Supplemental Materials for "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

1 Supplementary data

Additional data relevant to this work can be found at `https://drive.google.com/open?id=0Bzn36Iqm8hZscHFJcVh5aC1mZFU`. The link provides access to the following files: (i) A readme file, (ii) names of the molecules that failed conversion from coordinates to rational SMILES strings for the MG representation, (iii) all representations used in this work, (iv) the bins and histograms used to create the HD, HDA and HDAD representations, (v) all splits used for training, validation and test set, and (vi) properties of all molecules in the

QM9 data set.[?]

2 MARAD

This section will discuss how the MARAD (Molecular atomic radial angular distribution) is used to represent a molecule, and how we discretize it. However, first we will introduce an atomic version of MARAD, atomic radial angular distribution (ARAD), which we later use to generate MARAD.

ARAD is an atomic radial distribution function based representation. The environment of an atom I is represented by three functions: $A_r(I)$, $A_\perp(I)$ and $A_\parallel(I)$, see Eq. S1.

$$A_k(I) = \mathcal{Z}(R_I, \sigma_R; \chi_1)\mathcal{Z}(C_I, \sigma_C; \chi_2) \sum_i^{n_I} \Phi_i^k(I) \exp\left[-\frac{(\chi_3 - d_{i,I})^2}{2\sigma_d^2}\right] \mathcal{Z}(R_i, \sigma_R; \chi_4)\mathcal{Z}(C_i, \sigma_C; \chi_5)\xi(d_{i,I}) \tag{S1}$$

χ is integrated out when comparing two atoms (or molecules), or when discretizing the representation; σ_d, σ_R, σ_C are hyper parameters; $d_{i,I}$ is the distance between atom I and nearby atoms i; R_i and C_i correspond respectively to the row and column of atom i in the periodic table; $\xi(d_{i,I})$ is a scaling function that is used to give a higher importance to smaller distances, as chemical bonds in molecules are mostly affected by nearby atoms; and $\mathcal{Z}(R, \sigma; \chi)$ is used to introduce a chemical similarity between two atoms of different, or the same, elemental type. $\Phi_i^k(I)$, is equal to 1, $\sum_j \cos(\theta_{i,j}^I)\xi(d_{i,I})$ and $\sum_j \sin(\theta_{i,j}^I)\xi(d_{i,I})$ for $k = r$, $\parallel$ and $\perp$ respectively. $\theta_{i,j}(I)$ is the unsigned angle between the vector spanning from atom I to atom i and the vector spanning from atom I to atom j.

MARAD M_k is generated by summing $A_k(I)$ over all n atoms I in the molecules, which we discretize by calculating the scalar product between M_k and a grid.

The grid points $\mathcal{G}_{i,a,b}$ are placed with uniform spacing σ_d along the interatomic distances d, on the row R and column C in the periodic table for each element pair.

$$\mathcal{S}(M_k, \mathcal{G}_{j,a,b}) \equiv \sum_{I}^{n} \int_{-\infty}^{\infty} \cdots \int_{-\infty}^{\infty} A_k(I) \mathcal{G}_{j,a,b}^{proj} d\chi_1 \cdots d\chi_5 \tag{S2}$$

$$\mathcal{G}_{j,a,b} = \mathcal{Z}(R_a, \sigma_R; \chi_1)\mathcal{Z}(C_a, \sigma_C; \chi_2) \exp(-\frac{(\chi_3 - \sigma_d j)^2}{2\sigma_d^2})\mathcal{Z}(R_b, \sigma_R; \chi_4)\mathcal{Z}(C_b, \sigma_C; \chi_5)\xi(\sigma_d j) \tag{S3}$$

$$\begin{aligned}\int_{-\infty}^{\infty} \cdots \int_{-\infty}^{\infty} A_k(I) \mathcal{G}_{i,a,b}^{proj} d\chi_1 \cdots d\chi_5 = \frac{\sqrt{\pi}\sigma_d \sigma_R^4 \sigma_C^4}{(\sigma_R^2 + (R_I - R_a)^2)(\sigma_C^2 + (C_I - C_a)^2)} \\ \sum_{i}^{n_I} \Phi_i^k(I) \frac{\exp(-\frac{(\sigma_d j - d_{i,I})^2}{4\sigma_d^2})\xi(d_{i,I})\xi(\sigma_d j)}{(\sigma_R^2 + (R_i - R_b)^2)(\sigma_C^2 + (C_i - C_b)^2)}\end{aligned} \tag{S4}$$

Throughout this work, the hyperparameters σ_R, σ_C and σ_d were set equal to 1, 0.5 and 0.2 respectively, and a sinusoidal scaling function with a hard cutoff was used: $\xi(d) = 1 - \sin(\pi \frac{d}{2D})$ if d $< D$ and 0 otherwise, with a cutoff distance $D = 6$ Å. The chemical similarity was set equal to $\frac{\sigma^{3/2}}{\sqrt{\pi}2[(\sigma/2)^2 + (\chi - R)^2]}$

3 Graph Convolutions

Notationally, let A^x be the value of a particular atom layer x and P^y be the value of a particular pair layer y. The inputs that produce those values should be clear from the context. A_a^x refers to the value of atom a in atom layer x and $P_{(a,b)}^y$ refers to the value of pair (a, b) in pair layer y. Throughout, f represents an arbitrary function and g represents an arbitrary *commutative* function (g returns the same result regardless of the order the arguments are presented). In this work, f is a learned linear operator with a rectified linear (ReLU) activation function and g is a sum.

We removed the "Pair order invariance" property. The original design of the architecture maintained the property "For all pair layers y, $P^y_{(a,b)} = P^y_{(b,a)}$" We changed the architecture to allow (a, b) and (b, a) to have different values. The key change is a simplification of making a pair layer from an atom layer $(A \to P)$. The original definition for constructing pair layer P^y from atom layer A^x was

$$P^y_{ab} = g(f(A^x_a, A^x_b), f(A^x_b, A^x_a)) \tag{S5}$$

where the pair order invariance is maintained by providing both orders and combining them with the commutative function g. In this work, we simplify to

$$P^y_{ab} = f(A^x_a, A^x_b) \tag{S6}$$

In fact, no other operations relied on the pair order invariance property so this is a strict simplification of the model.

We changed the model to use distance in the $(P \to A)$ operation. The original definition was

$$A^y_a = g(f(P^x_{(a,b)}), f(P^x_{(a,c)}), f(P^x_{(a,d)}), \ldots) \tag{S7}$$

where g was a sum and f was a convolution with ReLU activation and batch normalization.[?] Rewriting Equation S7 for W a shared weight matrix and BN the ReLU and batch normalization operator:

$$\begin{aligned} A^y_a = \mathrm{BN}(h(a,b)) + \mathrm{BN}(h(a,c)) + \mathrm{BN}(h(a,d)) + \ldots \\ h(a,b) = W P^x_{(a,b)} \end{aligned} \tag{S8}$$

In this work, we scale the convolutions on each pair by multiple distance exponentials. For a weight matrix W, $d_{(a,b)}$ the euclidean distance between atoms a and b, and $\bigodot$ vector

concatenation.

$$A_a^y = \mathrm{BN}(h'(a,b)) + \mathrm{BN}(h'(a,c)) + \mathrm{BN}(h'(a,d)) + \ldots$$
$$h'(a,b) = \bigodot_{k\in\{0,1,2,3,6\}} \frac{WP^x_{(a,b)}}{d^k_{(a,b)}} \tag{S9}$$

Regarding the training, we have performed the regression optimization using the ADAM optimizer[?] with 10 simultaneous replicas, a learning rate of 0.01 (decayed by 0.96 every 2 epochs), and a batch size of 96 for 250k steps. We have tested the model's performance at various points during training and selected the step with the lowest error on the validation set. Hyperparameters are specified in Table 1.

4 Gated Graphs

The GG model assumes that the input adjacency matrix has entries in a discrete alphabet, where the size of the alphabet is k (represented by an integer in the set {0, 1, ..., k-1}. The distances between atoms in a molecule are real valued. In order to incorporate distances into the model we bin the real valued distance into one of 10 bins. Using these bins the discrete entry in the adjacency matrix between two atoms depends on whether or not the atom pair participates in a bond. For bonded pairs we use an integer between 0 and 3 to describe the bond type. For non-bonded pairs we use an integer between 4 and 13 to indicate the distance bin that the pairwise distance falls into. Note that this means that the model does not see the Euclidean distance between bonded atoms. We chose to do it this way as to keep k reasonably small, and because the distance seemed almost completely determined by bond type.

Each model and target combination has been trained using a random hyper parameter search with 50 trials. The hidden dimension of the nodes, d, has been chosen randomly in the set $\{33, 43, 73, 113\}$. T has been chosen uniformly in the range $3 \leq T \leq 8$. All models have been trained using SGD with the ADAM optimizer[?] with batch size 20 for 2 million

steps. Regarding the training, the initial learning rate has been chosen uniformly between $1e-5$ and $5e-4$. The learning rate has then been decreased linearly to a final learning rate $l * F$ where the decay factor F was chosen uniformly in the range $[.01, 1]$, i.e. l is the initial learning rate. The linear decay starts at a randomly chosen step between 10% and 90% of training.

Table 1: Hyperparameters for the graph convolutions that were optimized. Note that for the molecule level reductions, if multiple reductions are used the results are just concatenated together.

Feature	Description	Range	Chosen value
Number of weave modules	The number of repetitions of the weave module stacked on top of each other.	2–5	3
Convolution depths	Depths of all the convolutions in the weave modules	16–128	128
Final atom convolution depth	Depth of the final $(A \to A)$ convolution before molecule level reduction	16–512	16
Reduction: Sum molecule reduction	Whether to use a sum reduction across atoms	bool	true
Reduction: Gaussian histogram	Whether to use a Gaussian histogram reduction	bool	false
Reduction: Sum of softmaxes	Whether to use a sum of 10 softmax reductions	bool	true
FC layer sizes 0	Number of nodes in the 2 fully connected layers after the molecule feature reduction.	50–3000	339, 226

5 Random Forest Hyperparameters

The random forest regressor was quite insensitive to the number of trees chosen. The trend is consistently that more trees are better, but the total difference between 20 and 180 trees is generally less than 5% and the maximum observed is 7.8% for MARAD on C_{v}. We used a value of 120 trees for all other reported results.

We did not explore the other hyperparameters of random forests.

Table 2: MAE on out of sample molecules for different choices of the "num trees" hyperparmater for random forests.

features	num_trees	μ	α	$\varepsilon_{\text{HOMO}}$	$\varepsilon_{\text{LUMO}}$	$\Delta\varepsilon$	ZPVE	U_0	C_{v}	ω_1
CM	20	0.626	1.083	0.216	0.313	0.387	0.021	0.449	0.811	14.032
	40	0.614	1.056	0.211	0.307	0.379	0.020	0.438	0.792	13.513
	60	0.612	1.047	0.209	0.304	0.376	0.020	0.434	0.784	13.374
	90	0.609	1.041	0.208	0.303	0.374	0.020	0.432	0.780	13.269
	120	0.608	1.037	0.208	0.302	0.373	0.020	0.431	0.777	13.244
	180	0.606	1.034	0.207	0.301	0.372	0.020	0.429	0.775	13.184
BOB	20	0.464	0.656	0.125	0.142	0.170	0.012	0.214	0.472	3.696
	40	0.455	0.636	0.122	0.139	0.166	0.011	0.207	0.454	3.610
	60	0.453	0.629	0.121	0.138	0.165	0.011	0.205	0.449	3.580
	90	0.451	0.625	0.120	0.137	0.164	0.011	0.203	0.445	3.560
	120	0.450	0.623	0.120	0.137	0.164	0.011	0.202	0.443	3.547
	180	0.450	0.621	0.119	0.136	0.163	0.011	0.201	0.440	3.537
BAML	20	0.448	0.673	0.111	0.124	0.148	0.014	0.214	0.483	2.836
	40	0.440	0.652	0.109	0.120	0.144	0.013	0.206	0.462	2.771
	60	0.437	0.645	0.108	0.120	0.143	0.013	0.203	0.456	2.739
	90	0.435	0.640	0.107	0.119	0.142	0.013	0.201	0.452	2.721
	120	0.434	0.638	0.107	0.118	0.141	0.013	0.200	0.451	2.711
	180	0.433	0.636	0.106	0.118	0.141	0.013	0.199	0.448	2.701
ECFP4	20	0.492	3.770	0.146	0.149	0.170	0.247	3.735	1.603	15.047
	40	0.486	3.727	0.144	0.147	0.168	0.244	3.692	1.586	14.819
	60	0.485	3.716	0.143	0.146	0.167	0.243	3.676	1.581	14.769
	90	0.484	3.706	0.143	0.146	0.166	0.242	3.663	1.576	14.732
	120	0.483	3.699	0.143	0.145	0.166	0.242	3.659	1.574	14.710
	180	0.482	3.698	0.142	0.145	0.166	0.241	3.654	1.573	14.685
HDAD	20	0.470	1.806	0.122	0.144	0.164	0.056	1.513	0.936	3.634
	40	0.461	1.751	0.118	0.139	0.159	0.054	1.467	0.911	3.525
	60	0.458	1.728	0.117	0.138	0.157	0.053	1.454	0.903	3.492
	90	0.456	1.722	0.117	0.137	0.156	0.053	1.441	0.897	3.464
	120	0.454	1.709	0.116	0.136	0.156	0.053	1.437	0.895	3.445
	180	0.454	1.705	0.116	0.136	0.155	0.052	1.433	0.893	3.437
HD	20	0.471	1.749	0.131	0.146	0.157	0.053	1.460	0.920	4.383
	40	0.463	1.696	0.128	0.142	0.153	0.051	1.418	0.895	4.265
	60	0.460	1.683	0.127	0.140	0.152	0.050	1.403	0.887	4.218
	90	0.457	1.668	0.126	0.140	0.151	0.050	1.393	0.881	4.195
	120	0.457	1.660	0.126	0.139	0.150	0.050	1.387	0.879	4.178
	180	0.456	1.653	0.125	0.138	0.150	0.049	1.384	0.876	4.167
MARAD	20	0.623	0.707	0.184	0.253	0.323	0.011	0.225	0.336	20.223
	40	0.614	0.689	0.180	0.247	0.316	0.010	0.217	0.321	19.707
	60	0.610	0.682	0.179	0.245	0.313	0.010	0.213	0.316	19.542
	90	0.608	0.679	0.178	0.244	0.311	0.010	0.211	0.313	19.441
	120	0.607	0.676	0.178	0.243	0.311	0.010	0.210	0.311	19.353
	180	0.606	0.674	0.177	0.243	0.310	0.010	0.210	0.310	19.297

6 Learning Curves

Learning curves for all properties and all representations studied are shown in Figs. S1 - S6 using regressors EN, BR, RF, KRR, GC, and GG, respectively. These learning curves have been generated using the same settings and training set sizes as Fig. 2 in the main text.

EN

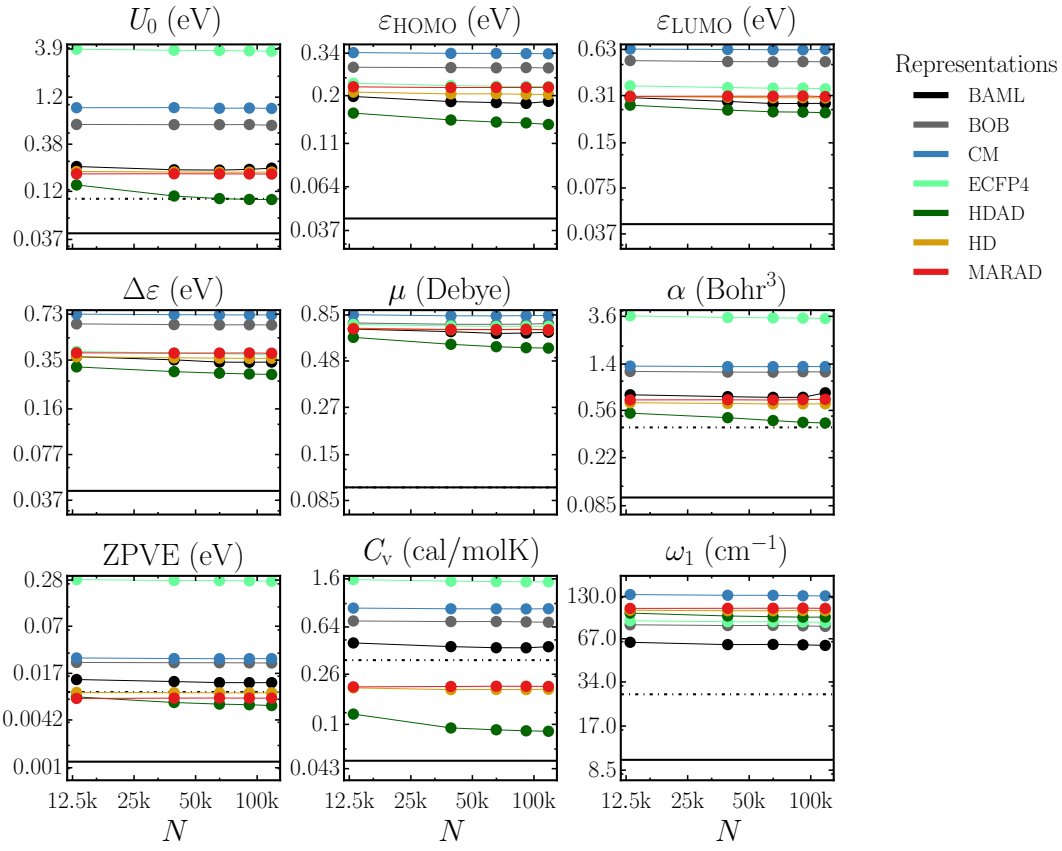

Figure S1: Learning curves for all properties and representations (except MG), using EN as regressor. Out-of-sample MAE as a function of training set size for QM9 molecules with the property and unit given in the title of each figures. Horizontal solid lines corresponds to target accuracies and horizontal dotted lines corresponds to B3LYP accuracies.

BR

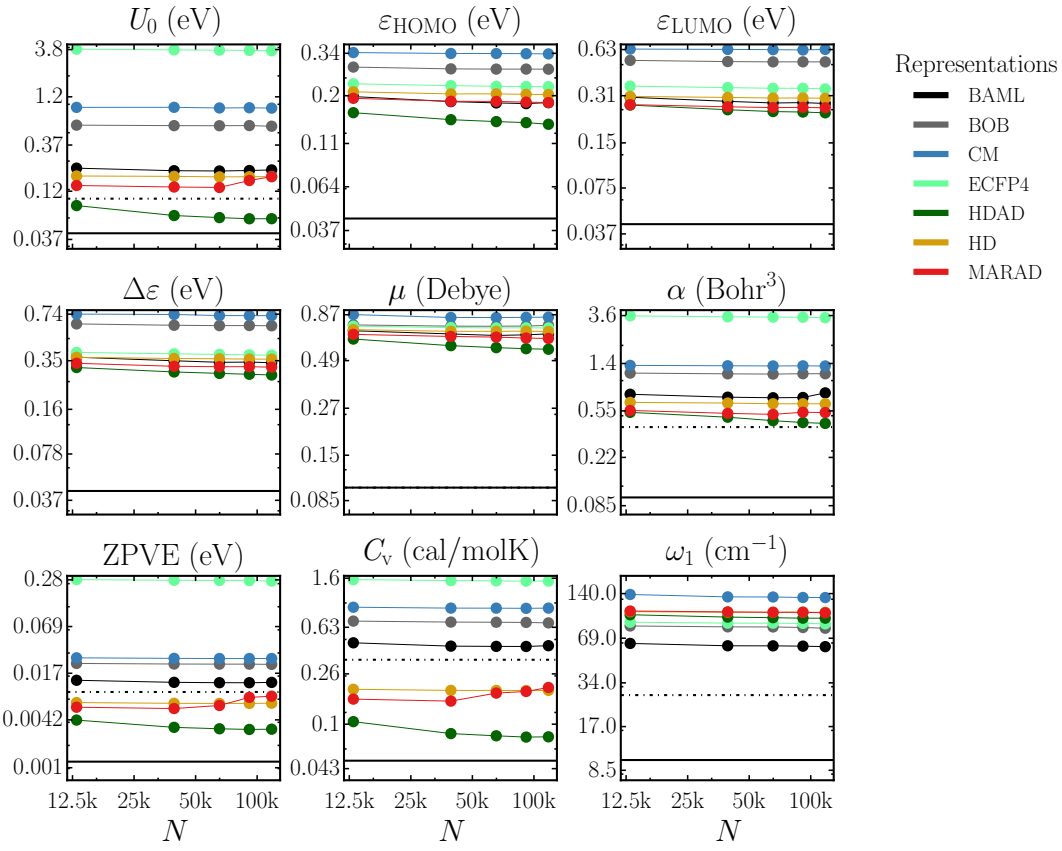

Figure S2: Learning curves for all properties and representations (except MG), using BR as regressor. Out-of-sample MAE as a function of training set size for QM9 molecules with the property and unit given in the title of each figures. Horizontal solid lines corresponds to target accuracies and horizontal dotted lines corresponds to B3LYP accuracies.

RF

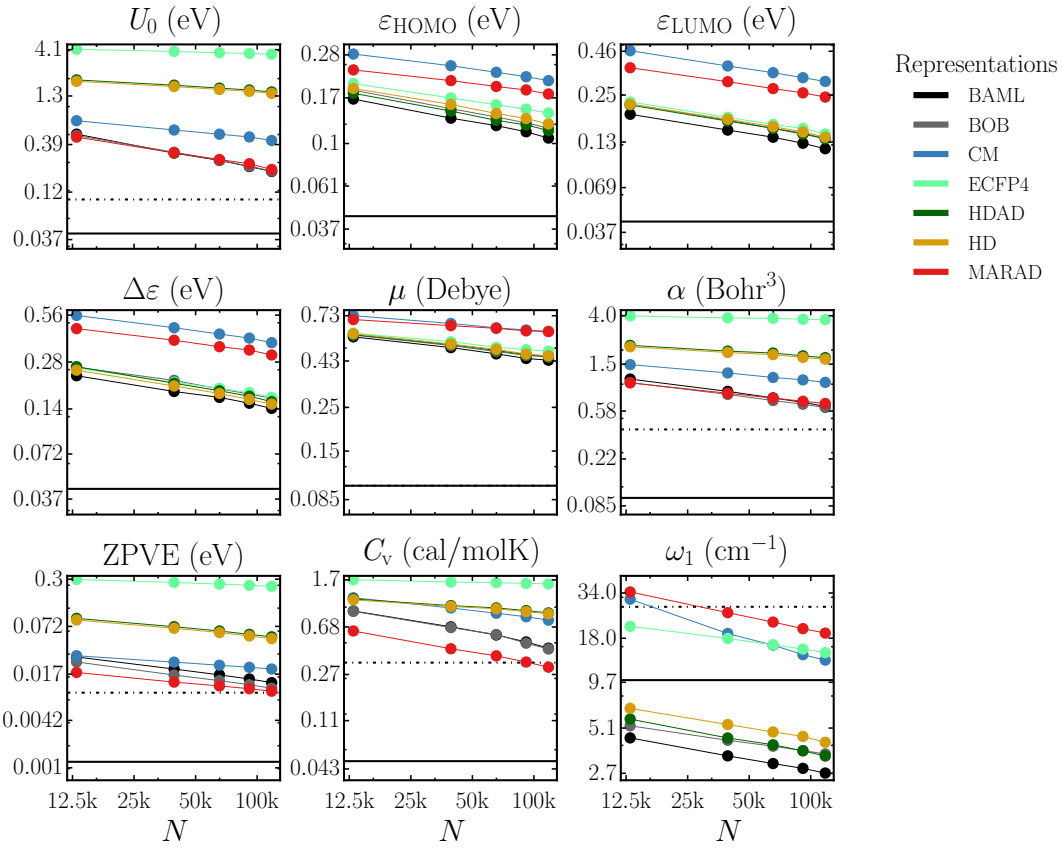

Figure S3: Learning curves for all properties and representations (except MG), using RF as regressor. Out-of-sample MAE as a function of training set size for QM9 molecules with the property and unit given in the title of each figures. Horizontal solid lines corresponds to target accuracies and horizontal dotted lines corresponds to B3LYP accuracies.

KRR

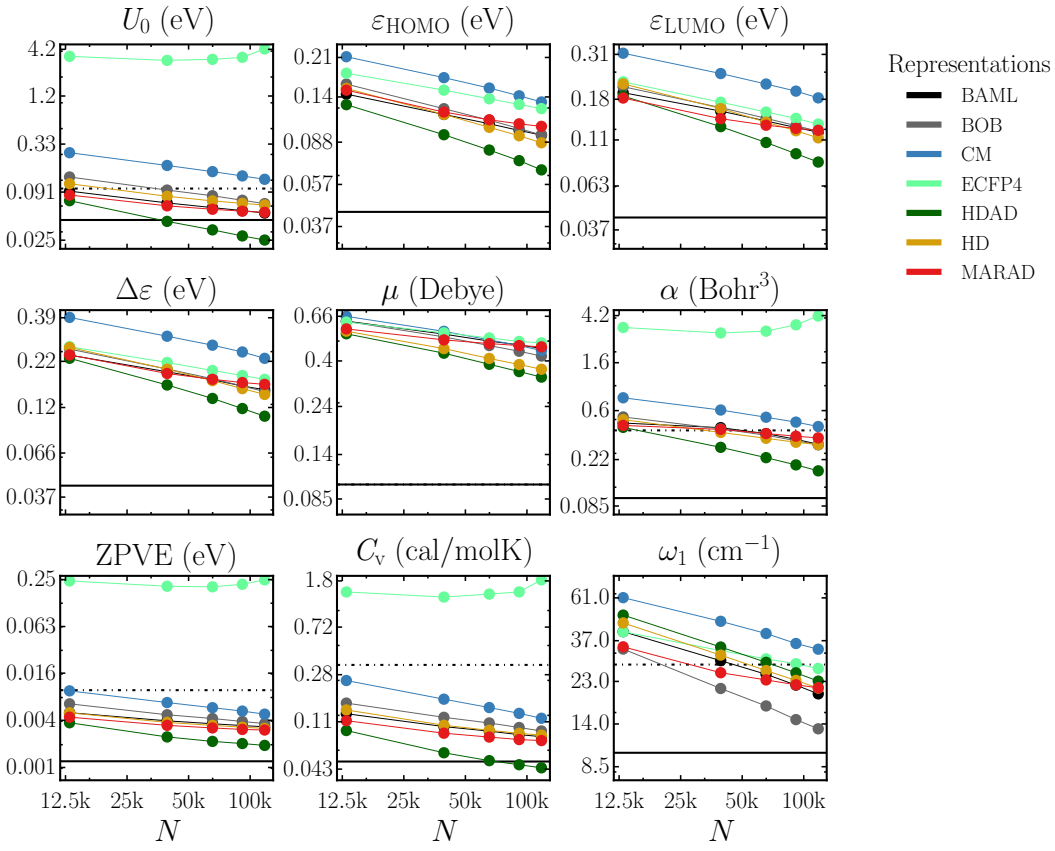

Figure S4: Learning curves for all properties and representations (except MG), using KRR as regressor. Out-of-sample MAE as a function of training set size for QM9 molecules with the property and unit given in the title of each figures. Horizontal solid lines corresponds to target accuracies and horizontal dotted lines corresponds to B3LYP accuracies.

GC

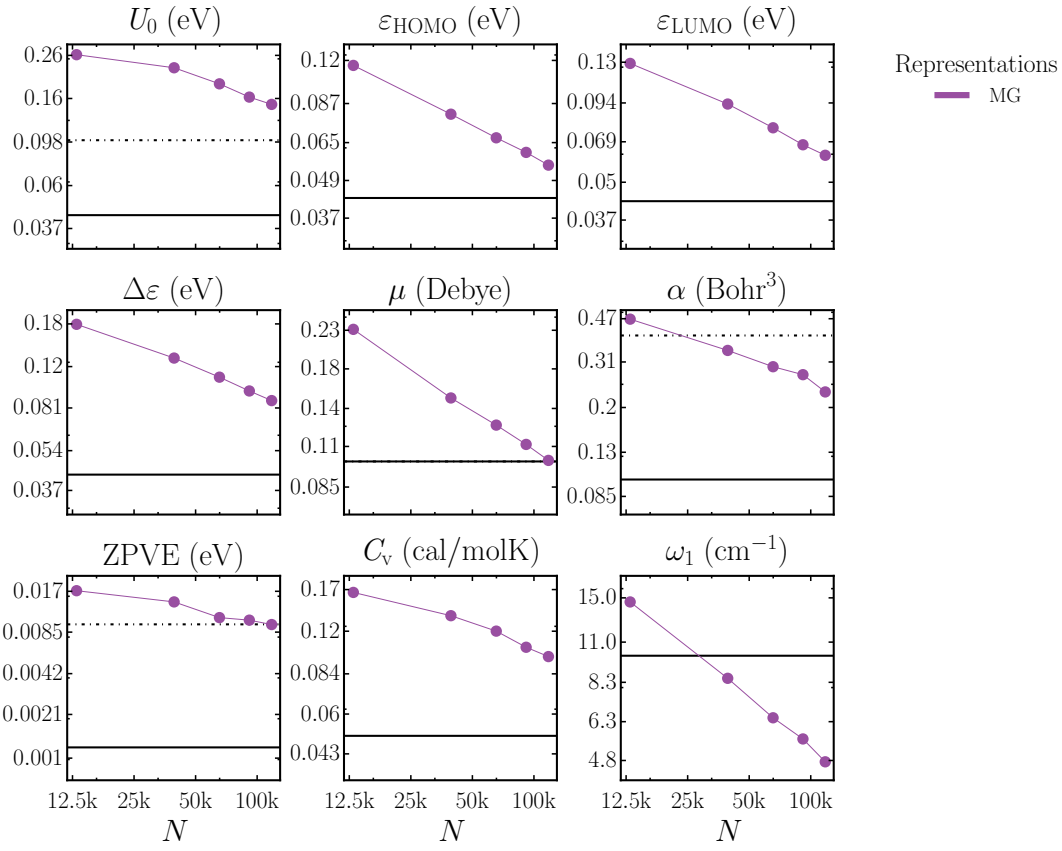

Figure S5: Learning curves for all properties, with MG as representation and CG as regressor. Out-of-sample MAE as a function of training set size for QM9 molecules with the property and unit given in the title of each figures. Horizontal solid lines corresponds to target accuracies and horizontal dotted lines corresponds to B3LYP accuracies.

GG

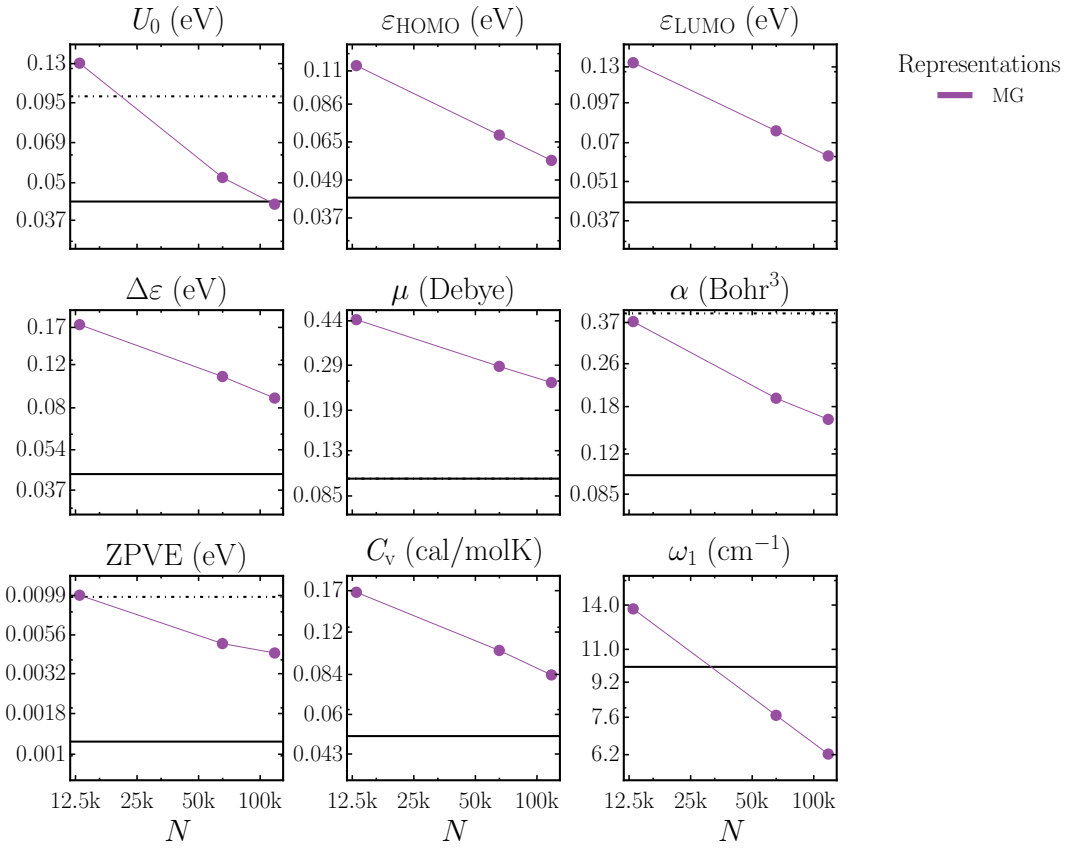

Figure S6: Learning curves for all properties, with MG as representation and GG as regressor. Out-of-sample MAE as a function of training set size for QM9 molecules with the property and unit given in the title of each figures. Horizontal solid lines corresponds to target accuracies and horizontal dotted lines corresponds to B3LYP accuracies.

Table 3: RMSE on out-of-sample data of all representations for all regressors and properties at ∼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 the mean RMSE of representations for each property and regressor and normalized (by RMSD) mean RMSE (NRMSE) over all properties for each regressor/representation combination.

		U_0	$\varepsilon_{\mathrm{HOMO}}$	$\varepsilon_{\mathrm{LUMO}}$	$\Delta\varepsilon$	μ	α	ZPVE	C_{v}	ω_1	
		eV	eV	eV	eV	Debye	Bohr^3	eV	cal/molK	cm^{-1}	
EN	CM	1.28	0.459	0.78	0.903	1.19	2.28	0.0366	1.17	166.0	0.451
	BOB	0.782	0.373	0.654	0.772	1.1	2.03	0.0292	0.904	101.0	0.372
	BAML	1.82	0.749	0.717	0.916	1.28	9.63	0.0229	1.63	91.7	0.569
	ECFP4	4.99	0.295	0.45	0.497	0.971	4.76	0.354	2.08	105.0	0.482
	HDAD	0.149	0.186	0.316	0.369	0.763	1.16	0.008 92	0.133	117.0	0.203
	HD	0.264	0.27	0.39	0.469	0.973	1.4	0.0124	0.269	126.0	0.256
	MARAD	0.246	0.29	0.393	0.5	1.01	1.49	0.0105	0.267	130.0	0.277
	Mean	1.36	0.375	0.529	0.632	1.04	3.25	0.0678	0.922	120.0	
BR	CM	1.28	0.459	0.781	0.904	1.19	2.28	0.0366	1.17	167.0	0.451
	BOB	0.764	0.368	0.652	0.771	1.09	1.97	0.028	0.884	101.0	0.365
	BAML	1.31	0.439	0.643	0.842	1.34	9.81	0.042	1.78	83.4	0.525
	ECFP4	4.98	0.295	0.451	0.497	0.971	4.75	0.354	2.08	105.0	0.481
	HDAD	0.0985	0.186	0.316	0.368	0.765	1.16	0.004 37	0.15	117.0	0.202
	HD	0.243	0.27	0.389	0.467	0.973	1.39	0.009 14	0.256	126.0	0.255
	MARAD	0.223	0.241	0.337	0.412	0.896	1.3	0.0109	0.257	125.0	0.245
	Mean	1.27	0.323	0.51	0.609	1.03	3.23	0.0693	0.939	118.0	
RF	CM	0.609	0.289	0.442	0.526	0.928	1.85	0.0264	1.04	35.2	0.28
	BOB	0.377	0.169	0.206	0.239	0.694	1.36	0.0176	0.666	6.27	0.172
	BAML	0.399	0.156	0.179	0.209	0.668	1.41	0.0208	0.667	4.91	0.171
	ECFP4	5.24	0.209	0.227	0.25	0.715	5.33	0.34	2.27	26.1	0.396
	HDAD	2.08	0.172	0.21	0.236	0.692	2.66	0.0805	1.26	6.37	0.229
	HD	2.0	0.18	0.208	0.221	0.69	2.59	0.0761	1.22	8.08	0.225
	MARAD	0.324	0.248	0.348	0.435	0.913	1.48	0.0147	0.446	34.4	0.234
	Mean	1.58	0.203	0.26	0.302	0.757	2.38	0.0822	1.08	17.3	
KRR	CM	0.185	0.181	0.245	0.309	0.664	1.14	0.006 82	0.161	49.5	0.159
	BOB	0.0969	0.129	0.165	0.204	0.612	0.965	0.005 01	0.122	22.9	0.117
	BAML	0.075	0.126	0.162	0.204	0.644	0.996	0.004 41	0.111	30.6	0.122
	ECFP4	5.46	0.18	0.187	0.249	0.701	5.33	0.32	2.37	37.2	0.395
	HDAD	0.0631	0.093	0.12	0.151	0.484	0.826	0.0029	0.116	36.8	0.0985
	HD	0.0937	0.121	0.156	0.198	0.523	0.956	0.0043	0.117	33.9	0.112
	MARAD	0.0741	0.137	0.165	0.217	0.66	1.03	0.003 95	0.101	34.6	0.130
	Mean	0.864	0.138	0.172	0.219	0.612	1.61	0.0496	0.442	35.1	
GG	MG	0.307	0.0867	0.103	0.146	0.382	0.288	0.021	0.148	13.0	0.0801
GC	MG	0.217	0.0766	0.0926	0.119	0.145	0.342	0.017	0.133	9.87	0.0565

Table 4: Slopes of the learning curves in Figs. S1 to S6, determined by a linear regression of the two models with largest training set size in each learning curve for all representations for all regressors and properties. The slopes are estimated under the assumption that the error asymptotically follow power-law decay βN^{α} with training set size (N) number of training samples, where α is the slope.

		U_0	$\varepsilon_{\mathrm{HOMO}}$	$\varepsilon_{\mathrm{LUMO}}$	$\Delta\varepsilon$	μ	α	ZPVE	C_{v}	ω_1
EN	CM	−0.04	−0.01	0.03	0.0	0.01	0.0	0.01	0.02	0.0
	BOB	−0.08	−0.01	0.0	−0.01	0.03	−0.01	−0.02	−0.03	−0.03
	BAML	0.06	0.05	−0.05	−0.04	0.05	0.36	0.04	0.06	−0.03
	ECFP4	−0.03	−0.01	−0.04	−0.02	−0.01	−0.03	−0.05	0.01	−0.03
	HDAD	0.01	−0.08	−0.05	−0.05	−0.03	−0.07	0.03	0.02	0.0
	HD	0.01	−0.01	−0.01	0.01	−0.02	0.01	0.01	0.02	−0.01
	MARAD	−0.01	0.01	−0.04	−0.02	−0.03	0.03	0.0	−0.01	−0.02
BR	CM	−0.04	−0.01	0.03	0.0	0.01	−0.01	0.0	0.02	0.0
	BOB	−0.07	0.0	0.0	−0.02	0.03	0.0	−0.03	−0.04	−0.05
	BAML	0.05	0.36	0.05	−0.05	−0.04	0.04	0.06	0.06	−0.03
	ECFP4	−0.01	−0.03	−0.01	−0.04	−0.02	−0.05	−0.03	0.01	−0.03
	HDAD	−0.03	−0.07	−0.08	−0.05	−0.05	0.03	0.01	0.02	0.00
	HD	0.01	−0.01	−0.02	0.0	−0.02	0.0	0.03	0.02	−0.02
	MARAD	0.38	−0.02	−0.02	−0.03	−0.02	−0.02	0.14	0.3	−0.04
RF	CM	−0.35	−0.18	−0.21	−0.27	−0.07	−0.2	−0.2	−0.26	−0.31
	BOB	−0.48	−0.24	−0.3	−0.26	−0.08	−0.27	−0.43	−0.45	−0.19
	BAML	−0.49	−0.29	−0.3	−0.3	−0.08	−0.33	−0.44	−0.48	−0.26
	ECFP4	−0.08	−0.18	−0.31	−0.28	−0.07	−0.03	−0.12	−0.03	−0.2
	HDAD	−0.19	−0.26	−0.26	−0.32	−0.08	−0.15	−0.29	−0.18	−0.29
	HD	−0.2	−0.25	−0.29	−0.27	−0.08	−0.14	−0.32	−0.17	−0.33
	MARAD	−0.57	−0.19	−0.24	−0.29	−0.03	−0.18	−0.3	−0.4	−0.25
KRR	CM	−0.36	−0.25	−0.33	−0.33	−0.22	−0.36	−0.36	−0.39	−0.26
	BOB	−0.36	−0.26	−0.27	−0.29	−0.21	−0.34	−0.25	−0.27	−0.43
	BAML	−0.26	−0.19	−0.22	−0.2	−0.16	−0.41	−0.14	−0.16	−0.41
	ECFP4	0.9	−0.17	−0.28	−0.2	−0.06	0.72	0.51	0.95	−0.22
	HDAD	−0.44	−0.38	−0.4	−0.4	−0.25	−0.48	−0.22	−0.24	−0.39
	HD	−0.17	−0.29	−0.34	−0.3	−0.21	−0.19	−0.14	−0.15	−0.36
	MARAD	−0.11	−0.1	−0.09	−0.09	−0.06	−0.14	−0.07	−0.08	−0.14
GG	MG	−0.36	−0.32	−0.35	−0.34	−0.26	−0.31	−0.23	−0.35	−0.36
GC	MG	−0.33	−0.38	−0.33	−0.36	−0.4	−0.66	−0.3	−0.31	−0.64

Table 5: Offsets of the learning curves in Figs. S1 to S6, determined by a linear regression of the two models with largest training set size in each learning curve for all representations for all regressors and properties. The slopes are estimated under the assumption that the error asymptotically follow power-law decay βN^{α} with training set size (N) number of training samples, where c would be the offset.

		U_0	$\varepsilon_{\mathrm{HOMO}}$	$\varepsilon_{\mathrm{LUMO}}$	$\Delta\varepsilon$	μ	α	ZPVE	C_{v}	ω_1
		eV	eV	eV	eV	Debye	Bohr^3	eV	cal/molK	cm^{-1}
EN	CM	1.41	0.379	0.421	0.701	0.757	1.39	0.0246	0.753	139
	BOB	1.59	0.333	0.548	0.73	0.553	1.38	0.0288	0.952	122
	BAML	0.0512	0.0573	0.27	0.283	0.431	0.0126	0.0131	0.167	92.2
	ECFP4	5.72	0.242	0.555	0.456	0.812	5.28	0.48	1.47	121
	HDAD	0.105	0.347	0.367	0.387	0.701	0.797	0.0208	0.135	89.6
	HD	0.165	0.217	0.352	0.336	0.915	0.581	0.008 86	0.148	121
	MARAD	0.206	0.194	0.463	0.472	0.95	0.479	0.0083	0.239	132
BR	CM	1.4	0.38	0.423	0.705	0.746	1.5	0.0251	0.749	138
	BOB	1.25	0.291	0.545	0.749	0.531	1.19	0.0309	1.04	139
	BAML	0.102	0.105	0.5	0.554	0.4	0.0114	0.008 35	0.222	87
	ECFP4	5.43	0.244	0.547	0.456	0.796	5.03	0.478	1.41	120
	HDAD	0.0568	0.366	0.409	0.477	0.779	0.94	0.002 24	0.0637	96.8
	HD	0.145	0.221	0.368	0.369	0.917	0.634	0.004 88	0.16	124
	MARAD	0.002 05	0.234	0.337	0.463	0.818	0.697	0.001 73	0.005 93	157
RF	CM	26.9	1.73	3.35	8.66	1.31	10.8	0.212	16.7	470
	BOB	57.1	2.03	4.45	3.51	1.11	13.8	1.59	82.9	31.5
	BAML	59.8	3.17	4.09	4.83	1.14	28.4	2.19	121.0	58.2
	ECFP4	9.31	1.16	5.28	4.16	1.04	4.96	0.994	2.3	154
	HDAD	13.6	2.51	2.68	6.55	1.12	9.93	1.61	7.15	98.0
	HD	14.4	2.34	4.13	3.73	1.22	8.3	1.99	6.64	188
	MARAD	155.0	1.56	3.79	9.28	0.871	5.53	0.328	33.1	353
KRR	CM	8.57	2.55	8.17	11.2	6.13	29.7	0.312	10.7	739
	BOB	4.48	2.09	2.83	4.42	5.04	15.3	0.0643	2.22	1950
	BAML	1.08	0.868	1.65	1.49	3.07	35.3	0.0173	0.501	2320
	ECFP4	0.000 115	0.5918	3.39	1.86	1.01	0.000 943	0.000 622	0.000 028 3	355
	HDAD	4.22	5.39	9.14	11.4	6.05	50.1	0.0242	0.765	2080
	HD	0.483	2.49	5.96	4.87	4.21	2.63	0.016	0.466	1400
	MARAD	0.202	0.347	0.36	0.492	0.927	1.74	0.006 48	0.194	112
GG	MG	2.95	2.36	3.67	4.85	5.0	5.88	0.0642	5.01	411
GC	MG	6.83	4.55	3.03	6.08	11.2	521.0	0.312	3.65	8140