all datasets were drawn from the GDB universe.[32,33] Links to all data sets are available at http://quantum-machine.org.

Log-log plots of BAML learning curves are shown in Figs. 2(a) and 2(b) for the $C_7H_{10}O_2$ isomers as well as for QM9. Mean absolute errors (MAEs) for out of sample predictions of nine properties are shown as a function of training set size for the BAML model. Properties studied include enthalpies and free energies of atomization at room temperature, heat-capacity at room temperature C_V, zero-point-vibrational-energy (ZPVE), norms of dipole moments μ and polarizability α, as well as HOMO/LUMO eigenvalues and gap. For any given property, we find near identical learning rates among BAML models based on bonds ($\mathbf{M}^{\mathrm{B}}$), bonds+angles ($\mathbf{M}^{\mathrm{A}}$), and bonds+angles+torsions ($\mathbf{M}^{\mathrm{T}}$), respectively. Note how the learning off-set a decreases systematically as target similarity to energy grows—for all properties, and for both data sets. We note that BAML can reach chemical accuracy (MAE ~1 kcal/mol with respect to reference) for atomization enthalpies of $C_7H_{10}O_2$ isomers after training on only 5k molecules, and MAE ~2.4 and ~1.6 kcal/mol for 134k organic molecules in QM9 after training on 10k and 40k molecules, respectively. Despite its simplicity, such predictive power has not yet been achieved by any other ML model, to the best of our knowledge. This observation confirms the expectations raised based on the aforementioned arguments.

In Fig. 2(c) individual contributions to the atomization energy are on display, resulting from bonds, angles, and torsion representations. For illustration, we have selected outliers, i.e., three constitutional isomers of $C_7H_{10}O_2$ for which the out-of-sample prediction error is maximal. In all three cases, these molecules experience high internal strain through few membered or joint hetero cycles. As such, it is reassuring to observe that substantial lowering of the error occurs as soon as the representation accounts explicitly for angular and torsional degrees of freedom. Fig. 2(d) indicates averaged changes obtained for the entire constitutional isomer testing set due to addition of higher order terms to the representation. More specifically, going from $\mathbf{M}^{\mathrm{D}}$ to $\mathbf{M}^{\mathrm{B}}$ (bonds) contributes on average ~15 kcal/mol; going from $\mathbf{M}^{\mathrm{B}}$ to $\mathbf{M}^{\mathrm{A}}$ (angles) contributes on average another ~2 kcal/mol; while going from $\mathbf{M}^{\mathrm{A}}$ to $\mathbf{M}^{\mathrm{T}}$ (torsion) improves things by merely ~0.5 kcal/mol, on average. Note that the last change might be small on average, however, for some molecules it can be consequential if high accuracy shall be achieved, such as for the aforementioned outliers (Fig. 2(c)). It is encouraging to see that these contributions decrease systematically. This suggests that ML models of energies converge rapidly in the interatomic many-body expansion. It should therefore be possible to construct local yet accurate ML models which scale linearly with system size.

The choice to use bags of interatomic many-body potentials as representation is not obvious, many representations used in the literature rely on the use of other properties, such as HOMO/LUMO eigenvalues or atomic radii and spectra. For two reasons we believe an energy based representation to be advantageous. First, energy is the very observable associated to the Hamiltonian which defines the system: The potential energy surface of a given molecular electronic spin-state is an equally unique representation of the system, two different systems will always differ in their potential energy surface. Secondly, energy is well studied and there is a conveniently large choice of energy models, including UFF, which can be used as representations.

The construction of representations for modeling other properties which at the same time also meet the uniqueness

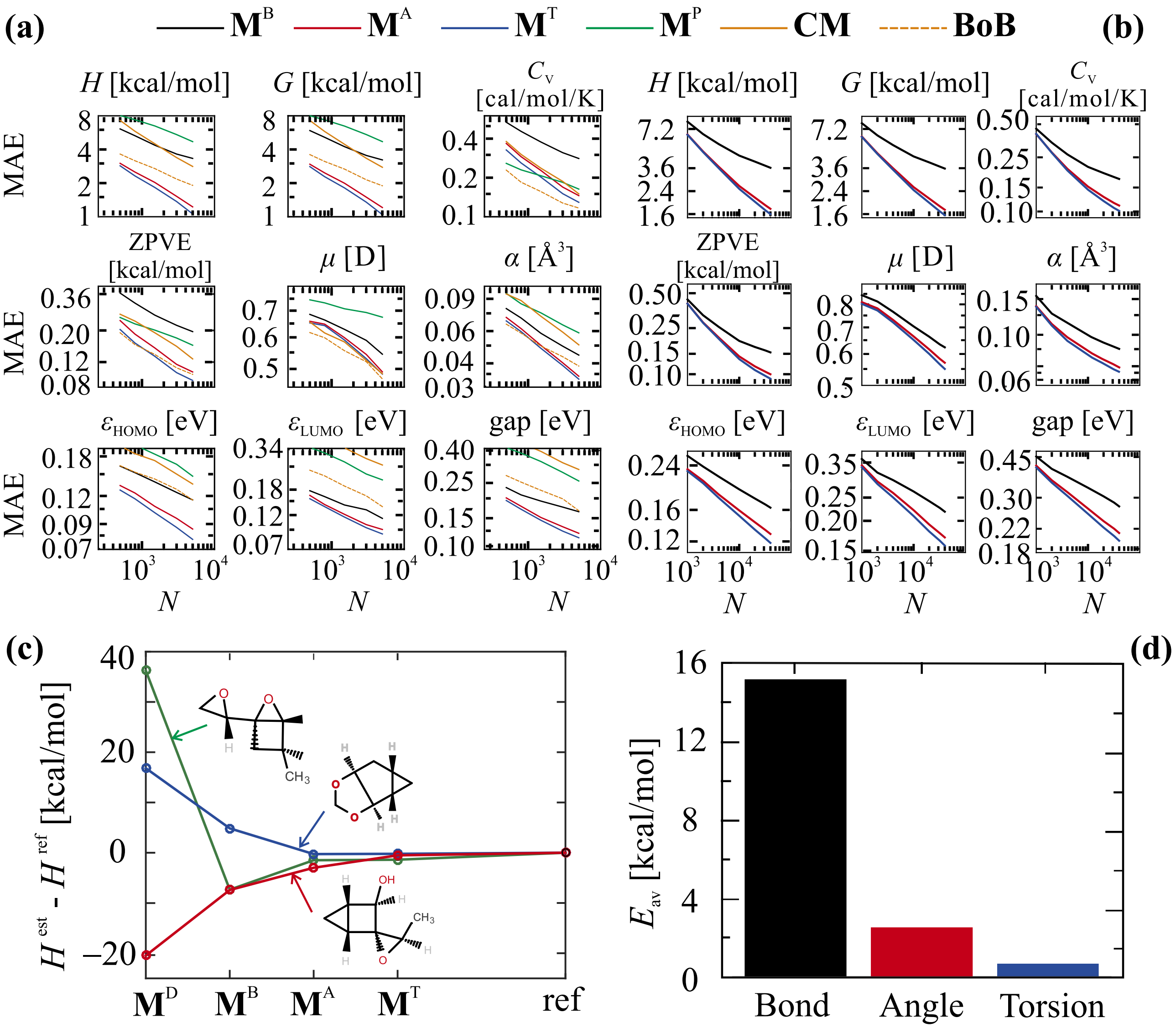

FIG. 2. (a) BAML and polarizability representation based ML learning curves for 9 molecular properties of 6k constitutional isomers of formula $C_7H_{10}O_2$.[31] Results for Coulomb matrix (CM)[7] and bag-of-bonds (BoB)[23] are shown for comparison. (b) BAML learning curves for 134k QM9 molecules for the same 9 molecular properties.[31] Property models cover H, G, C_V, ZPVE, μ, α, $\varepsilon_{\rm HOMO}$ and $\varepsilon_{\rm LUMO}$, i.e., enthalpy, free energy, heat capacity, zero point vibrational energy, dipole moment, polarizability, HOMO and LUMO energy, respectively. Panel (c) shows convergence of estimated enthalpy values to reference values (all shifted to zero) for three most extreme outlier isomers in the $C_7H_{10}O_2$ using BAML models trained on 5k molecules, and panel (d) is the averaged "contribution" of each order type in a many-body potential, i.e., the bond, angle, and torsion parts.

criterion is not obvious. To illustrate this aspect, we constructed a molecular representation ($\mathbf{M}^{\rm P}$) with high target similarity to another property, namely polarizability. Reasonable atomic polarizabilities can easily be obtained from Cartesian coordinates of a molecule, i.e., without electronic structure calculations, through the use of Voronoi polyhedra.[34] Unfortunately, this representation violates the uniqueness criterion. $\mathbf{M}^{\rm P}$ has high similarity to molecular polarizability but it is not unique: Any other molecule which happens to have the same set of atomic volumes, irrespective of differences in geometry, will result in the same representation. Learning curves obtained for $\mathbf{M}^{\rm P}$ based ML models are shown together with BAML in Fig. 2(a) for all constitutional isomers. All BAML models have steeper learning curves for all properties except ZPVE and C_V for which the bond based BAML model performs slightly worse. In the case of the latter, and for very small unconverged training set sizes, the polarizability ML model is even better than any BAML model, however, as training set size grows the lack of uniqueness kicks in with a flatter learning rate leading to worse performance. This observation underscores the importance of taking the convergence behavior into consideration: Learning behavior can differ in a and b. We note the tendency of the polarizability based ML model towards a smaller slope (C_V, ZPVE, μ, α), indicating the expected lack of uniqueness. Surprisingly, even for the target property polarizability $\mathbf{M}^{\rm P}$ based ML models perform worse than BAML. These numerical results support the idea that Hamiltonian/Ψ/potential energy surface is "special" in that all other molecular properties can be derived from them, in direct analogy to the wavefunction Ψ, necessary to calculate the corresponding expectation values.

TABLE I. Mean absolute errors and root mean square errors (in brackets) for the ML predictions of 9 molecular properties of molecules in the QM7b data set.[8] Results from this work (BAML, BoB, BoL, CM, LM) are shown together with a previously published estimation (SOAP,[21] rand CM[8]) for the same dataset. Errors are measured on test set of 2200 randomly selected configurations, while the remaining compounds in QM7b were used for training. Labels specify property and level of theory: Atomization energy (E), averaged molecular polarizability (α), HOMO and LUMO eigenvalues, ionization potential (IP), electron affinity (EA), first excitation energy ($E^*_{1\text{st}}$), excitation frequency of maximal absorption (E^*_{max}), and the corresponding maximal absorption intensity (I_{max}). Expected averaged deviation from experiment is specified in the last column. Bold and bold-italic numbers indicate respective best performance.

Property	SD	BAML	BoB	BoL	CM	LM	SOAP[21]	rand CM[8]	Accuracy
E (PBE0) (kcal/mol)	223.69	1.15 (2.54)	1.84 (4.15)	1.77 (4.07)	3.69 (5.77)	2.84 (4.94)	**0.92** (***1.61***)	3.69 (8.30)	3.46,[a] 5.30,[b] 2.08-5.07[c]
α (PBE0) (Å^3)	1.34	0.07 (0.12)	0.09 (0.13)	0.10 (0.15)	0.13 (0.19)	0.15 (0.20)	**0.05** (***0.07***)	0.11 (0.18)	0.05-0.27,[d] 0.04-0.14[e]
HOMO (GW) (eV)	0.70	**0.10** (***0.16***)	0.15 (0.20)	0.15 (0.20)	0.22 (0.29)	0.20 (0.26)	0.12 (0.17)	0.16 (0.22)	…
LUMO (GW) (eV)	0.48	**0.11** (***0.16***)	0.16 (0.22)	0.16 (0.22)	0.21 (0.27)	0.19 (0.25)	0.12 (0.17)	0.13 (0.21)	…
IP (ZINDO) (eV)	0.96	**0.15** (***0.24***)	0.20 (0.28)	0.20 (0.28)	0.33 (0.44)	0.31 (0.41)	0.19 (0.28)	0.17 (0.26)	0.20, 0.15[d]
EA (ZINDO) (eV)	1.41	**0.07** (***0.12***)	0.17 (0.23)	0.18 (0.24)	0.31 (0.40)	0.25 (0.33)	0.13 (0.18)	0.11 (0.18)	0.16,[f] 0.11[d]
$E^*_{1\text{st}}$ (ZINDO) (eV)	1.87	**0.13** (0.51)	0.21 (***0.30***)	0.22 (0.31)	0.42 (0.57)	0.35 (0.46)	0.18 (0.41)	**0.13** (0.31)	0.18,[f] 0.21[g]
E^*_{max} (ZINDO) (eV)	2.82	1.35 (1.98)	1.40 (1.91)	1.47 (2.02)	1.58 (2.05)	1.68 (2.20)	1.56 (2.16)	**1.06** (***1.76***)	…
I_{max} (ZINDO)	0.22	**0.07** (***0.11***)	0.08 (0.12)	0.08 (0.12)	0.09 (0.13)	0.09 (0.13)	0.08 (0.12)	**0.07** (0.12)	…

[a] MAE of formation enthalpy for the G3/99 set by PBE0.[25,26]
[b] MAE of atomization energy (AE) for 6 small molecules[27,28] by PBE0.
[c] MAE of AE from various studies[29] by B3LYP.
[d] MAE from various studies[29] by B3LYP.
[e] MAE from various studies by MP2.[29]
[f] MAE for the G3/99 set by PBE0.[25,26]
[g] MAE for a set of 17 retinal analogues by TD-DFT(PBE0).[30]

Finally, to place BAML into a broader perspective, we compare out-of-sample errors to Coulomb, London, and literature representation results all obtained for the same molecular data set, QM7b.[8] Table I displays MAEs and RMSEs of the $\mathbf{M}^{\mathrm{T}}$ based BAML model trained on 5k molecules, as well as London Matrix (LM), Coulomb Matrix (CM), BoB (bag of Coulomb matrix elements), bag of London (BoL) matrix elements, SOAP,[21] and randomized CM based neural network model.[8] The SOAP numbers correspond to a recently introduced sophisticated convolution of kernel, metric, and representation. BAML yields a MAE for atomization energy of ∼1.15 kcal/mol, only slightly worse than SOAP's ~0.92 kcal/mol. We consider such small differences to be negligible for most intents and purposes. We also note, however, in Table I that BAML has a considerably larger RMSE (2.5 kcal/mol) for the atomization energy than SOAP (1.6 kcal/mol). Similar observations hold for polarizability. For HOMO/LUMO eigenvalues, ionization potential, electron affinity, and the intensity of the most intense peak, BAML yields lowest MAE and lowest RMSE. BAML also has the lowest MAE for predicting the first excitation energy (together with the randomized neural network based Coulomb matrix model). The lowest RMSE for this property, however, is obtained using a BOB based ML model. The excitation energy of the most intense peak in the model is predicted with the lowest MAE and RMSE when using the randomized neural network based Coulomb matrix model. BAML is second for MAE, and third for RMSE (after BoB). To further illustrate the effect of target similarity, we also report BoL vs. BoB and LM vs. CM based results. Interestingly, for most properties, not only the energy, the corresponding London variant outperforms the Coulomb element based ML models. This supports the above observation that (a) London is more similar to atomization energy than Coulomb and (b) the more similar the representation to energy, the more transferable and applicable it is also for other properties. We note that this table does not represent a comprehensive assessment. It would have been preferable to compare learning curves, such as in Figs. 2(a) and 2(b). Overall, we believe that BAML emerges as the most appealing model: It (i) outperforms all previously established models on average, (ii) is based on understanding the role of target similarity and uniqueness, (iii) is simple and easy to interpret, and (iv) is computationally efficient.

In conclusion, we have presented arguments and numerical evidence in support of the notion that off-set and rate in learning curves are influenced, if not determined, by the employed representation's target similarity and uniqueness, respectively. For molecules, defined by their Hamiltonian which produces their wavefunction which produces the observables, the BAML approach appears to offer uniqueness as well as considerable similarity to energy. Consequently, BAML performs universally well for predicting *any* simple scalar global quantum mechanical observable. Higher-order contributions in the form of bonds, angles, and torsional degrees of freedom enable the systematic lowering of learning curve off-set a, resulting in BAML models with unprecedented accuracy, transferability, and speed. Use of UFF parameters has resulted in consistent and remarkable numerical performance, but other potentials could have been used just as well. We note that the converged optimized learning curves can be seen as a 3D Pareto frontier, spanned between error, chemical space, and energy similarity. This frontier negotiates the trade-off between training set size (or CPU budget), acceptable error, and our understanding of the molecule. This insight is akin to basis set or electron correlation convergence plots common in quantum chemistry, and it could be relevant for generalized automated generation of QM derived property models with predefined uncertainty and transferability.

Technical details pertinent to ML model generation have been summarized in the supplementary material.

O.A.v.L. acknowledges funding from the Swiss National Science Foundation (Nos. PP00P2_138932 and 310030_160067). This research was partly supported by the NCCR MARVEL, funded by the Swiss National Science Foundation. Calculations were performed at sciCORE (http://scicore.unibas.ch/) scientific computing core facility at University of Basel.

[1] A. Szabo and N. S. Ostlund, *Modern Quantum Chemistry: Introduction to Advanced Electronic Structure Theory* (Courier Corporation, 1989).

[2] K.-R. Müller, M. Finke, N. Murata, K. Schulten, and S. Amari, "A numerical study on learning curves in stochastic multilayer feedforward networks," Neural Comput. **8**, 1085–1106 (1996).

[3] P. Kirkpatrick and C. Ellis, "Chemical space," Nature **432**, 823 (2004).

[4] O. A. von Lilienfeld and M. E. Tuckerman, "Molecular grand-canonical ensemble density functional theory and exploration of chemical space," J. Chem. Phys. **125**, 154104 (2006).

[5] O. A. von Lilienfeld, "Accurate *ab initio* gradients in chemical compound space," J. Chem. Phys. **131**, 164102 (2009).

[6] O. A. von Lilienfeld, "First principles view on chemical compound space: Gaining rigorous atomistic control of molecular properties," Int. J. Quantum Chem. **113**, 1676–1689 (2013).

[7] M. Rupp, A. Tkatchenko, K.-R. Müller, and O. A. von Lilienfeld, "Fast and accurate modeling of molecular atomization energies with machine learning," Phys. Rev. Lett. **108**, 058301 (2012).

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

[9] J. Behler, "Atom-centered symmetry functions for constructing high-dimensional neural network potentials," J. Chem. Phys. **134**, 074106 (2011).

[10] A. P. Bartók, R. Kondor, and G. Csányi, "On representing chemical environments," Phys. Rev. B **87**, 184115 (2013).

[11] K. T. Schütt, H. Glawe, F. Brockherde, A. Sanna, K. R. Müller, and E. K. U. Gross, "How to represent crystal structures for machine learning: Towards fast prediction of electronic properties," Phys. Rev. B **89**, 205118 (2014).

[12] O. A. von Lilienfeld, R. Ramakrishnan, M. Rupp, and A. Knoll, "Fourier series of atomic radial distribution functions: A molecular fingerprint for machine learning models of quantum chemical properties," Int. J. Quantum Chem. **115**, 1084–1093 (2015).

[13] L. M. Ghiringhelli, J. Vybiral, S. V. Levchenko, C. Draxl, and M. Scheffler, "Big data of materials science: Critical role of the descriptor," Phys. Rev. Lett. **114**, 105503 (2015).

[14] F. Faber, A. Lindmaa, O. A. von Lilienfeld, and R. Armiento, "Crystal structure representations for machine learning models of formation energies," Int. J. Quantum Chem. **115**, 1094–1101 (2015).

[15] F. Faber, A. Lindmaa, O. A. von Lilienfeld, and R. Armiento, "Machine learning energies of 2 million elpasolite (ABC_2D_6) crystals," Phys. Rev. Lett. **117**, 135502 (2016).

[16] V. Botu and R. Ramprasad, "Adaptive machine learning framework to accelerate *ab initio* molecular dynamics," Int. J. Quantum Chem. **115**, 1074–1083 (2015).

[17] B. M. Axilrod and E. Teller, "Interaction of the van der Waals type between three atoms," J. Chem. Phys. **11**, 299–300 (1943).

[18] Y. Muto, "Force between nonpolar molecules," J. Phys. Math. Soc. Jpn. **17**, 629–631 (1943).

[19] A. K. Rappe, C. J. Casewit, K. S. Colwell, W. A. Goddard III, and W. M. Skiff, "UFF, a full periodic table force field for molecular mechanics and molecular dynamics simulations," J. Am. Chem. Soc. **114**, 10024–10035 (1992).

[20] P. Hohenberg and W. Kohn, "Inhomogeneous electron gas," Phys. Rev. **136**, B864 (1964).

[21] S. De, A. P. Bartok, G. Csanyi, and M. Ceriotti, "Comparing molecules and solids across structural and alchemical space," Phys. Chem. Chem. Phys. **18**, 13754–13769 (2016).

[22] M. Hirn, S. Mallat, and N. Poilvert, "Wavelet scattering regression of quantum chemical energies," e-print arXiv:1605.04654 (2016).

[23] K. Hansen, F. Biegler, R. Ramakrishnan, W. Pronobis, O. A. von Lilienfeld, K.-R. Müller, and A. Tkatchenko, "Machine learning predictions of molecular properties: Accurate many-body potentials and nonlocality in chemical space," J. Phys. Chem. Lett. **6**, 2326–2331 (2015).

[24] This simple example also underscores the qualitative importance of many-body contributions in interatomic energy decompositions which play a role even in effects as weak as London dispersion forces—in addition to the already established quantitative role they play in nature.[34,35]

[25] V. N. Staroverov, G. E. Scuseria, J. Tao, and J. P. Perdew, "Comparative assessment of a new nonempirical density functional: Molecules and hydrogen-bonded complexes," J. Chem. Phys. **119**, 12129–12137 (2003).

[26] L. A. Curtiss, K. Raghavachari, P. C. Redfern, and J. A. Pople, "Assessment of Gaussian-3 and density functional theories for a larger experimental test set," J. Chem. Phys. **112**, 7374–7383 (2000).

[27] Y. Zhao, J. Pu, B. J. Lynch, and D. G. Truhlar, "Tests of second-generation and third-generation density functionals for thermochemical kinetics," Phys. Chem. Chem. Phys. **6**, 673–676 (2004).

[28] B. J. Lynch and D. G. Truhlar, "Small representative benchmarks for thermochemical calculations," J. Phys. Chem. A **107**, 8996–8999 (2003).

[29] W. Koch and M. C. Holthausen, *A Chemist's Guide to Density Functional Theory* (John Wiley & Sons, 2015).

[30] C. S. López, O. N. Faza, S. L. Estévez, and A. R. de Lera, "Computation of vertical excitation energies of retinal and analogs: Scope and limitations," J. Comput. Chem. **27**, 116–123 (2006).

[31] R. Ramakrishnan, P. O. Dral, M. Rupp, and O. A. von Lilienfeld, "Quantum chemistry structures and properties of 134 kilo molecules," Sci. Data **1**, 140022 (2014).

[32] L. Ruddigkeit, R. van Deursen, L. C. Blum, and J.-L. Reymond, "Enumeration of 166 billion organic small molecules in the chemical universe database GDB-17," J. Chem. Inf. Model. **52**, 2864–2875 (2012).

[33] T. Fink, H. Bruggesser, and J.-L. Reymond, "Virtual exploration of the small-molecule chemical universe below 160 daltons," Angew. Chem., Int. Ed. **44**, 1504–1508 (2005).

[34] T. Bereau and O. A. von Lilienfeld, "Toward transferable interatomic van der Waals interactions without electrons: The role of multipole electrostatics and many-body dispersion," J. Chem. Phys. **141**, 034101 (2014).

[35] R. A. DiStasio, O. A. von Lilienfeld, and A. Tkatchenko, "Collective many-body van der Waals interactions in molecular systems," Proc. Natl. Acad. Sci. U. S. A. **109**, 14791–14795 (2012).