Materials Advances

ROYAL SOCIETY OF CHEMISTRY

PAPER

View Article Online
View Journal | View Issue

Cite this: *Mater. Adv.*, 2022, 3, 8306

Selected machine learning of HOMO–LUMO gaps with improved data-efficiency†

Bernard Mazouin, [a] Alexandre Alain Schöpfer [b] and O. Anatole von Lilienfeld [cde]*

Despite their relevance for organic electronics, quantum machine learning (QML) models of molecular electronic properties, such as HOMO–LUMO-gaps, often struggle to achieve satisfying data-efficiency as measured by decreasing prediction errors for increasing training set sizes. We demonstrate that partitioning training sets into different chemical classes prior to training results in independently trained QML models with overall reduced training data needs. For organic molecules drawn from previously published QM7 and QM9-data-sets we have identified and exploited three relevant classes corresponding to compounds containing either aromatic rings and carbonyl groups, or single unsaturated bonds, or saturated bonds The selected QML models of band-gaps (considered at GW and hybrid DFT levels of theory) reach mean absolute prediction errors of ~0.1 eV for up to an order of magnitude fewer training molecules than for QML models trained on randomly selected molecules. Comparison to Δ-QML models of band-gaps indicates that selected QML exhibit superior data-efficiency. Our findings suggest that selected QML, *e.g.* based on simple classifications prior to training, could help to successfully tackle challenging quantum property screening tasks of large libraries with high fidelity and low computational burden.

Received 27th June 2022,
Accepted 12th September 2022

DOI: 10.1039/d2ma00742h

rsc.li/materials-advances

I. Introduction

Machine Learning (ML) based surrogate models of quantum properties have gained a lot of traction in recent years.[1–5] This rise in interest is partly driven by the computational efficiency of ML algorithms that typically outpace the conventional quantum chemistry methods which attempt to numerically solve sophisticated approximations to the electronic Schrödinger equation. The application of these algorithms to Chemical Compound Space (CCS) is commonly referred to as Quantum Machine Learning (QML). During training, QML models get parameterized in terms of a heuristic functional form which encodes a statistical relation between sample training molecules and their corresponding labels (quantum property). The resulting QML model can subsequently be used to make quantum property predictions throughout CCS, *i.e.* for unknown out-of-sample molecules. Since its inception in 2012,[6] QML has already been applied to a variety of chemical classes including, among others, organic molecules,[7–9] amino acids,[9] polymers,[10] or solids.[11–15] Within these applications, it has been used to predict *ab initio* thermodynamic properties such as atomization energies,[6–8,16] energy above convex hull,[14] or free energy of solvation,[17] as well as electronic properties such as HOMO and LUMO energies or dipole moments.[7,8,18–21] Some state-of-the-art QML models can reach an accuracy on par with quantum chemistry algorithms already for modest training set sizes,[8] and are thus well positioned for their direct application in computational materials design efforts.[22–27]

Not surprisingly, the importance of rapid yet accurate QM property predictions has inspired the development of specialized ML methods. For example, optimized representations or Neural Network architectures have been designed just for this purpose.[8,9,18,20,30–34] In particular, one can adjust the QML procedure to the property of interest by including more information about the underlying physics in it. In order to obtain QML models with higher data-efficiency for atomization energies, more descriptive representations such as SLATM[35] and FCHL,[36,37] which include 3-body-terms and physically motivated power laws, yield better results than the more heuristic CM[16] or BoB[38] representations, which merely encode the nuclear repulsion terms. The integration of gradients in KRR has led to reduced errors for

[a] *University of Vienna, Faculty of Physics and Vienna Doctoral School in Physics, Kolingasse 14-16, 1090 Vienna, Austria*
[b] *Department of Chemistry, University of Basel, Klingelbergstrasse 70, 4056 Basel, Switzerland. E-mail: anatole.vonlilienfeld@utoronto.ca*
[c] *Departments of Chemistry, Materials Science and Engineering, and Physics, University of Toronto, St. George Campus, Toronto, ON, Canada*
[d] *Vector Institute for Artificial Intelligence, Toronto, ON, M5S 1M1, Canada*
[e] *Machine Learning Group, Technische Universität Berlin and Institute for the Foundations of Learning and Data, 10587 Berlin, Germany*

† Electronic supplementary information (ESI) available. See DOI: https://doi.org/10.1039/d2ma00742h

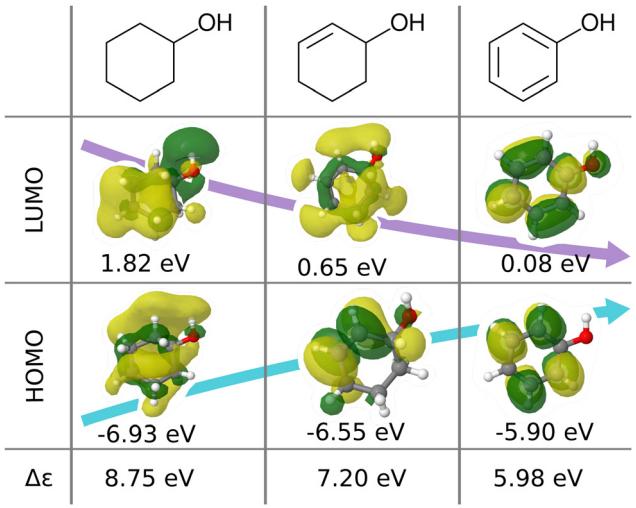

Fig. 1 Illustration of frontier molecular orbitals and eigenvalues being dominated by simple features such as bond-saturation: compositionally and structurally similar molecules (cyclohexanol, cyclohex-2-enol and phenol) exhibit vast differences in HOMOs, LUMOs, and eigenvalues. The orbitals are visualized with Jmol[28] using results from B3LYP/6-31G(2df,p) calculations performed with ORCA 4.0.1.[29]

response properties such as the dipole moment or forces.[21,39–41] Furthermore, a biased selection of training samples will also lead to QML models with improved accuracy.[35,42]

Among the various QM properties frequently evaluated, the eigenvalues of the frontier orbitals, *i.e.* highest occupied and lowest unoccupied molecular orbitals (HOMO and LUMO), are of special interest. These MO energies are intimately related to chemical reactions, polarizability, the optical gap and excitation energies. Their prediction often plays an important role for design decisions in the development of technological applications such as synthesis planning, electrochromic devices, light-emission diodes or photovoltaic solar panels.[43–48] Interestingly, the generation of accurate QML models of frontier orbital eigenvalues proves more difficult than for other quantum properties—even when using molecular training sets of considerable size. Consequently, significant research efforts are currently being made in order to devise QML models of MO energies with improved data-efficiency.

We believe that this difficulty is partly, if not mostly, due to the intensive nature of MO energies. Molecules with very similar stoichiometry and geometry do not necessarily have similar HOMO–LUMO gap values (see *e.g.* the molecules drawn in Fig. 1), whereas structurally dissimilar molecules can have very close values. While the latter can be resolved easily by allowing for QML models which are not monotonic in CCS, the former point represents the actual challenge since all ML models are based on similarity arguments and smoothness assumptions. HOMO–LUMO gaps suffer from a lack of smoothness (as on display in Fig. 1), which indicates the presence of additional dimensions that are not properly reflected by conventional QML representations.[3] An inspection of the HOMO–LUMO gaps (see Fig. 2) reveals a superposition of multiple groups, such as aliphatic and aromatic ones, that possibly accounts for the aforementioned 'hidden' dimensions. In this

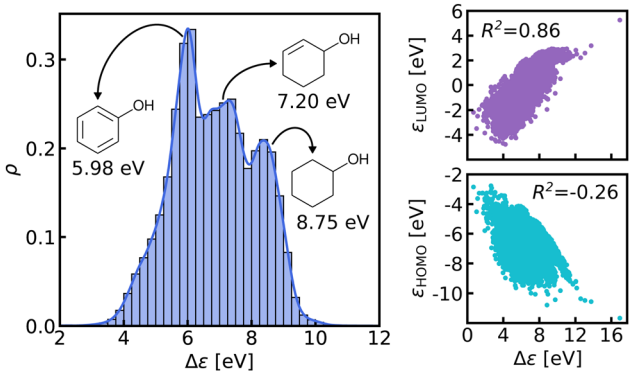

Fig. 2 Left: Normalized histogram and kernel density estimate (solid line) of HOMO–LUMO gaps in QM9 data-set. Gaps of 3 similar molecules (cyclohexanol, cyclohex-2-enol and phenol) are indicated. Right: The HOMO and LUMO energies plotted against the HOMO–LUMO gaps with the respective correlation coefficients.

work, we have investigated in depth how one can use this information – the existence of multiple subgroups – to improve the data-efficiency of ML models for HOMO–LUMO gaps, without inventing a new representation.

In this work, we study selected machine learning (SML) models of MO energies. SML relies on a divide-and-conquer-like strategy applied to training selection prior to training which turns out to improve the data-efficiency. More specifically, before training, we partition the training data into smaller classes, and we train QML models separately for each class. The idea for such a classification is based on the peculiar shape of the distribution of HOMO–LUMO-gaps obtained from B3LYP in QM9 or ZINDO in QM7b: it is multi-modal and appears to be composed of 3 sub-distributions, one per peak (see Fig. 2 and 3). According to a frequency analysis the molecules can be easily classified into three groups, solely based on simple structural features. The three example molecules indicated in Fig. 2, that are each located close to a different peak of the distribution, indicate such features (aliphatic chain, unsaturated bond, aromatic ring) that encode important information about the gap. Fig. 1 also

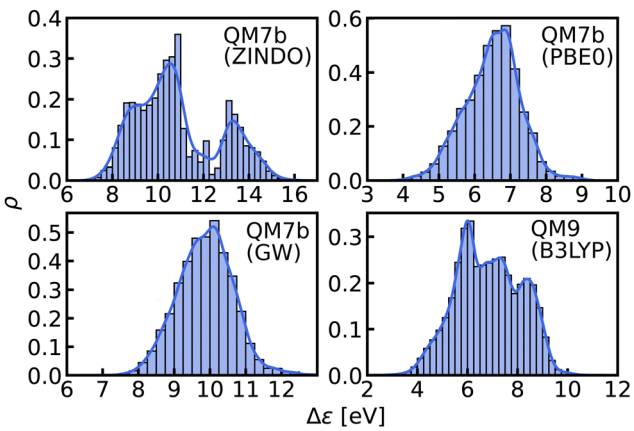

Fig. 3 Histograms and KDEs of the HOMO–LUMO gaps of all molecules from QM7b at ZINDO, PBE0 and GW levels of theory and of all molecules from QM9 at B3LYP level of theory.

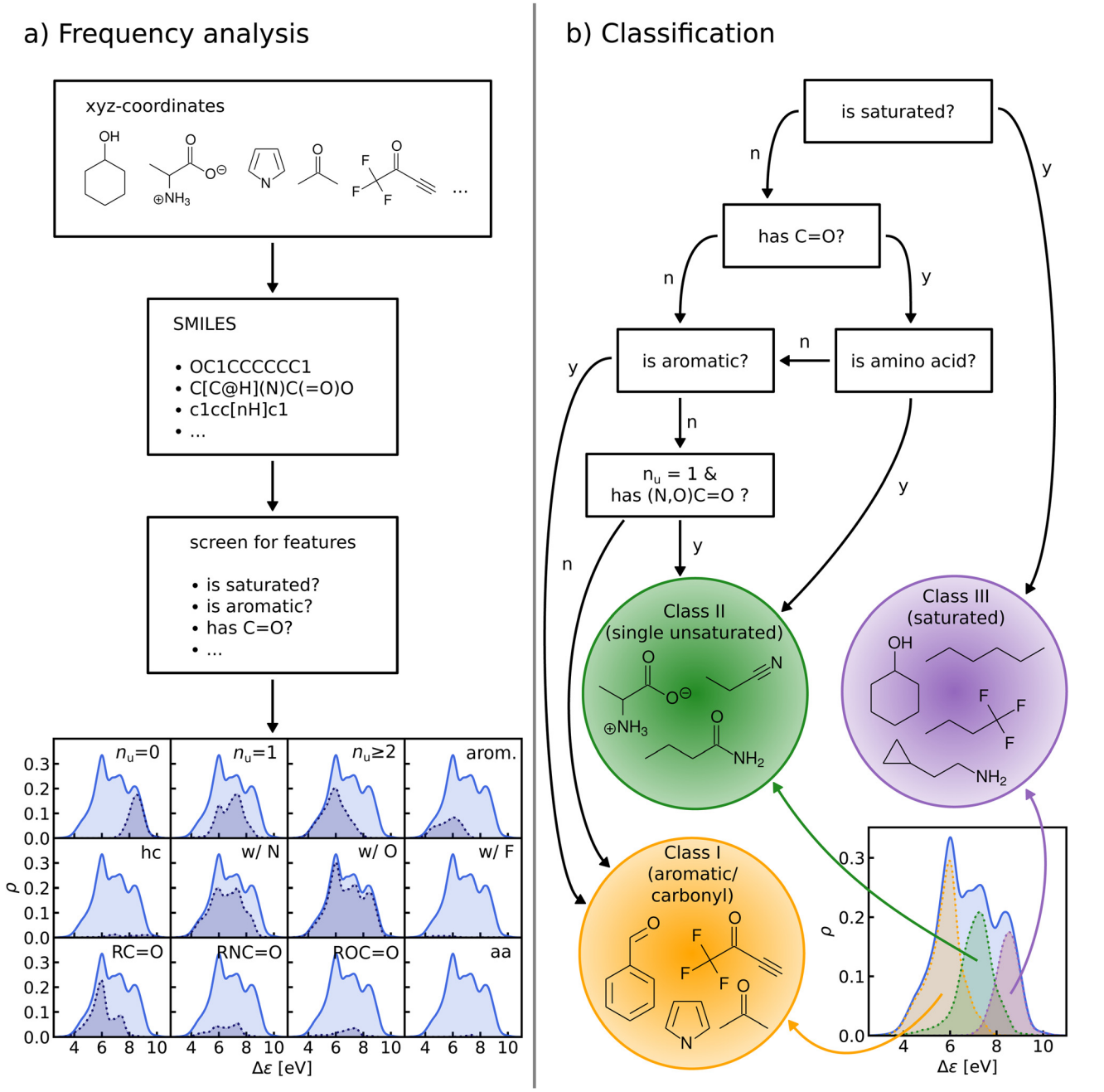

Fig. 4 (a) General procedure of the frequency analysis. We use SMILES strings obtained from QM9 molecules to screen for features such as double bonds, aromatic rings, carbonyl groups and so on. The plot at the bottom shows the results of this frequency analysis. Each quadrant shows the distribution of HOMO–LUMO-gaps of a subgroup of QM9 matching a given criterion. The top row shows the effect of saturation (n_u: number of unsaturated bonds, *i.e.* any double, triple or aromatic bond), the middle row shows the effect of elemental constitution (hc: hydrocarbons, w/N: with nitrogen and so on) and the bottom row shows various kinds of carbonyl groups (columns 1–3) and amino acids (column 4). (b) Flowchart detailing the sequence of decisions that results in our final classification. The distributions of the classes are highlighted in plot on the bottom left.

illustrates their dramatic effect on the character of their frontier orbitals, and thereby on eigenvalues and their gap. Based on the QM9 analysis, we have defined a set of simple rules for classification which (*vide infra*) results in subsequently trained QML models, henceforth dubbed SML, with much improved learning curves.

II. Data and methods

A. Data

The QM7b data-set[18,49] contains properties of 7211 organic molecules with up to 7 heavy atoms (C, O, N, S and Cl). These molecules were derived originally from the GDB-13 data-set. Thermodynamic and electronic properties are available at

different levels of theory, which makes the data-set suitable for Δ-ML applications.[50] For this work, we use the HOMO and LUMO energies as obtained from ZINDO[51,52] and GW[53,54] calculation for direct and Δ-ML. The HOMO–LUMO gaps correspond to the differences of LUMO and HOMO energies.

The QM9 data-set, published in 2014 by Ramakrishnan *et al.*,[55,56] consists of more than 133k organic molecules with up to 9 heavy atoms (C, N, O and F) with corresponding geometries, thermodynamic and electronic properties. These molecules were obtained from the GDB-17 data-set which contains over 166 billion molecular graphs. The properties were computed using DFT/B3LYP[57–59] with a 6-31G(2d,f) basis set. Over the last years, it has become an increasingly popular data-set in the QML community as it has been used as a staple to benchmark new QML models.[9,36,60–65]

B. Frequency analysis and classification

We perform a frequency analysis to identify functional groups that relate to the HOMO–LUMO gap in our molecules (see Fig. 4, panel a)). By screening for a set of structural features and functional groups such as double bonds, aromatic rings or carbonyl groups *e.g.* using SMILES[66–68] strings and substructure matching as implemented in RDKit,[69] we tag the molecules in the data-set. For each tag, we compute the Kernel Density Estimation (KDE) of the HOMO–LUMO-gaps of the matching molecules, normalize it with respect to the total number of molecules in the entire data-set, and draw it over its KDE. By visual inspection of the resulting plots we have detected those functional groups which govern the assignment to one of the classes in the HOMO–LUMO-gap distribution.

In the next step, based on the frequency analysis, we define simple rules to separate the molecules into disjoint classes. We make sure that the class distributions have a unimodal shape and coincide with the peaks of the total distribution (see Fig. 4, panel b)). For example, the distribution of saturated molecules in QM9 fits closely underneath the right peak, so that we can assign all saturated molecules to the class corresponding to that peak. The distribution of all carbonyl compounds however has two peaks that coincide with the left and middle peaks. Therefore we have further subdivided the group of carbonyl compounds by distinguishing, for instance, between those with aromatic rings from those without, until one ends up with unimodal subdistributions.

C. Kernel ridge regression

The main idea behind supervised learning is to establish and exploit statistical relations between inputs $\mathbf{X}_i$ and corresponding target property label outputs y_i. In our case, the inputs are molecular representations which, in strict correspondence to Schrödinger's equation, encode stoichiometry and geometry. We have relied on the SLATM representation,[35] which describes a molecule as a spectrum of atomic, 2-body and 3-body terms. The target labels are the properties of interest, *i.e.* the HOMO–LUMO gaps and the individual frontier orbital energies. A training set $\{\mathbf{X}_i y_i\}_{i=1}^{N_{\text{tr}}}$ is a sample for which both the inputs and target values are known, whereas for the test set only the inputs $\{\mathbf{X}_j\}_{j=1}^{N_{\text{te}}}$ are known, but the target values unknown. The ML model uses the training data to infer a statistical model that relates the input $\mathbf{X}_i$ to the output y_i. This statistical model can then be applied to the molecules in the test set in order to produce a prediction error estimate for the corresponding properties. As such, ML circumvents numerically solving the Schrödinger equation, and provides instead statistical estimates which are computationally more efficient than state-of-the-art quantum chemistry calculations.

We are dealing with a regresssion problem where the task is to predict continuous target values. Our method of choice is KRR[70–73] due to its ease of implementation and interpretability. Moreover, it has worked successfully in numerous applications.[1] We note, however, that the first QML models of frontier orbital eigenvalues were presented using neural networks,[18] and that we believe that the choice of the specific regressor is rather secondary, *i.e.* our procedure could be used in combination with any other regressors just as well. In the following, we briefly outline the KRR methodology only for the sake of completeness.

Within KRR, the prediction of a given property $\hat{y}_i$ is given by a sum over kernel matrix elements $k_{ij} = k(\mathbf{X}_i,\mathbf{X}_j)$ multiplied by regression coefficients α_j:

$$\hat{y}_i = \sum_{j=1}^{N_{\text{tr}}} k\left(\mathbf{X}_i, \mathbf{X}_j\right)\alpha_j. \tag{1}$$

The α-coefficients are obtained by solving the following system of linear equations:

$$\mathbf{y}^{\text{tr}} = (\mathbf{K} + \lambda\mathbf{I})\boldsymbol{\alpha}. \tag{2}$$

The parameter λ is a regularization coefficient, a.k.a. noise-level, that smooths out the noise. However, since we are dealing with computed target values that are noiseless to machine precision, we can fix λ to correspond to a small value such as 10^{-12}. $\mathbf{I}$ is the identity matrix. The kernel matrix $\mathbf{K}$, for which we employ the Laplacian kernel function ($k_{ij} = \exp(-|\mathbf{X}_i - \mathbf{X}_j|_1/\sigma)$), quantifies the similarity between any two representations of the i-th and j-th molecules. By virtue of this kernel matrix, each test molecule is compared to all the training molecules in order to make a prediction. The parameter σ modulates the sensitivity of the kernel and is optimized *via* grid search cross validation within each training set in this work. To evaluate the performance of our method, we use the target values of the test set to calculate the mean absolute error (MAE) between reference and predicted values. The logarithm of the prediction error generally decreases linearly with the logarithm of the training set size ($\log(E) \propto -\log(N_{\text{tr}})$),[74,75] which is shown numerically in terms of so-called learning curves. We have employed the QML package[76] to perform our calculations.

D. Δ Delta-machine learning

In Δ-ML,[50] correlations between different levels of theory are exploited to obtain better predictions of properties calculated at higher levels of theory for fewer training molecules. We consider

two levels of theory, a lower baseline, at which we know the output, and a higher target line, for which we want to obtain predictions. A machine is trained on the differences between the two levels. In other words, a QML model of a correction to the baseline model is being generated.

After that, these predictions are added to the baseline to generate estimates of the property at the higher level of theory:

$$\hat{y}_i^{\text{target}} = y_i^{\text{baseline}} + \sum_{j=1}^{N_{\text{tr}}} k(\mathbf{X}_i, \mathbf{X}_j)\alpha_j \quad (3)$$

The better the correlation between the levels of theory, the easier it is to learn the difference between them. In a more generalized version of this method called Multilevel-ML,[77] one can exploit the correlations between more than 2 levels of theory and basis sets to improve predictions. In this work, we combine the SML method with Δ-ML using data from the QM7b dataset, namely the ZINDO energies as baseline, and the GW energies as target.

E. Selected machine learning

In order to compare SML to generic QML training set selection, we follow the procedure visualized in Fig. 5. We train a model on all molecules drawn at random across the data-set (generic QML), then 3 different machines, each only with molecules from a single class (SML). Moreover, we generate 3 separate test sets, one for each class, while making sure that there is no overlap between any of the training and test sets. For each test set, we produce two predictions: one obtained from generic QML – with training molecules from all over the data-set – and a second one from SML – with training molecules from the corresponding class only. We expect the prediction errors of SML to be lower than those of generic QML for each class. By applying two different machines on exactly the same test set, their performances can be properly compared to one another.

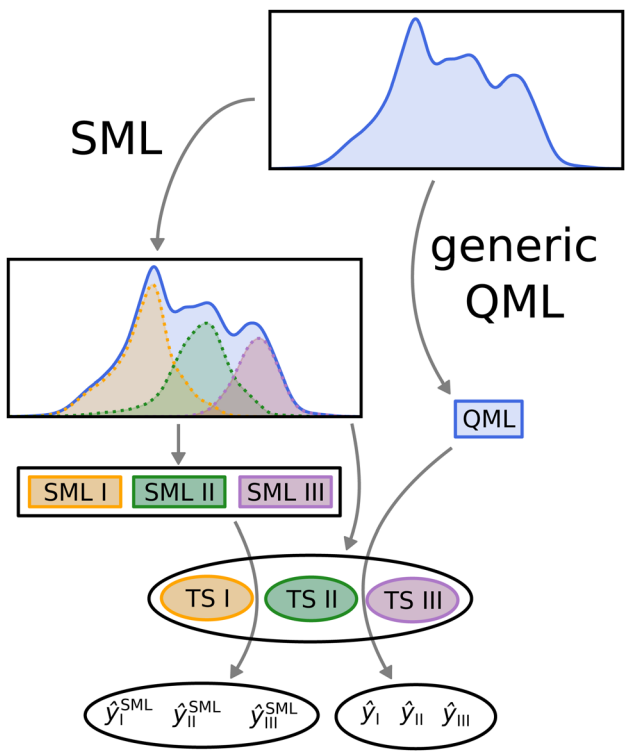

Fig. 5 Visual representation of the SML method. The data-set is first classified into separate classes – in our case 3. For each class, a machine is trained on its molecules (models SML I, II and III) and a disjoint test set is put aside (TS I, II and III). In addition, one machine is trained on molecules form all across the data-set (here dubbed QML). Eventually, for each test set, two predictions are computed, one using generic QML ($\hat{y}_{\text{I}}$, $\hat{y}_{\text{II}}$ and $\hat{y}_{\text{III}}$) and the other using SML ($\hat{y}_{\text{I}}^{\text{SML}}$, $\hat{y}_{\text{II}}^{\text{SML}}$ and $\hat{y}_{\text{III}}^{\text{SML}}$). These two predictions are then compared for each test set.

III. Results and discussion

A. Frequency analysis and classification

The graph at the bottom of panel (a) in Fig. 4 shows the frequency analysis of HOMO–LUMO gaps of QM9 molecules. The first row illustrates how different degrees of saturation affect the gap. The more unsaturated a molecule, the lower its HOMO–LUMO gap, with aromatic and fully saturated molecules having the smallest and largest gaps, respectively. This observation was to be expected since in unsaturated molecules, the frontier orbitals are often π-orbitals, which are closer in energy. The second row compares molecules with differing elemental composition. These distributions indicate that the composition alone is a relatively poor predictor of the location of the gap, and has thus been ignored for the classification. The third row illustrates the impact of the presence of common functional group signatures including carbonyl, ester, amide bonds, and amino acids. Carbonyl containing compounds mostly have lower gaps, but their distribution is bimodal, with both peaks coinciding with the left and mid peak of the reference distribution. Therefore more specific distinctions between different types of carbonyl compounds are required. The next two distributions suggest that amides and carboxylates (with an N- and O-atom linked to the C-atom of the carbonyl group) can be considered separately from the other carbonyl molecules, since their distributions are slightly more localized. Albeit rather rare in the set considered, amino acids appear to be located closer to the middle peak as well. In conclusion, the most relevant features for the classification are saturation *vs.* aromaticity, and the presence *vs.* absence of a carbonyl group, since they lead to well localized sub-distributions. Note that HOMO–LUMO-gap distributions from the QM7b data-set at ZINDO level of theory exhibit similar structures across different groups of molecules (Fig. S3 of ESI†).

The graph at the bottom of panel (b) of Fig. 4 showcases the resulting classification rules of QM9 molecules used for this study. First, we separate all saturated molecules from the rest and assign them to one class that we call 'saturated', which corresponds to the right peak. The remaining molecules are then subdivided into carbonyl and non-carbonyl molecules, which we further separate into aromatic and non-aromatic ones, amino acids and more specific carbonyl compounds

(amides and carboxylates). Finally, we end up with aromatic and carbonyl molecules with more than one unsaturated bond in one class that we name 'aromatic/carbonyl', that overlaps with the left peak. We put the remaining molecules together with amino acids and other carbonyl compounds into the last class, which we call 'single unsaturated', because most molecules of that class have only one unsaturated bond.

These rules lead to a classification which results in three molecular classes exhibiting well-behaved unimodal distributions for the data-sets considered here. To facilitate comparison in the following, we have numbered the classes (from left to right): class I – saturated, class II – single unsaturated, and class III – aromatic/carbonyl. We note again that classification rules for the ZINDO gaps of QM7b (Fig. S4 of ESI†) are similar. We have also performed two consistency checks of the classification with a Linear Discriminant Analysis (LDA) projection (Fig. S5–S8 of ESI†) and a Decision Tree Classification (Fig. S10 of ESI†). As shown in the ESI,† these checks confirm the validity of our classification scheme.

Our analysis suggests that a classification based on these simple rules suffices to mitigate the lack of smoothness of HOMO–LUMO gaps. Indeed, the classes are more homogeneous in terms of functional groups and the HOMO–LUMO gaps are more well-behaved within the classes, as is reflected by the unimodal shape of their distribution. Applying a Gaussian Mixture model on the gap values also results in a similar classification, however, it only partially captures the underlying structural features dominating the different classes. Fig. S13 and S14 of the ESI† demonstrate this observation for the classification of the QM7b data set. The final class distribution look similar, but exhibit some inconsistencies in the class labels attributed to some molecules. At this point we want to emphasize that our model explicitly differentiates between functional groups and eventually leads to a classification protocol with simple rules that allow an intuitive interpretation.

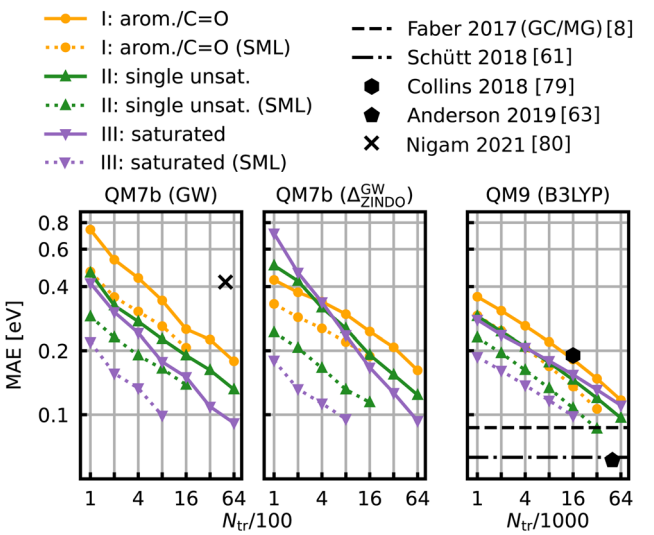

Fig. 6 Learning curves for the HOMO–LUMO gap of QM7b (GW) (left), Δ-ML on QM7b with ZINDO as baseline and (GW) as targetline (middle) and QM9 (B3LYP) (right). The points show the MAE averaged over 10 iterations with a different selection of training set molecules each. The average deviations of the MAE are not displayed since they are too small to be meaningful. The solid lines are the learning curves obtained from training with molecules from all over the data-sets, whereas the dotted lines are obtained from Selected ML. Reference results from the literature[8,61,63,79,80] are shown in black. The results for QM9 indicated with horizontal lines were obtained with training set sizes of ∼110k molecules.

B. Learning curves

The learning curves for the HOMO–LUMO gaps are presented in Fig. 6. They show the prediction errors w.r.t. increasing training set size on a log–log scale. In all cases, KRR with prior classification *via* SML (dotted lines) performs better than without classification (solid lines). While the slopes of the learning curves remain the same, there is a significant drop for the offsets. The largest drop can be observed for the class of saturated molecules, which is around 0.077 eV for 800 training molecules in QM7b (GW), 0.142 eV for Δ-ML in QM7b and 0.055 eV in QM9. Moreover, the prediction error for the class of saturated molecules is the lowest of all 3 classes, followed by the errors of the class of singly unsaturated molecules, and the highest errors are for the class of aromatic and carbonyl molecules. This trend is consistent for both data-sets. Only the learning curves of the saturated class reaches an error lower than 0.1 eV. For QM9, SML reaches this error with 16k training samples already, while more than 64k training samples would

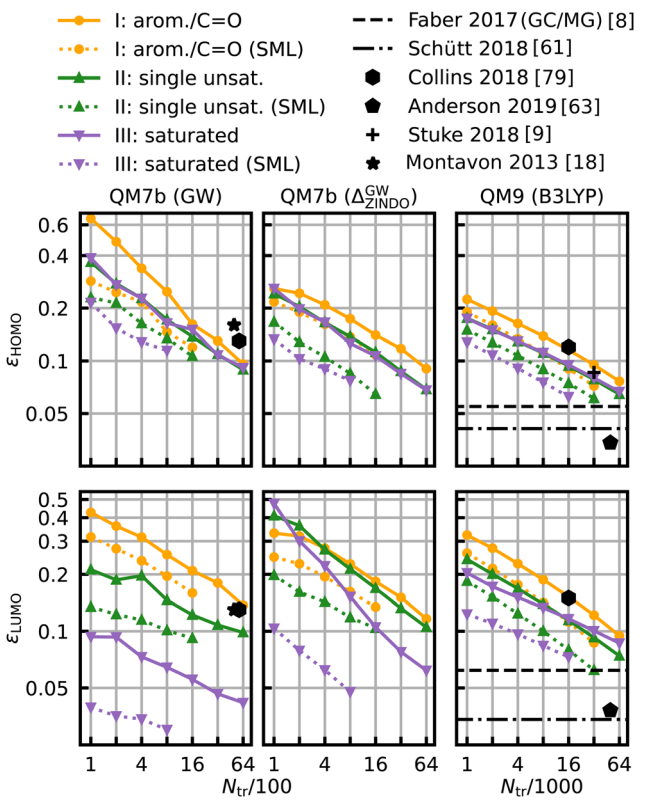

Fig. 7 Learning curves for the HOMO (first row) and LUMO energies (second row). As in Fig. 6, results are shown for QM7b (GW), Δ-ML (ZINDO to GW) and QM9 (B3LYP) from left to right and references are shown in black.[8,9,18,61,63,79]

be required without classification. Note that this prediction error is also on par with neural network prediction errors by Faber *et al.*[8] which had required training on 110k training samples. It is still far from the errors obtained by more recently developed NNs,[61,63,65] such as, for example, a NN by Liu *et al.*[78] reaches errors as low as 0.032 eV, however with training sets of ~100k molecules. Linear extrapolation of SML learning curves predicts an MAE of 0.065 eV for the class of saturated molecules with 100k training molecules, close to the NN by Schütt *et al.* (0.63 eV).[61] All in all, a prior classification systematically improves prediction errors of the HOMO–LUMO gaps.

The same model applied to the HOMO and LUMO energies results in the learning curves shown in Fig. 7. We can see that the learning curves generally follow the same trends as those for the HOMO–LUMO gaps: same slopes in both models, lower offsets for SML, lowest errors for the class of saturated molecules and highest errors for the class of aromatic and carbonyl molecules. A noteworthy difference between the results for HOMO and LUMO is the extent of the improvement in the prediction errors: the errors for the HOMO energies drop less than those for the LUMO energies. The learning curves for the LUMO energies of QM7b (GW) stand out since the error for the saturated molecules (0.030 eV) is much lower than for the other classes. These results demonstrate that our classification can also be transferred to other related properties, such as individual HOMO and LUMO energies, even though it has been derived from gaps only.

In Table S1 of the ESI† we compare MAEs obtained using the SML protocol with our classification rules to those with a Gaussian Mixture based classification. The results indicate no clear advantage of the GM classification. Therefore, due to its simplicity and basis in chemical bonding patterns, we prefer our classification.

C. Scatter plots

In Fig. 8 scatter plots of prediction *vs.* reference values are shown for QM9 and in Fig. S11 and S12 of the ESI† those for QM7b. The scatter plots reveal that the energies in single unsaturated and aromatic/carbonyl classes (I, and II) span a much wider range of values than the saturated molecules (class III), which explains the higher complexity and offsets of the learning curves. Some striking outliers are labelled in the Figures. The most noticeable ones in the QM9 data-set are small saturated ones such as C_2H_6 or CF_4 (Fig. 8, right column). The HOMO and LUMO energies as well as gaps of these molecules already stand out compared to the rest, which is

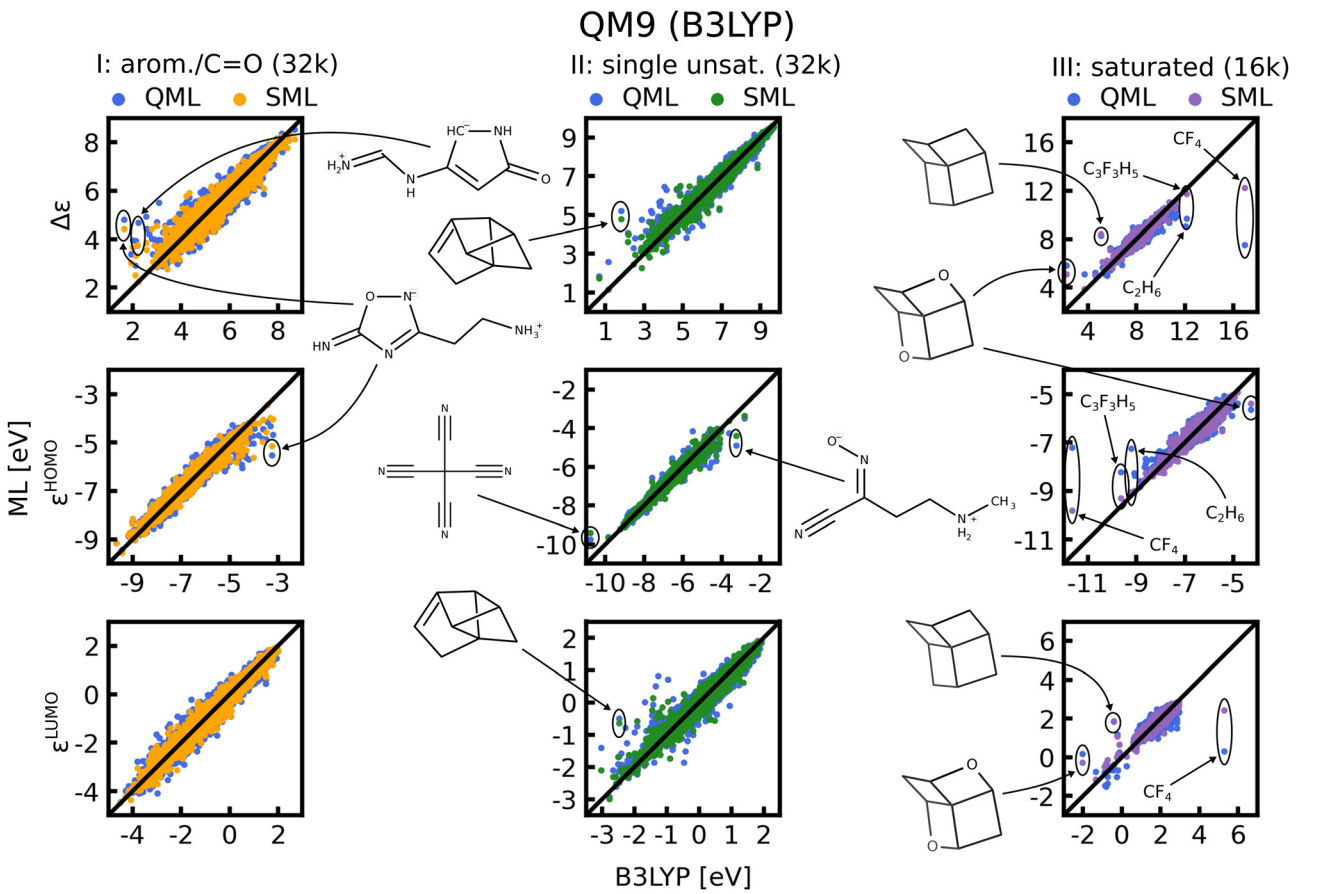

Fig. 8 Scatter plots of predicted ML *vs.* reference QM energies of QM9 for the largest training set size possible for each class, indicated in brackets. Some striking outliers are indicated in the figure, and if possible, predictions from both models (with and without classification) are highlighted for comparison.

why ML predictions for these molecules have a large MAE. Moreover, molecules with several rings and cage-like geometries have large prediction errors as well. An explanation for these outliers may be that similar molecules are scarce, such that they are not necessarily well represented in the training set. This issue could be resolved by always including such molecules in the training set. For the other outliers highlighted we could not find a pattern that explains the large errors, we nevertheless included them for the sake of completeness. It is worth noting that the predictions with models based on SML are in most cases closer to the exact reference values than without classification. An exception is for instance C_9H_{12} in the third column of Fig. 8, where the error from SML is larger compared to generic QML. In Fig. S11 and S12 of the ESI† more outlier examples are indicated. In general, the predictions within the classes from SML are closer to the reference values then those from QML after random training set selection.

D. Interpretation

The systematic improvements of the prediction errors across different data-sets and properties may be explained by the reduction of effective dimensionality achieved within each of the classes. Indeed, the lowest errors are obtained for the least diverse: the class of saturated molecules. Many functional groups such as carbonyl groups, aromatic rings or amides imply the presence of unsaturated bonds, indicating more chemical diversity in the other two classes. Because of the classification, the model parameters can more easily adapt since the HOMO–LUMO gaps are smoother within each class, such that the similarity between molecules better reflects the similarity between their gap values. In other words, in order to predict the gap for an unsaturated molecule, the model did not need to account for the correlations of gaps with saturated molecules. In this sense, no fitting coefficients have been 'wasted' on suboptimal correlations and can contribute instead to further lower the prediction error by exploiting more effective correlations within a given class.

It is also interesting to note that in the case of the QM7b results, even though the classification was based only on the distribution of ZINDO HOMO–LUMO gaps, the prediction errors also drop for predicting the GW gaps. As such, there seems to be a certain transferability of the classification scheme in the sense that it can be used across different levels of theory. But the classification is also transferable between different properties, as shown in Fig. 7. The greater improvement for LUMO energies than for HOMO energies is likely due to correlation between gaps and LUMO energies being stronger than between gaps and HOMO energies (see Fig. 3).

IV. Conclusion

We have found that simple classification protocols, prior to training, can dramatically improve the data-efficiency of QML models of HOMO–LUMO gaps in the QM9 and QM7b data sets. The classification is based on chemical bonding rules that allow us to define molecular classes based on structural input features alone. Our frequency analysis reveals that the presence of functional groups, such as aromatic rings and carbonyl groups, dominate sub-distributions of HOMO–LUMO gaps, and can therefore be exploited for classification. After classification, conventional kernel ridge regression based QML models afford learning curves with systematically lower offsets than without classifications. As a result, significantly fewer training molecules are required to reach competitive prediction errors ($\sim$0.1 eV), *e.g.* 16k for saturated molecules as compared to more than 64k training molecules necessary when drawing at random from QM9. We have also shown that our SML approach can be applied to related individual properties, *i.e.* the HOMO and LUMO energies alone. Further analysis has indicated, that the scheme is robust across different levels of theory for the labels, *i.e.* classification based on the distribution of ZINDO gaps was shown to be transferable to train more efficient QML models of GW gaps. Comparison to Δ-ML results on the same data set (QM7b) indicates that for HOMO–LUMO gaps, the classification approach presented here within offers substantially more improvement.

The additional step of prior classification alone can already lower the prediction errors in QML. The exploitation of simple relations between molecular structure and the HOMO–LUMO gap was enough to improve learning curves consistently. Our method addresses the lack-of-smoothness-problem by splitting the data set into classes that reflect the structural differences responsible for differing gap values. The results corroborate the idea that an adequate curation of the data can help optimize the performance of QML using already established representations. Nevertheless, there is no universally applicable classification that would work for any data-set. In our case, the classifications for QM9 and QM7b are only similar because they consist both of small organic molecules, but in general, such a classification depends on the chemical space a given data-set covers, the property of interest and also the level theory at which the property is calculated. Note that, to the best of our knowledge, there is no generic theoretical framework which would allow us to predict, rather than to detect, the minimal set of the most relevant features required for the classification. Studying the extension of this approach to chemistries that bear little resemblance with the organic chemistry represented by QM9 or QM7b will be the subject of future efforts.

Similar to HOMO–LUMO gaps based on B3LYP or ZINDO level of theory in QM9 and QM7b respectively, other properties with multimodal distributions could also be investigated. These could include properties related to the gap, such as excitation energies,[19] but also properties of entirely different origin such a highest vibrational frequencies[7] or NMR shifts. Similar to the gaps, one should then identify the structural features that govern these properties (well established for IR and NMR spectroscopy) in order to define molecular classes within which these distributions become unimodal and for which equal improvements in the data-efficiency of resulting QML models should be expected. Recently related work was made accessible in the context of local learning for improving

decision making within experimental design problems.[81] Other future work could also involve the use of more sophisticated unsupervised ML methods to find new and potentially better classification rules, based on more complex combinations of functional groups, or other molecular features. It is not obvious to us if it is generally possible to identify advantageous structural features (leading to similar improvements in QML model accuracy) for any arbitrary property, or if our findings are rather restricted to an exclusive list of observables.

Code availability

The code used in this work is freely available from **https://github.com/b3rn4rdm/SelectedML**.

Conflicts of interest

There are no conflicts to declare.

Acknowledgements

We acknowledge support from the European Research Council (ERC-CoG grant QML and H2020 projects BIG-MAP). This project has received funding from the European Union's Horizon 2020 research and innovation programme under Grant Agreements #957189. This project has received funding from the European Research Council (ERC) under the European Union's Horizon 2020 research and innovation programme (grant agreement no. 772834). This result only reflects the author's view and the EU is not responsible for any use that may be made of the information it contains. This work was partly supported by the NCCR MARVEL, funded by the Swiss National Science Foundation. Some of the computational results presented have been achieved using the Vienna Scientific Cluster (VSC).

References

1 R. Ramakrishnan and O. A. von Lilienfeld, Machine learning, quantum chemistry, and chemical space, *Rev. Comput. Chem.*, 2017, **30**, 225–256.
2 O. A. von Lilienfeld, Quantum machine learning in chemical compound space, *Angew. Chem., Int. Ed.*, 2018, **57**(16), 4164–4169.
3 O. A. von Lilienfeld, K.-R. Müller and A. Tkatchenko, Exploring chemical compound space with quantum-based machine learning, *Nat. Rev. Chem.*, 2020, 1–12.
4 O. A. von Lilienfeld and K. Burke, Retrospective on a decade of machine learning for chemical discovery, *Nat. Commun.*, 2020, **11**(1), 4895.
5 B. Huang and O. A. von Lilienfeld, *Ab initio* machine learning in chemical compound space, *arXiv*, 2021, preprint, arXiv:2012.07502.
6 M. Rupp, A. Tkatchenko, K.-R. Müller and O. A. von Lilienfeld, Fast and accurate modeling of molecular atomization energies with machine learning, *Phys. Rev. Lett.*, 2012, **108**(5), 058301.
7 R. Ramakrishnan and O. A. von Lilienfeld, Many molecular properties from one kernel in chemical space, *Chimia*, 2015, **69**(4), 182–186.
8 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, *J. Chem. Theory Comput.*, 2017, **13**(11), 5255–5264.
9 A. Stuke, M. Todorović, M. Rupp, C. Kunkel, K. Ghosh, L. Himanen and P. Rinke, Chemical diversity in molecular orbital energy predictions with kernel ridge regression, *J. Chem. Phys.*, 2019, **150**(20), 204121.
10 A. Patra, R. Batra, A. Chandrasekaran, C. Kim, T. D. Huan and R. Ramprasad, A multi-fidelity information-fusion approach to machine learn and predict polymer bandgap, *Comput. Mater. Sci.*, 2020, **172**, 109286.
11 F. A. Faber, A. Lindmaa, O. A. von Lilienfeld and R. Armiento, Machine learning energies of 2 million elpasolite (A B C 2 D 6) Crystals, *Phys. Rev. Lett.*, 2016, **117**(13), 135502.
12 G. Pilania, J. E. Gubernatis and T. Lookman, Multi-fidelity machine learning models for accurate bandgap predictions of solids, *Comput. Mater. Sci.*, 2017, **129**, 156–163.
13 O. Isayev, C. Oses, C. Toher, E. Gossett, S. Curtarolo and A. Tropsha, Universal fragment descriptors for predicting properties of inorganic crystals, *Nat. Commun.*, 2017, **8**, 15679.
14 W. Li, R. Jacobs and D. Morgan, Predicting the thermodynamic stability of perovskite oxides using machine learning models, *Comput. Mater. Sci.*, 2018, **150**, 454–463.
15 K. T. Butler, D. W. Davies, H. Cartwright, O. Isayev and A. Walsh, Machine learning for molecular and materials science, *Nature*, 2018, **559**(7715), 547–555.
16 K. Hansen, G. Montavon, F. Biegler, S. Fazli, M. Rupp, M. Scheffler, O. A. von Lilienfeld, A. Tkatchenko and K.-R. Muller, Assessment and validation of machine learning methods for predicting molecular atomization energies, *J. Chem. Theory Comput.*, 2013, **9**(8), 3404–3419.
17 J. Weinreich, N. J. Browning and O. A. von Lilienfeld, Machine learning of free energies in chemical compound space using ensemble representations: Reaching experimental uncertainty for solvation, *J. Chem. Phys.*, 2021, **154**(13), 134113.
18 G. Montavon, M. Rupp, V. Gobre, A. Vazquez-Mayagoitia, K. Hansen, A. Tkatchenko, K.-R. Müller and O. A. von Lilienfeld, Machine learning of molecular electronic properties in chemical compound space, *New J. Phys.*, 2013, **15**(9), 095003.
19 R. Ramakrishnan, M. Hartmann, E. Tapavicza and O. A. von Lilienfeld, Electronic spectra from TDDFT and machine learning in chemical space, *J. Chem. Phys.*, 2015, **143**(8), 084111.
20 K. Schütt, M. Gastegger, A. Tkatchenko, K.-R. Müller and R. J. Maurer, Unifying machine learning and quantum chemistry with a deep neural network for molecular wavefunctions, *Nat. Commun.*, 2019, **10**(1), 1–10.

21 A. S. Christensen, F. A. Faber and O. A. von Lilienfeld, Operators in quantum machine learning: Response properties in chemical space, *J. Chem. Phys.*, 2019, **150**(6), 064105.
22 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, *Adv. Funct. Mater.*, 2015, **25**(41), 6495–6502.
23 M. A. Shandiz and R. Gauvin, Application of machine learning methods for the prediction of crystal system of cathode materials in lithium-ion batteries, *Comput. Mater. Sci.*, 2016, **117**, 270–278.
24 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 and T. Wu, *et al.*, Design of efficient molecular organic light-emitting diodes by a high-throughput virtual screening and experimental approach, *Nat. Mater.*, 2016, **15**(10), 1120–1127.
25 A. D. Sendek, E. D. Cubuk, E. R. Antoniuk, G. Cheon, Y. Cui and E. J. Reed, Machine learning-assisted discovery of solid li-ion conducting materials, *Chem. Mater.*, 2019, **31**(2), 342–352.
26 A. Zunger, Inverse design in search of materials with target functionalities, *Nat. Rev. Chem.*, 2018, **2**(4), 1–16.
27 P. B. Jørgensen, M. Mesta, S. Shil, J. M. Garca Lastra, K. W. Jacobsen, K. S. Thygesen and M. N. Schmidt, Machine learning-based screening of complex molecules for polymer solar cells, *J. Chem. Phys.*, 2018, **148**(24), 241735.
28 Jmol: an open-source Java viewer for chemical structures in 3D, **https://www.jmol.org/**.
29 F. Neese, The ORCA program system, *Wiley Interdiscip. Rev.: Comput. Mol. Sci.*, 2012, **2**(1), 73–78.
30 J. Behler and M. Parrinello, Generalized neural-network representation of high-dimensional potential-energy surfaces, *Phys. Rev. Lett.*, 2007, **98**(14), 146401.
31 K. Schütt, H. Glawe, F. Brockherde, A. Sanna, K. Müller and E. Gross, How to represent crystal structures for machine learning: Towards fast prediction of electronic properties, *Phys. Rev. B: Condens. Matter Mater. Phys.*, 2014, **89**(20), 205118.
32 F. Pereira, K. Xiao, D. A. 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, *J. Chem. Inf. Model.*, 2017, **57**(1), 11–21.
33 O. T. Unke and M. Meuwly, PhysNet: A neural network for predicting energies, forces, dipole moments, and partial charges, *J. Chem. Theory Comput.*, 2019, **15**(6), 3678–3693.
34 J. Westermayr, M. Gastegger and P. Marquetand, Combining SchNet and SHARC: The SchNarc machine learning approach for excited-state dynamics, *J. Phys. Chem. Lett.*, 2020, **11**(10), 3828–3834.
35 B. Huang and O. A. von Lilienfeld, Quantum machine learning using atom-in-molecule-based fragments selected on the fly, *Nat. Chem.*, 2020, 1–7.
36 F. A. Faber, A. S. Christensen, B. Huang and O. A. von Lilienfeld, Alchemical and structural distribution based representation for universal quantum machine learning, *J. Chem. Phys.*, 2018, **148**(24), 241717.
37 A. S. Christensen, L. A. Bratholm, F. A. Faber and O. Anatole von Lilienfeld, FCHL revisited: Faster and more accurate quantum machine learning, *J. Chem. Phys.*, 2020, **152**(4), 044107.
38 K. Hansen, F. Biegler, R. Ramakrishnan, W. Pronobis, O. A. von Lilienfeld, K.-R. Müller and A. Tkatchenko, Machine learning predictions of molecular properties: Accurate many-body potentials and nonlocality in chemical space, *J. Phys. Chem. Lett.*, 2015, **6**(12), 2326–2331.
39 S. Chmiela, A. Tkatchenko, H. E. Sauceda, I. Poltavsky, K. T. Schütt and K.-R. Müller, Machine learning of accurate energy-conserving molecular force fields, *Sci. Adv.*, 2017, **3**(5), e1603015.
40 S. Chmiela, H. E. Sauceda, K.-R. Müller and A. Tkatchenko, Towards exact molecular dynamics simulations with machine-learned force fields, *Nat. Commun.*, 2018, **9**(1), 3887.
41 A. S. Christensen and O. A. von Lilienfeld, On the role of gradients for machine learning of molecular energies and forces, *Mach. Learn.: Sci. Technol.*, 2020, **1**(4), 45018.
42 N. J. Browning, R. Ramakrishnan, O. A. von Lilienfeld and U. Roethlisberger, Genetic optimization of training sets for improved machine learning models of molecular properties, *J. Phys. Chem. Lett.*, 2017, **8**(7), 1351–1359.
43 S. Kubatkin, A. Danilov, M. Hjort, J. Cornil, J.-L. Bredas, N. Stuhr-Hansen, P. Hedegård and T. Bjørnholm, Single-electron transistor of a single organic molecule with access to several redox states, *Nature*, 2003, **425**(6959), 698–701.
44 J. Roncali, Molecular engineering of the band gap of π-conjugated systems: Facing technological applications, *Macromol. Rapid Commun.*, 2007, **28**(17), 1761–1775.
45 M. Jurow, A. E. Schuckman, J. D. Batteas and C. M. Drain, Porphyrins as molecular electronic components of functional devices, *Coord. Chem. Rev.*, 2010, **254**(19-20), 2297–2310.
46 P. M. Beaujuge and J. R. Reynolds, Color control in π-conjugated organic polymers for use in electrochromic devices, *Chem. Rev.*, 2010, **110**(1), 268–320.
47 S. X. Tao, X. Cao and P. A. Bobbert, Accurate and efficient band gap predictions of metal halide perovskites using the DFT-1/2 method: GW accuracy with DFT expense, *Sci. Rep.*, 2017, **7**(1), 1–9.
48 A. Stoliaroff and C. Latouche, Accurate *ab initio* calculations on various PV-based materials: Which functional to be used?, *J. Phys. Chem. C*, 2020, **124**(16), 8467–8478.
49 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.*, 2009, **131**(25), 8732–8733.
50 R. Ramakrishnan, P. O. Dral, M. Rupp and O. A. von Lilienfeld, Big data meets quantum chemistry approximations: The Δ-machine learning approach, *J. Chem. Theory Comput.*, 2015, **11**(5), 2087–2096.
51 J. Ridley and M. Zerner, An intermediate neglect of differential overlap technique for spectroscopy: Pyrrole and the azines, *Theor. Chim. Acta*, 1973, **32**(2), 111–134.
52 M. C. Zerner, Semiempirical molecular orbital methods, *Rev. Comput. Chem.*, 1991, **2**, 313–365.

53 L. Hedin, New method for calculating the one-particle Green's function with application to the electron-gas problem, *Phys. Rev.*, 1965, **139**(3A), 796–823.
54 F. Aryasetiawan and O. Gunnarsson, The GW method, *Rep. Prog. Phys.*, 1998, **61**(3), 237–312.
55 L. Ruddigkeit, R. Van Deursen, L. C. Blum and J.-L. Reymond, Enumeration of 166 billion organic small molecules in the chemical universe database GDB-17, *J. Chem. Inf. Model.*, 2012, **52**(11), 2864–2875.
56 R. Ramakrishnan, P. O. Dral, M. Rupp and O. A. von Lilienfeld, Quantum chemistry structures and properties of 134 kilo molecules, *Sci. Data*, 2014, **1**(1), 1–7.
57 P. Hohenberg and W. Kohn, Inhomogeneous electron gas, *Phys. Rev.*, 1964, **136**(3B), B864.
58 W. Kohn and L. J. Sham, Self-consistent equations including exchange and correlation effects, *Phys. Rev.*, 1965, **140**(4A), A1133.
59 A. D. Becke, Density-functional thermochemistry. III. The role of exact exchange, *J. Chem. Phys.*, 1993, **98**(7), 5648–5652.
60 J. Gilmer, S. S. Schoenholz, P. F. Riley, O. Vinyals and G. E. Dahl, Neural message passing for quantum chemistry, International Conference on Machine Learning, 1263–1272, PMLR, 2017.
61 K. T. Schütt, H. E. Sauceda, P.-J. Kindermans, A. Tkatchenko and K.-R. Müller, SchNet-A deep learning architecture for molecules and materials, *J. Chem. Phys.*, 2018, **148**(24), 241722.
62 C. Chen, W. Ye, Y. Zuo, C. Zheng and S. P. Ong, Graph networks as a universal machine learning framework for molecules and crystals, *Chem. Mater.*, 2019, **31**(9), 3564–3572.
63 B. Anderson, T.-S. Hy and R. Kondor, Cormorant: Covariant molecular neural networks, *arXiv*, 2019, preprint, arXiv: 1906.04015.
64 C. Lu, Q. Liu, C. Wang, Z. Huang, P. Lin and L. He, Molecular property prediction: A multilevel quantum interactions modeling perspective, in Proceedings of the AAAI Conference on Artificial Intelligence, 2019, vol. 33, pp. 1052–1060.
65 Z. Liu, L. Lin, Q. Jia, Z. Cheng, Y. Jiang, Y. Guo and J. Ma, Transferable multi-level attention neural network for accurate prediction of quantum chemistry properties *via* multitask learning, *ChemRxiv*, 2020, **12588170**, v1.
66 D. Weininger, SMILES, a chemical language and information system. 1. Introduction to methodology and encoding rules, *J. Chem. Inf. Comput. Sci.*, 1988, **28**(1), 31–36.
67 D. Weininger, A. Weininger and J. L. Weininger, SMILES. 2. Algorithm for generation of unique SMILES notation, *J. Chem. Inf. Comput. Sci.*, 1989, **29**(2), 97–101.
68 D. Weininger, SMILES. 3. DEPICT. Graphical depiction of chemical structures, *J. Chem. Inf. Comput. Sci.*, 1990, **30**(3), 237–243.
69 RDKit: Open-source cheminformatics, **http://www.rdkit.org**, 2006.
70 K.-R. Müller, S. Mika, G. Ratsch, K. Tsuda and B. Schölkopf, An introduction to kernel-based learning algorithms, *IEEE Trans. Neural Netw.*, 2001, **12**(2), 181–201.
71 B. Schölkopf, A. J. Smola and F. Bach, *et al.*, *Learning with kernels: support vector machines, regularization, optimization, and beyond*, MIT Press, 2002.
72 T. Hastie, R. Tibshirani and J. Friedman, *The Elements of Statistical Learning*, Springer, 2009.
73 C. E. Rasmussen and C. K. I. Williams, *Gaussian Processes for Machine Learning*, The MIT Press, 2006.
74 C. Cortes, L. D. Jackel, S. A. Solla, V. Vapnik and J. S. Denker, Learning curves: Asymptotic values and rate of convergence, in Advances in Neural Information Processing Systems, 1994, pp. 327–334.
75 K.-R. Müller, M. Finke, N. Murata, K. Schulten and S.-i Amari, A numerical study on learning curves in stochastic multilayer feedforward networks, *Neural Comput.*, 1996, **8**(5), 1085–1106.
76 A. S. Christensen, F. A. Faber, B. Huang, L. A. Bratholm, A. Tkatchenko, K.-R. Müller and O. A. von Lilienfeld, QML: A Python Toolkit for Quantum Machine Learning, 2017, **https://github.com/qmlcode/qml**.
77 P. Zaspel, B. Huang, H. Harbrecht and O. A. von Lilienfeld, Boosting quantum machine learning models with a multilevel combination technique: Pople diagrams revisited, *J. Chem. Theory Comput.*, 2018, **15**(3), 1546–1559.
78 Z. Liu, L. Lin, Q. Jia, Z. Cheng, Y. Jiang, Y. Guo and J. Ma, Transferable multilevel attention neural network for accurate prediction of quantum chemistry properties *via* multitask learning, *J. Chem. Inf. Model.*, 2021, **61**(3), 1066–1082.
79 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, *J. Chem. Phys.*, 2018, **148**(24), 241718.
80 J. Nigam, M. Willatt and M. Ceriotti, Equivariant representations for molecular Hamiltonians and N-center atomic-scale properties, *arXiv*, 2021, preprint, arXiv:2109.12083.
81 D. Lemm, G. F. von Rudorff and O. A. von Lilienfeld, Improved decision making with similarity based machine learning, *arXiv*, 2022, preprint, arXiv:2205.05633.