JCTC Journal of Chemical Theory and Computation

Article

Cite This: J. Chem. Theory Comput. 2017, 13, 5255-5264

pubs.acs.org/JCTC

Prediction Errors of Molecular Machine Learning Models Lower than Hybrid DFT Error

Felix A. Faber,[†] Luke Hutchison,[‡] Bing Huang,[†] Justin Gilmer,[‡] Samuel S. Schoenholz,[‡] George E. Dahl,[‡] Oriol Vinyals,[¶] Steven Kearnes,[‡] Patrick F. Riley,[‡] and O. Anatole von Lilienfeld*[,†]

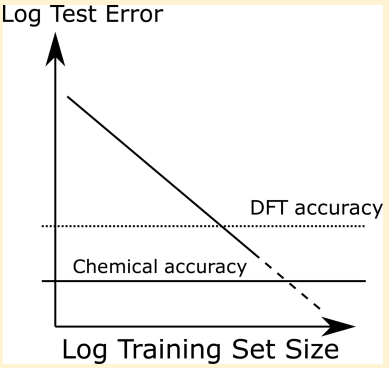
[†]Institute of Physical Chemistry and National Center for Computational Design and Discovery of Novel Materials, Department of Chemistry, University of Basel, Klingelbergstrasse 80, CH-4056 Basel, Switzerland

[‡]Google, 1600 Amphitheatre Parkway, Mountain View, California 94043, United States

[¶]Google, 5 New Street Square, London EC4A 3TW, U.K.

Supporting Information

ABSTRACT: We investigate the impact of choosing regressors and molecular representations for the construction of fast machine learning (ML) models of 13 electronic ground-state properties of organic molecules. The performance of each regressor/representation/property combination is assessed using learning curves which report out-of-sample errors as a function of training set size with up to ~118k distinct molecules. Molecular structures and properties at the hybrid density functional theory (DFT) level of theory come from the QM9 database [Ramakrishnan et al. *Sci. Data* **2014** *1*, 140022] and include enthalpies and free energies of atomization, HOMO/LUMO energies and gap, dipole moment, polarizability, zero point vibrational energy, heat capacity, and the highest fundamental vibrational frequency. Various molecular representations have been studied (Coulomb matrix, bag of bonds, BAML and ECFP4, molecular graphs (MG)), as well as newly developed distribution based variants including histograms of distances (HD), angles (HDA/MARAD), and dihedrals (HDAD). Regressors include linear models (Bayesian ridge regression (BR) and linear regression with elastic net regularization (EN)), random forest (RF), kernel ridge regression (KRR), and two types of neural networks, graph convolutions (GC) and gated graph networks (GG). Out-of sample errors are strongly dependent on the choice of representation *and* regressor *and* molecular property. Electronic properties are typically best accounted for by MG and GC, while energetic properties are better described by HDAD and KRR. The specific combinations with the lowest out-of-sample errors in the ~118k training set size limit are (free) energies and enthalpies of atomization (HDAD/KRR), HOMO/LUMO eigenvalue and gap (MG/GC), dipole moment (MG/GC), static polarizability (MG/GG), zero point vibrational energy (HDAD/KRR), heat capacity at room temperature (HDAD/KRR), and highest fundamental vibrational frequency (BAML/RF). We present numerical evidence that ML model predictions deviate from DFT (B3LYP) less than DFT (B3LYP) deviates from experiment for all properties. Furthermore, out-of-sample prediction errors with respect to hybrid DFT reference are on par with, or close to, chemical accuracy. The results suggest that ML models could be more accurate than hybrid DFT if explicitly electron correlated quantum (or experimental) data were available.



1. INTRODUCTION

Due to its favorable trade-off between accuracy and computational cost, density functional theory (DFT)[1,2] is the workhorse of quantum chemistry[3] — despite its well-known shortcomings regarding spin-states, van der Waals interactions, and chemical reactions.[4,5] Failures to predict reaction profiles are particularly worrisome,[6] and recent analysis casts even more doubts on the usefulness of DFT functionals obtained through parameter fitting.[7] The prospect of universal and computationally much more efficient machine learning (ML) models, trained on data from experiments or generated at higher levels of electronic structure theory such as post-Hartree−Fock or quantum Monte Carlo (e.g., exemplified in ref 8), therefore represents an appealing alternative strategy. Not surprisingly, a great deal of recent effort has been devoted to developing ever more accurate ML models of properties of molecular and condensed phase systems.

Several ML studies have already been published using a data set called QM9,[9] consisting of molecular quantum properties for the ~134k smallest organic molecules containing up to 9 heavy atoms (C, O, N, or F; not counting H) in the GDB-17 universe.[10] Some of these studies have developed or used representations we consider in this work, such as BAML (bonds, angles, machine learning),[11] bag of bonds (BOB),[12,13] and the Coulomb matrix (CM).[13,14] Atomic variants of the CM have also been proposed and tested on QM9.[15] Other representations have also been benchmarked on QM9 (or

Received: June 7, 2017
Published: September 19, 2017

DOI: 10.1021/acs.jctc.7b00577
J. Chem. Theory Comput. 2017, 13, 5255−5264

QM7 which is a smaller but similar data set), such as Fourier series of radial distance distributions,[16] motifs,[17] the smooth overlap of atomic positions (SOAP)[18] in combination with regularized entropy match,[19] and constant size descriptors based on connectivity and encoded distance distributions.[20] Ramakrishnan et al.[8] introduced a Δ-ML approach, where the difference between properties calculated at coarse/accurate quantum level of theories is being modeled. Furthermore, neural network models, as well as deep tensor neural networks, have recently been proposed and tested on the same or similar data sets.[21,22] Dral et al.[23] use such data to machine learn optimal molecule specific parameters for the OM2[24] semi-empirical method, and orthogonalization tests are benchmarked in ref 25.

However, limited work has yet been done in systematically assessing *various* methods *and* properties on large sets of the exact same chemicals.[26] In order to unequivocally establish if ML has the potential to replace hybrid DFT for the screening of properties, one has to demonstrate that ML test errors are systematically lower than estimated hybrid DFT accuracies for all the properties available. This study accomplishes that through a large scale assessment of unprecedented scale: (i) In order to approximate large training set sizes N, we included 13 quantum properties from up to ∼118k molecules in training (90% of QM9). (ii) We tested multiple regressors (Bayesian ridge regression (BR), linear regression with elastic net regularization (EN), random forest (RF), kernel ridge regression (KRR), neural network (NN) models graph convolutions (GC),[27] and gated graphs (GG)[28]) and (iii) multiple representations including BAML, BOB, CM, extended connectivity fingerprints (ECFP4), histograms of distance, angle, and dihedral (HDAD), molecular atomic radial angular distribution (MARAD), and molecular graphs (MG). (iv) We investigated *all* combinations of regressors and representations, except for MG/NN which was exclusively used together because GC and GG depend fundamentally on the input representation being a graph instead of a flat feature vector.

The best models for the various properties are atomization energy at 0 K (HDAD/KRR), atomization energy at room temperature (HDAD/KRR), enthalpy of atomization at room temperature (HDAD/KRR), atomization of free energy at room temperature (HDAD/KRR), HOMO/LUMO eigenvalue and gap (MG/GC), dipole moment (MG/GC), static polarizability (MG/GG), zero point vibrational energy (HDAD/KRR), heat capacity at room temperature (HDAD/KRR), and the highest fundamental vibrational frequency (BAML/RF). For training set size of ∼118k (90% of data set) we have found the additional out-of-sample error added by machine learning to be lower or as good as DFT errors at the B3LYP level of theory relative to experiment for all properties and that chemical accuracy (see Table 3 and Table 4) is reached or in sight.

This paper is organized as follows: First we will briefly describe the methods, including data set, model validation protocols, representations, and regressors. In Section 3III, we present the results and discuss them, and Section 4IV concludes the paper.

2. METHOD

2.1. Data Set. We have used the QM9 data set consisting of ∼134k drug-like organic molecules.[9] Molecules in the data set consist of H, C, O, N, and F and contain up to 9 heavy atoms. For each molecule several properties, calculated at the DFT level of theory (B3LYP/6-31G(2df,p)), were included. We used the following: atomization energy at 0 K U_0 (eV); atomization energy at room temperature U (eV); enthalpy of atomization at room temperature H (eV); atomization of free energy at room temperature G (eV); HOMO eigenvalue $\varepsilon_{\mathrm{HOMO}}$ (eV); LUMO eigenvalue $\varepsilon_{\mathrm{LUMO}}$ (eV); HOMO−LUMO gap $\Delta\varepsilon$ (eV); norm of dipole moment $\mu = \sqrt{\sum_{r\in x,y,z}(\int d\mathbf{r} n(\mathbf{r}) r)^2}$ (Debye), where $n(\mathbf{r})$ is the molecular charge density distribution; static isotropic polarizability $\alpha = \frac{1}{3}\sum_{i\in x,y,z}\alpha_{\mathrm{ii}}$ (Bohr^3), where α_{ii} is the diagonal element of the polarizability tensor; zero point vibrational energy ZPVE (eV); heat capacity at room temperature C_v (cal/mol/K); and the highest fundamental vibrational frequency ω_1 (cm^{-1}). For energies of atomization (U_0, U, H, and G) all models yield very similar errors. We will therefore only discuss U_0 for the remainder. The 3053 molecules specified in ref 9 which failed SMILES consistency tests were excluded from our study, as well as two linear molecules, leaving ∼131k molecules.

2.2. Model Validation. Starting from the ∼131k molecules in QM9 after removing the ∼3k molecules (see above) we created a number of train-validation-test splits. We split the data set into test and nontest sets and varied the percentage of data in the test set to explore the effect of the amount of data on error rates. Inside the nontest set, we performed 10-fold cross validation for hyperparameter optimization. That is, for each model 90% (the training set) of the nontest set is used for training and 10% (the validation set) is used for hyper-parameter selection. For each test/nontest split, we have trained 10 models on different subsets of the nontest set, and we report the mean error on the test set across those 10 models. Note that the nontest set will be referred to as the training set in the Results section in order to simplify discussion.

In terms of CPU investments necessary for training the respective models, we note that EN/BR, RF/KRR, and GC/GG required minutes, hours, and multiple days, respectively. Using GPUs could dramatically reduce such timings.

2.3. DFT Errors. To place the quality of our prediction errors in the right context, experimental accuracy estimates of hybrid DFT become desirable. Here, we summarize literature results comparing DFT *at the B3LYP level of theory* to experiments for the various properties we study. Where data are available, the corresponding deviation from experiment is listed in Table 3, alongside our ML prediction errors (*vide infra*).

In order to also get an idea of hybrid DFT energy errors for organic molecules, such as the compounds studied herewithin, we refer to a comparison of PBE and B3LYP results for 6k constitutional isomers of $C_7H_{10}O_2$.[8] After centering the data by subtracting their mean shift from G4MP2 (177.8 (PBE) and 95.3 (B3LYP) kcal/mol), the remaining MAEs are roughly ∼2.5 and ∼3.0 kcal/mol for B3LYP and PBE, respectively. This is in agreement with what Curtiss et al.[29] found. They compared DFT to experimental values from 69 small organic molecules (of which 47 were substituted with F, Cl, and S), with up to 6 heavy atoms (not counting hydrogens), and calculated the energies using B3LYP/6-311+G(3df,2p). The resulting mean absolute deviation from experimental values was 2.3 kcal/mol.

Rough hybrid DFT error estimates for dipole moment and polarizability have been obtained from ref 30. The errors are

DOI: 10.1021/acs.jctc.7b00577
J. Chem. Theory Comput. 2017, 13, 5255−5264

estimated referenced to experimental values, for a data set consisting of 49 molecules with up to 7 heavy atoms (C, Cl, F, N, O, P, or S).

Frontier molecular orbital energies (HOMO, LUMO, and HOMO−LUMO gap) cannot be measured directly. However, for the exact (yet unknown) exchange-correlation potential, the Kohn−Sham HOMO eigenvalues correspond to the negative of the vertical ionization potential (IP).[31] Unfortunately, within hybrid DFT, the precise meaning of the frontier eigenvalues and the gap is less clear, and we therefore refrain from a direct comparison of B3LYP to experimental numbers. Nevertheless, we have included eigenvalues and the gap due to their widespread use for molecular and materials design applications.

Hybrid DFT RMSE estimates with respect to experimental values of ZPVE and ω_1 (the highest fundamental vibrational frequency) were published in ref 32 for a set of 41 organic molecules, with up to 6 heavy atoms (not counting hydrogen) and calculated using B3LYP/cc-pVTZ.

Normally distributed data have a constant ratio between RMSE and MAE,[33] which is roughly 0.8. We have used this ratio to approximate the MAE from the RMSE estimates reported for ZPVE and ω_1. Deviation of DFT (at the B3LYP/6-311g** level of theory) from experimental heat capacities was reported by DeTar[34] who obtained errors of 16 organic molecules, with up to 8 heavy atoms (not counting hydrogens).

Note, however, that one should be cautious when referring to these errors: Strictly speaking they cannot be compared since different basis sets, molecules, and experiments were used. We also note that all DFT errors in this paper are estimated from B3LYP, and using other functionals can yield very different errors.

Nevertheless, we feel that the quoted errors provide meaningful guidance as to what one can expect from DFT for each property.

2.4. Representations. The design of molecular representations is a long-standing problem in cheminformatics and materials informatics, and many interesting and promising variants have already been proposed. Below, we provide the details on the representations selected for this study. While finalizing our study, competitive alternatives were introduced[35,36] but have been tested only for energies (and polarizabilities).

2.4.1. CM and BOB. The Coulomb matrix (CM) representation[14] is a square atom by atom matrix, where off diagonal elements are the Coulomb nuclear repulsion terms between atom pairs. The diagonal elements approximate the electronic potential energy of the free atoms. Atom indices in the CM are sorted by the L^1 norm of each atom's row (or column). The Bag of Bonds (BOB)[12] representation uses exclusively CM elements, grouping them for different atom pairs into different bags and sorting them within each bag by their relative magnitude.

2.4.2. BAML. The recently introduced BAML (bonds, angles, machine learning) representation can be viewed as a many-body extension of BOB.[11] All pairwise nuclear repulsions are replaced by Morse/Lennard-Jones potentials for bonded/nonbonded atoms, respectively. Furthermore, three- and four-body interactions between covalently bonded atoms are included using angular and torsional terms, respectively. Parameters and functional forms are based on the universal force field (UFF).[37]

2.4.3. ECFP4. Extended connectivity fingerprints[38] (ECFP4) are a common representation of molecules in cheminformatics based studies. They are particularly popular for drug discovery.[39−41] The basic idea, typical also for other cheminformatics descriptors[42] (e.g., the *signature* descriptor[43,44]), is to represent a molecule as the set of subgraphs up to a fixed diameter (here we use ECFP4, which is a max diameter of 4 bonds). To produce a fixed length vector, the subgraphs can be hashed such that every subgraph sets one bit in the fixed length vector to 1. In this work, we use a fixed length vector of size 1024. Note that ECFP4 is based solely on the molecular graph specifying all covalent bonds, e.g. as encoded by SMILES strings.

2.4.4. MARAD. Molecular atomic radial angular distribution (MARAD) is an atomic radial distribution function (RDF) based representation. Per atom it consists of three RDFs using Gaussians of interatomic distances and parallel and orthogonal projections of distances in atom triplets, respectively. Distances between two molecules can be evaluated analytically. Unfortunately, most regressors evaluated in this work, such as BR, EN, and RF, do not rely on inner products and distances between representations. We resolve this issue by projecting MARAD onto bins in order to work with all regressors (apart for GG and GC which use MG exclusively). The three-body terms in MARAD contain information about both angles and distances of all atoms involved. This differs from HDA (see below), where distances and angles are decoupled and placed in separated bins. Note that unlike BAML or HDAD, there are only two- and three-body terms, no four-body terms (dihedral angles) have been included within MARAD.

Details about how the projected MARAD is calculated can be found in the Supporting Information.

Further details and characteristics of MARAD will also be discussed in a forthcoming separate in-depth study.

2.4.5. HD, HDA, and HDAD. BOB, BAML, and MARAD rely on computing functions for given interatomic distances, and/or angles, and/or torsions and then either project that value on to discrete bins or sort the values. As a straightforward alternative, we also investigated representations which account directly from pairwise distances, triple-wise angles, and quad-wise dihedral angles through manually generated bins in histograms. The resulting representations in increasing interatomic many-body order are called HD (histogram of distances), HDA (histogram of distances and angles), and HDAD (histogram of distances, angles, and dihedral angles). For any given molecule, one iterates through each atom a_i, producing a set of distances, angle, and dihedral angle features for a_i.

Distance features were produced by measuring the distance between a_i and a_j (for $i \neq j$) for each element pair. The distance features were assigned a label incorporating the atomic symbols of a_i and a_j sorted alphabetically (with H last), e.g. if a_i was a carbon atom and a_j was a nitrogen atom, the distance feature for the atom pair would be labeled C−N. These labels will be used to group all features with the same label into a histogram and allow us to only count each pair of atoms once.

Angle features were produced by taking the principal angles formed by the two vectors spanning from each atom a_i to every subset of 2 of its 3 nearest atoms, a_j and a_k. The angle features were labeled by the element type of a_i, followed by the alphabetically sorted element types (except for hydrogens, which were listed last) of a_j and a_k. The example where a_i is a carbon atom, a_j is a hydrogen atom, a_k is a nitrogen would be assigned the label C−N−H.

Dihedral angle features were produced by taking the principal angles between two planes. We take a_i as the origin, and for

DOI: 10.1021/acs.jctc.7b00577
J. Chem. Theory Comput. 2017, 13, 5255−5264

each of the four nearest neighbors in turn, labeling the neighbor atom a_j and forming a vector $V_{ij} = a_i \rightarrow a_j$. Then all $\binom{3}{2}$ subsets of the remaining three out of four nearest neighbors of a_i are chosen and labeled as a_k and a_l. This third and fourth atoms respectively form two triangular faces when paired with V_{ij}: $\langle a_k, a_i, a_j\rangle$ and $\langle a_l, a_i, a_j\rangle$. The dihedral angle between the two triangular faces was calculated. These dihedral angle features were labeled with the atomic symbol for a_i, followed by the atomic symbols for a_j, a_k, and a_l, sorted alphabetically, with the exception that hydrogens were listed last, e.g. C−C−N−H.

The features from all molecules have been aggregated for each label type to generate a histograms for each label type. Figure 1 illustrates this for C−N distances, C−C−C angles, and

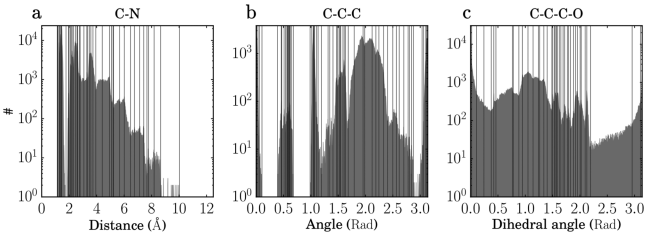


Figure 1. Illustration of select histograms of distances, angles, and dihedral angles in QM9. Vertical lines constitute placements of bin centers in the HD and/or HDAD representations. (a) All C−N distances. (b) All C−C−C angles. (c) All C−C−C−O dihedral angles.

C−C−C−O dihedrals for the entire QM9 data set. Certain typical molecular characteristics can be recognized upon mere inspection. For example, the CN histogram displays a strong and isolated peak between 1.1 and 1.5 Å, corresponding to occurrences of single, double, and triple bonds. For distances above 2 Å, peaks at typical radii of second and third covalent bonding shells around N can be recognized at 2.6 and 3.9 Å. Also C−C−C angles can be easily interpreted: The peak close to zero and π radians corresponds to geometries where three atoms are part of a linear (alkyne or nitrile) type of motif. The broad and largest peak corresponds to 120 and 109 degrees, typically observed in sp^2 and sp^3 hybridized atoms.

The morphology of each histogram has then been examined to identify apparent peaks and troughs, motivated by the idea that peaks indicate structural commonalities among molecules. Bin centers have been placed at each significant local minimum and maximum (shown as vertical lines in Figure 1). Approximately 15−25 bin centers have been chosen as a representation for each label type. All bin center values are provided in the Supporting Information. Bin boundaries were dened at the midpoint between each pair of adjacent bin centers. For each molecule, the collection of features has subsequently been rendered into a fixed-size representation, producing one vector component for each bin center, within each label type. This has been accomplished following a two-step process. (i) *Binning and interpolation:* Each feature value is projected on the two nearest bins. The relative amount projected on each bin uses linear projection between the two bins. For example: A feature with value 1.7 which lies between two bins placed at 1.0 and 2.0, respectively, contributes 0.3 and 0.7 to the first and second bin, respectively. (ii) *Reduction:* The collection of contributions within each bin of each molecule's feature vector is condensed to a single value by summing all contributions.

2.4.6. Molecular Graphs. We investigated several neural network models which are based on molecular graphs (MG) as representation. The inputs are real-valued vectors associated with each atom and with each pair of atoms. More specifically, we have used the featurization described in Kearnes et al.[27] with the removal of partial charge and the addition of Euclidean distances to the pair feature vectors. All elements of the feature vector are described in Tables 1 and 2.

Table 1. Atom Features for the MG Representation[a]

feature	description
atom type	H, C, N, O, F (one-hot)
chirality	R or S (one-hot or null)
formal charge	integer electronic charge
ring sizes	for each ring size (3−8), the number of rings that include this atom
hybridization	sp, sp^2, or sp^3 (one-hot or null)
hydrogen bonding	whether this atom is a hydrogen bond donor and/or acceptor (binary values)
aromaticity	whether this atom is part of an aromatic system

[a]Values are provided for each atom in the molecule.

Table 2. Atom Pair Features for the MG Representation[a]

feature	description
bond type	single, double, triple, or aromatic (one-hot or null)
graph distance	for each distance (1−7), whether the shortest path between the atoms in the pair is less than or equal to that number of bonds (binary values)
same ring	whether the atoms in the pair are in the same ring
spatial distance	the Euclidean distance between the two atoms

[a]Values are provided for each pair of atoms in the molecule.

The featurization process was unsuccessful for a small number of molecules (367) because of conversion failures from geometry to rational SMILES string when using OpenBabel[45] or RDKit[46] and was excluded from all results using the molecule graph features.

Note that within a previous draft of this study,[47] we reported biased results for GC/GG models due to the use of Mulliken partial charges within the MG representation. All MG results presented herewithin have been obtained without any Mulliken charges in the representation. Model hyperparameters for the GC model, however, still correspond to the previously obtained hyperparameter search.

2.5. Regressors. For all methods, we first standardized the property values so that all properties have zero mean and unit standard deviation.

2.5.1. Kernel Ridge Regression. KRR[48−51] is a type of regression with regularization[52] which uses kernel functions as basis set. A property p of a query molecule $\mathbf{m}$ is predicted by a sum of weighted kernel functions $K(\mathbf{m},\mathbf{m}_i^{\text{train}})$ between $\mathbf{m}$ and all molecules $\mathbf{m}_i^{\text{train}}$ in the training set

$$p(\mathbf{m}) = \sum_i^N \alpha_i K(\mathbf{m}, \mathbf{m}_i^{\text{train}}) \quad (1)$$

where α_i are regression coefficients, obtained by minimizing the Euclidean distance between the estimated and the reference property of all molecules in the training set. We used Laplacian and Gaussian kernels as implemented by scikit-learn[53] for all representations.

DOI: 10.1021/acs.jctc.7b00577
J. Chem. Theory Comput. 2017, 13, 5255−5264

Table 3. MAE on out-of-Sample Data of All Representations for All Regressors and Properties at ∼118k (90%) Training Set Size[a]

		U_0, eV	ε_{HOMO}, eV	ε_{LUMO}, eV	$\Delta\varepsilon$, eV	μ, Debye	α, Bohr^3	ZPVE, eV	C_ν, cal/molK	ω_1, cm^{-1}	NMMAE, arb. u.
EN	CM	0.9110	0.3380	0.6310	0.7220	0.844	1.330	0.02650	0.9060	131.00	0.4230
	BOB	0.6020	0.2830	0.5210	0.6140	0.763	1.200	0.02320	0.7000	81.40	0.3500
	BAML	0.2120	0.1860	0.2750	0.3390	0.686	0.793	0.01290	0.4390	60.40	0.2310
	ECFP4	3.680	0.2240	0.3440	0.3830	0.737	3.450	0.27000	1.5100	86.60	0.4620
	HDAD	0.0983	0.1390	0.2380	0.2780	0.563	0.437	0.00647	0.0876	94.20	0.1830
	HD	0.1920	0.2030	0.2990	0.3600	0.705	0.638	0.00949	0.1950	104.00	0.2360
	MARAD	0.1830	0.2220	0.3050	0.3910	0.707	0.698	0.00808	0.2060	108.00	0.2560
	mean	0.8400	0.2280	0.3730	0.4410	0.715	1.220	0.05090	0.5780	95.10	
BR	CM	0.9110	0.3380	0.6320	0.7230	0.844	1.330	0.02650	0.9070	131.00	0.4240
	BOB	0.5860	0.2790	0.5210	0.6140	0.761	1.140	0.02220	0.6840	80.90	0.3430
	BAML	0.2020	0.1830	0.2750	0.3390	0.685	0.785	0.01290	0.4440	60.40	0.2290
	ECFP4	3.6900	0.2240	0.3440	0.3830	0.737	3.450	0.27000	1.5100	86.70	0.4620
	HDAD	0.0614	0.1400	0.2380	0.2780	0.565	0.430	0.00318	0.0787	94.80	0.1820
	HD	0.1710	0.2030	0.2980	0.3590	0.705	0.633	0.00693	0.1900	104.00	0.2350
	MARAD	0.1710	0.1840	0.2570	0.3150	0.647	0.533	0.00854	0.2010	103.00	0.2260
	mean	0.8280	0.2210	0.3670	0.4300	0.706	1.190	0.05000	0.5740	94.50	
RF	CM	0.4310	0.2080	0.3020	0.3730	0.608	1.040	0.01990	0.7770	13.20	0.2390
	BOB	0.2020	0.1200	0.1370	0.1640	0.450	0.623	0.01110	0.4430	3.55	0.1420
	BAML	0.2000	0.1070	0.1180	0.1410	0.434	0.638	0.01320	0.4510	**2.71**	0.1410
	ECFP4	3.6600	0.1430	0.1450	0.1660	0.483	3.700	0.24200	1.5700	14.70	0.3490
	HDAD	1.4400	0.1160	0.1360	0.1560	0.454	1.710	0.05250	0.8950	3.45	0.1980
	HD	1.3900	0.1260	0.1390	0.1500	0.457	1.660	0.04970	0.8790	4.18	0.1970
	MARAD	0.2100	0.1780	0.2430	0.3110	0.607	0.676	0.01020	0.3110	19.40	0.1990
	mean	1.0800	0.1420	0.1740	0.2090	0.499	1.430	0.05690	0.7610	8.74	
KRR	CM	0.1280	0.1330	0.1830	0.2290	0.449	0.433	0.00480	0.1180	33.50	0.1360
	BOB	0.0667	0.0948	0.1220	0.1480	0.423	0.298	0.00364	0.0917	13.20	0.0981
	BAML	0.0519	0.0946	0.1210	0.1520	0.460	0.301	0.00331	0.0820	19.90	0.1050
	ECFP4	4.2500	0.1240	0.1330	0.1740	0.490	4.170	0.24800	1.8400	26.70	0.3830
	HDAD	**0.0251**	0.0662	0.0842	0.1070	0.334	0.175	**0.00191**	**0.0441**	23.10	0.0768
	HD	0.0644	0.0874	0.1130	0.1430	0.364	0.299	0.00316	0.0844	21.30	0.0935
	MARAD	0.0529	0.1030	0.1240	0.1630	0.468	0.343	0.00301	0.0758	21.30	0.1120
	mean	0.6620	0.1010	0.1260	0.1590	0.427	0.859	0.03830	0.3330	22.70	
GG	MG	0.0421	0.0567	0.0628	0.0877	0.247	**0.161**	0.00431	0.0837	6.22	0.0602
GC	MG	0.1500	**0.0549**	**0.0620**	**0.0869**	**0.101**	0.232	0.00966	0.0970	4.76	**0.0494**

[a]Regressors include linear regression with elastic net regularization (EN), Bayesian ridge regression (BR), random forest (RF), kernel ridge regression (KRR), and molecular graphs based neural networks (GG/GC). The best combination for each property is highlighted in bold. Additionally, the table contains mean MAE of representations for each property and regressor, normalized by MAD (see Table 4) and mean MAE (NMMAE) over all properties for each regressor/representation combination.

The level of noise in our data is very low, so strong regularization is not necessary. We have set the regularization parameter to 10^{-9}, and we note that prediction errors change negligibly when altering it to 10^{-10}. Kernel widths were chosen by screening values on a base-2 logarithmic grid for the 10% training set (from 0.25 to 8192 for Gaussian kernel and 0.1 to 16384 for Laplacian kernel). In order to simplify the width screening, prior to learning all feature vectors were normalized (scaling the input vector by the mean norm across the training set) by the Euclidean norm for the Gaussian kernel and the Manhattan norm for the Laplacian kernel.

2.5.2. Bayesian Ridge Regression. We use BR[54] as is implemented in scikit-learn.[53] BR is a linear model with a L_2 penalty on the coefficients. Unlike ridge regression where the strength of that penalty is a regularization hyperparameter which must be set, in Bayesian ridge regression the optimal regularizer is estimated from the data.

2.5.3. Elastic Net. Also EN[55] is a linear model. Unlike BR, the penalty on the weights is a mix of L_1 and L_2 terms. In addition to the regularization hyperparameter for the weight penalty, elastic net has an additional hyperparameter l1_ratio to control the relative strength of the L_1 and L_2 weight penalties. We used the scikit-learn[53] implementation and set l1_ratio = 0.5. We then did a hyperparameter search on regularizing the parameter in a base 10 logarithmic grid from 1e−6 to 1.0.

2.5.4. Random Forest. We use RF[56] as implemented in scikit-learn.[53] RF regressors produce a value by averaging many individual decision trees fitted on randomly resampled sets of the training data. Each node in each decision tree is a threshold of one input feature. Early experiments did not reveal strong differences in performance based on the number of trees used (see Supporting Information for details), once a minimal number was reached. We have used 120 trees for all regressions.

2.5.5. Graph Convolutions. We used the GC model as described in Kearnes et al.,[27] with several structural modifications and optimized hyperparameters. The graph convolution model is built on the concepts of "atom" layers (one real vector associated with each atom) and "pair" layers (one real vector associated with each pair of atoms). The graph

DOI: 10.1021/acs.jctc.7b00577
J. Chem. Theory Comput. 2017, 13, 5255−5264

Table 4. Mean and Mean Absolute Deviation (MAD) for All Properties in the QM9 Data Set, As Well As Target MAE and DFT (at the B3LYP Level of Theory) MAE Relative to the Experiment for Each Property, and the Number of Molecules Used To Estimate the Values[a]

	U_0, eV	ε_{HOMO}, eV	ε_{LUMO}, eV	$\Delta\varepsilon$, eV	μ, Debye	α, Bohr^3	ZPVE, eV	C_v, cal/mol K	ω_1, cm^{-1}
mean	−76.6	−6.54	0.322	6.86	2.67	75.3	4.06	31.6	3500
MAD	8.19	0.439	1.05	1.07	1.17	6.29	0.717	3.21	238
target	0.043	0.043	0.043	0.043	0.10	0.10	0.0012	0.050	10
DFT	0.10(69)	NA	NA	NA	0.10(49)	0.4(49)	0.0097(41)	0.34(16)	28(41)

[a]In parentheses of the DFT row. The target accuracies were taken from ref 13. Target accuracy for energies of atomization and orbital energies were set to 1 kcal/mol, which is generally accepted as (or close to) chemical accuracy within the chemistry community. Target accuracies used for μ and α are 0.1 D and 0.1 Bohr^3, respectively, which is within the error of CCSD relative to experiments.[30] Target accuracies used for ω_1 and ZPVE are 10 cm^{-1}, which is slightly larger than the CCSD(T) error for predicting frequencies.[60] Target accuracies used for C_v were not explained in the article.[13] Section 2.3 discusses how the errors for DFT were obtained.

convolution architecture defines operations to transform atom and pair layers to new atom and pair layers. There are three structural changes to the model used herewithin when compared to the one described in Kearnes et al.[27] We describe these briefly here with details in the Supporting Information. First, we removed the "pair order invariance" property by simplifying the $(A \rightarrow P)$ transformation. Since the model only uses the atom layer for the molecule level features, pair order invariance is not needed. Second, we used the Euclidean distance between atoms. In the $(P \rightarrow A)$ transformation, we divide the value from the convolution step by a series of distance exponentials. If the original convolution for an atom pair (a, b) with distance d produces a vector V, we concatenate the vectors V, $\frac{V}{d^1}$, $\frac{V}{d^2}$, $\frac{V}{d^3}$, and $\frac{V}{d^6}$ to produce the transformed value for the pair (a, b). Third, we followed other work on neural networks based on chemical graphs[57] which uses a sum of softmax operations to convert a real valued vector to a sparse vector and sums those sparse vectors across all the atoms. We use the same operation here along with a simple sum across the atoms to produce molecule level features from the top atom layer. We found that this works as well or better than the Gaussian histograms first used in GC.[27] To optimize the network, we searched the hyperparameter space using Gaussian Process Bandit Optimization[58] as implemented by HyperTune.[59] The hyperparameter search has been based on the evaluation of the validation set for a single fold of the data. Further details including parameters and search ranges chosen for this paper are listed in the Supporting Information.

2.5.6. Gated Graph Neural Networks. We used the GG neural networks model (GG) as described in Li et al.[28] Similar to the GC model, GG is a deep neural network whose input is a set of node features $\{x_v, v \in G\}$ and an adjacency matrix A with entries in a discrete set $S = \{0,1,\cdots,k\}$ to indicate different edge types. It has internal hidden representations for each node in the graph h_v^t of dimension d which it updates for T steps of computation. Its output is invariant to all graph isomorphisms, meaning the order of the nodes presented to the model does not matter. To include the most relevant distance information we distinguished four different covalent bonding types (single, double, triple, aromatic). For all remaining atom pairs we binned them by their interatomic distance [in Å] into 10 bins: [0, 2], [2,2.5], [2.5,3], [3,3.5], [3.5,4], [4,4.5], [4.5,5], [5,5.5], [5.5,6], and [6,∞]. Using these bins, the adjacency matrix was assigned entries with an alphabet of size 14 ($k = 14$), indicating bond type for covalently bonded atoms, and distance bin for all other atoms. We trained the GG model on each target property individually. Further technical details are specified in the Supporting Information.

3. RESULTS AND DISCUSSION

3.1. Overview. We present an overview of the most relevant numerical results in Table 3, which contains the test errors for all combinations of regressors and representations and properties for models trained on ∼118 k molecules. The best models for the respective properties are U_0 (HDAD/KRR), ε_{HOMO} (MG/GC), ε_{LUMO} (MG/GC), $\Delta\varepsilon$ (MG/GC), μ (MG/GC), α (MG/GG), ZPVE (HDAD/KRR), C_v (HDAD/KRR), and ω_1 (BAML/RF). We do not show results for the other three energies, $U(T = 298\ K)$, $H(T = 298\ K)$, $G(T = 298\ K)$, since identical observations as for U_0 can be made.

Overall, NN and KRR regressors perform well for most properties. The ML out-of-sample errors outperform DFT accuracy at the B3LYP level of theory and reach chemical (target) accuracy, both defined in Table 4, for U_0 (HDAD/KRR and MG/GG), μ (GC), C_v (HDAD/KRR), and ω_1 (BAML/KRR, MG/GC, HDAD/KRR, BOB/KRR, HD/KRR, and MG/GG). For the remaining properties (ε_{HOMO}, ε_{LUMO}, $\Delta\epsilon$, α, and ZPVE) the best models come within a factor 2 of target accuracy, while all (except ε_{HOMO}, ε_{LUMO}, and $\Delta\epsilon$, where we do not have reliable data) outperform DFT accuracy.

In Figure 2, out-of-sample errors as a function of training set size (learning curves) are shown for all properties and representations, along with the best corresponding regressor. It is important to note that *all* models on display systematically improve with training set size, exhibiting the typical linearly decaying behavior on a log−log plot.[11,61] Errors for most models decayed with roughly the same slopes, indicating similar exponents in the power-law of error decay. Notable exceptions, i.e. property models with considerably steeper learning curves (slopes and offsets of all learning curves can be found in Tables S4 and S5 in the Supporting Information), are MG/GC for μ, MG/GG and HDAD/KRR for α, HDAD/KRR and MG/GG for U_0, and MG/GG for ω_1. These results suggest that the specified representations capture particularly well the effective dimensionality of the corresponding property in chemical space.

3.2. Regressors. Inspection of Table 3 indicates that the regressors can roughly be ordered by performance, independent of property and representation: GC > GG > KRR > RF > BR > EN. It is noteworthy how EN, BR, and RF regressors perform substantially worse than GC/GG/KRR. The bad performance of EN and BR is due to their low order. This can also be seen from the learning curves of all regressors presented in Figures S1−S6 of the Supporting Information. The performance of BR and EN improves only slightly with increased training set size and even deteriorated for some property/representation combinations. These two regressors

DOI: 10.1021/acs.jctc.7b00577
J. Chem. Theory Comput. 2017, 13, 5255−5264

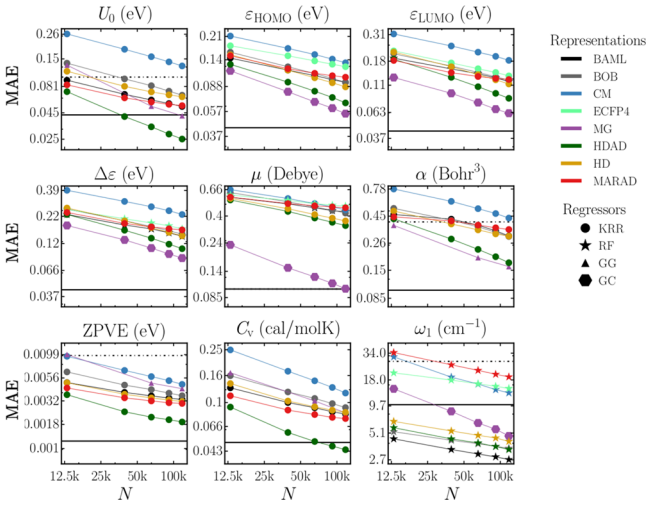


Figure 2. Learning curves (mean absolute error (MAE) as a function of training set size N) for 10 properties of QM9 molecules, displaying the best regressor for each representation and property. Horizontal solid lines correspond to target accuracies, and vertical dotted lines correspond to approximated B3LYP accuracies (unless off-chart), see also Table 3. Note that due to its poor performance ECFP4 results have been excluded for α, ZPVE, U, and C_v.

also exhibit very similar learning curves, and BR performs only slightly better than EN for most combinations. The only clear exception to this rule is for ZPVE and U_0 together with HDAD, where BR performs significantly better than EN. Also, BR and EN errors rapidly converge to a constant w.r.t. training set size for all representations and properties, except for HDAD, which is the only representation which has a noteworthy improvement with increased training set size for some properties. The constant learning rates are not surprising as (a) the number of free regression parameters in BR and EN is relatively small and does not grow with training set size and as (b) the underlying model is a linear combination with small flexibility. This behavior implies error convergence already for relatively small training sets.

RF performs poorly compared to GC, GG, and KRR for all properties except for ω_1, the highest lying fundamental vibrational frequency in each molecule. For this property RF yields an astounding performance with out-of-sample errors as small as single digit cm^{-1}. B3LYP achieves a mean absolute error of only 28 cm^{-1} with respect to experiment.[32] The distribution of ω_1, Figure 1 of ref 13, suggests a simple reason for this: There are three distinct peaks which correspond to typical C−H, N−H, and O−H stretch vibrations in increasing order. Therefore, the principal learning task in this property is to detect if there is an OH group, and if not if there is an NH group. If neither group is present, CH will yield the vibration with the highest frequency. As such, this is essentially about classifying which bonds are present in the molecule. RF works by fitting a decision tree to the target property. Each branch in the tree is based on an inequality of *one* entry in the representation. RF should therefore be able to identify which bonds are present in a molecule, simply by looking at the entries in the each element pair and/or triplet bin of the representations. For RF, a fractional importance can be assigned to each input feature (the sum of all importances is 1.0). Analyzing the importance of the bins in HDAD of the RF model reveals that the three bins with highest importance are O−H placed at 0.961 Å, N−H placed at at 1.01 Å, and C−C−H at 3.138 radians with an importance of 0.587, 0.305, and 0.064 ,respectively. These three first bins constitute ∼96% of the prediction of ω_1, and distances of the O−H and N−H bins are very similar to O−H and N−H bond lengths. C−C−H is placed on $\sim \pi$ radians which means that it has to correspond to a linear C−C−H (alkyne) chain which implies that the two carbons must be bonded by a triple bond (typically the C−H with the lowest pK_a and the highest C−H stretch vibration).

KRR performs remarkably well on average. For extensive energetic properties it yields the lowest overall errors in combination with HDAD and BOB. Its outstanding performance is not unsurprising in light of the multiple previous ML studies dealing with compositional as well as configurational spaces. The neural network flavors GC and GG, however, yield better performance on average, and the lowest errors for all electronic (mostly intensive) properties, i.e. dipole moment, polarizability, HOMO/LUMO eigenvalues and gaps. A possible explanation for this property dependent difference in performance between KRR and NN could be the inherent respective additive and multiplicative nature of these regressors. The energy being extensive, it is consistent with this picture that effective, quasi-particle based linear KRR based estimates have recently been reported to deliver very accurate predictions which can scale.[62]

3.3. Representations. As one would expect, HDAD contains more relevant information and thus it always outperforms HD when using KRR. Tests also showed that an HDA representation systematically yields errors in between HDAD and HD, and similar observations hold for BR and EN regressors. In the case of RF, however, we observe little difference between HDAD and HD, and HD can even yield slightly lower errors than HDAD. In our opinion, this is due to the additional bins of angles and dihedrals rather adding noise than signal. By contrast, the separation of distances, angles, and dihedral angles into different bins is not a problem for the KRR methods because the kernels used are purely distance based. This makes it possible for KRR to exploit the extra three- and four-body information in HDAD and to gain an advantage over HD. We note however that the remarkable performance of HDAD is possible despite its striking simplicity. As illustrated in Figure 1 and discussed above, characteristic chemical behavior can be directly obtained by human inspection of HDAD. As such, HDAD corresponds to a representation very much in the style of Occam's razor. Unfortunately, due to its discrete nature and its origin in sorting distances, HDAD will suffer from lack of differentiability, which might limit its applicability when modeling forces or other nonequilibrium properties.

MARAD, containing similar information as HDA, performs similarly to BAML — yet MARAD requires no prior knowledge about the physics encoded in the universal force field such as electronic hybridization states, bond order, or functional potential shapes (Morse, Lennard-Jones, harmonic angular potentials, or sinusoidal dihedrals). BOB and CM, previously state of the art, result in relatively poor performance. ECFP4 produces out-of-sample errors on par or slightly better than CM/KRR for intensive properties (μ, HOMO/LUMO eigenvalues and gap); however, the model produces errors that are off-the-chart for all extensive properties (α, ZPVE, U_0, and C_V).

DOI: 10.1021/acs.jctc.7b00577
J. Chem. Theory Comput. 2017, 13, 5255−5264

4. CONCLUSIONS

We benchmarked many combinations of regressors and representations on the same QM9 data set consisting of ∼131k organic molecules. For all properties, the best ML model prediction errors reach the accuracy of DFT at the B3LYP level with respect to experiment. For 7 out of 12 distinct properties (atomization energies, heat-capacity, ω_1, μ) out-of-sample errors reach levels on par with chemical accuracy, or better, using a training set size of ∼118k (90% of QM9 data set) molecules. For the remaining properties α, ε_{HOMO}, ε_{LUMO}, $\Delta\epsilon$, and ZPVE, errors of the best models come within a factor 2 of chemical accuracy.

Regressors EN, BR, and RF lead to rather high out-of-sample errors, while KRR and graph based NN regressors compete for the lowest errors. We have found that GC, GG, and KRR have the best performance across *all* properties, except for the highest vibrational frequency for which RF performs best. There is no single representation and regressor combination that works best for all properties (though forthcoming work with further improvements to the GG based models indicates best in class performance across all properties[63]). For intensive electronic properties (μ, HOMO/LUMO eigenvalues, and gap) we found MG/GC or MG/GG to yield the highest predictive power, while HDAD/KRR corresponds to the most accurate model for extensive properties (α, ZPVE, U_0, and C_V). The latter point is remarkable when considering the simplicity of KRR, being just a linear expansion in the feature space, and HDAD, being just histograms of distances, angles, and dihedrals. Using BR and EN is not recommended if accuracy is of importance, and both regressors perform worse across all properties. Apart from predicting highest fundamental vibrational frequency with the highest accuracy, RF-based models deliver rather unsatisfactory performance. The ECPF4-based models have shown poor general performance in comparison to all other representations studied; it is not recommended for investigations of molecular properties.

We should caution the reader that all our results refer to equilibrium structures of a set of only ∼131 k organic molecules. While ∼131k molecules might seem sufficiently large to be representative, this number is dwarfed by chemical space, i.e. the space populated by all theoretically stable molecules, estimated to exceed 10^{60} for medium sized organic molecules.[64] Furthermore, ML models for predicting properties of molecules in nonequilibrium or strained configurations might require substantially more training data. This point is also of relevance because some of the highly accurate models described herewithin (MG based) currently use bond-based graph connectivity in addition to distance, raising questions about the applicability to reactive processes.

In summary, for the organic molecules studied, we collected numerical evidence which suggesting that the out-of-sample error of ML is consistently better than the estimation error of DFT at the B3LYP level of accuracy i.e. ML models trained on DFT training data will predict true chemical values with an error no worse than twice that of DFT when trained on DFT data, and with a prediction error that may be lower than DFT error if trained on higher levels of theory, or on experimental data. While this is no guarantee that ML models would reach the same error levels if more accurate, explicitly electron correlated or experimental reference data were used, previous studies indicate that a similar performance can be expected when using higher levels of theory.[8] More specifically, one might intuitively expect that going beyond hybrid DFT to higher quality data (either wave function based-QM or experiment) in terms of reference methods would represent a more challenging learning problem and therefore imply the need for larger training set sizes. Results in ref 8, however, suggest that ML models can predict the differences between HF and MP2, CCSD, and CCSD(T) equally well using the same training set.

As such, we conclude that future reference data sets for training state-of-the-art machine learning models of molecular properties should preferably use reference levels of theory which go beyond DFT at the B3LYP level of accuracy. While it seems unlikely that for each class of molecules, hundreds of thousands of experimental training data points will become available in the foreseeable future, it might well be possible to reach such a scale using efficient implementations of explicit electron correlated methods within high-performance computing campaigns. Finally, we note that future work could deal with improving representations and regressors, with the goal of reaching similar predictive power using less data.

■ ASSOCIATED CONTENT

Supporting Information

The Supporting Information is available free of charge on the ACS Publications website at DOI: 10.1021/acs.jctc.7b00577.

Raw data, MARAD representation, graph convolutions, gated graphs, random forests, and learning curves, as well as root-mean-square errors for ML predictions after training on the largest training set (PDF)

■ AUTHOR INFORMATION

Corresponding Author

*E-mail: anatole.vonlilienfeld@unibas.ch.

ORCID

O. Anatole von Lilienfeld: 0000-0001-7419-0466

Author Contributions

F.A.F. and L.H. contributed equally.

Funding

O.A.v.L. acknowledges support from the Swiss National Science foundation (No. PP00P2_138932, 310030_160067), the research fund of the University of Basel, and from Google. This material is based upon work supported by the Air Force Office of Scientific Research, Air Force Material Command, USAF under Award No. FA9550-15-1-0026. This research was partly supported by the NCCR MARVEL, funded by the Swiss National Science Foundation. Some calculations were performed at sciCORE (http://scicore.unibas.ch/) scientific computing core facility at University of Basel.

Notes

The authors declare no competing financial interest.

■ ACKNOWLEDGMENTS

The authors thank Dirk Bakowies for helpful comments and Adrian Roitberg for pointing out an issue with the use of partial charges in the neural net models in an earlier version of this paper.

■ REFERENCES

(1) Hohenberg, P.; Kohn, W. Inhomogeneous Electron Gas. *Phys. Rev.* **1964**, *136*, B864.

DOI: 10.1021/acs.jctc.7b00577
J. Chem. Theory Comput. 2017, 13, 5255−5264

(2) Kohn, W.; Sham, L. J. Self-Consistent Equations Including Exchange and Correlation Effects. *Phys. Rev.* **1965**, *140*, A1133.
(3) Burke, K. Perspective on density functional theory. *J. Chem. Phys.* **2012**, *136*, 150901.
(4) Koch, W.; Holthausen, M. C. *A Chemist's Guide to Density Functional Theory*; Wiley-VCH, 2002; DOI: 10.1002/3527600043.
(5) Cohen, A. J.; Mori-Sánchez, P.; Yang, W. Challenges for Density Functional Theory. *Chem. Rev.* **2012**, *112*, 289−320.
(6) Plata, R. E.; Singleton, D. A. A Case Study of the Mechanism of Alcohol-Mediated Morita Baylis-Hillman Reactions. The Importance of Experimental Observations. *J. Am. Chem. Soc.* **2015**, *137*, 3811−3826.
(7) Medvedev, M. G.; Bushmarinov, I. S.; Sun, J.; Perdew, J. P.; Lyssenko, K. A. Density functional theory is straying from the path toward the exact functional. *Science* **2017**, *355*, 49−52.
(8) Ramakrishnan, R.; Dral, P. O.; Rupp, M.; von Lilienfeld, O. A. Big Data Meets Quantum Chemistry Approximations: The Δ-Machine Learning Approach. *J. Chem. Theory Comput.* **2015**, *11*, 2087−2096.
(9) Ramakrishnan, R.; Dral, P. O.; Rupp, M.; von Lilienfeld, O. A. Quantum chemistry structures and properties of 134 kilo molecules. *Sci. Data* **2014**, *1*, 140022.
(10) Ruddigkeit, L.; van Deursen, R.; Blum, L. C.; Reymond, J.-L. Enumeration of 166 Billion Organic Small Molecules in the Chemical Universe Database GDB-17. *J. Chem. Inf. Model.* **2012**, *52*, 2864−2875.
(11) Huang, B.; von Lilienfeld, O. A. Communication: Understanding molecular representations in machine learning: The role of uniqueness and target similarity. *J. Chem. Phys.* **2016**, *145*, 161102.
(12) Hansen, K.; Biegler, F.; Ramakrishnan, R.; Pronobis, W.; von Lilienfeld, O. A.; Müller, K.-R.; Tkatchenko, A. Machine Learning Predictions of Molecular Properties: Accurate Many-Body Potentials and Nonlocality in Chemical Space. *J. Phys. Chem. Lett.* **2015**, *6*, 2326−2331.
(13) Ramakrishnan, R.; von Lilienfeld, O. A. Many Molecular Properties from One Kernel in Chemical Space. *Chimia* **2015**, *69*, 182−186.
(14) Rupp, M.; Tkatchenko, A.; Müller, K.-R.; von Lilienfeld, O. A. Fast and Accurate Modeling of Molecular Atomization Energies with Machine Learning. *Phys. Rev. Lett.* **2012**, *108*, 058301.
(15) Barker, J.; Bulin, J.; Hamaekers, J.; Mathias, S. *Localized Coulomb Descriptors for the Gaussian Approximation Potential*. arXiv preprint arXiv:1611.05126, 2016.
(16) von Lilienfeld, O. A.; Ramakrishnan, R.; Rupp, M.; Knoll, A. Fourier series of atomic radial distribution functions: A molecular fingerprint for machine learning models of quantum chemical properties. *Int. J. Quantum Chem.* **2015**, *115*, 1084−1093.
(17) Huan, T. D.; Mannodi-Kanakkithodi, A.; Ramprasad, R. Accelerated materials property predictions and design using motif-based fingerprints. *Phys. Rev. B: Condens. Matter Mater. Phys.* **2015**, *92*, 014106.
(18) Bartók, A. P.; Kondor, R.; Csányi, G. On representing chemical environments. *Phys. Rev. B: Condens. Matter Mater. Phys.* **2013**, *87*, 184115.
(19) De, S.; Bartók, A. P.; Csányi, G.; Ceriotti, M. Comparing molecules and solids across structural and alchemical space. *Phys. Chem. Chem. Phys.* **2016**, *18*, 13754−13769.
(20) Collins, C. R.; Gordon, G. J.; von Lilienfeld, O. A.; Yaron, D. J. *Constant Size Molecular Descriptors For Use With Machine Learning*. arXiv preprint arXiv:1701.06649, 2016.
(21) Smith, J. S.; Isayev, O.; Roitberg, A. E. ANI-1: An extensible neural network potential with DFT accuracy at force field computational cost. *Chem. Sci.* **2017**, *8*, 3192−3203.
(22) Schütt, K. T.; Arbabzadah, F.; Chmiela, S.; Müller, K. R.; Tkatchenko, A. Quantum-chemical insights from deep tensor neural networks. *Nat. Commun.* **2017**, *8*, 13890.
(23) Dral, P. O.; von Lilienfeld, O. A.; Thiel, W. Machine Learning of Parameters for Accurate Semiempirical Quantum Chemical Calculations. *J. Chem. Theory Comput.* **2015**, *11*, 2120−2125.
(24) Weber, W.; Thiel, W. Orthogonalization corrections for semiempirical methods. *Theor. Chem. Acc.* **2000**, *103*, 495−506.
(25) Dral, P. O.; Wu, X.; Spörkel, L.; Koslowski, A.; Thiel, W. Semiempirical Quantum-Chemical Orthogonalization- Corrected Methods: Benchmarks for Ground-State Properties. *J. Chem. Theory Comput.* **2016**, *12*, 1097−1120.
(26) Hansen, K.; Montavon, G.; Biegler, F.; Fazli, S.; Rupp, M.; Scheffler, M.; von Lilienfeld, O. A.; Tkatchenko, A.; Müller, K.-R. Assessment and Validation of Machine Learning Methods for Predicting Molecular Atomization Energies. *J. Chem. Theory Comput.* **2013**, *9*, 3404−3419.
(27) Kearnes, S.; McCloskey, K.; Berndl, M.; Pande, V.; Riley, P. Molecular Graph Convolutions: Moving Beyond Fingerprints. *J. Comput.-Aided Mol. Des.* **2016**, *30*, 595−608.
(28) Li, Y.; Tarlow, D.; Brockschmidt, M.; Zemel, R. Gated Graph Sequence Neural Networks. *Proceeding of ICLR'16*; 2016.
(29) Curtiss, L. A.; Raghavachari, K.; Redfern, P. C.; Pople, J. A. Assessment of Gaussian-2 and density functional theories for the computation of enthalpies of formation. *J. Chem. Phys.* **1997**, *106*, 1063−1079.
(30) Hickey, A. L.; Rowley, C. N. Benchmarking quantum chemical methods for the calculation of molecular dipole moments and polarizabilities. *J. Phys. Chem. A* **2014**, *118*, 3678−3687.
(31) Stowasser, R.; Hoffmann, R. What do the Kohn-Sham orbitals and eigenvalues mean? *J. Am. Chem. Soc.* **1999**, *121*, 3414.
(32) Sinha, P.; Boesch, S. E.; Gu, C.; Wheeler, R. A.; Wilson, A. K. Harmonic Vibrational Frequencies: Scaling Factors for HF, B3LYP, and MP2Methods in Combination with Correlation Consistent Basis Sets. *J. Phys. Chem. A* **2004**, *108*, 9213−9217.
(33) Geary, R. C. The Ratio of the Mean Deviation to the Standard Deviation as a Test of Normality. *Biometrika* **1935**, *27*, 310−332.
(34) DeTar, D. F. Calculation of Entropy and Heat Capacity of Organic Compounds in the Gas Phase. Evaluation of a Consistent Method without Adjustable Parameters. Applications to Hydrocarbons. *J. Phys. Chem. A* **2007**, *111*, 4464−4477.
(35) Huo, H.; Rupp, M. *Unified Representation for Machine Learning of Molecules and Crystals*. arXiv preprint arXiv:1704.06439, 2017.
(36) Bartok, A. P.; De, S.; Poelking, C.; Bernstein, N.; Kermode, J.; Csanyi, G.; Ceriotti, M. *Machine Learning Unifies the Modelling of Materials and Molecules*. arXiv preprint arXiv:1706.00179, 2017.
(37) Rappe, A. K.; Casewit, C. J.; Colwell, K. S., III; Goddard, W. A.; Skiff, W. M. UFF, a full periodic table force field for molecular mechanics and molecular dynamics simulations. *J. Am. Chem. Soc.* **1992**, *114*, 10024−10035.
(38) Rogers, D.; Hahn, M. Extended-connectivity fingerprints. *J. Chem. Inf. Model.* **2010**, *50*, 742−754.
(39) Besnard, J.; Ruda, G. F.; Setola, V.; Abecassis, K.; Rodriguiz, R. M.; Huang, X.-P.; Norval, S.; Sassano, M. F.; Shin, A. I.; Webster, L. A.; Simeons, F. R. C.; Stojanovski, L.; Prat, A.; Seidah, N. G.; Constam, D. B.; Bickerton, G. R.; Read, K. D.; Wetsel, W. C.; Gilbert, I. H.; Roth, B. L.; Hopkins, A. L. Automated design of ligands to polypharmacological profiles. *Nature* **2012**, *492*, 215−220.
(40) Lounkine, E.; Keiser, M. J.; Whitebread, S.; Mikhailov, D.; Hamon, J.; Jenkins, J. L.; Lavan, P.; Weber, E.; Doak, A. K.; Côté, S.; Shoichet, B. K.; Urban, L. Large-scale prediction and testing of drug activity on side-effect targets. *Nature* **2012**, *486*, 361−367.
(41) Huigens, R. W., III; Morrison, K. C.; Hicklin, R. W.; Flood, T. A., Jr; Richter, M. F.; Hergenrother, P. J. A Ring Distortion Strategy to Construct Stereochemically Complex and Structurally Diverse Compounds from Natural Products. *Nat. Chem.* **2013**, *5*, 195.
(42) Todeschini, R.; Consonni, V. *Handbook of Molecular Descriptors*; Wiley-VCH: Weinheim, 2009; DOI: 10.1002/9783527613106.
(43) Faulon, J.-L.; Visco, D. P., Jr; Pophale, R. S. The Signature Molecular Descriptor. 1. Using Extended Valence Sequences in QSAR and QSPR Studies. *J. Chem. Inf. Comp. Sci.* **2003**, *43*, 707.
(44) Visco, J.; Pophale, R. S.; Rintoul, M. D.; Faulon, J. L. Developing a methodology for an inverse quantitative structure activity relationship using the signature molecular descriptor. *J. Mol. Graphics Modell.* **2002**, *20*, 429−438.

DOI: 10.1021/acs.jctc.7b00577
J. Chem. Theory Comput. 2017, 13, 5255−5264

(45) O'Boyle, N. M.; Banck, M.; James, C. A.; Morley, C.; Vandermeersch, T.; Hutchison, G. R. Open Babel: An open chemical toolbox. *J. Cheminform.* **2011**, *3*, 33.
(46) Landrum, G. *RDKit: Open-source cheminformatics software*; 2014; Vol. 3, p 2012. http://www.rdkit.org (accessed Sept 26, 2017).
(47) Faber, F. A.; Hutchison, L.; Huang, B.; Gilmer, J.; Schoenholz, S. S.; Dahl, G. E.; Vinyals, O.; Kearnes, S.; Riley, P. F.; von Lilienfeld, O. A. *Fast machine learning models of electronic and energetic properties consistently reach approximation errors better than DFT accuracy*. arXiv preprint arXiv:1702.05532, 2017.
(48) Müller, K.-R.; Mika, S.; Rätsch, G.; Tsuda, K.; Schölkopf, B. An introduction to kernel-based learning algorithms. *IEEE transactions on neural networks* **2001**, *12*, 181−201.
(49) Schölkopf, B.; Smola, A. J. *Learning with kernels: support vector machines, regularization, optimization, and beyond*; MIT Press: 2002.
(50) Vovk, V. Kernel Ridge Regression. In *Empirical Inference: Festschrift in Honor of Vladimir N. Vapnik*; Schölkopf, B., Luo, Z., Vovk, V., Eds.; Springer Berlin Heidelberg: Berlin, Heidelberg, 2013; pp 105−116; DOI: 10.1007/978-3-642-41136-6_11.
(51) Hastie, T.; Tibshirani, R.; Friedman, J. *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*, 2nd ed.; Springer: New York, 2011; DOI: 10.1007/978-0-387-84858-7.
(52) Hoerl, A. E.; Kennard, R. W. Ridge Regression Biased Estimation for Nonorthogonal Problems. *Technometrics* **2000**, *42*, 80−86.
(53) Pedregosa, F.; Varoquaux, G.; Gramfort, A.; Michel, V.; Thirion, B.; Grisel, O.; Blondel, M.; Prettenhofer, P.; Weiss, R.; Dubourg, V.; Vanderplas, J.; Passos, A.; Cournapeau, D.; Brucher, M.; Perrot, M.; Duchesnay, E. scikit-learn: Machine Learning in Python. *J. Mach. Learn. Res.* **2011**, *12*, 2825−2830.
(54) Neal, R. M. *Bayesian Learning for Neural Networks*; Springer-Verlag New York, Inc.: Secaucus, NJ, USA, 1996; DOI: 10.1007/978-1-4612-0745-0.
(55) Zou, H.; Hastie, T. Regularization and variable selection via the elastic net. *J. R. Stat. Soc. Series. B Stat. Methodol.* **2005**, *67*, 301−320.
(56) Breiman, L. Random forests. *Machine learning* **2001**, *45*, 5−32.
(57) Duvenaud, D. K.; Maclaurin, D.; Iparraguirre, J.; Bombarell, R.; Hirzel, T.; Aspuru-Guzik, A.; Adams, R. P. Convolutional Networks on Graphs for Learning Molecular Fingerprints. *Advances in Neural Information Processing Systems*; 2015; pp 2215−2223.
(58) Desautels, T.; Krause, A.; Burdick, J. W. Parallelizing Exploration-Exploitation Tradeoffs in Gaussian Process Bandit Optimization. *J. Mach. Learn. Res.* **2014**, *15*, 4053−4103.
(59) Google HyperTune. https://cloud.google.com/ml/ (accessed 2016).
(60) Tew, D. P.; Klopper, W.; Heckert, M.; Gauss, J. Basis Set Limit CCSD(T) Harmonic Vibrational Frequencies. *J. Phys. Chem. A* **2007**, *111*, 11242−11248.
(61) Müller, K.-R.; Finke, M.; Murata, N.; Schulten, K.; Amari, S. A numerical study on learning curves in stochastic multilayer feedforward networks. *Neural Comput.* **1996**, *8*, 1085−1106.
(62) Huang, B.; von Lilienfeld, O. A. *The "DNA" of chemistry: Scalable quantum machine learning with "amons*. arXiv preprint arXiv:1707.04146, 2017.
(63) Gilmer, J.; Schoenholz, S. S.; Riley, P. F.; Vinyals, O.; Dahl, G. E. *Neural Message Passing for Quantum Chemistry*. Proceedings of the 34nd International Conference on Machine Learning, ICML 2017, 2017.
(64) Kirkpatrick, P.; Ellis, C. Chemical space. *Nature* **2004**, *432*, 823.

DOI: 10.1021/acs.jctc.7b00577
J. Chem. Theory Comput. 2017, 13, 5255−5264