

This document is confidential and is proprietary to the American Chemical Society and its authors. Do not copy or disclose without written permission. If you have received this item in error, notify the sender and delete all copies.

**Prediction errors of molecular machine learning models
lower than hybrid DFT error**

Journal:	<i>Journal of Chemical Theory and Computation</i>
Manuscript ID	ct-2017-00577j.R1
Manuscript Type:	Article
Date Submitted by the Author:	19-Sep-2017
Complete List of Authors:	Faber, Felix; University of Basel Hutchison, Luke; Google Huang, Bing; University of Basel Gilmer, Justin; Google Schoenholz, Samuel; Google Dahl, George; Google Vinyals, Oriol; Google Kearnes, Steven; Google Riley, Patrick; Google von Lilienfeld, O. Anatole; University of Basel

SCHOLARONE™
Manuscripts

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, CA, US - 94043 CA*

[¶]*Google, 5 New Street Square, London EC4A 3TW, UK*

[§]*Authors contributed equally*

E-mail: anatole.vonlilienfeld@unibas.ch

Abstract

We investigate the impact of choosing regressors and molecular representations for the construction of fast machine learning (ML) models of thirteen 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 ~ 118 k distinct molecules. Molecular structures and properties at hybrid density functional theory (DFT) level of theory come from the QM9 database [Ramakrishnan et al, *Scientific Data* **1** 140022 (2014)] 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), and angles (HDA/MARAD), and dihedrals (HDAD). Regressors include linear models (Bayesian ridge regression (BR)) and linear regression with elastic net regu-

larization (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 ~ 118 k 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 was 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 chemistry³—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 lot 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 ~ 134 k 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 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 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 OM2²⁴ semiempirical 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 ~ 118 k 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 Kelvin (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 ~ 118 k

(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 B3LYP level of theory relative to experiment for all properties, and that chemical accuracy (See table 3) 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 III, we present the results and discuss them, and section IV concludes the paper.

2 Method

2.1 Data set

We have used the QM9 data set consisting of ~ 134 k 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 DFT level of theory (B3LYP/6-31G(2df,p)), were included. We used: Atomization energy at 0 Kelvin 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 $\varepsilon_{\text{HOMO}}$ (eV); LUMO eigenvalue $\varepsilon_{\text{LUMO}}$ (eV); HOMO-LUMO gap $\Delta\varepsilon$ (eV); norm of dipole moment $\mu = \sqrt{\sum_{r \in x,y,z} (\int dr n(\mathbf{r}) r)^2}$ (Debye), where $n(\mathbf{r})$ is the molecular charge density distribution; static isotropic polarizability $\alpha = \frac{1}{3} \sum_{i \in x,y,z} \alpha_{ii}$ (Bohr³), 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 ~ 131 k molecules.

2.2 Model validation

Starting from the ~ 131 k molecules in QM9 after removing the ~ 3 k molecules (see above) we have created a number of train-validation-test splits. We have splitted the data set into test and non-test sets and varied the percentage of data in test set to explore the effect of amount of data in error rates. Inside the non-test set, we have performed 10 fold cross validation for hyperparameter optimization. That is, for each model 90% (the training set) of the non-test set is used for training and 10% (the validation set) is used for hyperparameter selection. For each test/non-test split, we have trained 10 models on different subsets of the non-test set, and we report the mean error on the test set across those 10 models. Note that the non-test set will be referred to as 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 B3LYP level of theory* to experiments for the various properties we study. Where data is 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 C₇H₁₀O₂.⁸ 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

1 to experimental values from 69 small organic
2 molecules (of which 47 were substituted with
3 F, Cl, and S), with up to 6 heavy atoms (not
4 counting hydrogens), and calculated the ener-
5 gies using B3LYP/6-311+G(3df,2p). The re-
6 sulting mean absolute deviation from experi-
7 mental values was 2.3 kcal/mol.

8
9 Rough hybrid DFT error estimates for dipole
10 moment and polarizability have been obtained
11 from Refs.³⁰. The errors are estimated refer-
12 enced to experimental values, for a data set
13 consisting of 49 molecules with up to 7 heavy atoms
14 (C, Cl, F, H, O, P, or S)

15
16 Frontier molecular orbital energies (HOMO,
17 LUMO and HOMO-LUMO gap) can not be
18 measured directly. However, for the exact (yet
19 unknown) exchange-correlation potential, the
20 Kohn-Sham HOMO eigenvalues correspond to
21 the negative of the vertical ionization poten-
22 tial (IP).³¹ Unfortunately, within hybrid DFT,
23 the precise meaning of the frontier eigenvalues
24 and the gap is less clear, and we therefore re-
25 frain from a direct comparison of B3LYP to
26 experimental numbers. Nevertheless, we have
27 included eigenvalues and the gap due to their
28 widespread use for molecular and materials de-
29 sign applications.

30
31 Hybrid DFT RMSE estimates with respect to
32 experimental values of ZPVE and ω_1 (the high-
33 est fundamental vibrational frequency) were
34 published in Ref.³² for a set of 41 organic
35 molecules, with up to 6 heavy atoms (not count-
36 ing hydrogen) and calculated using B3LYP/cc-
37 pVTZ.

38
39 Normally distributed data has a constant
40 ratio between RMSE and MAE,³³ which is
41 roughly 0.8. We have used this ratio to ap-
42 proximate the MAE from the RMSE estimates
43 reported for ZPVE and ω_1 . Deviation of DFT
44 (at the B3LYP/6-311g** level of theory) from
45 experimental heat capacities were reported by
46 DeTar³⁴ who obtained errors of 16 organic
47 molecules, with up to 8 heavy atoms (not count-
48 ing hydrogens).

49
50 Note, however, that one should be cautious
51 when referring to these errors: Strictly speak-
52 ing they can not be compared since different ba-
53 sis sets, molecules, and experiments were used.
54 We also note that all DFT errors in this pa-

per 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 chem-informatics and
materials informatics, and many interesting and
promising variants have already been proposed.
Below, we provide the details on the represen-
tations selected for this study. While finalizing
our study, competitive alternatives were intro-
duced^{35,36} but have been tested only for ener-
gies (and polarizabilities).

2.4.1 CM and BOB

The Coulomb matrix (CM) representation¹⁴ is
a square atom by atom matrix, where off diago-
nal elements are the Coulomb nuclear repulsion
terms between atom pairs. The diagonal ele-
ments approximate the electronic potential en-
ergy 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 rep-
resentation uses exclusively CM elements, group-
ing 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, an-
gles, machine learning) representation can
be viewed as a many-body extension of
BOB.¹¹ All pairwise nuclear repulsions are re-
placed by Morse/Lennard-Jones potentials for
bonded/non-bonded atoms respectively. Fur-
thermore, 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 Fingerprints³⁸ (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 descriptors⁴² (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 under in the Supplementary materials.

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 a Hydrogen atom, a_k 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 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 atom respectively form two triangular faces when

1 paired with V_{ij} : $\langle a_k, a_i, a_j \rangle$ and $\langle a_l, a_i, a_j \rangle$. The
 2 dihedral angle between the two triangular faces
 3 was calculated. These dihedral angle features
 4 were labeled with the atomic symbol for a_i , fol-
 5 lowed by the atomic symbols for a_j , a_k and a_l ,
 6 sorted alphabetically, with the exception that
 7 hydrogens were listed last, e.g. C-C-N-H.
 8

9 The features from all molecules have been ag-
 10 gregated for each label type to generate a his-
 11 tograms for each label type. Fig. 1 exemplifies
 12 this for C-N distances, C-C-C angles, and C-
 13 C-C-O dihedrals for the entire QM9 data set.
 14 Certain typical molecular characteristics can be
 15 recognized upon mere inspection. For example,
 16 the CN histogram displays a strong and iso-
 17 lated peak between 1.1 and 1.5 Å, correspond-
 18 ing to occurrences of single, double, and triple
 19 bonds. For distances above 2 Å, peaks at typ-
 20 ical radii of second and third covalent bonding
 21 shells around N can be recognized at 2.6 Å and
 22 3.9 Å. Also C-C-C angles can be easily inter-
 23 preted: The peak close to zero and π Rad cor-
 24 responds to geometries where three atoms are
 25 part of a linear (alkyne, or nitrile) type of mo-
 26 tif. The broad and largest peak corresponds to
 27 120 and 109 degrees, typically observed in sp^2
 28 and sp^3 hybridized atoms.
 29

30 The morphology of each histogram has then
 31 been examined to identify apparent peaks
 32 and troughs, motivated by the idea that
 33 peaks indicate structural commonalities among
 34 molecules. Bin centers have been placed at
 35 each significant local minimum and maximum
 36 (Shown as vertical lines in Fig. 1). Values at
 37 15-25 bin centers have been chosen as a rep-
 38 resentation for each label type. All bin center
 39 values are provided in the Supplementary Ma-
 40 terial. For each molecule, the collection of
 41 features has subsequently been rendered into a
 42 fixed-size representation, producing one vector
 43 component for each bin center, within each la-
 44 bel type. This has been accomplished following
 45 a two-step process. (i) *Binning and interpola-*
 46 *tion*: Each feature value is projected on the two
 47 nearest bins. The relative amount projected on
 48 each bin uses linear projection between the two
 49 bins. For example: A feature with value 1.7
 50 which lies between two bins placed at 1.0 and
 51 2.0 respectively, contributes 0.3 and 0.7 to the
 52
 53
 54
 55
 56
 57
 58
 59
 60

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 have 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 specifi-
 cally, 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.

The featurization process was unsuccessful
 for a small number of molecules (367) because
 of conversion failures from geometry to ratio-
 nal SMILES string when using OpenBabel⁴⁵ or
 RDKit,⁴⁶ and were excluded from all results us-
 ing the molecule graph features.

Table 1: Atom features for the MG represen-
 tation: Values provided for each atom in the
 molecule.

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.

Note that within a previous draft of this
 study,⁴⁷ we reported biased results for GC/GG
 models due to use of Mulliken partial charges
 within the MG representation. All MG re-
 sults presented herewithin have been obtained
 without any Mulliken charges in the represen-
 tation. Model hyper parameters for the GC

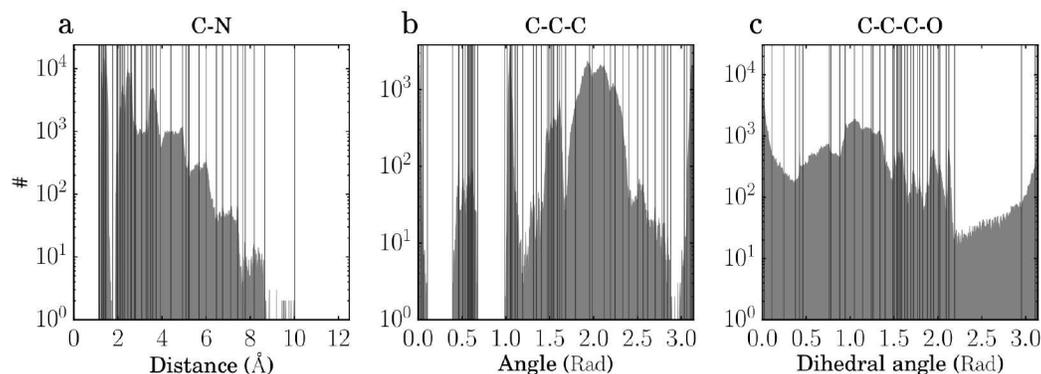


Figure 1: Illustration of select histograms of distances, angles and dihedral angles in QM9. Vertical lines constitutes placements of the bins in the HD and/or HDAD representations. (a) All C-N distances. (b) All C-C-C angles. (c) All C-C-C-O dihedral angles.

model, however, still correspond to the previously obtained hyper parameter 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

KRR⁴⁸⁻⁵¹ is a type of regression with regularization⁵² 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-learn⁵³ for all representations.

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

Table 2: Atom pair features for the MG representation: Values provided for each pair of atoms in the molecule.

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.

1 logarithmic grid for the 10 percent training set
2 (from 0.25 to 8192 for Gaussian kernel and 0.1
3 to 16384 for Laplacian kernel). In order to sim-
4 plify the width screening, prior to learning all
5 feature vectors were normalized (scaling the in-
6 put vector by the mean norm across the train-
7 ing set) by the Euclidean norm for the Gaussian
8 kernel and the Manhattan norm for the Lapla-
9 cian kernel.

12 2.5.2 Bayesian Ridge Regression

14 We use BR⁵⁴ as is implemented in scikit-
15 learn.⁵³ BR is a linear model with a L^2 penalty
16 on the coefficients. Unlike Ridge Regression
17 where the strength of that penalty is a regu-
18 larization hyperparameter which must be set,
19 in Bayesian Ridge Regression the optimal reg-
20 ularizer is estimated from the data.

24 2.5.3 Elastic Net

26 Also EN⁵⁵ is a linear model. Unlike BR, the
27 penalty on the weights is a mix of L^1 and L^2
28 terms. In addition to the regularization hyper-
29 parameter for the weight penalty, Elastic net
30 has an additional hyperparameter `l1_ratio` to
31 control the relative strength of the L^1 and L^2
32 weight penalties. We used the scikit-learn⁵³ im-
33 plementation and set `l1_ratio` = 0.5. We then
34 did a hyperparameter search on regularizing pa-
35 rameter in a base 10 logarithmic grid from $1e-6$
36 to 1.0.

42 2.5.4 Random Forest

44 We use RF⁵⁶ as implemented in scikit-learn.⁵³
45 RF regressors produce a value by averaging
46 many individual decision trees fitted on ran-
47 domly resampled sets of the training data. Each
48 node in each decision tree is a threshold of one
49 input feature. Early experiments did not reveal
50 strong differences in performance based on the
51 number of trees used, once a minimal number
52 was reached. We have used 120 trees for all
53 regressions.

2.5.5 Graph Convolutions

We have used the GC model as described
in Kearnes et al.²⁷, with several structural mod-
ifications and optimized hyperparameters. The
graph convolution model is built on the con-
cepts of “atom” layers (one real vector asso-
ciated with each atom) and “pair” layers (one
real vector associated with each pair of atoms).
The graph convolution architecture defines op-
erations 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 de-
tails in the Supplementary Material. First, we
have removed the “Pair order invariance” prop-
erty by simplifying the ($A \rightarrow P$) transfor-
mation. Since the model only uses the atom layer
for the molecule level features, pair order in-
variance is not needed. Second, we have used
the Euclidean distance between atoms. In the
($P \rightarrow A$) transformation, we divide the value
from the convolution step by a series of dis-
tance exponentials. If the original convolution
for an atom pair (a, b) with distance d pro-
duces 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 have fol-
lowed other work on neural networks based on
chemical graphs⁵⁷ which uses a sum of softmax
operations to convert a real valued vector to
a sparse vector and sum those sparse vectors
across all the atoms. We use the same oper-
ation here along with a simple sum across the
atoms to produce molecule level features from
the top atom layer. We have found that this
works as well or better than the Gaussian his-
tograms first used in GC.²⁷ To optimize the
network, we have searched the hyperparame-
ter space using Gaussian Process Bandit Op-
timization⁵⁸ as implemented by HyperTune.⁵⁹
The hyperparameter search has been based on
the evaluation of the validation set for a sin-
gle fold of the data. Further details including
parameters, and search ranges chosen for this
paper are listed in the Supplementary materi-
als.

2.5.6 Gated Graph Neural Networks

We have used the GG Neural Networks model (GG) as described in Li et al.²⁸. Similar to the GC model, it 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 $S = \{0, 1, \dots, k\}$ 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 distinguish four different covalent bonding types (single, double, triple, aromatic). For all remaining atom-pairs we bin 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 has entries in an alphabet of size 14 ($k=14$), indicating bond type for covalently bonded atoms, and distance bin for all other atoms. We have trained the GG model on each target property individually. Further technical details are specified in the Supplementary materials.

3 Results and discussion

3.1 Overview

We present an overview of the most relevant numerical results in Table 3. It 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), $\varepsilon_{\text{HOMO}}$ (MG/GC), $\varepsilon_{\text{LUMO}}$ (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, $U(T = 298K)$, $H(T = 298K)$, $G(T = 298K)$ 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 B3LYP level of theory and reach chemical (target)

accuracy, both defined alongside in table 3, 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 ($\varepsilon_{\text{HOMO}}$, $\varepsilon_{\text{LUMO}}$, $\Delta\varepsilon$, α , and ZPVE) the best models come within a factor 2 of target accuracy, while all (except $\varepsilon_{\text{HOMO}}$, $\varepsilon_{\text{LUMO}}$ and $\Delta\varepsilon$) where we don't have reliable data. outperforming DFT accuracy.

In Fig. 2 out-of-sample errors as a function of training set size (learning curves) are shown for all properties and representations 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 shown decay 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 off-sets of all learning curves can be found in Tables S4 and S5 in the Supplementary Material), are MG/GC for μ , MG/GG and HDAD/KRR for α , CM/KRR and BOB/KRR for $\langle R^2 \rangle$, 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 model capacities. This can also be seen from the learning curves of all regressors presented in Figures S1 to S6 of the Supplementary Material. The performance of BR and EN improves only slightly with increased training set size and even gets worse for some property/representation combinations. These two regressors also exhibit very similar learn-

Table 3: MAE on out-of-sample data of all representations for all regressors and properties at $\sim 118k$ (90%) training set size. 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 are highlighted in bold. Additionally, the table contains mean MAE of representations for each property and regressor; and normalized, by MAD (See Table 4), mean MAE (NMMAE) over all properties for each regressor/representation combination.

		U_0	$\varepsilon_{\text{HOMO}}$	$\varepsilon_{\text{LUMO}}$	$\Delta\varepsilon$	μ	α	ZPVE	C_v	ω_1	NMMAE
		eV	eV	eV	eV	Debye	Bohr ³	eV	cal/molK	cm ⁻¹	arb. u.
EN	CM	0.911	0.338	0.631	0.722	0.844	1.33	0.0265	0.906	131.0	0.423
	BOB	0.602	0.283	0.521	0.614	0.763	1.2	0.0232	0.7	81.4	0.35
	BAML	0.212	0.186	0.275	0.339	0.686	0.793	0.0129	0.439	60.4	0.231
	ECFP4	3.68	0.224	0.344	0.383	0.737	3.45	0.27	1.51	86.6	0.462
	HDAD	0.0983	0.139	0.238	0.278	0.563	0.437	0.00647	0.0876	94.2	0.183
	HD	0.192	0.203	0.299	0.36	0.705	0.638	0.00949	0.195	104.0	0.236
	MARAD	0.183	0.222	0.305	0.391	0.707	0.698	0.00808	0.206	108.0	0.256
	Mean	0.84	0.228	0.373	0.441	0.715	1.22	0.0509	0.578	95.1	
BR	CM	0.911	0.338	0.632	0.723	0.844	1.33	0.0265	0.907	131.0	0.424
	BOB	0.586	0.279	0.521	0.614	0.761	1.14	0.0222	0.684	80.9	0.343
	BAML	0.202	0.183	0.275	0.339	0.685	0.785	0.0129	0.444	60.4	0.229
	ECFP4	3.69	0.224	0.344	0.383	0.737	3.45	0.27	1.51	86.7	0.462
	HDAD	0.0614	0.14	0.238	0.278	0.565	0.43	0.00318	0.0787	94.8	0.182
	HD	0.171	0.203	0.298	0.359	0.705	0.633	0.00693	0.19	104.0	0.235
	MARAD	0.171	0.184	0.257	0.315	0.647	0.533	0.00854	0.201	103.0	0.226
	Mean	0.828	0.221	0.367	0.43	0.706	1.19	0.05	0.574	94.5	
RF	CM	0.431	0.208	0.302	0.373	0.608	1.04	0.0199	0.777	13.2	0.239
	BOB	0.202	0.12	0.137	0.164	0.45	0.623	0.0111	0.443	3.55	0.142
	BAML	0.2	0.107	0.118	0.141	0.434	0.638	0.0132	0.451	2.71	0.141
	ECFP4	3.66	0.143	0.145	0.166	0.483	3.7	0.242	1.57	14.7	0.349
	HDAD	1.44	0.116	0.136	0.156	0.454	1.71	0.0525	0.895	3.45	0.198
	HD	1.39	0.126	0.139	0.15	0.457	1.66	0.0497	0.879	4.18	0.197
	MARAD	0.21	0.178	0.243	0.311	0.607	0.676	0.0102	0.311	19.4	0.199
	Mean	1.08	0.142	0.174	0.209	0.499	1.43	0.0569	0.761	8.74	
KRR	CM	0.128	0.133	0.183	0.229	0.449	0.433	0.0048	0.118	33.5	0.136
	BOB	0.0667	0.0948	0.122	0.148	0.423	0.298	0.00364	0.0917	13.2	0.0981
	BAML	0.0519	0.0946	0.121	0.152	0.46	0.301	0.00331	0.082	19.9	0.105
	ECFP4	4.25	0.124	0.133	0.174	0.49	4.17	0.248	1.84	26.7	0.383
	HDAD	0.0251	0.0662	0.0842	0.107	0.334	0.175	0.00191	0.0441	23.1	0.0768
	HD	0.0644	0.0874	0.113	0.143	0.364	0.299	0.00316	0.0844	21.3	0.0935
	MARAD	0.0529	0.103	0.124	0.163	0.468	0.343	0.00301	0.0758	21.3	0.112
	Mean	0.662	0.101	0.126	0.159	0.427	0.859	0.0383	0.333	22.7	
GG	MG	0.0421	0.0567	0.0628	0.0877	0.247	0.161	0.00431	0.0837	6.22	0.0602
GC	MG	0.15	0.0549	0.062	0.0869	0.101	0.232	0.00966	0.097	4.76	0.0494

ing 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 representa-

tion 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

Table 4: Mean and mean absolute deviation (MAD) for all properties in the QM9 data set, as well as target MAE, and DFT (at B3LYP level of theory) MAE relative to experiment for each property, and the number of molecules used to estimate the values (In parentheses of DFT row). The target accuracies 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 Bohr³ 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 CCSD(T) error for predicting frequencies.⁶⁰ Target accuracies used for C_v were not explained in article.¹³ Section 2.3 discusses how the errors for DFT were obtained.

	U_0	ϵ_{HOMO}	ϵ_{LUMO}	$\Delta\epsilon$	μ	α	ZPVE	C_v	ω_1
	eV	eV	eV	eV	Debye	Bohr ³	eV	cal/molK	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)

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 , Fig. 1 of reference,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 impor-

tance are: O-H placed at 0.961 Å, N-H placed 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, respectively. 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

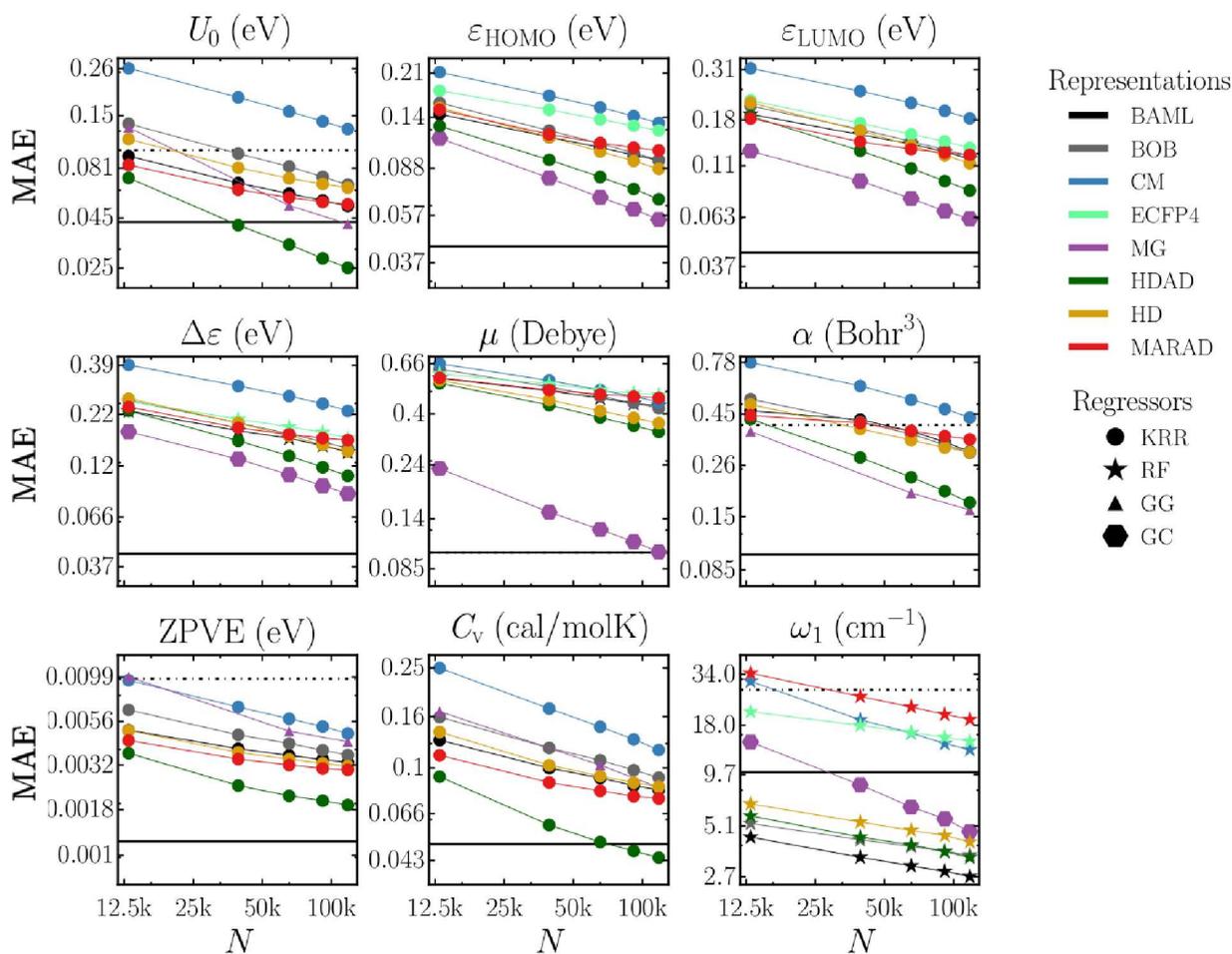


Figure 2: Learning curves (mean absolute error (MAE) as a function of training set size N) for 10 properties of QM9 molecules, displaying the best regressor for each representation and property. Horizontal solid lines correspond to target accuracies, vertical dotted lines correspond to approximated B3LYP 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 .

predictions which can scale.⁶²

3.3 Representations

As one would expect, HDAD contains more relevant information and thus it always out-

performs 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 regressor. In the case of RF, however, we observe little difference between HDAD and HD,

1 and HD can even yield slightly lower errors than
2 HDAD. In our opinion, this is due to the addi-
3 tional bins of angles and dihedrals rather adding
4 noise than signal. By contrast, the separation
5 of distances, angles and dihedral angles into dif-
6 ferent bins is not a problem for the KRR meth-
7 ods because the kernels used are purely distance
8 based. This makes it possible for KRR to ex-
9 ploit the extra three- and four-body informa-
10 tion in HDAD and to gain an advantage over
11 HD. We note however that the remarkable per-
12 formance of HDAD is possible despite its strik-
13 ing simplicity. As illustrated in Fig. 1 and dis-
14 cussed above, characteristic chemical behavior
15 can be directly obtained by human inspection of
16 HDAD. As such, HDAD corresponds to a rep-
17 resentation very much "Occam's razor style".
18 Unfortunately, due to its discrete nature and
19 its origin in sorting distances, HDAD will suf-
20 fer from lack of differentiability, which might
21 limit its applicability when modeling forces or
22 other non-equilibrium properties.

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

48 4 Conclusions

49 We have benchmarked many combinations of
50 regressors and representations on the same
51 QM9 data set consisting of ~ 131 k organic
52 molecules. For all properties, the best ML
53 model prediction errors reach the accuracy of
54 DFT at B3LYP level with respect to experi-
55 ment. For 7 out of 12 distinct properties (at-
56 omization energies, heat-capacity, ω_1 , μ) out-

of-sample errors reach levels on par with chem-
ical accuracy, or better, using a training set size
of ~ 118 k (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 low-
est errors. We have found that GC, GG, and
KRR have best performance across *all* prop-
erties, except for the highest vibrational fre-
quency for which RF performs best. There is no
single representation and regressor combination
that works best for all properties (though forth-
coming work with further improvements to the
GG based models indicates best in class per-
formance across all properties⁶³). For inten-
sive electronic properties (μ , HOMO/LUMO
eigenvalues and gap) we have found MG/NN
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 of property in training set,
and HDAD, being just histograms of distances,
angles, dihedrals. Using BR and EN is not rec-
ommended if accuracy is of importance, both
regressors perform worse across all properties.
Apart from predicting highest fundamental vi-
brational frequency best, RF based models de-
liver rather unsatisfactory performance. The
ECFP4 based models have shown poor general
performance in comparison to all other repre-
sentations studied; it is not recommended for
investigations of molecular properties.

We should caution the reader that all our re-
sults refer to equilibrium structures of a set of
only ~ 131 k organic molecules. While ~ 131 k
molecules might seem sufficiently large to be
representative, this number is dwarfed in com-
parison to chemical space, i.e. the space popu-
lated by all theoretically stable molecules, es-
timated to exceed 10^{60} for medium sized or-
ganic molecules.⁶⁴ Furthermore, ML models
for predicting properties of molecules in non-
equilibrium or strained configurations might re-
quire substantially more training data. This

1 point is also of relevance because some of the
2 highly accurate models described herewithin
3 (MG based) currently use bond based graph
4 connectivity in addition to distance, raising
5 questions about the applicability to reactive
6 processes.
7

8 In summary, for the organic molecules stud-
9 ied, we have collected numerical evidence which
10 suggests that the out-of-sample error of ML
11 is consistently better than estimated DFT at
12 B3LYP level accuracy. While this is no guar-
13 antee that ML models would reach same error
14 levels if more accurate, explicitly electron corre-
15 lated or experimental reference data was used,
16 previous studies indicate that similar perfor-
17 mance can be expected when using higher levels
18 of theory.⁸ More specifically, one might intu-
19 itively expect that going beyond hybrid DFT to
20 higher quality data (either wavefunction based-
21 QM or experiment) in terms of reference meth-
22 ods would represent a more challenging learn-
23 ing problem, and therefore imply the need for
24 larger training set sizes. Results in Ref.,⁸ how-
25 ever, suggest that ML models can predict the
26 differences between HF and MP2, CCSD, and
27 CCSD(T) equally well using the same training
28 set.
29

30 As such, we conclude that future reference
31 datasets for training state-of-the-art machine
32 learning models of molecular properties should
33 preferably use reference levels of theory which
34 go beyond DFT at B3LYP level of accuracy.
35 While it seems unlikely that for each class of
36 molecules, hundreds of thousands of experimen-
37 tal training data points will become available
38 in the foreseeable future, it might well be pos-
39 sible to reach such scale using efficient imple-
40 mentations of explicit electron correlated meth-
41 ods within high-performance computing cam-
42 paigns. Finally, we note that future work could
43 deal with improving representations and regres-
44 sors, with the goal of reaching similar predictive
45 power using less data.
46

47 **Acknowledgement** The authors thank Dirk
48 Bakowies for helpful comments, and Adrian
49 Roitberg for pointing out an issue with the use
50 of partial charges in the neural net models in
51 an earlier version of this paper. O.A.v.L. ac-
52
53
54
55
56
57
58
59
60

knowledges support from the Swiss National
Science foundation (No. PP00P2_138932,
310030_160067), the research fund of the Uni-
versity of Basel, and from Google. This mate-
rial 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 comput-
ing core facility at University of Basel.

5 SI

Supplementary information regarding raw data,
MARAD representation, graph convolutions,
gated graphs, random forests, and learning
curves are reported, as well as root mean square
errors for ML predictions after training on the
largest training set.

References

- 1
2
3 (1) Hohenberg, P.; Kohn, W. Inhomogeneous
4 Electron Gas. *Phys. Rev.* **1964**, *136*, B864.
- 5
6 (2) Kohn, W.; Sham, L. J. Self-Consistent
7 Equations Including Exchange and Cor-
8 relation Effects. *Phys. Rev.* **1965**, *140*,
9 A1133.
- 10
11 (3) Burke, K. Perspective on density func-
12 tional theory. *J. Chem. Phys.* **2012**, *136*,
13 150901.
- 14
15 (4) Koch, W.; Holthausen, M. C. *A Chemist's*
16 *Guide to Density Functional Theory*;
17 Wiley-VCH, 2002.
- 18
19 (5) Cohen, A. J.; Mori-Sánchez, P.; Yang, W.
20 Challenges for Density Functional Theory.
21 *Chem. Rev.* **2012**, *112*, 289–320.
- 22
23 (6) Plata, R. E.; Singleton, D. A. A Case
24 Study of the Mechanism of Alcohol-
25 Mediated Morita Baylis-Hillman Reac-
26 tions. The Importance of Experimental
27 Observations. *J. Am. Chem. Soc.* **2015**,
28 *137*, 3811–3826.
- 29
30 (7) Medvedev, M. G.; Bushmarinov, I. S.;
31 Sun, J.; Perdew, J. P.; Lyssenko, K. A.
32 Density functional theory is straying from
33 the path toward the exact functional. *Sci-*
34 *ence* **2017**, *355*, 49–52.
- 35
36 (8) Ramakrishnan, R.; Dral, P. O.; Rupp, M.;
37 von Lilienfeld, O. A. Big Data Meets
38 Quantum Chemistry Approximations:
39 The Δ -Machine Learning Approach.
40 *J. Chem. Theory Comput.* **2015**, *11*,
41 2087–2096.
- 42
43 (9) Ramakrishnan, R.; Dral, P. O.; Rupp, M.;
44 von Lilienfeld, O. A. Quantum chem-
45 istry structures and properties of 134 kilo
46 molecules. *Sci. Data* **2014**, *1*, 140022.
- 47
48 (10) Ruddigkeit, L.; van Deursen, R.;
49 Blum, L. C.; Reymond, J.-L. Enum-
50 eration of 166 Billion Organic Small
51 Molecules in the Chemical Universe
52 Database GDB-17. *J. Chem. Inf. Model.*
53 **2012**, *52*, 2864–2875.
- 54
55 (11) Huang, B.; von Lilienfeld, O. A. Commu-
56 nication: Understanding molecular repre-
57 sentations in machine learning: The role of
58 uniqueness and target similarity. *J. Chem.*
59 *Phys.* **2016**, *145*, 161102.
- 60 (12) Hansen, K.; Biegler, F.; Ramakrish-
nan, R.; Pronobis, W.; von Lilien-
feld, O. A.; Müller, K.-R.; Tkatchenko, A.
Machine Learning Predictions of Molec-
ular 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 Ker-
nel in Chemical Space. *chimia* **2015**, *69*,
182–186.
- (14) Rupp, M.; Tkatchenko, K.-R., Alexandre
haand Müller; von Lilienfeld, O. A. Fast
and Accurate Modeling of Molecular At-
omization Energies with Machine Learn-
ing. *Phys. Rev. Lett.* **2012**, *108*, 058301.
- (15) Barker, J.; Bulin, J.; Hamaekers, J.;
Mathias, S. Localized Coulomb Descrip-
tors for the Gaussian Approximation Po-
tential. *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 **2015**, *115*, 1084–1093.
- (17) Huan, T. D.; Mannodi-Kanakithodi, A.;
Ramprasad, R. Accelerated materials
property predictions and design using
motif-based fingerprints. *Phys. Rev. B*
2015, *92*, 014106.
- (18) Bartók, A. P.; Kondor, R.; Csányi, G. On
representing chemical environments. *Phys.*
Rev. B **2013**, *87*, 184115.
- (19) De, S.; Bartók, A. P.; Csányi, G.; Ceri-
otti, M. Comparing molecules and solids
across structural and alchemical space.

- 1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
17
18
19
20
21
22
23
24
25
26
27
28
29
30
31
32
33
34
35
36
37
38
39
40
41
42
43
44
45
46
47
48
49
50
51
52
53
54
55
56
57
58
59
60
- 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* **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* **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. *ICLR* **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 MP2 Methods 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, W. A. G.; 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 III, R. W.; Morrison, K. C.; Hicklin, R. W.; Flood Jr, T. A.; Richter, M. F.; Hergenrother, P. J. A Ring Distortion Strategy to Construct Stereochemically Complex and Structurally Diverse Compounds from Natural Products. *Nature chemistry* **2013**, *5*, 195.
- (42) Todeschini, R.; Consonni, V. *Handbook of Molecular Descriptors*; Wiley-VCH, Weinheim, 2009.
- (43) Faulon, J.-L.; Visco, Jr., D. P.; 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. Graph. Model.* **2002**, *20*, 429–438.
- (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*, 1–14.
- (46) Landrum, G. RDKit: Open-source cheminformatics. <http://www.rdkit.org> **2014**, *3*, 2012.
- (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. 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.
- (51) Hastie, T.; Tibshirani, R.; Friedman, J. *The Elements of Statistical Learning: Data Mining, Inference, and Prediction, Second Edition*, 2nd ed.; Springer: New York, 2011.
- (52) Hoerl, Arthur, E.; Kennard, Robert, W. Ridge Regression Biased Estimation for Nonorthogonal Problems. *Technometrics* **2000**, *80*.

- 1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
17
18
19
20
21
22
23
24
25
26
27
28
29
30
31
32
33
34
35
36
37
38
39
40
41
42
43
44
45
46
47
48
49
50
51
52
53
54
55
56
57
58
59
60
- (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.
- (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 multi-layer 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 34th International Conference on Machine Learning, ICML 2017*. 2017.
- (64) Kirkpatrick, P.; Ellis, C. Chemical space. *Nature* **2004**, *432*, 823.

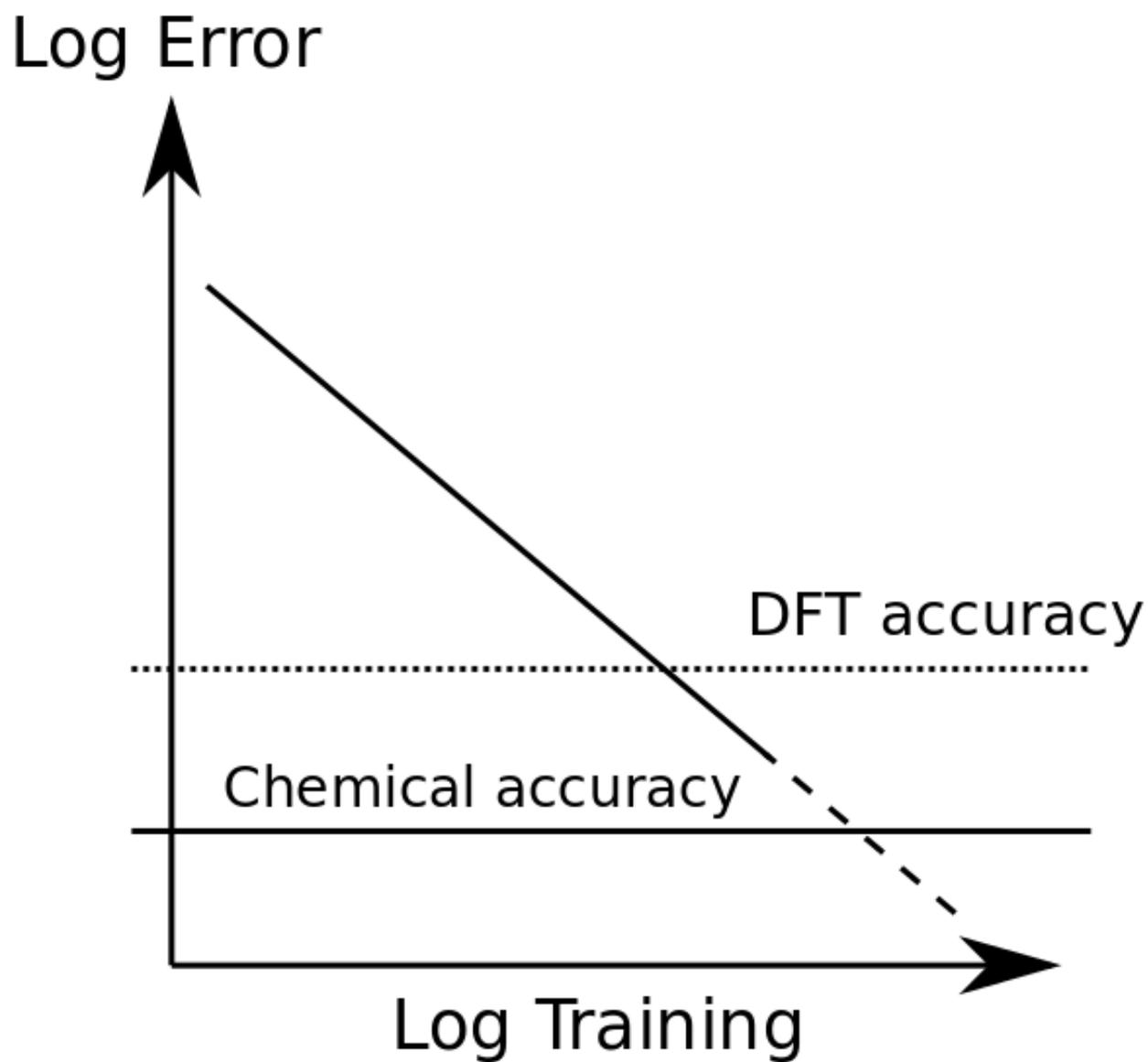


Figure 3: TOC