Chemical Science

ROYAL SOCIETY OF CHEMISTRY

EDGE ARTICLE

View Article Online
View Journal | View Issue

Cite this: *Chem. Sci.*, 2021, **12**, 10755

All publication charges for this article have been paid for by the Royal Society of Chemistry

Received 17th March 2021
Accepted 29th June 2021

DOI: 10.1039/d1sc01542g

rsc.li/chemical-science

Physically inspired deep learning of molecular excitations and photoemission spectra†

Julia Westermayr and Reinhard J. Maurer *

Modern functional materials consist of large molecular building blocks with significant chemical complexity which limits spectroscopic property prediction with accurate first-principles methods. Consequently, a targeted design of materials with tailored optoelectronic properties by high-throughput screening is bound to fail without efficient methods to predict molecular excited-state properties across chemical space. In this work, we present a deep neural network that predicts charged quasiparticle excitations for large and complex organic molecules with a rich elemental diversity and a size well out of reach of accurate many body perturbation theory calculations. The model exploits the fundamental underlying physics of molecular resonances as eigenvalues of a latent Hamiltonian matrix and is thus able to accurately describe multiple resonances simultaneously. The performance of this model is demonstrated for a range of organic molecules across chemical composition space and configuration space. We further showcase the model capabilities by predicting photoemission spectra at the level of the GW approximation for previously unseen conjugated molecules.

1 Introduction

The photoelectric effect[1] describes the response of molecules and materials to electromagnetic radiation by emission of electrons. This effect plays a fundamental role in daily life, but also in cutting-edge technology, such as optoelectronic devices,[2,3] regenerative electron sources for free-electron lasers,[4] or photovoltaics, for instance to design artificial ion pumps that mimic nature.[5]

Novel functional materials in modern optoelectronic devices are often characterized by their molecular charge transport properties between acceptor and donor molecules. Such devices include organic diodes and transistors, which crucially depend on the subtle alignment of molecular acceptor and donor levels of different compounds with respect to each other. These fundamental molecular resonances associated with electron addition and removal in matter can be studied with photoemission and inverse photoemission spectroscopy.[6,7] However, the search for optimal materials combinations is limited by the speed at which organic materials combinations can be spectroscopically characterized. This is exacerbated by the challenge of interpreting macroscopically averaged photoemission data for complex molecules.[8–11]

First-principles simulation of photoemission signatures have the potential to dramatically accelerate high throughput screening of organic materials, but the high computational cost associated with accurate many-body excited-state calculations limits their applicability to small molecular systems.[12,13] Machine learning (ML) methods have the ability to overcome the gap between experiment and theory for spectroscopic characterization by reducing the computational effort of spectroscopic simulations without sacrificing prediction accuracy.[14,15]

ML methods in the context of spectroscopy have previously focused on predicting single energy levels,[15–19] oscillator strengths,[20,21] dipole moments,[22–24] highest occupied molecular orbital (HOMO) and lowest unoccupied molecular orbital (LUMO) energies[25–28] or band gaps.[29–31] They have also been applied successfully to identify and characterize structures from X-ray absorption spectra.[32–34] Electronic excitations of molecules across chemical compound space show crossings of states with different character and discontinuous behaviour. For ML models based on smooth features to capture this behaviour while simultaneously predicting multiple electronic excitations is a formidable challenge.[15,35] By predicting spectral lineshapes[36,37] or continuous densities-of-states[38] directly, some of these problems can be circumvented as spectral signatures are smooth. Furthermore spectra can be represented by basis functions or discrete grids providing a consistent representation that is independent of the number of energy levels or the size of the molecule.[39–41] However, a consequence of this simplification is that direct information on the number and character of the molecular resonances is lost.

In this work, we develop a deep convolutional neural network that accurately predicts molecular resonances across

Department of Chemistry, University of Warwick, Gibbet Hill Road, Coventry, CV4 7AL, UK. E-mail: r.maurer@warwick.ac.uk

† Electronic supplementary information (ESI) available. See DOI: 10.1039/d1sc01542g

a wide range of organic molecular compounds. We encode the fundamental physics of molecular resonances by representing them *via* a Hamiltonian matrix associated with a closed set of secular equations. In contrast to previous efforts,[42–45] this matrix representation is not based on local atomic orbital features and the elements of this matrix have no direct physical correspondence beyond the fact that the matrix eigenvalues correspond to the learned molecular resonances. As we are only training on rotationally invariant quantities, the model achieves this without the need to explicitly encode vectorial[46–49] or tensorial equivariance properties[23,25] beyond the rotationally invariant representation of the input molecular coordinates.[28] The simple algebraic modification of describing vectorial targets by diagonalization of a matrix output leads to increased learning rates, reduced prediction errors, and increased transferability in predicting electron addition and removal energies across molecular composition space. We showcase the capabilities of this model by predicting photoemission spectra of previously unseen organic electronics precursor molecules at the level of Density Functional Theory (DFT). We further show that the model can be augmented to account for solvation effects or many-body electron correlation effects using only a small fraction of the original training data. Correlation effects are described at the level of GW many-body perturbation theory, which provides spectroscopic predictions of large, complex molecules in close agreement with experiment.

2 Results

2.1 Scalar, vectorial, and matrix-valued deep learning representations of molecular resonances

The deep convolutional neural network we propose is based on the SchNet framework[28,50] and its architecture is illustrated in Fig. 1.

In order to learn n molecular resonances with the conventional scalar SchNet model, n ML models, one for every electronic state or resonance i need to be trained. In the following, we refer to this as a one-state (1S) model (panel a). Similarly, a vector of n molecular resonances can be represented using one ML model with a single vectorial output, which we refer to as multi-state (MS) model (panel b).[51] This is identical to a previously proposed model in the context of photochemistry.[35] The pseudo-Hamiltonian model (SchNet + H), which we propose here is shown in panel c and internally builds an ML basis that satisfies the properties of a quantum mechanical Hamiltonian, *i.e.*, it is symmetric and has eigenvalues that correspond to electron addition/removal energies. The dimension of the effective Hamiltonian output layer scales with the number of eigenvalues defined by the user. This is in contrast to a full quantum mechanical Hamiltonian, which scales with the size of the molecular system. This advantage makes it feasible to learn a large set of molecular resonances in a defined energy range for molecules of arbitrary size. The eigenvalues are

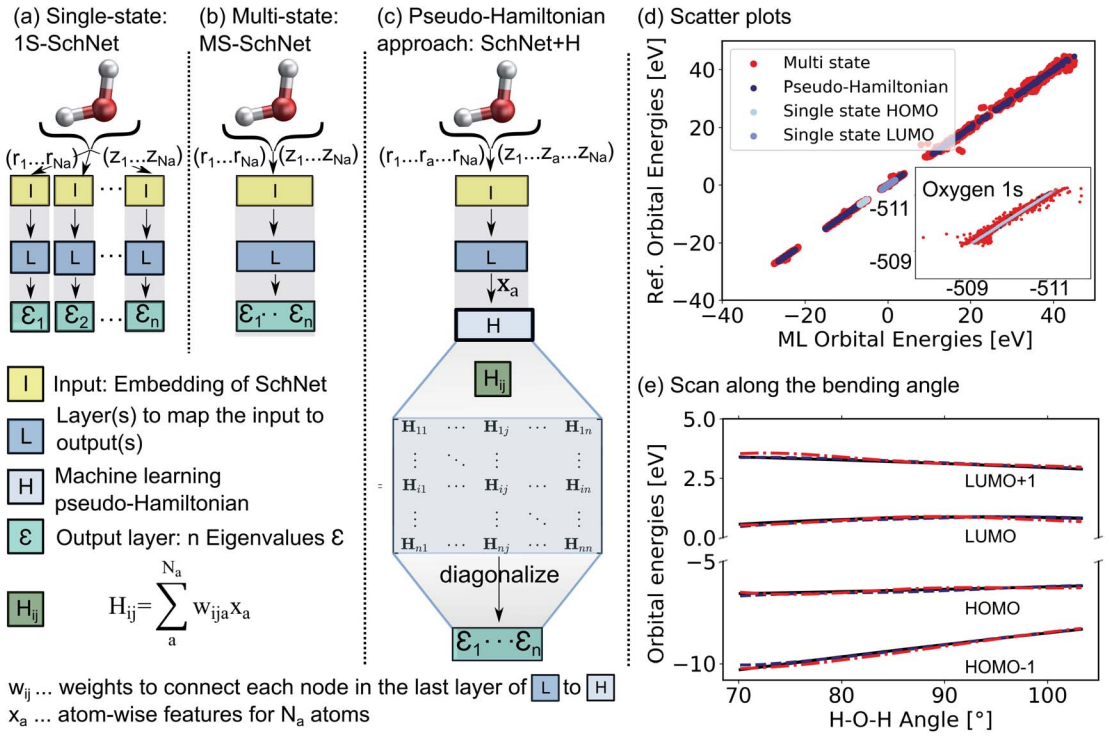

Fig. 1 Comparison of the architecture of (a) a conventional single-state ML model (1S-SchNet), (b) a multi-state ML model (MS-SchNet), and (c) the proposed pseudo-Hamiltonian model (SchNet + H) along with the prediction accuracy for fitting 15 eigenvalues of the H_2O molecule. The elements of the Hamiltonian matrix, H_{ij}, are obtained by pooling atomic features, $\mathbf{x}_a$, from the last layer of the network **L**. (d) Scatter plots show the ML-fitted eigenvalues of a test set plotted against the reference eigenvalues. (e) Orbital energies around the HOMO–LUMO gap are plotted along the bending mode of the molecule using the MS-SchNet and SchNet + H models.

obtained after diagonalization of the ML pseudo-Hamiltonian. Further details on the model training are given in the Methods Section 4.

The prediction accuracy of the three models is first analyzed by training on the 15 lowest Kohn–Sham DFT eigenvalues of 1000 configurations of the H_2O molecule generated by *ab initio* molecular dynamics (for details on the training data, see ESI†) as shown in panels d and e of Fig. 1. As can be seen from the scatter plots in Fig. 1d and the prediction errors reported in Table S1,† the set of 15 1S models shows an accurate prediction of eigenvalues compared to the reference values with mean absolute errors (MAEs) ranging from 0.6 meV up to 5.5 meV for a given orbital energy. This is known and expected as each model only has to cover a small energy range.[28] A single deep neural network with multi-variate outputs to predict all 15 eigenvalues shows substantial deviation between reference and prediction across all energies, *i.e.*, for low-lying semi-core as well as for valence and virtual eigenstates (panel e) with MAEs of up to 300 meV. The MS model is about twenty times less accurate in terms of MAEs of the HOMO energy than the 1S models (52 meV *vs.* 2 meV). This finding is in line with similar models reported in the literature.[17,18,22,26,27,35,39,42]

The lack of prediction accuracy of the MS model can be understood as the model has to cover a large range of energies while having to capture the dependence of each eigenvalue as a function of input. In contrast, our proposed model, SchNet + H, which learns eigenvalues indirectly *via* the pseudo Hamiltonian matrix, faithfully reproduces orbital energies across the whole energy range. The maximum MAE is 67 meV and the HOMO orbital energy can be predicted with 26 meV accuracy. Analysis of the learning behaviour shows that the prediction error decreases faster with the number of data points for the SchNet + H model compared to the MS model (see ESI Fig. S1†). In Fig. 1e, the predicted and reference eigenvalue energies of frontier orbitals around the HOMO energy are plotted as a function of the bending angle in H_2O. While all models provide a qualitatively correct description of the smooth dependence, the MS model shows larger deviations with respect to the reference values compared to the SchNet + H model.

2.2 Predicting molecular resonances across chemical space

One might be able to attribute the improved performance of the SchNet + H model compared to MS-SchNet simply to the increased size of the output layer which provides more flexibility. We note that both MS-SchNet and SchNet + H have almost the same number of parameters and even a further increase of the number of nodes and layers in the MS-SchNet model does not yield a better prediction (see ESI† for more details). Instead, we attribute the improved accuracy of SchNet + H to the fact that the matrix elements of the pseudo-Hamiltonian are much smoother functions in chemical space than the molecular resonances on which the model is trained. By decoupling the algebraic diagonalization that gives rise to avoided crossings and non-differential behaviour of molecular resonances from the ML model, we train an effective representation with smoother coordinate dependence. This can be seen in Fig. 2

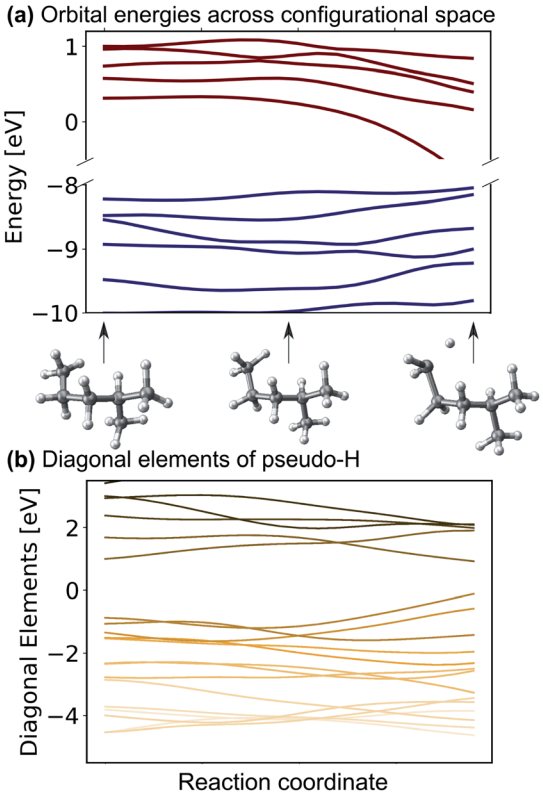

Fig. 2 (a) Eigenvalues and (b) diagonal matrix elements of the pseudo-Hamiltonian of the SchNet + H model trained on molecules of the QM7-X data set[52] along a trajectory of conformational change in 2-methylpentane.

where the orbital energies and diagonal matrix elements predicted by the SchNet + H model are shown along a reaction coordinate of 2-methylpentane. The structures are part of the first subset of the QM7-X data set[52] on which the SchNet + H model has been trained. The QM7-X data set is an extension of QM7 (ref. 53) that contains 4.2 M equilibrium and non-equilibrium structures of a large number of molecules across chemical compound space. The quantum machine data sets[54] are often used as a benchmark in ML studies,[28,39,55–60] which we have also done here (plots reporting model accuracy are given in ESI Fig. S3c†). The diagonal elements of the internally formed ML basis shown in panel b vary more continuously with molecular composition than the orbital energies shown in panel a. The diagonal elements show numerous crossings along the coordinate, which is reminiscent of the behaviour of quasi-diabatic representations often used to represent multiple electronic states in computational photochemistry.[61,62] The smooth functional form is found for different elements of the pseudo-Hamiltonian matrix and is not only true for the diagonal elements. This finding also holds for variation across chemical composition space. In ESI Fig. S3,† we show the behaviour of eigenvalues and Hamiltonian matrix elements predicted by the ML model along a coordinate of molecules with increasing number of atoms. The smooth functional behaviour of Hamiltonian matrix elements is also discernible in this case. It can be seen that the matrix elements are randomly distributed in terms

of value and position in the matrix with slightly more weight on diagonal elements for larger molecules. It is noticeable that the model makes effective use of all matrix elements.

To further validate the accuracy of the model, we train it to represent 12 Kohn–Sham eigenvalues of ethanol[42,54] along a molecular dynamics trajectory. Scatter plots are shown in ESI Fig. S2† and errors on a hold-out test set are reported in the ESI Table S2† along with other models reported in the literature for comparison. By comparing broadly across literature, we find that SchNet + H provides the same or better accuracy for the prediction of multiple resonances (between 12 and 53 across different training sets) compared to what most other models achieve for a single molecular resonance (*e.g.* the HOMO).[17,18,26,35,39,63] The exception to this is the atomic-orbital-based SchNOrb Hamiltonian model,[42] which predicts an average MAE for the same 12 eigenvalues of about 0.02 eV. However, we note that SchNOrb is a much larger and more flexible model, which is trained on eigenvalues and Hamiltonian matrices to predict all molecular eigenvalues (with a total averaged MAE of 0.48 eV). SchNOrb in its current form can only predict eigenvalues as a function of atomic positions for a fixed molecular composition.

Encouraged by the promising performance of SchNet + H, we have trained a transferable model of molecular electronic states based on the OE62 data base.[66] This data set is especially challenging as it features greater elemental diversity and more heteroatoms and functional groups than there are in the QM9 or QM7-X data bases.[26,66] The 62k molecules in OE62 are selected from known molecular crystal structures in the Cambridge Structural Database.[67] For each equilibrium structure, the data set reports Kohn–Sham orbital eigenvalues calculated at the PBE + vdW and hybrid PBE (PBE0) functional level of DFT. The SchNet + H model trained on the PBE0 orbital energies is termed ML(PBE0). The predicted orbital energies against reference values of a test set are shown in Fig. 3a in light blue. The SchNet + H model is trained to capture up to 53 electronic states between −10 eV up to and including the LUMO+1 state. The model error for each data point in the whole training set shows a very large deviation for some systems with particularly high structural complexity. One such outlier is shown in panel a, which contains an 8-membered nitrogen cage in the center (see also Fig. S4 in the ESI†). We note that these data points do not influence the model accuracy and its ability to generalize across chemical compound space, which we have tested by removing outliers and retraining the model. Training errors are further reported along with the number of training data in ESI Table S2.† The model error (MAE of 0.13 eV) is quite convincing with few prominent deviations at low orbital energies that are associated with a small number of outlier molecules of particularly high structural complexity.

For a subset of 30 876 molecules, the OE62 set further reports PBE0 (ref. 68) eigenvalues calculated with the Multipole

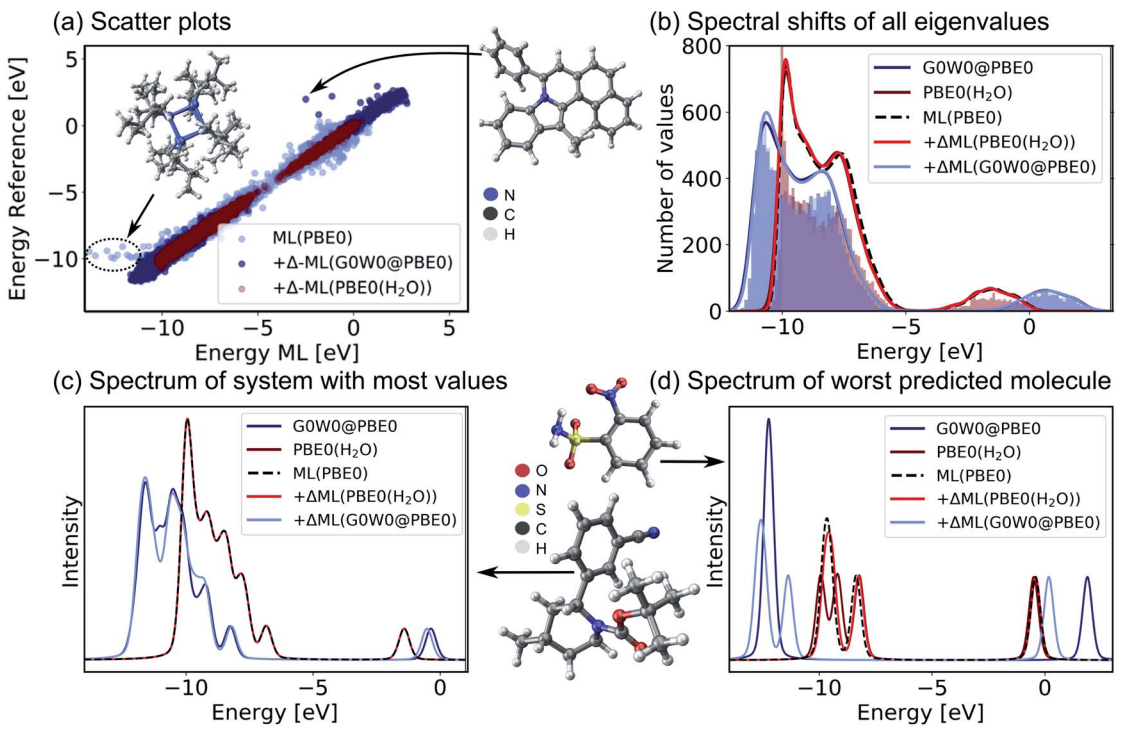

Fig. 3 Validation of the SchNet + H model to predict PBE0 eigenvalues of the OE62 data base and the Δ-ML model that corrects the PBE0 fitted eigenvalues to G0W0@PBE0 accuracy or to PBE0 + implicit water solvation. (a) Scatter plots of a test set show the accuracy of each model. (b) Histograms of orbital eigenvalue (quasiparticle) energies for PBE0 in implicit water solvation and G_0W_0@PBE0 are shown for the GW5000 data set. A Gaussian envelope with 0.5 eV width is placed over each peak to depict the energy shifts between data sets and ML models. The eigenvalues of (c) the molecule with most eigenvalues within the modelled energy range and with (d) the worst predicted eigenvalues in the test set are shown using a Pseudo-Voigt lineshape[64,65] based on a 30% Lorentzian and 70% Gaussian ratio with 0.5 eV width.

Expansion (MPE) implicit solvation method.[69] For a further subset of 5239 molecules in vacuum (termed GW5000), the data set reports quasiparticle energies calculated at the many-body perturbation theory in the G_0W_0@PBE0 approximation.[70–72] With the exception of the HOMO, Kohn–Sham orbital energies lack a physical meaning[73] and important properties of optoelectronic materials, such as donor and acceptor levels[20,39] or band gaps are often incorrectly described.[70] In order to obtain charged excitations in molecules and materials, the GW method[13,71] can be used to correct artifacts that arise from approximations in the exchange–correlation functional in DFT. The computation of quasiparticle energies is computationally unfeasible for the full OE62 data set and for much larger molecular systems with potential relevance in organic electronics. The electronic resonances that include solvation effects and correlation effects captured in the two data subsets should principally deviate from the PBE0 energies of the full data set in relatively systematic ways. We therefore apply a Δ-ML approach[20,74] to train ML models to capture the difference in orbital energy and quasiparticle energy between PBE0 in vacuum and in water and PBE0 and G_0W_0@PBE0, respectively. Our Δ-ML approach is explained in more detail in the Methods section. Briefly, the SchNet + H model of the PBE0 eigenvalues learns a baseline for the full 62k data set (50k training data points), whereas the Δ-ML models learn the difference with respect to this ML(PBE0) baseline from a much smaller training data set (4k).

Test errors of orbital (quasiparticle) energies predicted by the two Δ-ML models are also reported in Fig. 3a. We note that the error distribution is narrower for the Δ-ML-corrected models than for ML(PBE0). Fig. 3b shows that the ML(PBE0) and the two Δ-ML models predict eigenenergies with high fidelity and accurately represent the data sets with a MAE (RMSE) as low as 2 and 4 meV for PBE0(H2O) and G0W0@PBE0, respectively. On closer inspection, we find that the excitation spectrum of the molecule in the test set with the most eigenvalues in the represented energy range shows quantitative agreement with the reference spectrum and a MAE (RMSE) of 29 (52) meV in the vicinity of the peaks (see Fig. 3c). The spectrum for the molecule with the highest prediction error (Fig. 3d) shows noticeable deviations only for the Δ-ML(G0W0@PBE0) model. Here the model predicts a splitting of the HOMO levels and underestimates the energy of the LUMO compared to the reference data with a MAE of 0.51 meV and a RMSE of 0.94 meV on the spectrum in the vicinity of the peaks. We note that this molecule is a rare case in the data base that contains more heteroatoms than carbon atoms, which could be a reason for the increased prediction errors.

The Δ-ML(G0W0@PBE0) is only trained on a subset of 4k datapoints of the GW5000 data set as no quasiparticle energies are available for the full 62k data points of the OE62 data set. By applying the SchNet + H ML(PBE0) and Δ-ML(G0W0@PBE0) models to predict the quasiparticle energies of the full OE62 data set, we can gauge the transferability of the models across chemical space. We find that the models predict the same vertical shift of occupied and unoccupied states between PBE0 and G0W0@PBE0 levels of theory for the full OE62 data set that we have shown in Fig. 3b for the GW5000 set (see ESI Fig. S4b†). In addition, the predictions show a linear correlation of the Kohn–Sham HOMO and LUMO orbital energies with the corresponding quasiparticle energies (Fig. S4a†). This linear relation has previously been identified for HOMO energies of the smaller GW5000 subset in ref. 66, which we can now extend for all orbitals in the OE62 set. Not surprisingly, the application of the Δ-ML(G0W0@PBE0) induces a downward shift of occupied PBE0 energies and an upward shift in energy for unoccupied orbitals to create electron removal and addition quasiparticle energies. Hardly any shift can be found for the eigenenergies obtained from the implicit solvation model indicating that solvation has a minor impact on the molecular resonances.

The combined SchNet + H ML(PBE0) and Δ-ML(G0W0@PBE0) models can predict (inverse) photoemission spectra, ionization potentials and electron affinities of large and complex organic molecules which are well out of reach for *ab initio* calculations at this level of theory. Previous works have predicted individual HOMO and LUMO quasiparticle energies of the GW5000 (ref. 27) and GW100 (ref. 63 and 78) data sets. Our model is able to predict many quasiparticle resonances over a wide energy range and is therefore able to simulate photoemission spectra.

2.3 Prediction of energy levels and photoemission spectra of functional organic molecules

In the following, we report the ML-based prediction of the photoemission spectra of a range of organic molecules which are commonly used as acceptor and donor compounds in organic electronics applications. To showcase the wide applicability of our model, three different types of functional organic molecules are selected: azenes, derivatives of azulenes, and other polycyclic aromatic hydrocarbons. Azulenes are particularly interesting as they exhibit unusually low HOMO–LUMO gaps for molecules of such small conjugation length due to their topological properties.[79,80] Polycyclic aromatic hydrocarbons are often considered for the design of new organic light-emitting diode materials, field-effect transistors or photovoltaics.[3,7,81] Their electronic properties make these molecules not only relevant for optoelectronic applications, but also for other research areas such as astrochemistry[82] and atmospherical chemistry.[83]

The excitation spectra are predicted with the ML model trained on PBE0 orbital energies of the OE62 data set (denoted as ML(PBE0)) and the Δ-ML model trained on the difference of the ML(PBE0) model and the G0W0@PBE0 values of 4k datapoints of the GW5000 data set. The combination of both models is denoted as ML(G0W0@PBE0) in the following. All photoemission spectra shown in Fig. 4a–d and ESI Fig. S6–S8† are ML predictions of molecules the model has not seen before. In addition to the photoemission spectra, the LUMO energies are plotted and the spectra obtained from Kohn–Sham eigenvalues are shown to highlight the Δ-ML quasiparticle correction. The spectra obtained with ML(G0W0@PBE0) are in excellent agreement with experiment. Compared to spectra based on Kohn–Sham orbital energies, they accurately reflect the

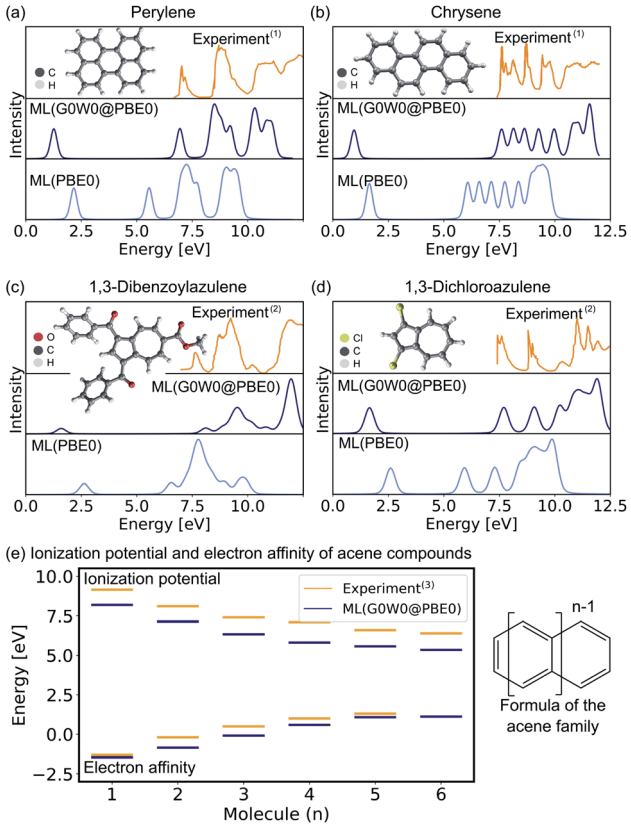

Fig. 4 Experimental and ML predicted photoemission spectra along with the LUMO (quasiparticle) orbital energies at the PBE0 (G0W0@PBE0) level for (a) perylene, (b) chrysene, (c) 1,3-dibenzoylazulene, and (d) 1,3-dichloroazulene. A Pseudo-Voigt lineshape[64,65] based on a 30% Lorentzian and 70% Gaussian ratio with 0.3 eV width was used. (e) Electron affinities and ionization potentials of acene molecules are plotted with increasing ring size. (1)Experimental photoemission spectra have been extracted from ref. 75, (2) ref. 76, and (13) ref. 77.

positions and intensities of photoemission features. In addition, the model correctly predicts the spectral fingerprints of similar molecules and accurately describes substituent effects. For instance, the model accurately predicts the differences of 1,3-dibromoaculene and 1,3-dichloroaculene (see panel d and ESI† for details). Even a highly complex molecule such as 1,3-dibenzoylazulene with 48 atoms (see Fig. 4d), is predicted with high accuracy with respect to the experimental spectrum.

In addition to the photoemission spectra, we predict the electron affinities and ionization potentials of molecules of the acene family. As can be seen in Fig. 4d, acenes are built from linearly condensed benzene rings and are often referred to as "1d graphene strips". Acenes are especially interesting as they are relevant in electronic devices due to their narrow HOMO–LUMO gaps that can result in generally high conductivity.[2,77] The predicted ionization potentials and electron affinities fit well to experimental values although the HOMO–LUMO gaps are slightly underestimated. This underestimation is not an artifact of the ML model, but is a well known limitation of the G0W0 method for acene molecules.[77] Due to the instability of hexacene ($n = 6$), the experimental prediction of charged excitations is challenging, hence no electron affinity value is available to which the ML predictions can be compared.[2] The respective photoemission spectra are reported in ESI Fig. S8† and are in qualitatively good agreement with experimental spectra reported in literature.[77]

3 Conclusion

In this work, we have developed a machine learning model that can be used to predict orbital energies of large and complex molecules in various configurations during molecular dynamics and orbital and quasiparticle energies across chemical compound space in general. By using physical relations and building an internal ML basis that exploits the fundamental symmetries of a quantum chemical Hamiltonian, but does not scale with system size, molecular resonances such as orbital and quasiparticle energies can be predicted with high accuracy. The developed model is accurate enough to be used in combination with a Δ-ML model trained on the difference between the ML predicted orbital energies of DFT and quasiparticle energies from many-body perturbation theory. This provides an extremely data-efficient way to eliminate errors in spectral signatures that arise from exchange–correlation approximations in Kohn–Sham DFT and to achieve close to experimental accuracy in the prediction of photoemission spectra, ionization potentials, and electron affinities. We evidence this by predicting these quantities with high accuracy compared to experiment for unseen azulene-like molecules, acenes, and polyaromatic hydrocarbons that are often targeted for the design of new organic electronic materials.[3] The model clearly has the ability to distinguish between functional groups and predict trends as a function of molecule size in conjugated systems. The results demonstrate the transferability and scalability of the model. While we have only shown the application of this model for frontier orbital and quasiparticle energies, we are confident that it will be similarly applicable to the prediction of core-levels and X-ray photoemission signatures.[6,41]

The ability to efficiently predict molecular resonances at high accuracy is key to enable large-scale computational screening of novel acceptor and donor molecules to be used in organic electronics and thin film device applications.[7,81,84] We expect that the presented method will be very useful in this context. It will likely be especially powerful in combination with generative ML[85,86] or reinforcement learning models[87] that can recommend new molecular structures with specific tailored properties. In this way, a fully automated search algorithm for new molecules with optimally tuned acceptor and donor levels could be created.[81,88,89]

4 Methods

The underlying ML model used in this work is SchNet.[28,90] As the network architecture of SchNet is explained in the original references in details, we will only briefly describe it here: SchNet is a convolutional message-passing neural network that was originally developed to model scalar valued properties and their derivatives[91] and has recently been extended to model multiple

energy levels and multi-state properties in the context of molecular excited states. This model was previously termed SchNarc and we call it MS-SchNet for consistency in this work.[35,92]

4.1 SchNet + H

(MS-)SchNet combines a network that learns the molecular representation in an end-to-end fashion with a network that maps this tailored representation to the targeted outputs. The first part of the network, the input layer **I** in Fig. 1, takes atomic positions, r_1 to r_{Na}, with N_a being the number of atoms in a system, and elemental charges, z_1 to z_{Na}, as an input. It transforms this information into atomistic descriptors using filter-generating networks and atom-wise layers to optimize the representation. This representation enters into the network, **L** in Fig. 1, which itself contains layers that learn atomistic features x_a. These features are sum-pooled and usually form (excitation) energies. The SchNet + H model developed here is an adaption of MS-SchNet, in which the architecture of the network is altered such that the final fully-connected layer represents a symmetric matrix, $\mathbf{H}^{ML}$ (H in Fig. 1), that returns a diagonal matrix of n eigenvalues ε_i^{ML} after diagonalization:

$$\mathrm{diag}(\{\varepsilon_i^{ML}\}) = \mathbf{U}^T\mathbf{H}^{ML}\mathbf{U}. \quad (1)$$

As SchNet learns the molecular representation, the need for extensive hyperparameter search is reduced. As illustrated in Fig. 1, Hamiltonian elements for states i and j, H_{ij}, are obtained by sum-pooling of atomic features, $\mathbf{x}_a$. w_{ija} denotes the weights that connect the last layer of the standard SchNet network to the pseudo-Hamiltonian layer.

$$H_{ij} = \sum_a^{N_a} w_{ija} x_a \quad (2)$$

Diagonalization of the pseudo-Hamiltonian matrix is carried out after each pass trough the network and the eigenvalues predicted by the ML model enter the loss function, L_2:

$$L_2 = \frac{1}{N}\sum_i^n \left(\varepsilon_i^{ML} - \varepsilon_i^{ref}\right) \quad (3)$$

where ε_i^{ref} indicate reference eigenvalues in the training data set. Due to the fact that we backpropagate through the diagonalization, the atom-wise features are connected and form a global molecular representation of the orbital energies.

SchnNet + H models consistently provide better accuracy than MS-SchNet models. While the accuracy of direct training in MS-SchNet can be improved by placing a Gaussian function on top of the orbital energies in the loss function, this did not lead to more accurate results than the SchNet + H model. Our goal was to develop a model that predicts molecular resonances across chemical space and does not scale with system size. We therefore define an energy range within which we represent all orbital energies up to a maximum number of values that defines the size of $\mathbf{H}^{ML}$. The energy range that was fitted for each data set is reported in ESI Table S2.† A varying number of orbital energies are used for training with the maximum number of eigenvalues being 53 for the OE62 and GW5000 training sets.[66] Every molecule that contains fewer orbital energies than the maximum amount of fitted values can be predicted by using a mask in the loss function that makes sure only relevant values are included.

4.2 Δ-MS-SchNet

The GW5000 training set contains 5k data points and represents a subset of the OE62 data set with G0W0@PBE0 quasiparticle energies. Due to the complexity of the data set with molecules up to 100s of atoms, 5k data points are not enough to train a model directly on quasiparticle energies (MAEs of 0.3 eV). To circumvent this problem, Δ-ML[20] was applied. This approach can be used to train the difference between a baseline method and a higher accuracy method. In this case, we trained a model on the difference between the orbital energies obtained from DFT as predicted by the SchNet + H model, ε^{ML}(DFT), and the quasiparticle energies of G0W0@PBE0, ε^{QC}(G0W0):

$$\Delta\varepsilon^{ML}(\mathrm{G0W0} - \mathrm{DFT}) = \varepsilon^{ref}(\mathrm{G0W0}) - \varepsilon^{ML}(\mathrm{DFT}) \quad (4)$$

For the Δ-ML model, a conventional MS model is sufficient as the differences in DFT (predicted by the SchNet + H model) and G0W0 vary less strongly as a function of input than the actual targets.[93,94] The architecture of the Δ-ML model is identical to panel (b) in Fig. 1. The Δ-ML model is trained separately from the SchNet + H model and is not combined in an end-to-end fashion. Nevertheless, the models depend on each other as the SchNet + H models provides the baseline for the Δ-ML model and predictions of both models need to be combined to obtain reliable quasiparticle energies.

Although the accuracy of the Δ-models can be improved by using DFT reference values as the baseline for Δ-models (MAE of 0.02 eV are obtained with DFT baseline models compared to MAEs of 0.16 eV with SchNet + H(PBE0) baseline models), the ML predicted DFT values are chosen as a baseline to circumvent the use of DFT reference calculations for new predictions altogether. This provides an ML prediction that is independent of electronic structure calculations and practical for large-scale screening studies. The predicted G0W0@PBE0 values are obtained by using the following equation:

$$\varepsilon^{ML}(\mathrm{G0W0}) = \varepsilon^{ML}(\mathrm{DFT}) + \Delta\varepsilon^{ML}(\mathrm{G0W0} - \mathrm{DFT}). \quad (5)$$

For the prediction of G0W0@PBE0 values, we thus use two ML models, one SchNet + H model trained on DFT orbital energies and one MS-SchNet model trained on the difference between quasiparticle and orbital energies. Further details on model size, training and test set split, and model parameters can be found in the ESI.† The chosen model parameters are reported in ESI Table S3.†

4.3 Spectra predictions

The comparison to experimental photoemission spectra shown in Fig. 4 and ESI Fig. S5–S7† is obtained by convolution of the orbital energies to account for electronic lifetime broadening,

instrument response, and many-body effects, such as inelastic losses. For the broadening we use a Pseudo-Voigt lineshape[64,65] with 30% Lorentzian and 70% Gaussian and varying widths of 0.3–0.5 eV. The spectral shifts of all eigenvalues of molecules across chemical compound space given in Fig. 3 and ESI Fig. S4 and S7† are obtained by Gaussian convolution with a width of 0.5 eV and subsequent summation.

Author contributions

R. J. M. proposed and supervised the project. J. W. designed and implemented the model. J. W. performed the model training, data acquisition, and analysis. J. W. and R. J. M. discussed and interpreted the data and wrote the manuscript.

Data availability

The extracted experimental data and the data shown in the figures are available on figshare at DOI: 10.6084/m9.figshare.14212595. All code developed in this work is available on http://www.github.com/schnarc. The QM9 data were provided by Adam McSloy and will be published along with the relevant publication for which they were generated.

Conflicts of interest

There are no conflicts to declare.

Acknowledgements

This work was funded by the Austrian Science Fund (FWF) [J 4522-N] (J. W.) and the UKRI Future Leaders Fellowship programme (MR/S016023/1) (R. J. M.). We are grateful for use of the computing resources from the Northern Ireland High Performance Computing (NI-HPC) service funded by EPSRC (EP/T022175/1). Further computing resources were provided *via* the Scientific Computing Research Technology Platform of the University of Warwick and the EPSRC-funded HPC Midlands + computing centre (EP/P020232/1). The authors want to thank Benedikt Klein, Kristof Schütt, and Adam McSloy for helpful discussion regarding this manuscript and Adam McSloy for providing the data for training the QM9 orbital energies.

Notes and references

1 J. Pendry, *Nature*, 1979, **277**, 351–352.
2 M. Watanabe, Y. J. Chang, S.-W. Liu, T.-H. Chao, K. Goto, M. M. Islam, C.-H. Yuan, Y.-T. Tao, T. Shinmyozu and T. J. Chow, *Nat. Chem.*, 2012, **4**, 574–578.
3 H. Okamoto, in *Organic Chemistry of π-Conjugated Polycyclic Aromatic Hydrocarbons: Acenes and Phenacenes*, ed. Y. Kubozono, Springer Singapore, Singapore, 2019, pp. 211–228.
4 F. Liu, *et al.*, *Nat. Commun.*, 2021, **12**, 673.
5 K. Xiao, L. Chen, R. Chen, T. Heil, S. D. C. Lemus, F. Fan, L. Jiang and M. Antonietti, *Nat. Commun.*, 2019, **10**, 74.
6 B. Klein, S. Hall and R. Maurer, *J. Phys.: Condens. Matter*, 2021, **33**, 15.
7 H. Ishii, K. Sugiyama, E. Ito and K. Seki, *Adv. Mater.*, 1999, **11**, 605–625.
8 O. Hofmann, E. Zojer, L. Hörmann, A. Jeindl and R. Maurer, *Phys. Chem. Chem. Phys.*, 2021, **23**(14), 8132–8180.
9 P. Norman and A. Dreuw, *Chem. Rev.*, 2018, **118**, 7208–7248.
10 C.-G. Zhan, J. A. Nichols and D. A. Dixon, *J. Phys. Chem. A*, 2003, **107**, 4184–4195.
11 P. Puschnig, E.-M. Reinisch, T. Ules, G. Koller, S. Soubatch, M. Ostler, L. Romaner, F. S. Tautz, C. Ambrosch-Draxl and M. G. Ramsey, *Phys. Rev. B: Condens. Matter Mater. Phys.*, 2011, **84**, 235427.
12 *Quantum Chemistry and Dynamics of Excited States: Methods and Applications*, ed. L. González and R. Lindh, John Wiley & Sons, 2020.
13 L. Reining, *Wiley Interdiscip. Rev. Comput. Mol. Sci.*, 2018, **8**, e1344.
14 J. Westermayr, M. Gastegger, K. T. Schütt and R. J. Maurer, *J. Chem. Phys.*, 2021, **154**, 230903.
15 J. Westermayr and P. Marquetand, *Chem. Rev.*, 2020, DOI: 10.1021/acs.chemrev.0c00749.
16 J. Behler, *Angew. Chem., Int. Ed.*, 2017, **56**, 12828–12840.
17 T. Zubatyuk, B. Nebgen, N. Lubbers, J. S. Smith, R. Zubatyuk, G. Zhou, C. Koh, K. Barros, O. Isayev and S. Tretiak, arXiv:1909.12963, 2019.
18 J. Westermayr, M. Gastegger, M. F. S. J. Menger, S. Mai, L. González and P. Marquetand, *Chem. Sci.*, 2019, **10**, 8100–8107.
19 W. Pronobis, K. R. Schütt, A. Tkatchenko and K.-R. Müller, *Eur. Phys. J. B*, 2018, **91**, 178.
20 R. Ramakrishnan, M. Hartmann, E. Tapavicza and O. A. von Lilienfeld, *J. Chem. Phys.*, 2015, **143**, 084111.
21 B.-X. Xue, M. Barbatti and P. O. Dral, *J. Phys. Chem. A*, 2020, **124**, 7199–7210.
22 J. Westermayr and P. Marquetand, *J. Chem. Phys.*, 2020, **153**, 154112.
23 Y. Zhang, S. Ye, J. Zhang, C. Hu, J. Jiang and B. Jiang, *J. Phys. Chem. B*, 2020, **124**, 7284–7290.
24 M. Gastegger, J. Behler and P. Marquetand, *Chem. Sci.*, 2017, **8**, 6924–6935.
25 K. T. Schütt, O. T. Unke and M. Gastegger, arXiv:2102.03150, 2021.
26 A. Stuke, M. Todorović, M. Rupp, C. Kunkel, K. Ghosh, L. Himanen and P. Rinke, *J. Chem. Phys.*, 2019, **150**, 204121.
27 G. Tirimbó, O. Caylak and B. Baumeier, arXiv:2012.01787, 2020.
28 K. T. Schütt, H. E. Sauceda, P.-J. Kindermans, A. Tkatchenko and K.-R. Müller, *J. Chem. Phys.*, 2018, **148**, 241722.
29 Y. Zhuo, A. Mansouri Tehrani and J. Brgoch, *J. Phys. Chem. Lett.*, 2018, **9**, 1668–1673.
30 G. Pilania, J. Gubernatis and T. Lookman, *Comput. Mater. Sci.*, 2017, **129**, 156–163.
31 O. Isayev, c. Oses, c. Toher, E. Gossett, S. Curtarolo and A. Tropsha, *Nat. Commun.*, 2017, **8**, 15679.
32 C. Zheng, K. Mathew and C. Chen, *npj Comput. Mater.*, 2018, **4**, 12.

33 J. Timoshenko, D. Lu, Y. Lin and A. I. Frenkel, *J. Phys. Chem. Lett.*, 2017, **8**, 5091–5098.
34 J. Timoshenko, A. Anspoks, A. Cintins, A. Kuzmin, J. Purans and A. I. Frenkel, *Phys. Rev. Lett.*, 2018, **120**, 225502.
35 J. Westermayr, M. Gastegger and P. Marquetand, *J. Phys. Chem. Lett.*, 2020, **11**, 3828–3834.
36 A. A. Kananenka, K. Yao, S. A. Corcelli and J. L. Skinner, *J. Chem. Theory Comput.*, 2019, **15**, 6850–6858.
37 A. Sanchez-Gonzalez, *et al.*, *Nat. Commun.*, 2017, **8**, 15461.
38 V. Fung, G. Hu, P. Ganesh and B. G. Sumpter, *Nat. Commun.*, 2021, **12**, 88.
39 K. Ghosh, A. Stuke, M. Todorović, P. B. Jørgensen, M. N. Schmidt, A. Vehtari and P. Rinke, *Adv. Sci.*, 2019, **6**, 1801367.
40 C. Ben Mahmoud, A. Anelli, G. Csányi and M. Ceriotti, *Phys. Rev. B: Condens. Matter Mater. Phys.*, 2020, **102**, 235130.
41 C. D. Rankine, M. M. M. Madkhali and T. J. Penfold, *J. Phys. Chem. A*, 2020, **124**, 4263–4270.
42 K. Schütt, M. Gastegger, A. Tkatchenko, K.-R. Müller and R. J. Maurer, *Nat. Commun.*, 2019, **10**, 1–10.
43 Z. Qiao, M. Welborn, A. Anandkumar, F. R. Manby and T. F. Miller, *J. Chem. Phys.*, 2020, **153**, 124111.
44 M. Welborn, L. Cheng and T. F. Miller, *J. Chem. Theory Comput.*, 2018, **14**, 4772–4779.
45 L. Cheng, M. Welborn, A. S. Christensen and T. F. Miller, *J. Chem. Phys.*, 2019, **150**, 131103.
46 S. Chmiela, H. E. Sauceda, I. Poltavsky, K.-R. Müller and A. Tkatchenko, *Comput. Phys. Commun.*, 2019, **240**, 38–45.
47 S. Batzner, T. E. Smidt, L. Sun, J. P. Mailoa, M. Kornbluth, N. Molinari and B. Kozinsky, arxiv:2101.03164, 2021.
48 B. K. Miller, M. Geiger, T. E. Smidt and F. Noé, arXiv:2008.08461, 2020.
49 N. Thomas, T. Smidt, S. Kearnes, L. Yang, L. Li, K. Kohlhoff and P. Riley, arXiv:1802.08219, 2018.
50 K. T. Schütt, H. Glawe, F. Brockherde, A. Sanna, K.-R. Müller and E. K. Gross, *Phys. Rev. B: Condens. Matter Mater. Phys.*, 2014, **89**, 205118.
51 J. Westermayr, F. A. Faber, A. S. Christensen, O. A. von Lilienfeld and P. Marquetand, *Mach. Learn.: Sci. Technol.*, 2020, **1**, 025009.
52 J. Hoja, L. M. Sandonas, B. G. Ernst, A. Vazquez-Mayagoitia, R. A. DiStasio Jr and A. Tkatchenko, *Sci. Data*, 2021, **8**, 43.
53 M. Rupp, A. Tkatchenko, K.-R. Müller and O. A. von Lilienfeld, *Phys. Rev. Lett.*, 2012, **108**, 058301.
54 http://quantum-machine.org/datasets/.
55 A. S. Christensen, F. A. Faber and O. A. von Lilienfeld, *J. Chem. Phys.*, 2019, **150**, 064105.
56 S. Chmiela, A. Tkatchenko, H. E. Sauceda, I. Poltavsky, K. T. Schütt and K.-R. Müller, *Sci. Adv.*, 2017, **3**, 5.
57 M. Gastegger, L. Schwiedrzik, M. Bittermann, F. Berzsenyi and P. Marquetand, *J. Chem. Phys.*, 2018, **148**, 241709.
58 A. S. Christensen, L. A. Bratholm, F. A. Faber and O. Anatole von Lilienfeld, *J. Chem. Phys.*, 2020, **152**, 044107.
59 H. Kim, J. Park and S. Choi, *Sci. Data*, 2019, **6**, 109.
60 M. Veit, D. M. Wilkins, Y. Yang, R. A. DiStasio and M. Ceriotti, *J. Chem. Phys.*, 2020, **153**, 024113.
61 H. Köppel, J. Gronki and S. Mahapatra, *J. Chem. Phys.*, 2001, **115**, 2377–2388.
62 Y. Shu and D. G. Truhlar, *J. Chem. Theory Comput.*, 2020, **16**, 6456–6464.
63 O. Rahaman and A. Gagliardi, *J. Chem. Inf. Model.*, 2020, **60**, 5971–5983.
64 M. Schmid, H.-P. Steinrück and J. M. Gottfried, *Surf. Interface Anal.*, 2014, **46**, 505–511.
65 M. Schmid, H.-P. Steinrück and J. M. Gottfried, *Surf. Interface Anal.*, 2015, **47**, 1080.
66 A. Stuke, C. Kunkel, D. Golze, M. Todorović, J. T. Margraf, K. Reuter, P. Rinke and H. Oberhofer, *Sci. Data*, 2020, **7**, 58.
67 F. H. Allen, *Acta Crystallogr. B*, 2002, **58**, 380–388.
68 C. Adamo and V. Barone, *J. Chem. Phys.*, 1999, **110**, 6158–6170.
69 M. Sinstein, C. Scheurer, S. Matera, V. Blum, K. Reuter and H. Oberhofer, *J. Chem. Theory Comput.*, 2017, **13**, 5582–5603.
70 D. Golze, M. Dvorak and P. Rinke, *Front. Chem.*, 2019, **7**, 377.
71 L. Hedin, *Phys. Rev.*, 1965, **139**, A796–A823.
72 F. Aryasetiawan and O. Gunnarsson, *Rep. Prog. Phys.*, 1998, **61**, 237.
73 R. Stowasser and R. Hoffmann, *J. Am. Chem. Soc.*, 1999, **121**, 3414–3420.
74 M. Bogojeski, L. Vogt-Maranto, M. Tuckerman, K.-R. Müller and K. Burke, *Nat. Commun.*, 2020, **11**, 5223.
75 D. Dougherty, J. Lewis, R. Nauman and S. McGlynn, *J. Electron Spectrosc. Relat. Phenom.*, 1980, **19**, 21–33.
76 M. S. Deleuze, *J. Chem. Phys.*, 2002, **116**, 7012–7026.
77 T. Rangel, K. Berland, S. Sharifzadeh, F. Brown-Altvater, K. Lee, P. Hyldgaard, L. Kronik and J. B. Neaton, *Phys. Rev. B: Condens. Matter Mater. Phys.*, 2016, **93**, 115206.
78 M. J. van Setten, F. Caruso, S. Sharifzadeh, X. Ren, M. Scheffler, F. Liu, J. Lischner, L. Lin, J. R. Deslippe, S. G. Louie, C. Yang, F. Weigend, J. B. Neaton, F. Evers and P. Rinke, *J. Chem. Theory Comput.*, 2015, **11**, 5665–5687.
79 H. Xin and X. Gao, *ChemPlusChem*, 2017, **82**, 945–956.
80 Y. Chen, Y. Zhu, D. Yang, S. Zhao, L. Zhang, L. Yang, J. Wu, Y. Huang, Z. Xu and Z. Lu, *Chem.–Eur. J.*, 2016, **22**, 14527–14530.
81 Y. Yamaguchi, M. Takubo, K. Ogawa, K.-i. Nakayama, T. Koganezawa and H. Katagiri, *J. Am. Chem. Soc.*, 2016, **138**, 11335–11343.
82 A. K. Lemmens, D. B. Rap, J. M. Thunnissen, B. Willemsen and A. M. Rijs, *Nat. Commun.*, 2020, **11**, 1.
83 A. Cachada, P. Pato, T. Rocha-Santos, E. F. da Silva and A. Duarte, *Sci. Total Environ.*, 2012, **430**, 184–192.
84 J. Niskanen, C. J. Sahle, K. Gilmore, F. Uhlig, J. Smiatek and A. Föhlisch, *Phys. Rev. E*, 2017, **96**, 013319.
85 N. Gebauer, M. Gastegger and K. Schütt, *Adv. Neural Inf. Process. Syst.*, 2019, **32**, 7566–7578.
86 R. Mercado, T. Rastemo, E. Lindelöf, G. Klambauer, O. Engkvist, H. Chen and E. J. Bjerrum, *Mach. Learn.: Sci. Technol.*, 2020, **2**, 025023.
87 G. Simm, R. Pinsler and J. M. Hernandez-Lobato, *Proceedings of the 37th International Conference on Machine Learning*, 2020, pp. 8959–8969.

88 S.-P. Peng and Y. Zhao, *J. Chem. Inf. Model.*, 2019, **59**, 4993–5001.
89 D. C. Elton, Z. Boukouvalas, M. D. Fuge and P. W. Chung, *Mol. Syst. Des. Eng.*, 2019, **4**, 828–849.
90 K. T. Schütt, P. Kessel, M. Gastegger, K. A. Nicoli, A. Tkatchenko and K.-R. Müller, *J. Chem. Theory Comput.*, 2019, **15**, 448–455.
91 K. T. Schütt, P. J. Kindermans, H. E. Sauceda, S. Chmiela, A. Tkatchenko and K. R. Müller, *Advances in Neural Information Processing Systems*, 2017, pp. 992–1002.
92 *SchNarc*, https://github.com/schnarc/SchNarc.
93 M. R. Raghunathan Ramakrishnan, P. O. Dral and O. A. von Lilienfeld, *Sci. Data*, 2014, **1**, 140022.
94 P. O. Dral, A. Owens, A. Dral and G. Csányi, *J. Chem. Phys.*, 2020, **152**, 204110.