Chemical diversity in molecular orbital energy predictions with kernel ridge regression

Annika Stuke,[1] Milica Todorović,[1] Matthias Rupp,[2] Christian Kunkel,[3, 1]
Kunal Ghosh,[1, 4] Lauri Himanen,[1] and Patrick Rinke[1, 3]

[1]*Department of Applied Physics, Aalto University, P.O. Box 11100, Aalto FI-00076, Finland*[*]
[2]*Fritz Haber Institute of the Max Planck Society, Faradayweg 4-6, 14195 Berlin, Germany*
[3]*Chair for Theoretical Chemistry and Catalysis Research Center,*
Technische Universität München, Lichtenbergstr. 4, 85747 Garching, Germany
[4]*Department of Computer Science, Aalto University, P.O. Box 15400, Aaalto FI-00076, Finland*

(Dated: March 27, 2019)

Instant machine learning predictions of molecular properties are desirable for materials design, but the predictive power of the methodology is mainly tested on well-known benchmark datasets. Here, we investigate the performance of machine learning with kernel ridge regression (KRR) for the prediction of molecular orbital energies on three large datasets: the standard QM9 small organic molecules set, amino acid and dipeptide conformers, and organic crystal-forming molecules extracted from the Cambridge Structural Database. We focus on prediction of highest occupied molecular orbital (HOMO) energies, computed at density-functional level of theory. Two different representations that encode molecular structure are compared: the Coulomb matrix (CM) and the many-body tensor representation (MBTR). We find that KRR performance depends significantly on the chemistry of the underlying dataset and that the MBTR is superior to the CM, predicting HOMO energies with a mean absolute error as low as 0.09 eV. To demonstrate the power of our machine learning method, we apply our model to structures of 10k previously unseen molecules. We gain instant energy predictions that allow us to identify interesting molecules for future applications.

I. INTRODUCTION

Machine learning (ML) in molecular and materials science has gained increased attention in the last decade and its application domain is widening continuously.[1] Applications include the search for improved and novel materials, [2, 3] computational drug design, [4] battery development, [5, 6] identification of new molecules for organic light-emitting diodes [7] or catalyst advancements for greenhouse gas conversion. [8, 9] In the context of *ab initio* molecular science, ML has been applied to learn a variety of molecular properties such as atomization energies, [10–19] polarizabilities, [12, 14, 16, 19, 20] electron ionization energies and affinities, [12, 16, 19, 60] dipole moments, [12–14, 20–23] enthalpies, [12, 13, 18] band gaps, [12–14, 20], binding energies on surfaces [59] as well as heat capacities. [12–14] A few studies addressed the prediction of spectroscopically relevant observables, such as electronic excitations, [12, 16, 19, 24] ionization potentials, [12, 16, 19, 25] nuclear chemical shifts, [26] atomic core level excitations [26] or forces on atoms. [26] In this paper, we employ kernel ridge regression (KRR) to predict the energy of the highest occupied molecular orbital (HOMO), which is of particular current interest for the development of new substances and materials. Electronic devices based on organic compounds are widely used in the technological industry and frontier orbital energies give important information about the optoelectronic properties of possible candidate materials.

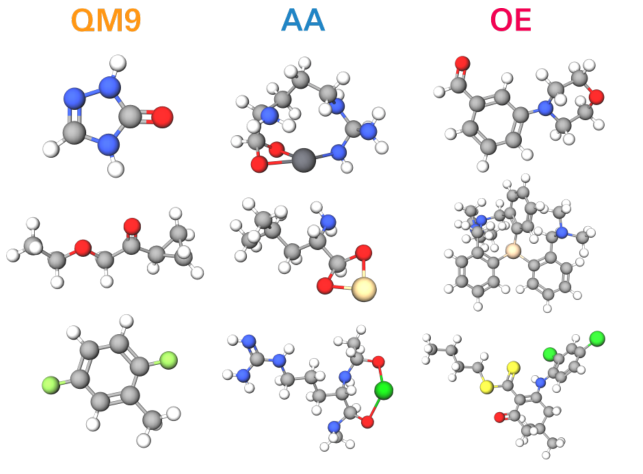

FIG. 1: Example molecules taken from the three datasets used in this work (see Appendix A for chemical names). Depicted elements are H (white), C (grey), N (blue), F (light green), O (red), Pb (dark grey), Cd (gold), Ba (dark green), Si (bronze), Cl (green) and S (yellow).

While the majority of ML studies focuses on ground state properties, a handful of studies have addressed the prediction frontier orbital energies, for example using neural networks, [13, 19, 20, 27, 60] random forest models, [28] or kernel ridge regression. [12, 14, 16, 17] To our knowledge, the best prediction accuracy reported for HOMO energy predictions on the well-known QM9 dataset [29] of small organic molecules was achieved with a deep neural network and featured a mean absolute error (MAE) of 0.041 eV. [20] QM9 was also employed by many other studies to explore the effect of molecular descriptors on prediction accuracy. [11–13, 20]

[*] annika.stuke@aalto.fi

While these results are quantitative and valuable, it is not clear to what extent the predictive power of the employed methodology is transferable to other molecular datasets. Motivated by optoelectronic applications, we are interested in HOMO predictions for large optically-active molecules with complex aromatic backbones and diverse functional groups which differ notably from the QM9 dataset.

We here employ KRR on three different datasets, two of which have not yet been used for the prediction of molecular orbital energies with ML. These two less-known datasets consist of 44 k conformers of proteinogenic amino acids from a public database of oligo-peptide structures [30] and 64 k organic molecules extracted from the organic crystals of the Cambridge Structural Database.[31, 48] In addition, we also use the well-known QM9 benchmark database of 134 k small organic molecules [29, 32] as a third dataset to compare results with previous studies in this field on a common basis. For all three datasets, we have calculated reference HOMO energies with density functional theory (DFT).

Moreover, we compare the performance of two different molecular representations: the well-studied Coulomb matrix (CM), which is simple, easy to compute and yields fast and inexpensive ML predictions. The second is a constant-size representation recently introduced by one of us [33] that relies on interatomic many-body functions including bonding and angular terms, called the many-body tensor representation (MBTR). While previous studies already demonstrated that the CM can easily be outperformed by more sophisticated molecular descriptors [12, 13, 16], we aim to analyze the degree of accuracy that can be achieved with the simple and cheap CM in comparison to the costlier MBTR.

The primary goal of our study is the comparison of KRR performance across three datasets with different chemical diversity. We show that the accuracy of HOMO energy predictions with KRR depends – besides the choice of molecular representation – crucially on the chemistry of the underlying dataset. Our measurable acceptable accuracy for HOMO energy predictions is 0.1 eV. Experiments for HOMO energy determination typically have a resolution of several tenth of eV, and prediction errors of state-of-the-art theoretical spectroscopy methods commonly range between 0.1 and 0.3 eV. We demonstrate how differences in model performance across chemically diverse settings can be related to certain dataset properties. Moreover, we quantify molecular orbital energy predictions that are presently available for realistic datasets of technological relevance.

Once trained, our KRR model can make instant HOMO energy predictions for numerous unknown molecules at no further cost. We demonstrate this by producing a spread of HOMO energy predictions for a new dataset of 10k organic molecules [18], whose original HOMO energies are unknown. With instant energy predictions for all 10k molecules we gain a rough estimate of the HOMO energy distribution for this dataset. We can further identify interesting molecules within a certain energy range for additional analysis. Hence, large numbers of new molecules, whose orbital energies have not yet been measured or computed, can be quickly screened for their usability in future applications. In this way, KRR can complement conventional theoretical and experimental methods to greatly accelerate the analysis of materials.

The manuscript is organized as follows: In Section II we introduce the three datasets used in this work. In Section III we briefly review the two descriptors that we employ in our ML framework, which is described in Section IV. In Section V we present our results and then discuss our findings in Section VI.

II. DATASETS

We train and evaluate our model on three different datasets, for which example molecules are depicted in Fig. 1. For all three datasets, we performed DFT calculations with the FHI-aims code. [34–37] We optimized the atomic structure of all molecules using the Perdew-Burke-Ernzerhof (PBE) functional [38] including Tkatchenko-Scheffler van der Waals corrections (PBE+vdW), [39] tight computational settings and the tier 2 basis sets of FHI-aims. For reasons of computational tractability, we also calculate HOMO energies with DFT, by taking the eigenvalue of the highest occupied molecular state from the PBE+vdW calculation.

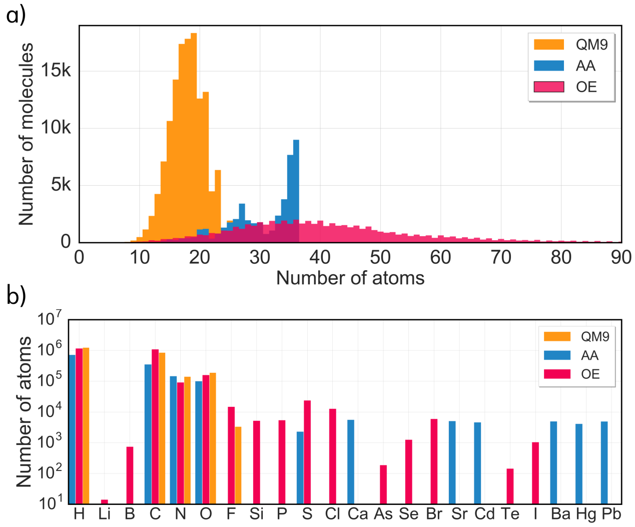

FIG. 2: Distribution of a) molecular size (including H atoms) and b) element type within QM9 (orange), AA (blue) and OE (red). OE molecules reach a size of up to 174 atoms. Since the occurence of large molecules with more than 90 atoms is very low (they amount to 1.4% of the entire dataset), we cut off the distribution at 90 atoms to facilitate a better comparison with the other two datasets.

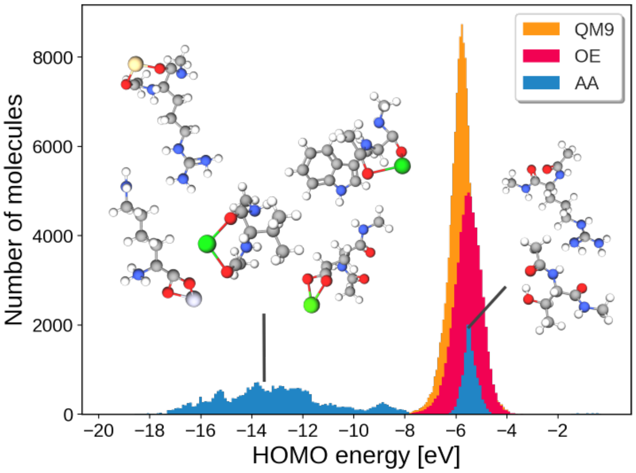

FIG. 3: Distributions of pre-computed HOMO energies for all three datasets. In QM9 and OE, HOMO energies are centered around -6 eV, while most HOMO energies in AA are distributed over a wider negative range. Shown as an inset are example molecules from the AA dataset and their location in the HOMO energy distribution. The HOMO energies in the deep negative range correspond to amino acids and dipeptides with one of six different metal cation additions: Ca^{2+}, Ba^{2+}, Sr^{2+}, Cd^{2+}, Pb^{2+} or Hg^{2+}. The HOMO energies centered around -6 eV correspond to bare amino acids and dipeptides.

Dataset	Mean [eV]	Std. dev. [eV]	Min [eV]	Max [eV]
QM9	-5.77	0.52	-10.45	-2.52
AA	-10.85	3.97	-19.63	-0.06
OE	-5.49	0.57	-12.70	-2.73

TABLE I: Mean value, standard deviation and ranges of DFT-computed HOMO energies for all three datasets as depicted in Fig. 3. QM9 and OE have similar mean values, ranges and standard deviations of HOMO energies, while the HOMO energies of the AA dataset spread out over a wide range of values due to metal cations with large atomic number that are attached to the amino acids and dipeptides.

Although DFT Kohn-Sham energies have limited accuracy, they provided us with a large, convenient and consistent dataset for developing our methodology. In the future, we will extend our study to HOMO energies computed with the more appropriate *GW* method.[40, 41] However, at present it is not possible to calculate hundreds of thousands of molecules with the *GW* method with reasonable computational resources. In the following, we describe the datasets in more detail.

II.1. QM9: 134 k small organic compounds

This dataset is extracted from the QM9 database and consists of 133,814 small organic molecules with up to 9 heavy atoms made up of C, N, O and F atoms [29]. It contains small amino acids and nucleobases, pharmaceutically relevant organic building blocks, for a total of 621 stoichiometries. The QM9 database has been used in a variety of ML studies [12–16, 18, 20, 24, 42–45] and has become the *drosophila* of ML in chemistry.

II.2. AA: 44 k amino acids and dipeptides

This dataset, denoted AA, contains 44,004 isolated and cation-coordinated conformers of 20 proteinogenic amino acids and their amino-methylated and acetylated (capped) dipeptides. [30] The molecular structures are made of up to 39 atoms including H, C, N, O, S, Ca, Sr, Cd, Ba, Hg and Pb. The amino acid conformers reveal different protonation states of the backbone and the sidechains. Furthermore, amino acids and dipeptides with divalent cations (Ca^{2+}, Ba^{2+}, Sr^{2+}, Cd^{2+}, Pb^{2+}, and Hg^{2+}) are included. Since all amino acids share a common backbone the complexity of this dataset lies in differing sidechains and differing dihedral angles. AA has been used to benchmark several ML models [15, 17, 46] and clustering techniques. [47]

II.3. OE: 64 k opto-electronically active molecules

This dataset, referred to as OE, consists of 64,710 large organic molecules with up to 174 atoms extracted from organic crystals in the Cambridge Structural Database (CSD). [48] Schober *et al.* have screened the CSD for monomolecular organic crystals with the objective to identify organic semiconductors with high charge carrier mobility. [31, 57, 58] For this study, we extracted molecules from the crystals and relaxed them in vacuum with the aforementioned computational parameters. The OE dataset is not yet publicly available. OE offers the largest chemical diversity among the sets in this work both in terms of size as well as number of different elements (Fig. 2). It contains the 16 different element types H, Li, B, C, N, O, F, Si, P, S, Cl, As, Se, Br, Te and I. The electronic structures are more complex than in QM9 and AA, containing, e.g. large conjugated systems and unusual functional groups.

II.4. Comparison of datasets

Figure 2 illustrates the chemical diversity present in our datasets. The molecular size distribution of the three datasets in Fig. 2 shows that QM9 and AA both contain molecules of similar sizes. The QM9 distribution exhibits a

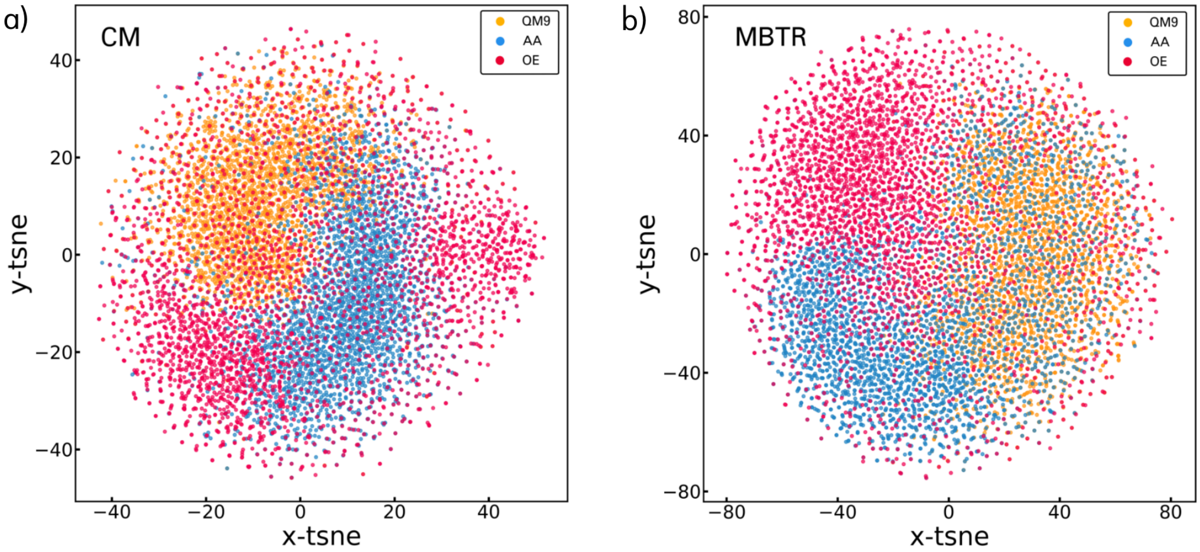

FIG. 4: t-SNE analysis of the three datasets, where molecules are represented by a) CM and b) MBTR. OE molecules are widely spread out in both dimensions, while AA and QM9 molecules form their own groups. The t-SNE algorithm was run on 9,000 randomly sampled molecules, i.e. 3,000 molecules from each dataset.

distinct peak at around 18 atoms, whereas AA has a bimodal distribution centered around 27 and 35 atoms. Conversely, the size distribution of OE is much broader, extending to molecules with as many as 174 atoms. In terms of element diversity, all datasets overlap on the 4 elements of organic chemistry: H, C, O, and N, as illustrated in Fig. 2. QM9 contains only F in addition, whereas AA branches out into common cations. OE offers the largest element diversity, including common semiconductor elements.

To compare our target property across the different datasets, we show distributions of the DFT pre-computed HOMO energies for each dataset in Figure 3. HOMO energies in QM9 and OE are centered around -6 eV, with a standard deviation of 0.5 eV and 0.6 eV, respectively. For AA, only a fraction of the HOMO energies are centered around -6 eV, while most HOMO energies are distributed over a wider range between -19.6 eV and -8 eV. The HOMO energies around -6 eV correspond to amino acids and dipeptides in bare organic configurations (free of cations). The HOMO energies between -19 and -8 eV correspond to amino acids and dipeptides with one of the six cations Ca^{2+}, Ba^{2+}, Sr^{2+}, Cd^{2+}, Pb^{2+} or Hg^{2+}. The metal ions in AA shift the HOMO energies of the amino acids towards lower values.

Judging by the distributions of molecular size and element types in Fig. 2 and by the distributions of HOMO energies in Fig. 3, we expect QM9 to be learned relatively easily. QM9 is clustered both in chemical space (in terms of molecular size and element types) and in target space (in terms of HOMO energies). We therefore expect KRR to benefit from mapping similar input structures to similar target values. For OE, the target energy distribution is similar to QM9, but the molecular structures are widely spread through chemical space. Similarity is therefore present only in one space, while the other is diverse. It might be challening for KRR to map from the diverse chemical space of OE to its confined target space and we expect learning to be slow. The AA set is spread out both in chemical space and target space. As it turns out, this will not be a problem for learning.

In Fig. 4 the three datasets are visualized with the t-distributed stochastic neighbor embedding (t-SNE) dimensionality reduction technique for both the their CM and MBTR molecular representations. [49] From each dataset, we randomly picked 3,000 molecules, which are mapped to a two dimensional space by t-SNE. Based on the pairwise similarities between descriptors, we can identify patterns within the datasets. In Fig. 4a), we can see that the AA and QM9 sets form distinct clusters and that the two clusters (orange and blue) have almost no overlap in the CM representation. The MBTR t-SNE analysis produces the same result, as indicated by Fig. 4b) (cluster orientation is arbitrary). This implies that the molecules in QM9 are very different from those in AA and that we increase the chemical diversity of our study by including the AA dataset. Fig. 4 further illustrates that the OE set is maximally diverse in itself. The corresponding red point cloud is evenly spread over the whole figure. In contrast to the QM9 and AA sets, OE offers a variety of rigid backbones (also denoted scaffolds) to which different functional groups are attached. Chemical diversity in part arises due to the rich combinatorial space emerging from scaffold-functional group pairings. For example, OE contains molecules with conjugated

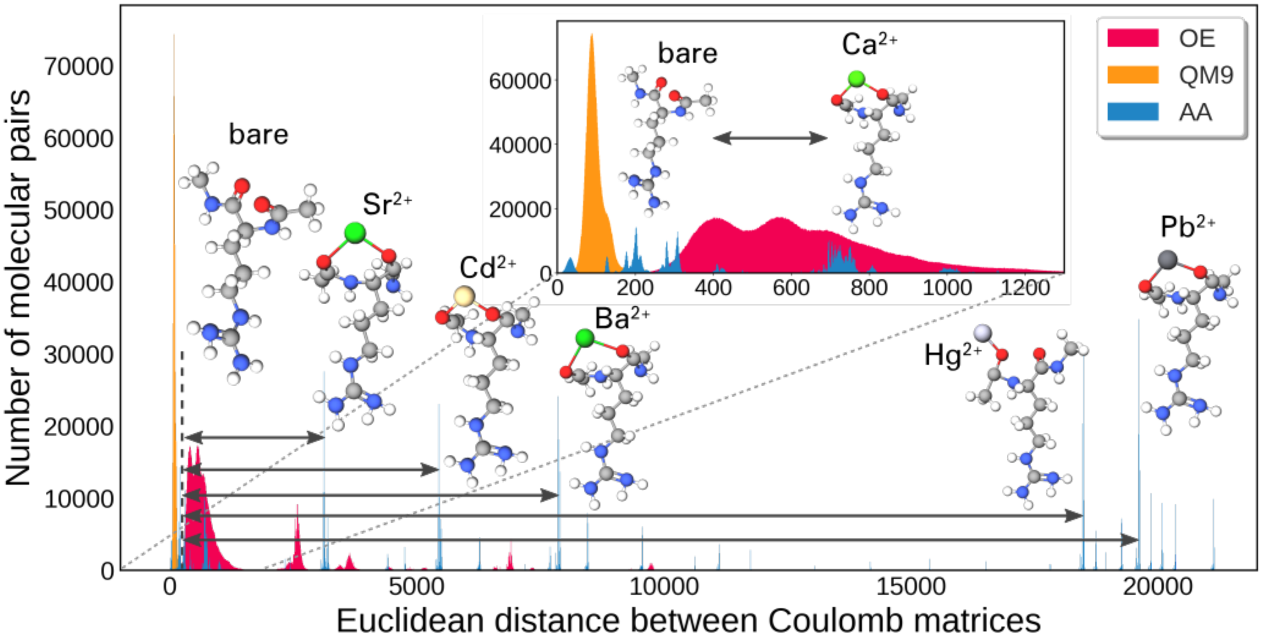

FIG. 5: Distributions of pairwise Euclidean distances computed for 3,000 randomly chosen molecules within each dataset. Molecules are represented by the CM. Molecular distances within QM9 are centered around small values, while OE distances are distributed evenly over a larger and wider range. The AA distribution consists of separate clusters, where each cluster corresponds to distances between amino acids with different metal ions attached to them. The inset shows an enlarged view of distances between 0 and 1300, where most of the distances within QM9 and OE are distributed. The two molecules are taken from the AA dataset and indicate the distance between bare structures and structures with a Ca^{2+} ion attached.

and aromatic backbones and electron withdrawing and donating functional groups that are of technological relevance and completely absent from QM9 and AA [57, 58].

To quantify the similarity between molecules in our three datasets, we computed Euclidean distances for molecular pairs – represented by the CM – for 3,000 randomly chosen molecules for each dataset, as shown in Fig. 5. Molecular distances within QM9 are small, indicating great similarity among QM9 molecules. The OE distances are distributed evenly over a larger and wider range of distances, indicating high dissimilarity among OE molecules. The AA distance distribution has several separated clusters up to very large distances. The distances are separated and ordered by the atomic number of the cations, where each cluster corresponds to amino acids with different metal ions attached. The first cluster includes distances from bare amino acids to amino acids with Ca^{2+}, followed by clusters including distances to amino acids with Sr^{2+}, Cd^{2+}, Ba^{2+}, Hg^{2+} and Pb^{2+}. Structural dissimilarity in the AA dataset mainly arises due to amino acids with different metal ions, while amino acids within the same cluster are highly similar to each other.

III. MOLECULAR REPRESENTATION

For the ML model to make accurate predictions, it is important to represent the molecules for the machine in an appropriate way. [33, 50–52] Cartesian (x,y,z)-coordinates, which are, for example, used for DFT calculations, are not applicable, since they are not invariant to translations, rotations, and reordering of atoms. In this work, we compare the performance of two different molecular representations.

III.1. Coulomb matrix (CM)

In the CM formalism, [11] each molecule is represented by a matrix C,

$$C_{ij} = \begin{cases} 0.5 Z_i^{2.4} & \text{if } i = j \\ \frac{Z_i Z_j}{\|\mathbf{R_i} - \mathbf{R_j}\|} & \text{if } i \neq j \end{cases} . \tag{1}$$

The CM encodes nuclear charges Z_i and corresponding Cartesian coordinates R_i, with off-diagonal elements representing Coulomb repulsion between atom pairs and diagonal elements encoding a polynomial fit of free-atom energies to Z. An example of the CM for a molecule of OE is shown in Fig. 6. To enforce permutational invariance, we simultaneously sort rows and columns of all CMs with respect to their ℓ^2-norm.

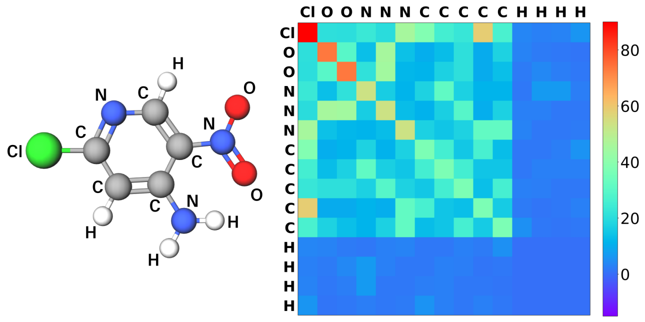

FIG. 6: Coulomb matrix representation (right) of the molecule 2-chloro-5-nitropyridin-4-amine (left) taken from OE.

III.2. Many-body tensor representation (MBTR)

The MBTR [33] can be viewed as a many-body expansion of the Bag of Bonds (BoB) representation, [53] which in turn is based on elements of the CM. One-body terms of the MBTR describe the atom types that are present in the molecule. Two-body terms encode inverse distances between all pairs of atoms (bonded and non-bonded), separately for each combination of atom types. The inverse distances are sorted by increasing order and broadened into a continuous Gaussian distribution of inverse distances as shown in Fig. 7a). Three-body terms encode angle distributions for any triplets of atoms present in the molecule, as shown in Fig. 7b). Each N-body term has a broadening parameter (in total ρ_1, ρ_2 and ρ_3) that controls the smearing of atom type distribution, inverse distance distribution and angle distribution, respectively, and need to be fine-tuned for optimal KRR performance. We use the *DScribe* package [54] to compute the MBTR and the *qmmlpack* package to refine the MBTR hyperparameters for small training set sizes of up to 4k molecules. Exponential weighting was employed for the computation of inverse distance and angle terms. Here, we apply only two-body and three-body terms in the MBTR, since we found that the inclusion of one-body terms does not improve the performance, but increases computational time (we refer to Appendix B for details).

IV. MACHINE LEARNING METHOD

IV.1. Kernel ridge regression

We employ kernel ridge regression [55] (KRR) to model the relationship between molecular structures and HOMO energies. In KRR, training samples are mapped into a high-dimensional space using nonlinear mapping, and the structure-HOMO relationship is learned in the high-dimensional space. This learning procedure is conducted implicitly by defining a kernel function, which measures the similarity of training samples in the high-dimensional space by employing a kernel function. In this work, we use two different kernel functions. The first kernel is the Gaussian kernel

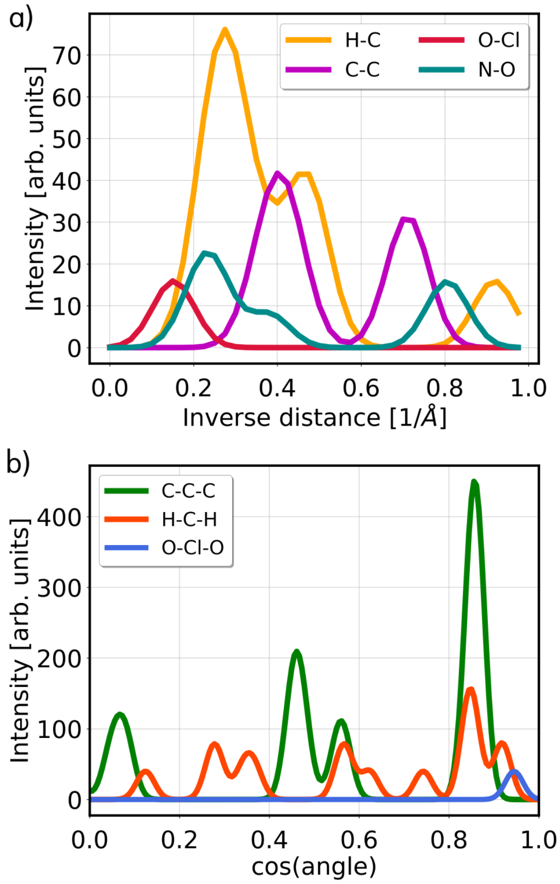

FIG. 7: Many-body tensor representation of the molecule from Fig. 6 (2-chloro-5-nitropyridin-4-amine). a) Inverse distance distribution of selected two-body terms with optimized broadening parameter ρ_2. b) Angle distribution of selected three-body terms with optimized broadening parameter ρ_3.

$$k_{\text{Gaussian}}(\boldsymbol{x}, \boldsymbol{x}') = e^{-\frac{||\boldsymbol{x}-\boldsymbol{x}'||_2^2}{2\sigma^2}}, \tag{2}$$

where

$$||\boldsymbol{x}-\boldsymbol{x}'||_2 = \sqrt{\sum_i |\boldsymbol{x_i}-\boldsymbol{x}'_{\boldsymbol{i}}|^2} \tag{3}$$

is the Euclidean distance and $\boldsymbol{x}$, $\boldsymbol{x}'$ are two training molecules represented by either the CM or the MBTR. We note that the Euclidean distance distribution between molecular pairs shown in Fig. 5 gives us a direct insight into the learning process of a Gaussian kernel function. The second kernel is the Laplacian kernel

$$k_{\text{Laplacian}}(\boldsymbol{x}, \boldsymbol{x}') = e^{-\frac{||\boldsymbol{x}-\boldsymbol{x}'||_1}{\sigma}}. \tag{4}$$

which uses the 1-norm as similarity measure,

$$||\boldsymbol{x}-\boldsymbol{x}'||_1 = \sum_i |\boldsymbol{x_i}-\boldsymbol{x}'_{\boldsymbol{i}}|. \tag{5}$$

In Eqs. (2) and (4), σ is the kernel width.

In the KRR training phase with N training molecules, the goal is to find a vector $\boldsymbol{\alpha} \in \mathbb{R}^N$ of regression weights α_i that solves the minimization problem

$$\min_{\alpha} \sum_{i=1}^{N} (E^{\text{pred}}(\boldsymbol{x}_i) - E_i^{\text{ref}})^2 + \lambda \boldsymbol{\alpha}^T \mathbf{K} \boldsymbol{\alpha}, \tag{6}$$

where the analytic solution for $\boldsymbol{\alpha}$ is given by

$$\boldsymbol{\alpha} = (\mathbf{K} + \lambda \mathbf{I})^{-1} \mathbf{E}^{\text{ref}}. \tag{7}$$

The matrix $\mathbf{K}$ is the kernel matrix, whose elements represent inner products between training samples in the high-dimensional space, calculated as kernel evaluations $K_{i,j} := k(\boldsymbol{x_i}, \boldsymbol{x_j})$. The scalar λ is the regularization parameter, which penalizes complex models with large regression weights over simpler models with small regression weights. The expression E^{ref} denotes the reference HOMO energy computed by DFT and E^{pred} is the predicted HOMO energy, which is obtained as sum over weighted kernel functions

$$E^{\text{pred}}(\boldsymbol{x}) = \sum_{i=1}^{N} \alpha_i k(\boldsymbol{x}, \boldsymbol{x}_i). \tag{8}$$

The sum runs over all molecules $\boldsymbol{x}_i$ in the training set with their corresponding regression weights α_i. After training, predictions are made for test molecules that were not used to train the model, employing eq. (8) to predict out-of-sample molecules and estimate performance of the model. No further scaling or normalization of the data was done to preserve the meaning of the HOMO energies.

IV.2. KRR training, cross-validation and error evaluation

For each dataset QM9, AA and OE, we randomly selected a subset of 32 k molecules for training and a further 10 k molecules for out-of-sample testing. In order to obtain statistically meaningful results for the training and testing performance of KRR, we repeated the random selection of training set and test set nine more times for each dataset after the data was reshuffled. As a result, we acquired 10 different training sets of 32k molecules and 10 different test sets of 10k molecules for each dataset.

From each training set of 32k, we randomly drew 6 different subsets of sizes 1k, 2k, 4k, 8k and 16k, where smaller sets were always subsets of larger ones. For each of these subsets, we trained and cross-validated a KRR model, as described in the following paragraph. We then evaluated the KRR model on the corresponding test set of 10k by predicting HOMO energies for 10k out-of-sample molecules and by computing the MAE $\frac{1}{n}\sum_{i=1}^{n} |E_i^{\text{pred}} - E_i^{\text{ref}}|$ between predictions and DFT reference energies. For each dataset and for each training set size, we computed the average MAE value and its standard deviation across the 10 randomly drawn training and test sets. We then plotted all of these average MAEs as a function of training set size and as a result, attain one learning curve for each dataset, as shown in Fig. 8 and Fig. 11. We note that for the two molecular representations, CM and MBTR, identical training and test sets were used.

In the scope of this study, there are 3 types of hyperparameters:

(i) the kernel function (Gaussian type or Laplacian type)

(ii) MBTR hyperparameters (broadening values ρ_1, ρ_2 and ρ_3)

(iii) KRR hyperparameters (kernel width σ and regularization parameter λ)

The large number of possible values for the hyperparameters leads to a wide range of possible KRR models. The choice of kernel function depends on the molecular descriptor. The Gaussian kernel performs best on the MBTR, whose N-body terms themselves consist of Gaussian distributions. The Laplacian kernel, on the other hand, can better model piecewise smooth functions, such as discontinuities of the sorted CM. Therefore, we chose the Laplacian kernel for the CM and the Gaussian kernel for the MBTR. We provide learning curves in Appendix B to prove that we picked the optimal kernel for each descriptor.

The MBTR and KRR hyperparameters were simultaneously optimized in a cross-validated grid search for each of the three smaller training sets of 1k, 2k and 4k. In particular, MBTR hyperparameters ρ_1, ρ_2 and ρ_3 were varied on a grid of 10 points between 10^{-4} and 10^{-1} and KRR hyperparameters σ and λ were varied on a grid of 12 points between 10^{-12} and 10^{-1}. The aim was to find the combination of best KRR and MBTR parameters. For more details on the cross-validated grid search we refer to Appendix C. The optimized model was then evaluated on 10k out-of-sample molecules from the test set. Finally, the MAE on the test set was reported as a data point on the learning curve for a given training set size.

We found that the values of the optimized MBTR hyperparameters do not change throughout the three small training sets. We therefore fixed the optimized MBTR hyperparameters for the larger larger training sets of 8k, 16k and 32k and optimized only the KRR hyperparameters (on a grid from 10^{-12} to 10^{-1}).

Another measurent for the performance of a ML model is the R^2 coefficient, which describes the proportion of variability in a dataset that can be explained by the model. Although we do not use the R^2 coefficient to optimize KRR hyperparameters, we employ this metric to interpret how well our models fit the data.

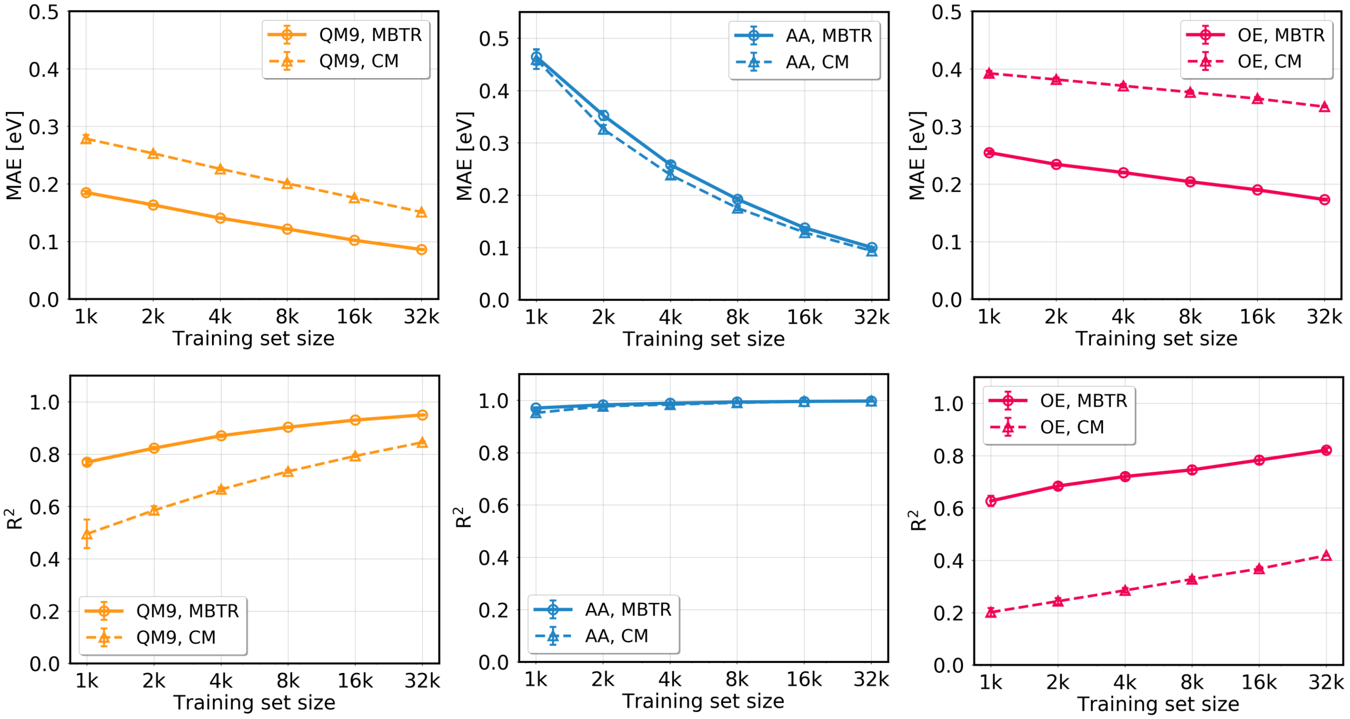

FIG. 8: Mean absolute errors (MAE) (upper panel) and squared correlation coefficient R^2 (lower panel) for out-of-sample predictions of the HOMO energy as a function of training set size for QM9 (orange), AA (blue) and OE (red). For each dataset, performances of MBTR (filled lines) and CM (dashed lines) are compared to each other. The Laplacian kernel was used in combination with the CM and the Gaussian kernel was used in combination with the MBTR. The MAE datapoint and its error bar correspond to the average value and its standard deviation computed over 10 instances of training, cross validating and out-of-sample testing.

	MAE [eV]		**RMSE [eV]**		**R^2**	
Dataset	**MBTR**	**CM**	**MBTR**	**CM**	**MBTR**	**CM**
QM9	0.086 ± 0.001	0.151 ± 0.001	0.118 ± 0.002	0.207 ± 0.003	0.950 ± 0.002	0.845 ± 0.004
AA	0.100 ± 0.001	0.094 ± 0.002	0.201 ± 0.006	0.194 ± 0.007	0.997 ± 0.001	0.998 ± 0.001
OE	0.173 ± 0.002	0.336 ± 0.003	0.239 ± 0.006	0.435 ± 0.005	0.821 ± 0.009	0.413 ± 0.005

TABLE II: Mean absolute errors (MAEs), root mean square errors (RMSEs) and R^2 coefficients for KRR predictions of molecules from the QM9, AA and OE datasets. Results are shown for the MBTR and CM representations. Errors and R^2 coefficients are measured on test sets of 10 k randomly selected out-of-sample molecules of each dataset, while 32 k molecules were used for training. We report errors and R^2 coefficients as the average over 10 repetitions, accompanied by the standard deviation of the mean. The corresponding reference-versus-predicted scatter plots are shown in Fig. 10 for the first run out of 10 repetitions.

V. RESULTS

Upper panels in Fig. 8 show out-of-sample MAEs as a function of training set size ("learning curves") for the different datasets and for CM and MBTR as descriptors. As expected, the MAE decreases for all datasets with increasing training set size (see also Fig. 11). The best MAEs for a training set size of 32k molecules are presented in Table II. The lowest MAE is achieved for QM9, closely followed by AA. In constrast, the MAEs for OE are approximately twice as high. The learning rate (slope of the MAE curves) is highest for AA and lowest for OE, and is independent of the descriptor. The MBTR performs significantly better than the CM for QM9 and OE, while for AA, the learning curves of the two descriptors are the same within statistical errors.

The lower panels of Fig. 8 show the squared correlation coefficient R^2 as a function of training set size. R^2 increases systematically with training set size, but its rate varies across the three datasets. For AA it is close to 1 already for small

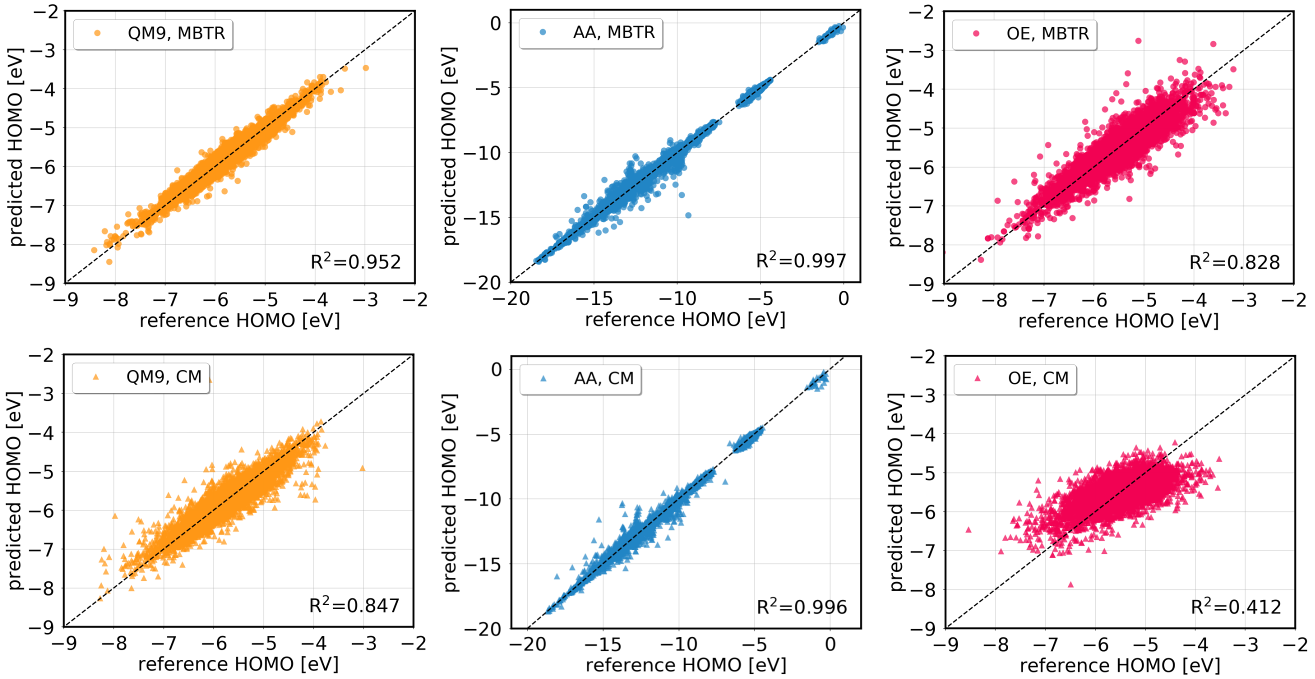

FIG. 9: Scatter plots of out-of-sample predictions of the HOMO energy on 10 k molecules of QM9 (orange), AA (blue) and OE (red). We show reference HOMO energies pre-calculated with DFT versus model predictions for the first run out of 10 repetitions. The training set size is 32 k. The upper panel shows results obtained with the MBTR and the lower panel shows results obtained with the CM. The Laplacian kernel was used in combination with the CM and the Gaussian kernel was used in combination with the MBTR.

training set sizes, whereas for QM9, R^2 starts off low and then approaches 1 for 32 k. For the OE dataset, R^2 steadily increases and reaches a value of 0.81 at 32 k.

Correlating R^2 and MAE, we observe that model fitting for OE is consistently poor (low R^2) and the prediction errors are consistently high (high MAEs). The CM appears to be less suitable for this prediction task.

Fig. 10 presents scatter plots for a fixed training set size of 32 k, which describes how well the KRR predictions correlate with the reference values in the test set. We observe the best correlation (i.e. predictive power) for AA and the worst for OE. The MBTR appears to have higher predictive power than CM for QM9 and OE, whereas MBTR and CM perform similarly for AA.

Fig. 11 summarizes the learning curves for the three datasets. The top panel shows the MBTR and the bottom panel the CM results. The MBTR-based KRR models generally produce faster learning rates and lower MAEs. The AA learning rate is particularly fast, and eventually leads to MAEs of 0.10 eV for MBTR and 0.09 eV for CM, which are comparable to those of QM9. The prediction quality for the OE dataset is notably worse in relation to the other datasets.

VI. DISCUSSION

VI.1. Dependence of KRR performance on dataset diversity

The learning success of KRR depends on the structural complexity of individual molecules (e.g. number of atoms, atom types, backbone types etc.) as well as on the diversity and redundancy within a dataset. Redundancy usually means that certain structural features occur frequently in a dataset, i.e. many data points are similar to each other. Diversity implies the opposite: Few instances in a dataset are similar to the rest. Redundant datasets are learned well with ML, even when trained on small portions. Diverse datasets can pose a problem for ML, even when applied to large data.

The differences in the learning curves we observe in Figs. 8 and 11 reflect the chemical differences in the three datasets employed in our study. QM9 is greatly redundant and includes molecules with simple structures. Therefore, it can be learned well even on small training set sizes, where redundancy is low.

The AA dataset has inbuilt redundancy, but also includes many different metal cations. For small training set sizes, where not enough similar structures per metal cation are present, the error is high. This situation then improves quickly with increasing traing set size. As a result, the learning rate of AA is faster than of QM9.

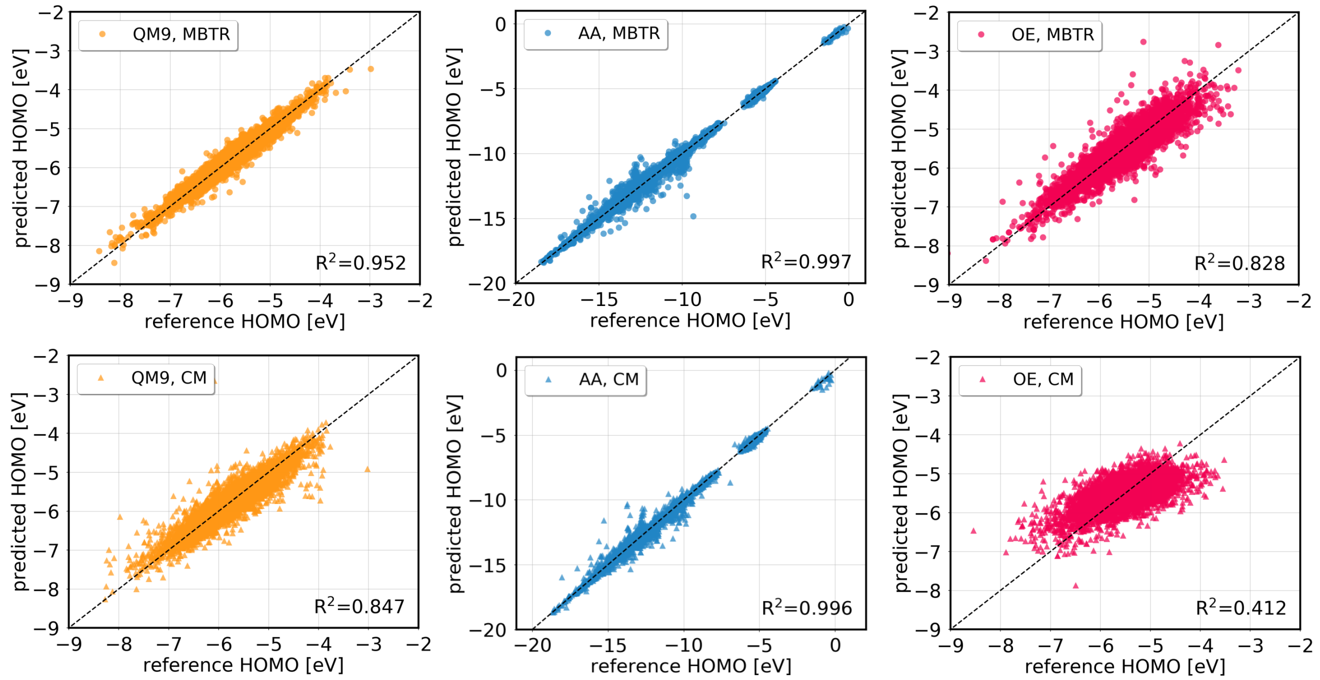

FIG. 10: Scatter plots of out-of-sample predictions of the HOMO energy on 10 k molecules of QM9 (orange), AA (blue) and OE (red). We show reference HOMO energies pre-calculated with DFT versus model predictions for the first run out of 10 repetitions. The training set size is 32 k. The upper panel shows results obtained with the MBTR and the lower panel shows results obtained with the CM. The Laplacian kernel was used in combination with the CM and the Gaussian kernel was used in combination with the MBTR.

OE is highly diverse and includes molecules with complex structures. It has similar chemical element range as AA with 16 different element types, but features larger molecules and more structural diversity, as discussed in Section II.4. The diversity explains the high errors throughout all training set sizes and the slower learning rate. The t-SNE analysis in Fig. 4 confirms that OE has little overlap with QM9, regardless of the molecular representation. OE is the most diverse of the three datasets, since the corresponding point cloud does not cluster in a particular region and instead fills the whole space of Fig. 4.

The KRR learning rate is slowest for OE, due to the aforementioned chemical and structural complexity in this dataset. At a training set size of 32 k, the MAE is still twice as high as for QM9 and AA. With 0.173 eV, the MAE for OE is still too high for the spectroscopic applications we intend, which typically require errors below 0.1 eV. Advanced machine learning methodology and more sophisticated materials descriptors both help to reduce overall prediction errors [13, 14, 20, 27, 56] and we may test alternative approaches in further work.

Lastly, our study of three different datasets illustrates that MAEs are not transferable across datasets, even if they are evaluated for the same machine learning method and the same descriptor. If we had based our predictive power expectations for the OE set on the KRR QM9 performance, we would have been disappointed to find much larger errors in reality. It is therefore paramount to further investigate the performance of machine learning methods across chemical space.

VI.2. Dependence of KRR performance on molecular descriptor

Next, we discuss the relative performance for the CM and MBTR molecular descriptors. Overall, MBTR outperforms the CM across the datasets, which is in line with previous findings. [12, 13, 16, 28] This is reasonable in light of the higher information content about atom types, their bond lenghts and angles encoded in the MBTR when compared to the CM matrix. The CM and the MBTR exhibit the same performance only for the AA dataset of molecular conformers, where complexity is dominated by the torsional angles, and bonding patterns are similar. This result is partly explained by the exclusion of torsional angle information into MBTR (four-body terms), while consistent chemical information in AA benefits the performance of the CM.

It is interesting that CM generally produces MAEs only twice as large as the MBTR with much smaller data structures and at a fraction of the computational cost. The CM representation is simple to compute, supplies benchmark results comparable to previous work and may prove a convenient tool for preliminary studies of large unknown datasets.

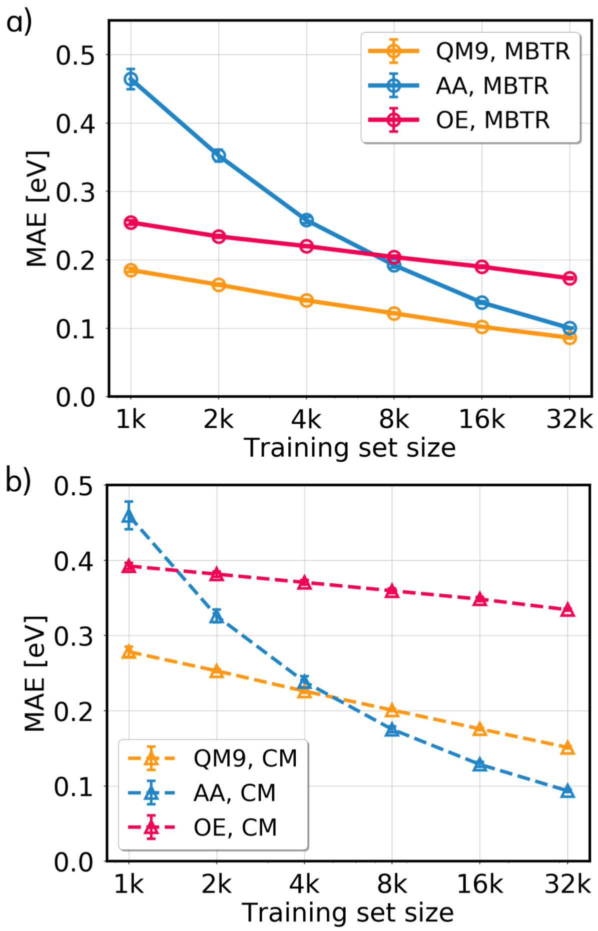

FIG. 11: Comparison of learning curves for QM9 (orange), AA (blue) and OE (red). KRR models are based on a) Gaussian kernel in combination with MBTR and b) Laplacian kernel in combination with CM. The error decay is shown on a logarithmic scale. The MAE datapoint and its error bar correspond to the average value and its standard deviation computed over 10 training instances.

VI.3. Application of QM9 model

Our results for the QM9 dataset allow us to compare our findings with previous studies. Given a KRR training set of HOMO energies for 32 k molecules, we obtained MAE values of 0.086 eV and 0.151 eV with the MBTR and CM descriptors respectively. In 2017, Faber *et al.* [13] reported KRR results on 118 k QM9 molecules, using a molecular descriptor based on interatomic many-body expansions including bonding, angular and higher-order terms, [12] which is comparable to the MBTR used in this work. The HOMO energy was predicted with an out-of-sample MAE of 0.095 eV, and the CM representation achieved an MAE of 0.133 eV. Our QM9 results are in very good agreement with this study, even if our KRR training set is much smaller. These errors are relatively small and comparable to experimental and computational errors in HOMO determinaton, which typically range in between several tenth of eV. Errors in HOMO energy predictions with machine learning may be further reduced by developing customized deep learning neural network architectures, which have been reported to produce an MAE of 0.041 eV after training on 110 k QM9 molecules. [20]

To showcase the value of our trained KRR model, we apply it to a dataset of 10k diastereomers of parent $C_7H_{10}O_2$ isomers. This dataset contains molecular structures, but no HOMO energies. Computing the HOMO energies with DFT would take considerable effort and time. With our KRR model – trained on 32k QM9 molecules represented by the MBTR – we gain an immediate overview of the HOMO energies that occur in the dataset. A histogram of all predicted HOMO energies is shown in Fig. 12. We can see that they are uniformly distributed between -6.8 and -3.2 eV for all diastereomers. The energetic scan allows us to quickly detect molecules of interest in a large collection of compounds. Individual molecules can be easily identified, for instance, the molecule with lowest HOMO energy, molecules with highest HOMO energy or those molecules with average HOMO energies. Various molecules of interest, e.g., structures with HOMO energies in a particular region, could subsequently be further investigated with first-principle methods or experiments to determine their functionality for certain applications. In this fashion, fast energy predictions of our KRR model can be analyzed for structures with desired HOMO energy.

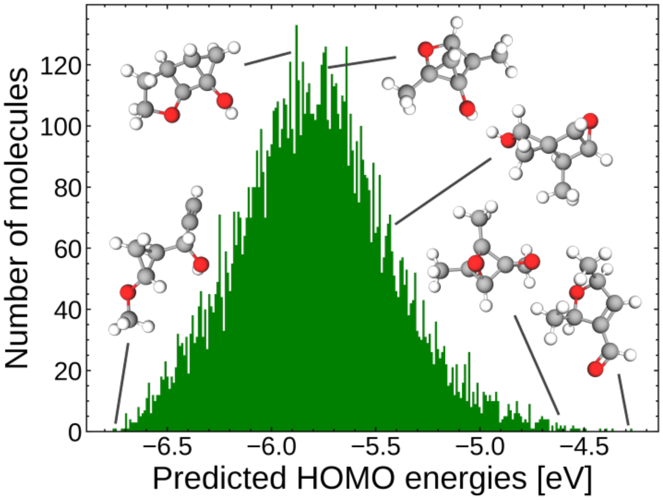

FIG. 12: HOMO energies predicted by our QM9-trained KRR model for a new dataset of 10k diastereomers, for which only molecular structures, but no pre-computed HOMO energies are available. Molecules that fall within a certain energy range can easily be identified and might be further assessed for potential applications, as illustrated for six example molecules.

VII. CONCLUSION

In this study, we have trained and tested KRR models on three molecular datasets with different chemical composition to predict molecular HOMO energies. Our comparison between two different molecular descriptors, the CM and the MBTR, shows that the MBTR outperforms the CM on both OE and QM9 due to its ability to encode complex information from the molecular structures. For AA we could find no significant difference in performance between the two representations.

Our work demonstrates that the predictive performance of KRR inherently depends on the complexity of the dataset it is applied to, in addition to the training set size and descriptor. Rapidly decreasing learning curves and low MAEs are achieved for QM9, which is known as a standard benchmark set for ML in molecular chemistry, containing pharmaceutically relevant compounds with rather simple bonding patterns. The same is true for AA which consists of a primitive and restricted collection of amino acids and peptides. The OE dataset, however, comprises large opto-electronic molecules with complicated electronic structures and unconventional functional groups, is much more difficult to learn. It yields almost flat learning curves and considerably higher MAEs. To further improve the predictive power for molecules of technological interest, such as the ones in the OE set, future work should focus on generating larger datasets, devising better descriptors or more sophisticated machine learning methods.

ACKNOWLEDGMENTS

We greatfully acknowledge the CSC-IT Center for Science, Finland, and the Aalto Science-IT project for generous computational resources. This study has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 676580 with The Novel Materials Discovery (NOMAD) Laboratory, a European Center of Excellence and from the Magnus Ehrnrooth Foundation. This work was furthermore supported by the Academy of Finland through its Centres of Excellence Programme 2015-2017 under project number 284621.

Appendix A: Chemical names of molecules

Figure 13 shows example molecules from QM9, AA and OE and their chemical names.

Appendix B: Choice of kernels and MBTR terms

In Fig. 14, we show learning curves for different parameter configurations of the CM and the MBTR. MAEs correspond to out-of-sample predictions made in a single experiment. We generated learning curves using the MBTR with one-, two- and three-body terms (k=[1,2,3]), two-and three-body terms (k=[2,3]) and only two-body-terms (k=2). For QM9 and OE, we the MBTR with k=[2,3] performs slightly better than the MBTR with k=[1,2,3], while for AA, the performance of k=[2,3] and k=[1,2,3] is equal for larger training set sizes. Therefore, we chose to employ the MBTR with two- and three body terms (k=[2,3]) in this study. Moreover, Fig. 14 reveals that, when the MBTR is used as molecular descriptor, the Gaussian kernel (GK) yields better results than the Laplacian kernel (LK). For the CM, on the other hand, the Laplacian kernel works better than the Gaussian kernel due to discontinuities of the sorted CM.

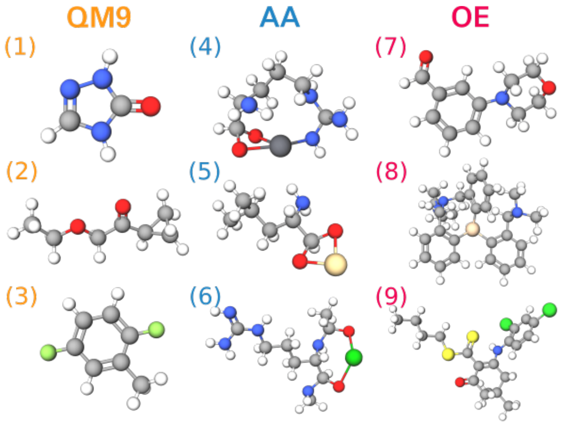

FIG. 13: Example molecules taken from the three datasets used in this work. Chemical names are (1) 1H-1,2,4-Triazol-3-ol, (2) 1-Cyclopropyl-2-ethoxyethanone, (3) 2,5-Difluorotoluene, (4) Arginine (uncapped), (5) Leucine (uncapped), (6) Ac-Arg-NMe (dipeptide), (7) 6-(4-Morpholinyl)-2-pyridinecarbaldehyde, (8) Tris2-[(dimethylamino)methyl]phenylsilane, (9) Butyl 2-[(2,4-dichlorophenyl)amino]-4,4-dimethyl-6-oxo-1-cyclohexene-1-carbodithioate. Depicted elements are H (white), C (grey), N (blue), F (light green), O (red), Pb (dark grey), Cd (gold), Ba (dark green), Si (bronze), Cl (green) and S (yellow).

Appendix C: Cross-validated grid search

For MBTR and KRR hyperparameter selection, we employed a 5-fold cross-validated grid search. In a 5-fold cross validation, a given *original training set* (1k, 2k, 4k, 8k, 16k or 32k) is shuffled randomly and split into 5 equally sized groups. One group (20% of the training set) is taken as a hold-out set for *validation*. The remaining groups (80% of the training set) are taken as *training data*. Then, a grid search is performed: Each possible combination of KRR and MBTR hyperparameter values is trained on the *training data*

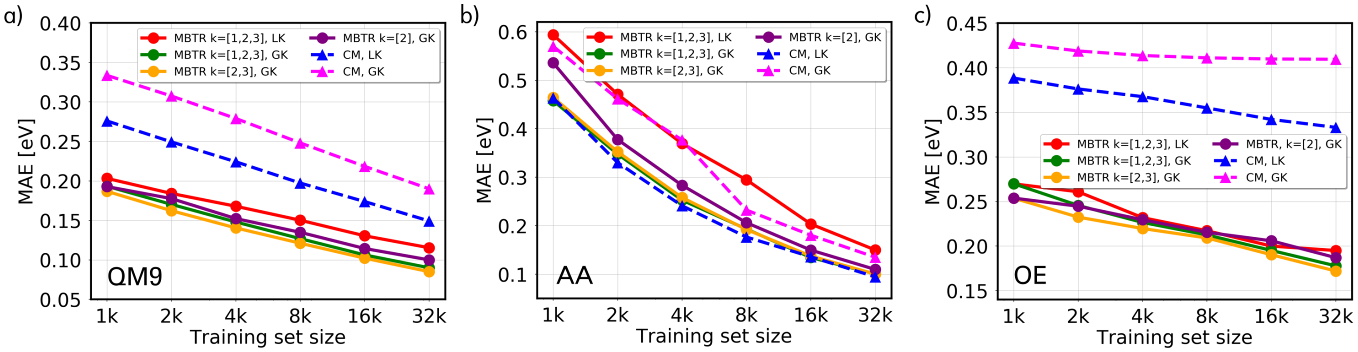

FIG. 14: Learning curves for different parameter configurations of the CM and the MBTR for the three datasets QM9, AA and OE. For the MBTR, the letter k refers to many-body terms, i.e. k=1 refers to one-body terms describing the atoms types that are present in a molecule, k=2 refers to two-body terms describing pairwise inverse distances between atoms and k=3 refers to three-body terms describing angle distributions for any triplets of atoms. Investigated combinations of many-body terms are one-, two- and three-body terms (k=[1,2,3]), two- and three-body terms (k=[2,3]) and only two-body terms. In addition, we consider two kernels, the Gaussian kernel (GK) and the Laplacian kernel (LK).

and evaluated on the *validation data*.

The assignment of the 5 groups into *training data* and *validation data* is repeated 5 times, until each group was used as *validation data* exactly once and used as *training data* exactly 4 times. As we repeat the process 5 times, we get 5 MAEs for each possible set of hyperparameters. We consider the average value over these 5 MAEs for each set of hyperparameters and choose the set with lowest average MAE. With the chosen set of optimal hyperparameters, we train the KRR model on the entire *original training set* (=*training data*+*validation data*). Finally, we evaluate the trained model on the test set of 10k out-of-sample molecules and report the MAE on the test set in the final learning curve for the *original training set* size.

[1] M. Rupp, O. A. von Lilienfeld, and K. Burke, "Guest editorial: Special topic on data-enabled theoretical chemistry," Journal of Chemical Physics **148**, 241401 (2018).

[2] T. Müller, A. G. Kusne, and R. Ramprasad, "Machine learning in materials science," in *Reviews in Computational Chemistry* (John Wiley & Sons, Ltd, 2016) Chap. 4, pp. 186–273, NoStop

[3] A. Zunger, "Inverse design in search of materials with target functionalities," Nature Reviews Chemistry **2**, 0121 EP – (2018), perspective.

[4] J. Ma, R. P. Sheridan, A. Liaw, G. E. Dahl, and V. Svetnik, "Deep neural nets as a method for quantitative structure–activity relationships," Journal of Chemical Information and Modeling **55**, 263–274 (2015), pMID: 25635324, https://doi.org/10.1021/ci500747n.

[5] A. D. Sendek, E. D. Cubuk, E. R. Antoniuk, G. Cheon, Y. Cui, and E. J. Reed, "Machine learning-assisted discovery of many new solid li-ion conducting materials," Tech. Rep. arXiv:1808.02470 [cond-mat.mtrl-sci] (ArXiV, 2018).

[6] M. A. Shandiz and R. Gauvin, "Application of machine learning methods for the prediction of crystal system of cathode materials in lithium-ion batteries," Computational Materials Science **117**, 270 – 278 (2016).

[7] R. Gómez-Bombarelli, J. Aguilera-Iparraguirre, T. D. Hirzel, D. Duvenaud, D. Maclaurin, M. A. Blood-Forsythe, H. S. Chae, M. Einzinger, D.-G. Ha, T. C.-C. Wu, G. Markopoulos, S. Jeon, H. Kang, H. Miyazaki, M. Numata, S. Kim, W. Huang, S. I. Hong, M. A. Baldo, R. P. Adams, and A. Aspuru-Guzik, "Design of efficient molecular organic light-emitting diodes by a high-throughput virtual screening and experimental approach." Nature materials **15 10**, 1120–7 (2016).

[8] B. R. Goldsmith, J. Esterhuizen, J.-X. Liu, C. J. Bartel, and C. Sutton, "Machine learning for heterogeneous catalyst design and discovery," AIChE Journal **64**, 2311–2323 (2018), https://onlinelibrary.wiley.com/doi/pdf/10.1002/aic.16198.

[9] B. Meyer, B. Sawatlon, S. Heinen, O. A. von Lilienfeld, and C. Corminboeuf, "Machine learning meets volcano plots: computational discovery of cross-coupling catalysts," Chem. Sci. **9**, 7069–7077 (2018).

[10] K. Hansen, G. Montavon, F. Biegler, S. Fazli, M. Rupp, M. Scheffler, O. A. von Lilienfeld, A. Tkatchenko, and K.-R. Müller, "Assessment and Validation of Machine Learning Methods for Predicting Molecular Atomization Energies," Journal of Chemical Theory and Computation **9**, 3404–3419 (2013).

[11] M. Rupp, A. Tkatchenko, K.-R. Müller, and O. A. von Lilienfeld, "Fast and Accurate Modeling of Molecular Atomization Energies with Machine Learning," Physical Review Letters **108**, 058301 (2012).

[12] B. Huang and O. A. von Lilienfeld, "Communication: Understanding molecular representations in machine learning: The role of uniqueness and target similarity," The Journal of Chemical Physics **145**, 161102 (2016), https://doi.org/10.1063/1.4964627.

[13] F. A. Faber, L. Hutchison, B. Huang, J. Gilmer,

S. S. Schoenholz, G. E. Dahl, O. Vinyals, S. Kearnes, P. F. Riley, and O. A. von Lilienfeld, "Prediction errors of molecular machine learning models lower than hybrid dft error," Journal of Chemical Theory and Computation **13**, 5255–5264 (2017), pMID: 28926232, https://doi.org/10.1021/acs.jctc.7b00577.

[14] F. A. Faber, A. S. Christensen, B. Huang, and O. A. von Lilienfeld, "Alchemical and structural distribution based representation for universal quantum machine learning," The Journal of Chemical Physics **148**, 241717 (2018), https://doi.org/10.1063/1.5020710.

[15] A. P. Bartók, S. De, C. Poelking, N. Bernstein, J. R. Kermode, G. Csányi, and M. Ceriotti, "Machine learning unifies the modeling of materials and molecules," Science Advances **3** (2017), 10.1126/sciadv.1701816, NoStop

[16] C. R. Collins, G. J. Gordon, O. A. von Lilienfeld, and D. J. Yaron, "Constant size descriptors for accurate machine learning models of molecular properties," The Journal of Chemical Physics **148**, 241718 (2018), https://doi.org/10.1063/1.5020441.

[17] S. De, A. P. Bartók, G. Csányi, and M. Ceriotti, "Comparing molecules and solids across structural and alchemical space," **18**, 13754–13769 (2016).

[18] R. Ramakrishnan, P. O. Dral, M. Rupp, and O. A. von Lilienfeld, "Big data meets quantum chemistry approximations: The δ-machine learning approach," Journal of Chemical Theory and Computation **11**, 2087–2096 (2015), pMID: 26574412, https://doi.org/10.1021/acs.jctc.5b00099.

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

[20] K. T. Schütt, H. E. Sauceda, P.-J. Kindermans, A. Tkatchenko, and K.-R. Müller, "Schnet – a deep learning architecture for molecules and materials," The Journal of Chemical Physics **148**, 241722 (2018), https://doi.org/10.1063/1.5019779.

[21] F. Pereira and J. A. de Sousa, "Machine learning for the prediction of molecular dipole moments obtained by density functional theory," in *J. Cheminformatics* (2018).

[22] T. Bereau, D. Andrienko, and O. A. von Lilienfeld, "Transferable atomic multipole machine learning models for small organic molecules," J Chem Theor Comput **11**, 3225–3233 (2015).

[23] T. Bereau, R. A. DiStasio, Jr., A. Tkatchenko, and O. A. von Lilienfeld, "Non-covalent interactions across organic and biological subsets of chemical space: Physics-based potentials parametrized from machine learning," J Chem Phys **148**, 241706 (2018).

[24] W. Pronobis, K. T. Schütt, A. Tkatchenko, and K.-R. Müller, "Capturing intensive and extensive dft/tddft molecular properties with machine learning," The European Physical Journal B **91**, 178 (2018).

[25] R. Ramakrishnan, M. Hartmann, E. Tapavicza, and O. A. von Lilienfeld, "Electronic spectra from tddft and machine learning in chemical space," The Journal of Chemical Physics **143**, 084111 (2015), https://doi.org/10.1063/1.4928757.

[26] M. Rupp, R. Ramakrishnan, and O. A. von Lilienfeld, "Machine Learning for Quantum Mechanical Properties of Atoms in Molecules," The Journal of Physical Chemistry Letters **6**, 3309–3313 (2015).

[27] E. O. Pyzer-Knapp, K. Li, and A. Aspuru-Guzik, "Learning from the harvard clean energy project: The use of neural networks to accelerate materials discovery," Advanced Functional Materials **25**, 6495–6502.

[28] F. Pereira, K. Xiao, D. A. R. S. Latino, C. Wu, Q. Zhang, and J. Aires-de Sousa, "Machine learning methods to predict density functional theory b3lyp energies of homo and lumo orbitals," Journal of Chemical Information and Modeling **57**, 11–21 (2017), pMID: 28033004, https://doi.org/10.1021/acs.jcim.6b00340.

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

[30] M. Ropo, M. Schneider, C. Baldauf, and V. Blum, "First-principles data set of 45,892 isolated and cation-coordinated conformers of 20 proteinogenic amino acids," Scientific Data **3** (2016), 10.1038/sdata.2016.9.

[31] C. Schober, K. Reuter, and H. Oberhofer, "Virtual screening for high carrier mobility in organic semiconductors," J. Phys. Chem. Lett. **7**, 3973–3977 (2016).

[32] L. Ruddigkeit, R. van Deursen, L. C. Blum, and J.-L. Reymond, "Enumeration of 166 billion organic small molecules in the chemical universe database gdb-17," Journal of Chemical Information and Modeling **52**, 2864–2875 (2012), pMID: 23088335, https://doi.org/10.1021/ci300415d.

[33] H. Huo and M. Rupp, "Unified Representation for Machine Learning of Molecules and Crystals," arXiv:1704.06439 [cond-mat, physics:physics] (2017), arXiv: 1704.06439.

[34] V. Blum, R. Gehrke, F. Hanke, P. Havu, V. Havu, X. Ren, K. Reuter, and M. Scheffler, "Ab initio molecular simulations with numeric atom-centered orbitals," Computer Physics Communications **180**, 2175–2196 (2009).

[35] V. Havu, V. Blum, P. Havu, and M. Scheffler, "Efficient $o(n)$ integration for all-electron electronic structure calculation using numeric basis functions," J. Comput. Phys. **228**, 8367 (2009).

[36] S. V. Levchenko, X. Ren, J. Wieferink, R. Johanni, P. Rinke, V. Blum, and M. Scheffler, "Hybrid functionals for large periodic systems in an all-electron, numeric atom-centered basis framework," Comput. Phys. Comm. **192**, 60 – 69 (2015).

[37] X. Ren, P. Rinke, V. Blum, J. Wieferink, A. Tkatchenko, S. Andrea, K. Reuter, V. Blum, and M. Scheffler, "Resolution-of-identity approach to hartree-fock, hybrid density functionals, rpa, mp2, and gw with numeric atom-centered orbital basis functions," New J. Phys. **14**, 053020 (2012).

[38] J. P. Perdew, K. Burke, and M. Ernzerhof, "Generalized gradient approximation made simple," Phys. Rev. Lett. **77**, 3865–3868 (1996).

[39] A. Tkatchenko and M. Scheffler, "Accurate molecular van der waals interactions from ground-state electron density and free-atom reference data," Phys. Rev. Lett. **102**, 073005 (2009).

[40] L. Hedin, Phys. Rev. **139**, A796 (1965).

[41] P. Rinke, A. Qteish, J. Neugebauer, C. Freysoldt, and M. Scheffler, "Combining *GW* calculations with exact-exchange density-functional theory: An analysis of valence-

band photoemission for compound semiconductors," New J. Phys. **7**, 126 (2005).

[42] R. Ramakrishnan and O. A. Lilienfeld, "Machine learning, quantum chemistry, and chemical space," in *Reviews in Computational Chemistry* (John Wiley & Sons, Ltd, 2017) Chap. 5, pp. 225–256, NoStop

[43] K. T. Schütt, F. Arbabzadah, S. Chmiela, K.-R. Müller, and A. Tkatchenko, "Quantum-Chemical Insights from Deep Tensor Neural Networks," Nature Communications **8**, 13890 (2017), arXiv: 1609.08259.

[44] N. Lubbers, J. S. Smith, and K. Barros, "Hierarchical modeling of molecular energies using a deep neural network," The Journal of Chemical Physics **148**, 241715 (2018), https://doi.org/10.1063/1.5011181.

[45] O. T. Unke and M. Meuwly, "A reactive, scalable, and transferable model for molecular energies from a neural network approach based on local information," The Journal of Chemical Physics **148**, 241708 (2018), https://doi.org/10.1063/1.5017898.

[46] N. Artrith, A. Urban, and G. Ceder, "Efficient and accurate machine-learning interpolation of atomic energies in compositions with many species," Phys. Rev. B **96**, 014112 (2017).

[47] S. De, F. Musil, T. Ingram, C. Baldauf, and M. Ceriotti, "Mapping and classifying molecules from a high-throughput structural database," Journal of Cheminformatics **9**, 6 (2017).

[48] C. R. Groom, I. J. Bruno, M. P. Lightfoot, and S. C. Ward, "The Cambridge Structural Database," Acta Crystallographica Section B **72**, 171–179 (2016).

[49] L. van der Maaten and G. Hinton, "Visualizing high-dimensional data using t-sne," Journal of Machine Learning Research **9**, 2579–2605 (2008).

[50] M. Rupp, "Machine learning for quantum mechanics in a nutshell," International Journal of Quantum Chemistry **115**, 1058–1073 (2015).

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

[52] O. A. von Lilienfeld, R. Ramakrishnan, M. Rupp, and A. Knoll, "Fourier series of atomic radial distribution functions: A molecular fingerprint for machine learning models of quantum chemical properties," International Journal of Quantum Chemistry **115**, 1084–1093 (2015), https://onlinelibrary.wiley.com/doi/pdf/10.1002/qua.24912.

[53] K. Hansen, F. Biegler, R. Ramakrishnan, W. Pronobis, O. A. von Lilienfeld, K.-R. Müller, and A. Tkatchenko, "Machine Learning Predictions of Molecular Properties: Accurate Many-Body Potentials and Nonlocality in Chemical Space," The Journal of Physical Chemistry Letters **6**, 2326–2331 (2015).

[54] "DScribe," `https://github.com/SINGROUP/dscribe`, accessed: 2018-11-21.

[55] T. Hastie, R. Tibshirani, and J. Friedman, *The elements of statistical learning: data mining, inference and prediction*, 2nd ed. (Springer, 2009).

[56] G. Montavon, K. Hansen, S. Fazli, M. Rupp, F. Biegler, A. Ziehe, A. Tkatchenko, O. A. von Lilienfeld, and K.-R. Müller, "Learning Invariant Representations of Molecules for Atomization Energy Prediction," Advances in Neural Information Processing Systems 25 , 440–448 (2012).

[57] C. Kunkel, C. Schober, J. T. Margraf, K. Reuter, and H. Oberhofer, "Finding the right bricks for molecular legos: A data mining approach to organic semiconductor design," Chem. Mater. **31**, 969–978 (2019).

[58] C. Kunkel, C. Schober, H. Oberhofer, and K. Reuter, "Knowledge discovery through chemical space networks: the case of organic electronics," J. Mol. Model. **25**, 87 (2019).

[59] M. Todorović, M. U. Gutmann, J. Corander, and P. Rinke, "Bayesian inference of atomistic structure in functional materials," npj Comp. Mat. **5**, 35 (2019).

[60] K. Ghosh, A. Stuke, M. Todorović, P. B. Jørgensen, M. N. Schmidt, A. Vehtari, and P. Rinke, "Deep learning spectroscopy: Neural networks for molecular excitation spectra," Adv. Sci. **0**, 1801367.

[61] Z. Wu, B. Ramsundar, E. N. Feinberg, J. Gomes, C. Geniesse, A. S. Pappu, K. Leswing, and V. Pande, "MoleculeNet: A Benchmark for Molecular Machine Learning," arXiv:1703.00564 [physics, stat] (2017), arXiv: 1703.00564.

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

[63] G. Pilania, J. E. Gubernatis, and T. Lookman, "Multi-fidelity machine learning models for accurate bandgap predictions of solids," Computational Materials Science **129**, 156–163 (2017).

[64] R. Gómez-Bombarelli, J. N. Wei, D. Duvenaud, J. M. Hernández-Lobato, B. Sánchez-Lengeling, D. Sheberla, J. Aguilera-Iparraguirre, T. D. Hirzel, R. P. Adams, and A. Aspuru-Guzik, "Automatic chemical design using a data-driven continuous representation of molecules," ACS Cent. Sci. **4**, 268–276 (2018).

[65] L. Li, J. C. Snyder, I. M. Pelaschier, J. Huang, U.-N. Niranjan, P. Duncan, M. Rupp, K.-R. Müller, and K. Burke, "Understanding machine-learned density functionals," International Journal of Quantum Chemistry **116**, 819–833 (2016).

[66] M. I. Jordan and T. M. Mitchell, "Machine learning: Trends, perspectives, and prospects," Science **349**, 255–260 (2015).

[67] "Wiley: Atomistic Computer Simulations: A Practical Guide - Veronika Brazdova, David R. Bowler," .

[68] NREL, "National Center for Photovoltaics, Research Cell Record Efficiency Chart," (2017), https://www.nrel.gov/pv/ Accessed 4.8.2017.

[69] S. Chu and A. Majumdar, "Opportunities and challenges for a sustainable energy future," Nature **488**, 294–303 (2012).

[70] W. Shockley and H. J. Queisser, "Detailed Balance Limit of Efficiency of p-n Junction Solar Cells," Journal of Applied Physics **32**, 510–519 (1961).

[71] B. Huskinson, M. P. Marshak, C. Suh, S. Er, M. R. Gerhardt, C. J. Galvin, X. Chen, A. Aspuru-Guzik, R. G. Gordon, and M. J. Aziz, "A metal-free organic–inorganic aqueous flow battery," Nature **505**, 195–198 (2014).

[72] M. Liu, Y. Pang, B. Zhang, P. D. Luna, O. Voznyy, J. Xu, X. Zheng, C. T. Dinh, F. Fan, C. Cao, F. P. G. de Arquer, T. S. Safaei, A. H. Mepham, A. Klinkova, E. Kumacheva, T. Filleter, D. Sinton, S. O. Kelley, and E. H. Sargent, "Enhanced electrocatalytic co2 reduction via field-induced reagent concentration," Nature **537**, 382–386 (2016).

[73] L. C. Blum and J.-L. Reymond, "970 million druglike small

molecules for virtual screening in the chemical universe database GDB-13," J. Am. Chem. Soc. **131**, 8732 (2009).

[74] O. A. von Lilienfeld, "Quantum machine learning in chemical compound space," Angewandte Chemie International Edition **57**, 4164–4169, https://onlinelibrary.wiley.com/doi/pdf/10.1002/anie.201709686.