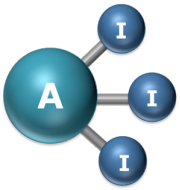

Artificial Intelligence and Augmented Intelligence for Automated Investigations for Scientific Discovery

A deep neural network for generation of functional organic molecules
Project Report
Project Dates: 28/06/2021- 10/09/2021
University of Warwick

Project Student: Rhyan Barrett, University of Warwick
Supervised by: Dr. Reinhard Maurer and Dr. Julia Westermayr, University of Warwick

Report Date: 15/09/2021

AI3SD-Intern-Series:Report-1_Barrett

A deep neural network for generation of functional organic molecules
AI3SD-Intern-Series:Report-1_Barrett
Report Date: 15/09/2021
DOI: 10.5258/SOTON/AI3SD0145
Published by University of Southampton

Network: Artificial Intelligence and Augmented Intelligence for Automated Investigations for Scientific Discovery
This Network+ is EPSRC Funded under Grant No: EP/S000356/1

Principal Investigator: *Professor Jeremy Frey*
Co-Investigator: *Professor Mahesan Niranjan*
Network+ Coordinator: *Dr Samantha Kanza*

Contents

1 Project Details

Title	A deep neural network for generation of functional organic molecules
Project Reference	AI3SD-FundingCall3_001
Supervisor Institution	University of Warwick
Project Dates	28/06/2021- 10/09/2021
Website	Maurer Group
Keywords	Generative, Functional molecule, Optoelectronics, HOMO-LUMO gap, orbital energies, quasiparticle energies, photoelectron spectroscopy, chemical design

2 Project Team

2.1 Project Student

Name and Title	Rhyan Barrett
Employer name / University Department Name	University of Warwick
Work Email	Rhyan.Barrett@warwick.ac.uk
Website Link (if available)	https://www.linkedin.com/in/rhyan-barrett-683707149/

2.2 Project Supervisor

Name and Title	Professor Reinhard Maurer
Employer name / University Department Name	University of Warwick
Work Email	R.Maurer@warwick.ac.uk
Website Link (if available)	https://warwick.ac.uk/rjmaurer

2.3 Researchers & Collaborators

Dr. Julia Westermayr, Postdoctal Research Fellow in Computational Chemistry,
Website: https://warwick.ac.uk/fac/sci/chemistry/research/maurer/maurergroup/people/julia/

3 Lay Summary

This project develops a framework to allow for the generation of new functional organic molecules with special optoelectronic properties that can foster the design of novel electronic devices. We will do this by using a combination of different deep learning models that reduces the time and computational costs needed to generate new molecules with potential relevance

to optoelectronics substantially. Our deep learning models are trained on a labelled data set containing the configurations of organic molecules extracted from a database of crystal-forming molecules and the corresponding excitation energies computed at a high level of accuracy. We first apply a generative model trained on this data set to create new molecules and then predict their optoelectronic properties with a previously developed model for excitation energies and photoemission spectra, which was trained on the same data set. We further develop an automated workflow that retrains the generative model to produce molecules that satisfy specific property targets.

4 Aims and Objectives

Machine learning (ML) methods have great utility in molecular simulation and computational chemistry. Computational predictions of molecular systems based on quantum theory are accurate but computationally extremely demanding. As a consequence they require large-scale shigh performance computing resources. ML methods have the ability to retain the predictive power of quantum theory but reduce the computational cost. [**?**] As large amounts of computational materials simulation data become available in public databases, sufficient data is available to train ML models to perform quantum theoretical predictions. In this project, we used public data to train a deep neural network to predict stable structures of functional organic molecules. The data set is called "OE62" and contains approximately 62,000 structures that are relaxed at Density Functional Theory (DFT) level and were extracted from organic crystals. [**?**]

The main aim of this project was to build a generative model which could produce molecules which had similar properties to our training set but were novel in terms of their molecular structures and were likely not considered before by experiments. The second goal to our project was to create an automated framework that combines this generative model with a recently developed deep learning model that can accurately predict molecular energy levels and photoemission resonances. [**?**] Both of these models, the generative model and the model to predict to energy levels, were validated using DFT reference calculations. Our goal was to combine them to create a loop that generates molecules and then predicts their desired properties. Molecules that satisfy certain target properties, are then used to retrain the existing generative model. The final framework can be used to generate new systems with desirable properties extremely fast. Desired properties that we have considered are a large ionization potential (highest occupied molecular orbital (HOMO) energy level), small electron affinities (lowest unoccupied molecular orbital (LUMO) energy level), and small HOMO-LUMO-gaps. This may be used combine different systems to design organic electronic devices.

5 Methodology

The main tools that were used in the project were FHIaims [**?**, **?**] for the DFT calculations and the two ML models, namely Gschnet, [**?**] the generative model, and Schnet + H, [**?**] a model for predicting orbital and quasiparticle energies. The latter is based on the continuous filter convolutional deep neural network SchNet. [**?**, **?**] We will briefly describe both the models here and the parameters we used in our DFT calculations.

The G-Schnet model is an autoregressive model designed to have an input of position vectors in three dimensional euclidean space as well as an array of scalar values corresponding to the atomic number of atoms in the input molecule. A key component to the G-Schnet model is that it is rotationally invariant which is done by placing atoms sequentially onto a generated molecule according to the learned conditional probability distribution. The model also contains a series of convolutional layers which allows it to capture local properties such as functional groups, rings and any other key features contained in the molecules.

The architecture of the G-Schnet model is very similar to the Schnet model which was originally developed to predict quantum mechanical properties of molecular systems. We will briefly describe the different elements here. First of all, the model represents the position and charge array as an atom representation containing all the relevant information about the system. This is done in two main parts, an embedding layer, mapping atom types to initial feature vectors, and interaction blocks that update the features according to the atomic environments using continuous-filter convolutions. We used nine interaction layers and one embedding layer, each with 128 atom features, similar to the parameters used for training the QM9 data set. The next step in the model is to predict the conditional probabilities of charges and position vectors. To do this we encode the charges of all the atom types using an embedding layer as well as a stop token, telling the model to stop building off the atom. The atomic representation in the first step of the model is then multiplied by the charge embeddings. This new vector is then processed by five atom-wise dense layers with shifted softplus non-linearity and a softmax activation functions. Both the charges and positions of predicted atoms are treated in a similar way. Figure **??** gives an overview of the process described. [**?**]

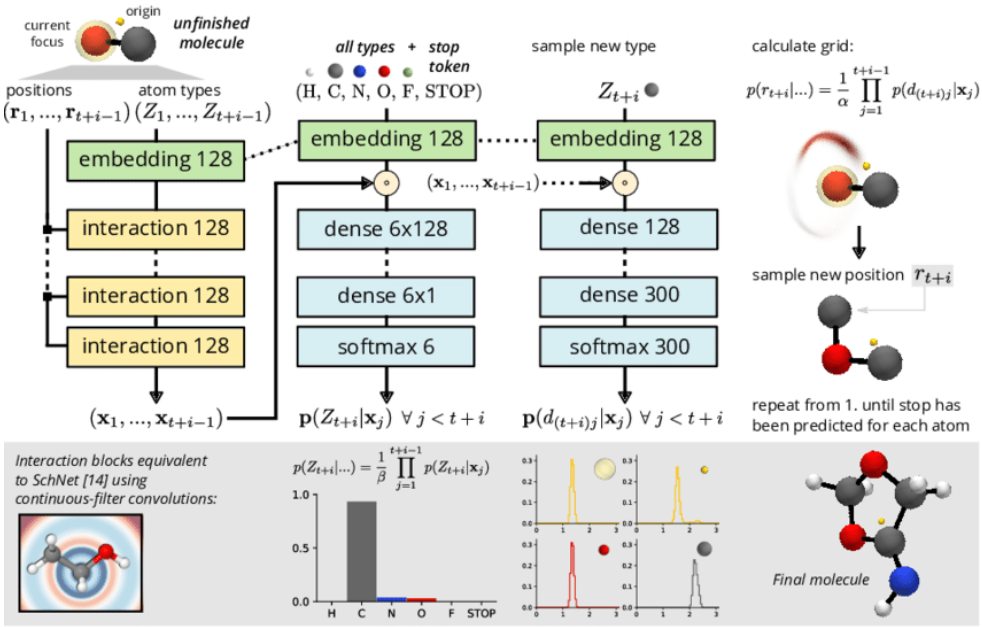

Figure 1: The G-SchNet architecture with an exemplary generation step where an representative atom is generated given two already placed atoms and two tokens; the origin and focus. The final molecule created by iterating the procedure in the lower right. Network layers are shown as blocks where the output shape is shown. The embedding layers have common weights which allow for element-wise multiplication of feature vectors. Figure taken from Ref. **?**

The second model we used in our project is the SchNet+H model. [**?**] As the network architecture of SchNet+H is explained in the original reference **?** in details, we will only briefly describe it here. The SchNet+H is an adaptation of Schnet [**?**, **?**] which is a convolutional message-passing neural network that was originally developed to model scalar valued properties and their derivatives and has recently been extended to model multiple energy levels and multi-state properties in the context of molecular excited states. This model takes the same input as SchNet, i.e., it learns a molecular representation based on interatomic distances. In our case we are using the Schnet+H model to predict the quasiparticle and orbital energies of our newly generated molecules. The OE62 training data set has limited information on the quasiparticle energies so for the prediction of quasiparticle energy values, as such calculations are extremely

expensive and were only provided for 5,000 structures out of 62,000 structures. Thus, we use two ML models, one SchNet+H model trained on DFT orbital energies and one multi-state (MS)-SchNet model, a variation of the SchNet model (further details can be found in references **?** and **?**), trained on the difference between quasiparticle and orbital energies, see equation **??**.

$$\epsilon^{ML}(G0W0) = \epsilon^{ML}(DFT) + \Delta\epsilon^{ML}(G0W0 - DFT) \tag{1}$$

Here the G0W0 corresponds to the quasiparticle energy. We used these predicted energies to then collect molecules with large HOMO energies and small LUMO energies. Prior to predicting orbital and quasiparticle energies we relaxed all molecular geometries at the PBE [**?**] level of density functional theory. This is done in two stages. The method is given below. We used the trust radius enhanced variant of the Broyden-Fletcher-Goldfarb-Shanno (BFGS) algorithm as implemented in FHI-aims with a maximum atomic residual force criterion of $f_{\max} < 0.01 eV\mathring{A}^{-1}$. The electronic wave functions were expanded in a Tier1 (light) basis set. Since our database only contains closed-shell molecules, we performed spin-restricted DFT calculations. Long-range dispersion forces were included in the geometry relaxations using the Tkatchenko-Scheffler method, [**?**] while relativistic effects were treated on the level of the atomic zero-order regular approximation (atomic ZORA). In the second stage, starting from these pre-relaxed structures, we obtained the final geometries by performing a new relaxation with Tier2 (tight) basis sets and a convergence criterion of $f_{\max} < 0.001 eV\mathring{A}^{-1}$. The orbital energies were then computed based on the relaxed structures using the PBE0 [**?**] functional and tight integration settings. The quasiparticle energies at the GW level [**?**] were computed using the setup of reference **?**, which is itself based on reference **?**. Due to the large costs of the simulations, we only computed 200 reference calculations to validate our models. We note that calculations for larger systems failed due to excessive memory requirements. These systems will be revisited in a future work.

6 Results

In order to assess the quality of our model, we compared the generated structures to the molecules in the original training set. We did this by computing bond lengths, bond angles, molecule lengths and ring frequency in both the generated and original datasets. We used the bond lengths and angles that were also used in the original G-SchNet publication. [**?**] In addition to this analysis, we also compared the root mean squared deviation (RMSD) of the atomic positions between our generated molecules and the fully relaxed generated molecules after two steps of DFT with light and tight settings using FHI aims.

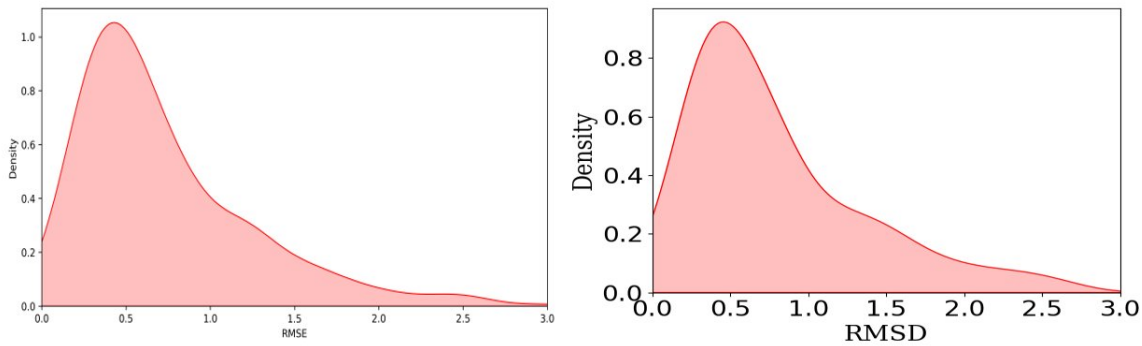

Figure 2: The left image corresponds to the RMSD of light optimised vs generated structures. The right image is the RMSD of tight optimised vs generated structures. We note that although the differences seem minor compared to G-SchNet generated structures, the differences in the DFT/PBE0 orbital energies were large between the light and tight optimised structures.

Above is the RMSD plot between the generated and fully relaxed structures. As can be seen above, the tight DFT optimisation after the light DFT optimisation has very little effect on geometry of the molecule. However, these small structural changes had huge effects on the DFT/PBE0 orbital energies, thus we referred to the tight optimisation settings for analysis of the deep learning models for orbital energies. The RMSD peaks at a value of 0.5 Å, but has a long tail which is mostly due to large molecules which make up a significant portion of the generated dataset. However the RMSDs are still sufficiently small to conclude that molecules are generated close to their most relaxed molecular state. This allows us to exclude any geometry optimisation from any further molecular generation for molecules of small to medium size. For large molecules geometry optimisation may still be necessary.

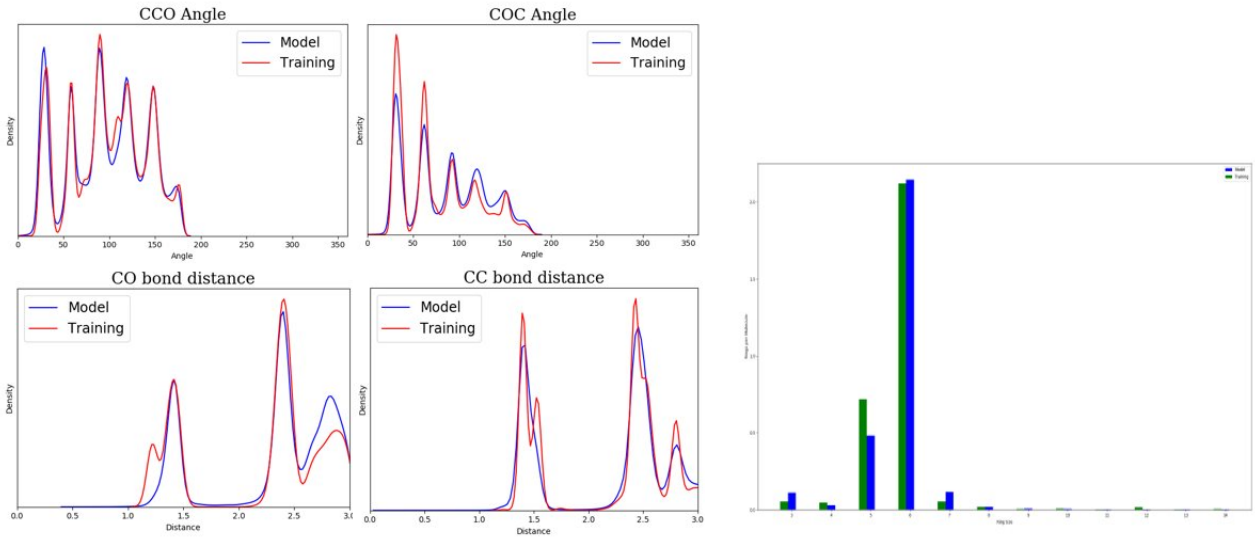

Figure 3: Graphs on the left show the bond angles and length for the generated data set and the original training set. The graph on the right show the rings per molecule for the training and generated data sets.

As proposed above we computed bond length, bond angles, and ring counts of the generated data set to see if G-Schnet captures the bonding patterns and structural distribution of the training set. For the bond length and bond angles, the atomic simulation environment (ASE) [**?**] supports the calculation of a connectivity matrix and distance matrix from which these two properties could be calculated. In this project we calculated the bond lengths of C-C and C-O since they occured frequently in molecules of any sizes and were also used in the original validation of G-SchNet. Similarly we calculated bond angles for C-O-C and C-C-O which is also in line with the original G-SchNet protocol. For ring counts, the canonical SMILES representation of generated molecules were read with RDKit [**?**] and the symmetrized smallest set of smallest rings was computed. The ring count was then divided through by the total number of molecules to get the number of each sized ring per molecule. This is done to give a fairer comparison between the generated and training data set of different sizes. As can be seen in Figure 3, the G-Schnet generated data base produces similar distributions for all desired properties leading us to conclude that G-Schnet has accurately learns the distribution of molecular structures in the training set.

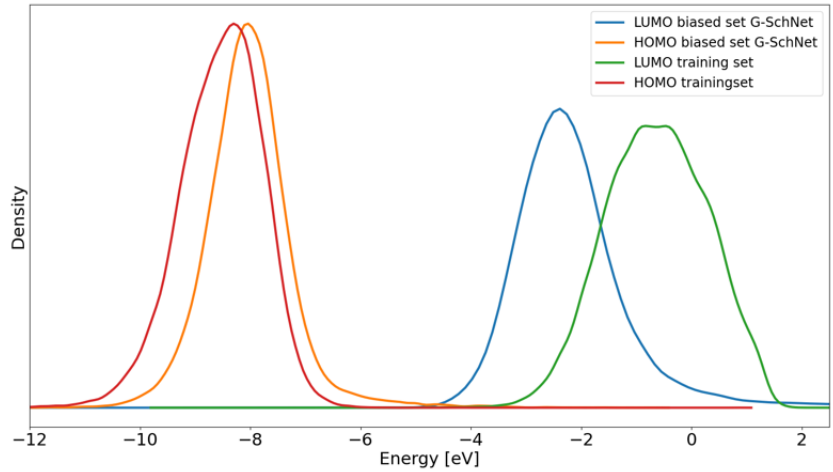

Figure 4: HOMO and LUMO energies given by the biased model

The final part to our project was the creation of biased models in an attempt to generate novel molecules. We were interested in molecules with a large highest occupied molecular orbital (HOMO), a small lowest unoccupied molecular orbital (LUMO) or a small HOMO-LUMO gap. We retrained our original G-Schnet model on generated molecules with each of these properties. Figure 4 shows the HOMO and LUMO energies of the training set and the data set generated by the biased model for small HOMO-LUMO gap. The peaks of the HOMO and LUMO energies can easily be seen to be moving closer, which shows the success of the approach carried out in this work.

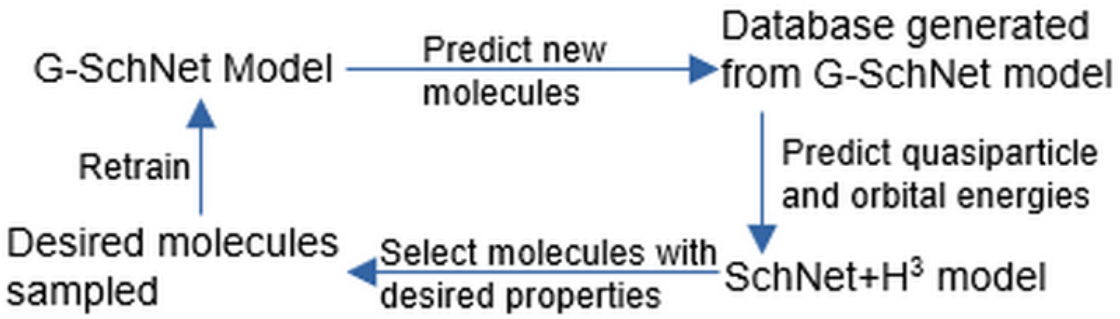

Figure 5: The figure describes the procedure of generating molecules with increasingly better properties. Here we generate molecules using G-SchNet then select molecules based on properties predicted by SchNet+H. Taking these molecules we retrain G-SchNet to give our improved model.

Figure 5 is an overview of how our biased models will improve with each loop. Once we have trained our initial biased model we will again select molecules generated, which have a small HOMO-LUMO gap and use these molecules to train another biased model. The second biased model should be more inclined to generate molecules with an even smaller HOMO-LUMO gap. We will run this loop many times in order to produce molecules with increasingly smaller HOMO-LUMO gap and in doing so satisfy our goal of generating novel molecules with tailored properties.

7 Conclusions & Future Work

We have developed a generative model that can create new large functional organic molecules up to 170 atoms. This is seen through G-Schnet generated databases replicating the distributions of different bond lengths and angles. The RMSD between G-Schnet and optimised structures shows that our model accurately generates molecules close to their most relaxed state. Furthermore, deep learning models trained on the energy levels of the used data set provide very similar quasiparticle energies for DFT-optimised structures and for G-SchNet structures, which gives us enough evidence that G-Schnet produces molecules close to their relaxed state allowing us to leave out any structural optimiser for screening purposes. Through previous work we have seen that SchNet accurately predicts the ionisation potential and electron affinity of these systems. This allowed us to select molecules with a small HOMO-LUMO gap. With these we showed that training a biased model with these molecules reduced the size of the HOMO-LUMO gap of a generated data base. We have thus fulfilled the main objectives we set out, to create a model which can accurately produce molecules of a particular class with useful properties, in our case functional organic molecules with small HOMO-LUMO gap.

In the near future we will build biased models which produce molecules with large HOMO energies and small LUMO energies. In addition we want to perform this optimisation loop many times, i.e., retrain the G-SchNet model with a small set of molecules with targeted properties, to hopefully produce optoelectronic molecules with an increasing smaller HOMO-LUMO gap. We hope that in the future our work will not only be used to suggest new optoelectronic molecules but also molecules for different cpurposes. This may have its applications to medicine and other areas where molecular development is vital.

8 Outputs, Data & Software Links

Github code: https://github.com/rhyan10/Automated-Molecular-Generation/tree/master
Note that this code will be private until the work will be published in a peer-reviewed journal.

References

[1] Carlo Adamo and Vincenzo Barone. Toward reliable density functional methods without adjustable parameters: The PBE0 model. *J. Chem. Phys.*, 110(13):6158–6170, 1999.

[2] Jörg Behler. Four generations of high-dimensional neural network potentials. *Chem. Rev.*, 121(16):10037–10072, 2021.

[3] Volker Blum, Ralf Gehrke, Felix Hanke, Paula Havu, Ville Havu, Xinguo Ren, Karsten Reuter, and Matthias Scheffler. Ab initio molecular simulations with numeric atom-centered orbitals. *Comput. Phys. Commun.*, 180(11):2175–2196, 2009.

[4] Fabio Caruso, Matthias Dauth, Michiel J. van Setten, and Patrick Rinke. Benchmark of gw approaches for the gw100 test set. *J. Chem. Theory Comput.*, 12(10):5076–5087, 2016.

[5] Matthias Ernzerhof and Gustavo E. Scuseria. Assessment of the perdew–burke–ernzerhof exchange-correlation functional. *J. Chem. Phys.*, 110(11):5029–5036, 1999.

[6] Niklas W. A. Gebauer, Michael Gastegger, and Kristof T. Schütt. Generating equilibrium molecules with deep neural networks. *arXiv:1810.11347*, 2018.

[7] Dorothea Golze, Marc Dvorak, and Patrick Rinke. The GW compendium: A practical guide to theoretical photoemission spectroscopy. *Front. Chem.*, 7:377, 2019.

[8] Greg Landrum. Rdkit: Open-source cheminformatics software. 2016.

[9] Ask Hjorth Larsen, Jens Jørgen Mortensen, Jakob Blomqvist, Ivano E Castelli, Rune Christensen, Marcin Dułak, Jesper Friis, Michael N Groves, Bjørk Hammer, Cory Hargus, Eric D Hermes, Paul C Jennings, Peter Bjerre Jensen, James Kermode, John R Kitchin, Esben Leonhard Kolsbjerg, Joseph Kubal, Kristen Kaasbjerg, Steen Lysgaard, Jón Bergmann Maronsson, Tristan Maxson, Thomas Olsen, Lars Pastewka, Andrew Peterson, Carsten Rostgaard, Jakob Schiøtz, Ole Schütt, Mikkel Strange, Kristian S Thygesen, Tejs Vegge, Lasse Vilhelmsen, Michael Walter, Zhenhua Zeng, and Karsten W Jacobsen. The atomic simulation environment—a python library for working with atoms. *J. Phys. Condens. Matter*, 29(27):273002, jun 2017.

[10] K. T. Schütt, P. Kessel, M. Gastegger, K. A. Nicoli, A. Tkatchenko, and K.-R. Müller. Schnetpack: A deep learning toolbox for atomistic systems. *J. Chem. Theory Comput.*, 15(1):448–455, 2019.

[11] 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.*, 148(24):241722, 2018.

[12] Annika Stuke, Christian Kunkel, Dorothea Golze, Milica Todorović, Johannes T. Margraf, Karsten Reuter, Patrick Rinke, and Harald Oberhofer. Atomic structures and orbital energies of 61,489 crystal-forming organic molecules. *Sci. Data*, 7:58, 2020.

[13] Alexandre Tkatchenko and Matthias Scheffler. Accurate molecular van der waals interactions from ground-state electron density and free-atom reference data. *Phys. Rev. Lett.*, 102:073005, Feb 2009.

[14] Julia Westermayr, Michael Gastegger, and Philipp Marquetand. Combining SchNet and SHARC: The SchNarc Machine Learning Approach for Excited-State Dynamics. *J. Phys. Chem. Lett.*, 11(10):3828–3834, 2020.

[15] Julia Westermayr and Reinhard J. Maurer. Physically inspired deep learning of molecular excitations and photoemission spectra. *Chem. Sci.*, 12:10755–10764, 2021.

[16] Igor Ying Zhang, Xinguo Ren, Patrick Rinke, Volker Blum, and Matthias Scheffler. Numeric atom-centered-orbital basis sets with valence-correlation consistency from H to Ar. *New J. Phys.*, 15(12):123033, 2013.