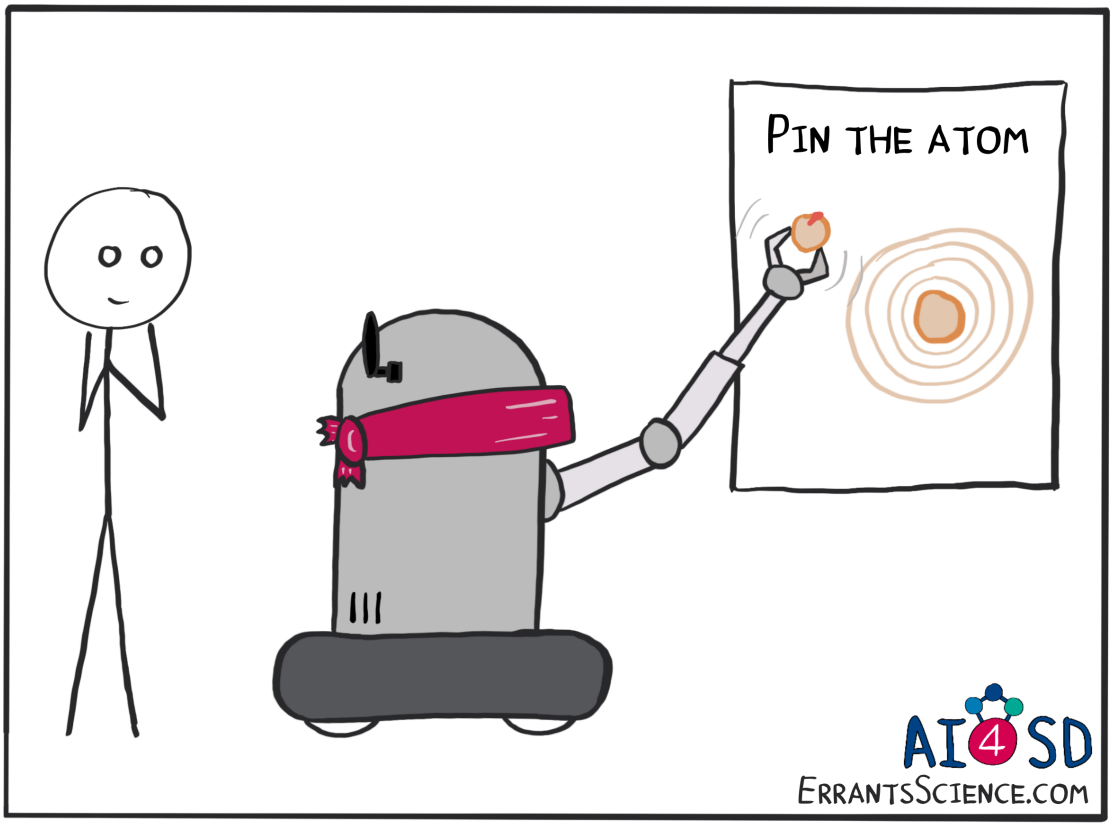
AI 4 Science Discovery Network+

Deep-Learning-Enhanced Quantum Chemistry: Pushing the Limits of Materials Discovery Final Report

Project Dates: 01/07/2019 - 31/12/2019

University of Warwick



Dr. Adam McSloy and Dr. Reinhard J. Maurer
University of Warwick

Report Date: 04/02/2020

AI4SD-Project-Series:Report-2_Maurer_Final

Deep-Learning-Enhanced Quantum Chemistry: Pushing the Limits of Materials Discovery
AI4SD-Project-Series:Report-2_Maurer_Final
Report Date: 04/02/2020
DOI: 10.5258/SOTON/P0041

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	Deep-Learning-Enhanced Quantum Chemistry: Pushing the Limits of Materials Discovery
Funding reference	AI4SD-FundingCall1_027
Lead Institution	University of Warwick
Project Dates	01/07/2019 - 31/12/2019
Website	https://www.warwick.ac.uk/maurergroup
Keywords	Density Functional Tight Binding, deep learning, hybrid metal-organic materials

2 Project Team

2.1 Principal Investigator

Name and Title	Dr. Reinhard Maurer
Employer name / University Department Name	University of Warwick, Department of Chemistry
Work Email	r.maurer@warwick.ac.uk
Website Link (if available)	www.warwick.ac.uk/maurergroup

2.2 Co-Investigators

Name and Title	Dr. Benjamin Hourahine
Employer name / University Department Name	University of Strathclyde, Department of Physics
Work Email	benjamin.hourahine@strath.ac.uk
Website Link (if available)	strath.ac.uk/staff/hourahinebenjamindr

Name and Title	Prof. David Yaron
Employer name / University Department Name	Carnegie Mellon University, Department of Chemistry
Work Email	yaron@cmu.edu
Website Link (if available)	cmu.edu/chemistry/people/faculty/yaron.html

2.3 Researchers & Collaborators

Dr. Balint Aradi (Project Advisor) is a staff scientist at the University of Bremen in the Bremen Centre for Computational Materials Science. His work primarily focuses on multiscale quantum mechanical modelling of semiconductor nanowires, combining *ab initio* and tight binding approaches. Dr. Balint Aradi serves as a project adviser and collaborator who has been engaging in fruitful discussions on the DFTB software implementation aspects of the project.

Dr. Adam McSloy (PDRA) is a research fellow working with RJM on the application of machine-learning methods in electronic structure theory, particularly for the simulation of chemistry at complex interfaces. He has received his PhD from Loughborough University in 2017 for his work on the computational analysis of interface stability within Li-Ion batteries and solid oxide fuel cells. He has developed software for automated interatomic-potential derivation and has extensive experience in the parametrization of interatomic materials potentials. He has been directly employed via the grant from July 2019 to December 2019

3 Publicity Summary

Discovery of new functional materials is central to achieving radical advances in societally important challenges (efficient energy materials, organic solar cells, etc). The vast space of possible materials combinations and compositions gives hope that useful materials exist, but finding ways to navigate this vast space remains a fundamental challenge to materials research. Quantum theoretical computational materials research based on density functional theory has revolutionised the search for new materials over the last twenty years with its ability to predict materials properties based on atomic structure. However, the computational cost of solving quantum mechanical equations at high-performance computing centres remains a severe bottleneck to its commonplace use in research. In this project, we use machine learning to develop an accurate quantum mechanical simulation method of structural, optical, and electronic properties of hybrid organic-metallic materials used in modern solar cells that is efficient enough to run on standard desktop computers for systems that include many thousands of atoms.

4 Executive Summary

Modern materials simulation has become an integral part of chemistry and materials research. Scientific exploration in these fields has become reliant on the ability to ***i***) rapidly explore the configuration and composition space of molecules and materials with molecular dynamics (MD) and ***ii***) accurately predict materials composition, reactivity, and electronic and spectroscopic properties from electronic structure theory and quantum chemical calculations. Machine learning (ML) methods, particularly ML-based construction of high-dimensional representations of energy landscapes and other properties, has hugely benefited our ability to tackle ***i*** at larger time and length scales. The same cannot be said about ***ii***, as the prediction of electronic and spectroscopic properties and chemical reactivity is heavily reliant on non-scalar quantum mechanical observables beyond the PES. This project presents an effort to develop a deep learning approach to the efficient construction of an approximate quantum chemical electronic structure method that satisfies both ***i*** and ***ii***. By training a model with data from an accurate but computationally costly method, we develop a computationally efficient approximate method, namely Density Functional Tight-Binding, with orders of magnitude faster computational prediction of molecule/materials properties and similar accuracy as the original method. We further outline how this approach will be incorporated in future work and how it can be integrated into existing software packages.

5 Aims and Objectives

Our primary goal is to develop a generalised ML-DFTB method that is capable of overcoming the dichotomy between accuracy and computational expense which plagues traditional computational methods. We use a highly efficient parametrised variant of DFT, namely DFTB, where interaction integrals have traditionally been parametrized against reference data. Here we use a machine-learned based representation of these interaction integrals, which can retain the accuracy and prediction capability of DFT calculations at the computational efficiency of a numerically less demanding method. The aim for this project can be broken down into the three proposed tasks:

1. Build upon the DFTB-NN deep learning layer published by Li *et al.*[1] to develop an AI engine that connects molecular structure and composition with electronic observables *via* the intermediate step of a neural-network representation of quantum mechanical interaction integrals. This will involve ***i)*** extending the current network to enable modelling of metallic and d-orbital containing systems, and ***ii)*** Implement additional cost functions to improve performance on larger hybrid systems.

   STATUS: The DFTB-NN infrastructure of Yaron *et al.* has been augmented to enable its use for mteals and metal-organic systems by incorporating d-orbitals, and finite temperature effects. New cost functions have been developed to enable robust and accurate model training of diverse and large training sets.

2. Provide a proof-of-principle application to address a pressing materials science challenge, namely the description of chemical reactivity and electronic properties of metal-organic interfaces as they appear in organic electronics and in heterogeneous catalysis. This will include the analysis of metrics of accuracy and transferability of the ML representations of parameters. It is during this task that we will design and construct a DFT-level dataset on which the model can be trained and a new parameter set derived. A benchmark database of hybrid organic-metallic materials recently published by the PI will then be used to validate said potential.[2]

   STATUS: A large DFT-based training data set of 223,488 structures of metal-nanoparticle-adsorbed molecules has been created and an ML model based on this dataset has been trained. Based on the progress in this project, we will now perform a detailed hyperparameter optimisation of the model and an independent model validation.

3. Integrate the ML-DFTB approach into the market-leading materials simulation software DFTB+, providing access to this work for many industrial and academic users.

   STATUS: We have developed a data pipeline to extract DFTB+ parameters from the ML-DFTB code in the standard DFTB+ file format. Strategies for an interface between the ML-DFTB framework and the DFTB+ have been discussed with the main developers and outlined in a recently submitted article on new features within DFTB+.[3] With the progress enabled by AI3SD funding, we will now proceed to develop the ML-DFTB and DFTB+ interface.

6 Methodology

6.1 Scientific Methodology

The two primary computational chemistry methods employed in this work are those of density functional theory and density functional tight binding theory. Both of which are discussed below. While the former is used to generate the dataset, the augmentation of the latter is the

main focus of this work.

Density Functional Theory (DFT) is an electronic structure method able to address increasingly large systems while accurately predicting a variety of properties for a wide range of materials. It is an effective one-particle theory based on the many-electron Schrödinger equation with the electron density serving as the central variable that defines all properties of the system. In a DFT calculation, the atomic positions define an external potential, for which the energetically most favorable distribution of electrons is found. This is done by calculating interaction integrals between electronic states at different atoms - the computational bottleneck of the calculation.

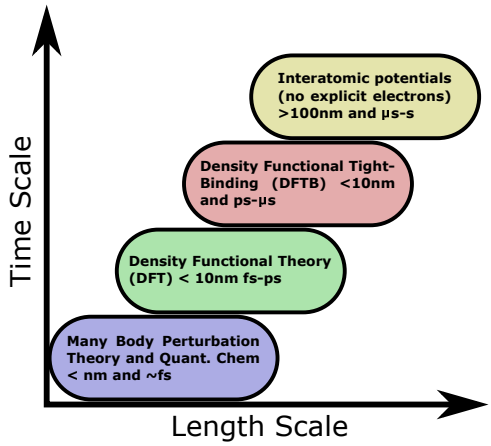


Figure 1. Typical time (fs/ps/μs) and length scale (nm) regimes of atomistic simulation methods.

Density-Functional Tight Binding (DFTB) is a semi-empirical method that bridges the divide between electronic structure methods and force fields (see figure 1).[4] This method provides a three order of magnitude increase in computational speed compared to DFT by replacing the complex quantum mechanical integral evaluation with a set of precalculated, and tabulated, system specific parameters and is commonly employed in organic and molecular material investigations. Thus, it delivers scalability close to interatomic potentials without sacrificing the ability to simulate electronic observables and reactive chemistry. Unfortunately, currently only few of the required integral parameter sets exist that are accurate enough to address modern composite and hybrid organic-inorganic materials such as organic perovskites or metal-organic nanostructures. Furthermore, these parameter sets are notoriously difficult and time-consuming to construct reliably for all but the simplest systems.

6.2 AI Methodology

In this project two distinct AI-based methodological approaches were pursued; the first, building upon the recent work by Li *et al.*[1], a neural network with a purpose-built DFTB-layer is used to enable a deep-learning-based construction of quantum mechanical interaction integrals by learning from DFT data. By constructing a pseudo-tensorial representation of DFTB we can build quantum mechanical intuition directly into the network structure itself (i.e the DFTB-layer), rather than as a guiding feature normally relegated to the realms of training data. This DFTB-layer will take as its inputs Hamiltonian matrix elements and have as output various electronic properties (e.g. band structure, free energy, etc.). Through the use of DFT level data and the application of backpropagation, we can train and improve upon standard DFTB parametrisation sets (such as the Au-org parameter set for

molecule-gold metal interaction) or generate them from scratch.[5] Furthermore, we can replace the standard spline model representation of the Slater-Koster integral parametrisation set to a neural network based one that better captures the complex atomic environment dependence. This shift will enable the creation of reactive parametrisation sets able to break bonds. Conventional DFTB parameter sets cannot accurately describe bond breaking.

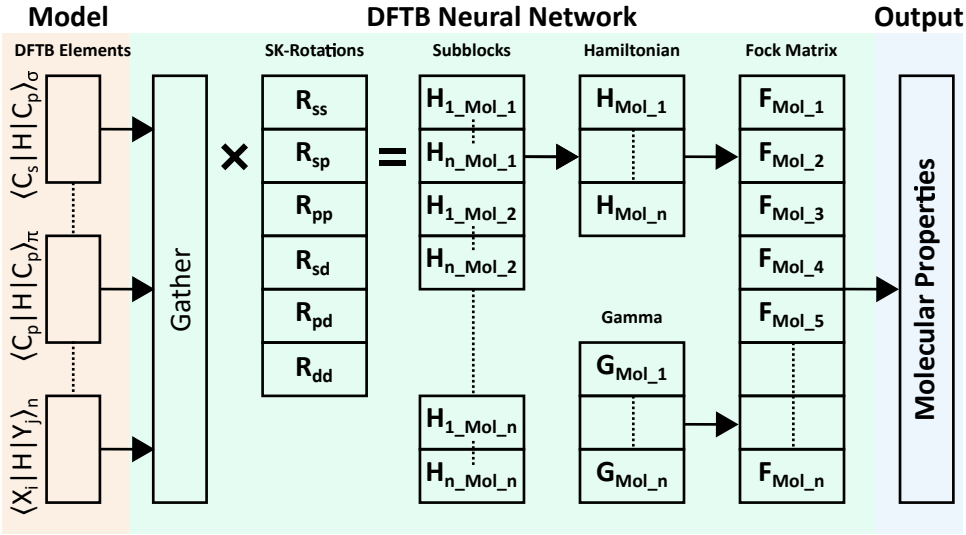


Figure 2. Schematic representation of the DFTB deep learning network's architecture.[1]

In a forward pass, a spline or latent neural network (orange) model, feeds a set of aligned matrix elements to the DFTB neural network (green). These are gathered and Slater-Koster rotated to yield subblocks of the Hamiltonian matrices, which then combine to construct full Hamiltonian matrices. Fock matrices are then constructed and used to predict various molecular properties. Charges, and other SCF related properties, are calculated externally to the network but are updated periodically. During backpropagation, cost functions evaluate the fitness of the predicted molecular properties and use this information to update and improve the fitness of the feed models (orange). This process results in iterative improvements to the feed model's ability to yield accurate aligned matrix elements.

The second approach taken leverages the power of SchNet, which is a deep-learning architecture for molecules and materials.[6] This framework, which is written in PyTorch, is unique in that it utilises a purpose built continuous-convolution filter as part of its "interaction" layer to enable a representation of the atoms to be learnt, rather than relying on a predefined descriptor like other methods.[7] We have recently published a model to predict molecular Hamiltonians in local basis representation called SchNorb (SchNet for Orbitals).[8] This model uses a SchNet input layer and a complex tensor network of repeating single body and pairwise interactions to encode the rotational equivariance properties of molecular wavefunctions. At the moment, ML models can only be generated for interaction integrals and Hamiltonian of a single molecule in different configuraitons. Creating such a SchNorb model for a large number of molecules would also provide a deep-learning quantum surrogate that can be incorporated into DFTB+.

7 Results

7.1 Dataset Construction

A substantial amount of time and thought has also been given to the form, composition and generation of the training set. In order for the neural-network to produce a transferable and well generalised parameter set, it must be trained on a dataset that not only spans the chemical space of interest, but samples it to a sufficient density. Using pair-wise interatomic distances as measure of chemical space, we can define a "good" dataset as one which features interatomic distance histograms that *i*) span a sufficient distance range, *ii*) are continuous over said range, *iii*) contain a sufficient number of samples, and *iv*) are well distributed.

Table 1. Au_n-molecule composite systems. Superscript c=clustered Au atoms

Au atom count	Structures per molecule	Structures per system	Cumulative
1	384	37248	37248
2	384	37248	74496
4^c	384	37248	111744
6^c	384	37248	148992
8^c	384	37248	186240
10^c	384	37248	223488

From the ANI-1 dataset[9] we have selected 97 molecules in non-equilibrium structures, with up to 4 heavy atoms, as the basis for our training data set. The six primary Au_n-molecule systems were then constructed, these have been tabulated in table 1. For each system, 384 geometries were generated for each molecule, yielding a total of 37248 geometries for each system. 74496 Au_n-molecule systems, where n = 1 & 2, were generated by a purpose built concurrent spherical intersection method, as shown in figure **??**. (CSIM). CSIM works by generating a series of concentric spheres, with radii $\vec{r} = [r_1, r_2, \ldots, r_n]$, about each atom's centre. The function used to generate $\vec{r}$ will also define the bond-distance-histogram's distribution, i.e. a linear function will yield a linear bond distance distribution, a $\log_{10}$ function a $\log_{10}$, and so on. The intersection points between two or more spheres are identified, and the smallest number of these points required to "check-off" each distance in $\vec{r}$ for each Au-x pair selected. This process is repeated, adding noise to the positions each time, until the target number of conformations is reached. This method reduces the number of conformations needed to sample a given chemical space by 66 % compared to random sampling, and offers a means to obtain well distributed data. These systems, shown in figure 4, serve the purpose of ensuring coverage of all Au-X distances (up to 6 Å), where X=C, N, H, O, Au. An additional 670464 systems, where n = 4, 6, 8 & 10, were generated *via* a more traditional molecular dynamics method. These where intended to help the model better learn the behaviour associated with clusters and cluster-molecule interactions. The training data set, which was used in the ML-DFTB approach contains the total energies, dipole moment, charges, fermi energies, and projected densities of states (PDoS)* for all of the generated structures.

*Technically only the variables needed to construct the PDoS distributions were collected were calculated. The actual distributions were calculated downstream to ensure method consistency.

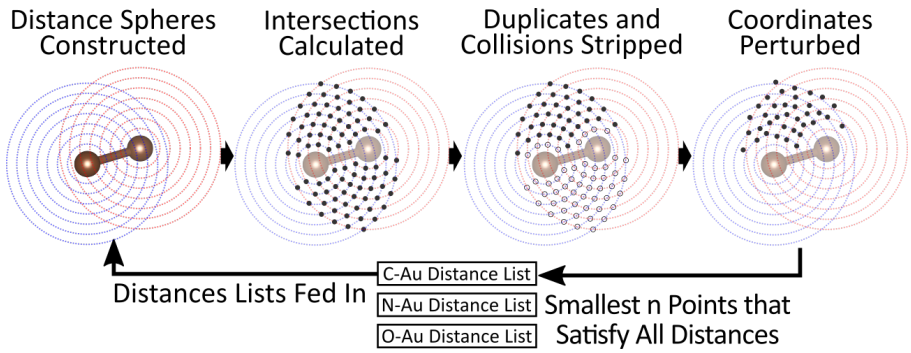


Figure 3. Diagrammatic representation of CISM selection method.

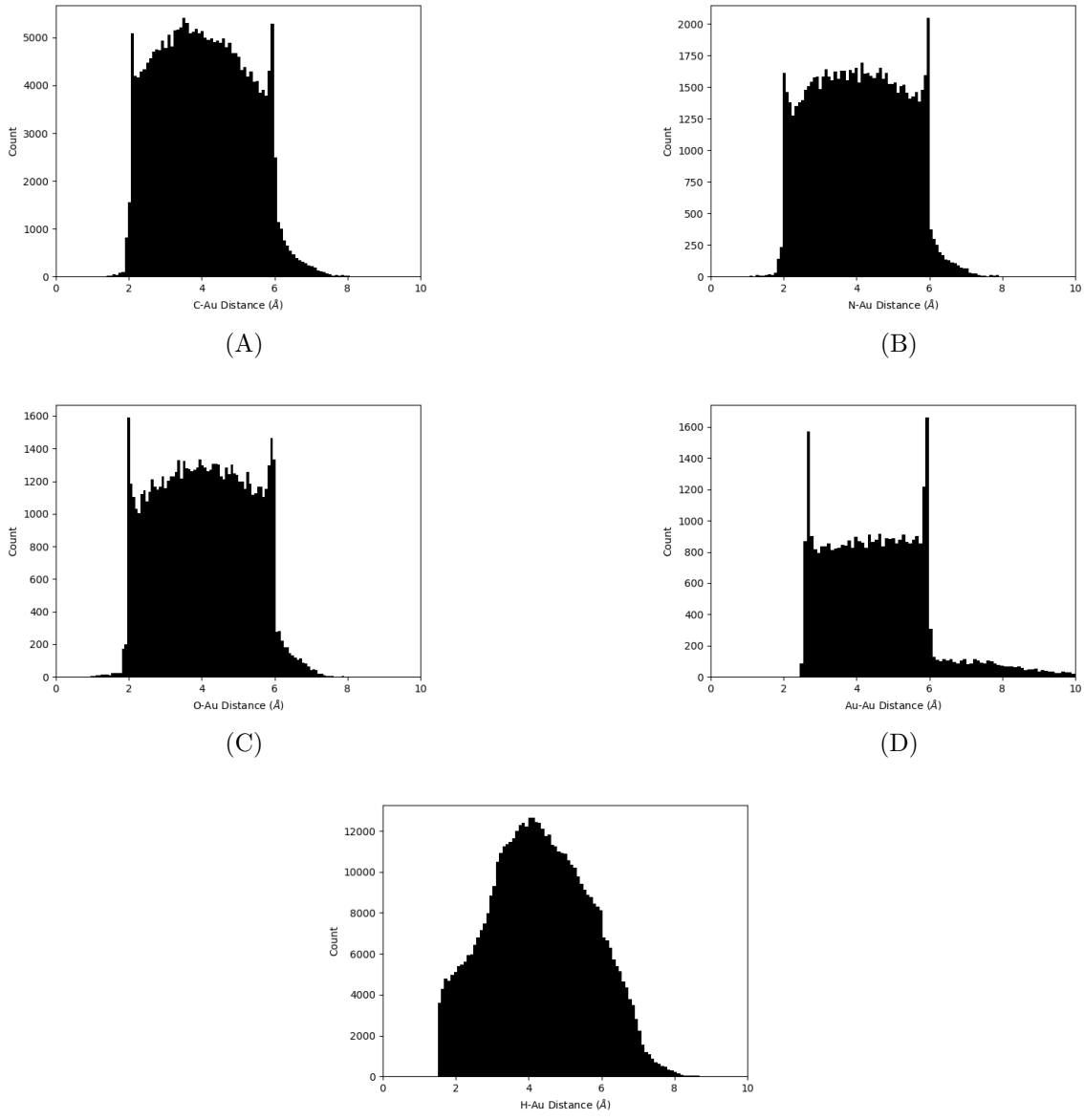


Figure 4. Au-X distance histograms for Au_1 & Au_2-molecule systems where, X = a) C, b) N, c) O, d) Au & e) H. Generated by CSIM targeting C, N, O & Au over a 6 Å range.

The strategy of ML-DFTB is to train a representation of the Hamiltonian and the interaction integrals by optimising the ability to predict certain target properties such as total energies of molecules. For the SchNorb network a second, molecule-centric, training dataset was constructed as the philosophy behind this approach is different. Rather than learning from final target properties, SchNorb directly trains on raw QM data such as Hamiltonian and overlap matrices. For this purpose, single point DFT-level calculations were carried out, using ORCA, on approximately 134k molecules taken from the QM9 dataset.[10,11] The QM9 dataset was employed as it is known to sample the C-H-N-O chemical space to a reasonable degree.

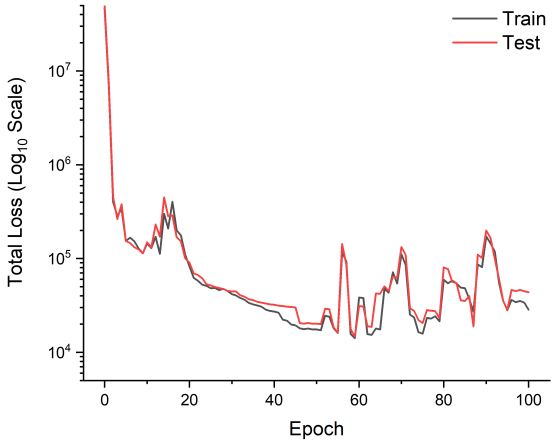
7.2 Network Structure

Thus far, we have made upgrades to the density functional tight binding neural network's architecture, which is implemented in TensorFlow. The improvements include the introduction of ***i)*** d-orbital modelling, ***ii)*** fermi & gaussian smearing, ***iii)*** finite temperature modelling and ***iv)*** augmentation of the cost functions (e.g. inclusion of Mermin free energy). Tests have also been carried out in order to ensure model stability and validity. A band-pass projected density of states (PDoS) cost function was also devised to ensure the model learns, and trains to, the correct electronic structure. The "band-pass" component refers the function's ability to evaluate only the n_{lumo} and n_{homo} states above and below the fermi level. This, in conjunction with the ability to mask out all but a selection of user defined basis functions, enables a meaningful comparison between minimal and non-minimal basis set derived PDoSs to be made (i.e. DFTB vs. DFT). The final evaluation of the predicted and reference PDoSs is made using the Wasserstein distance metric. These changes have made it possible for us to apply the neural network to systems containing metal atoms, such as gold and to ensure that the electronic energy levels that are predicted by DFTB are composed of the correct angular momentum contributions. The PDOS therefore serves as a cost function and a regularizer. Making use of in-house codes we can easily format datasets for training and extract the results in a format readable by DFTB+.

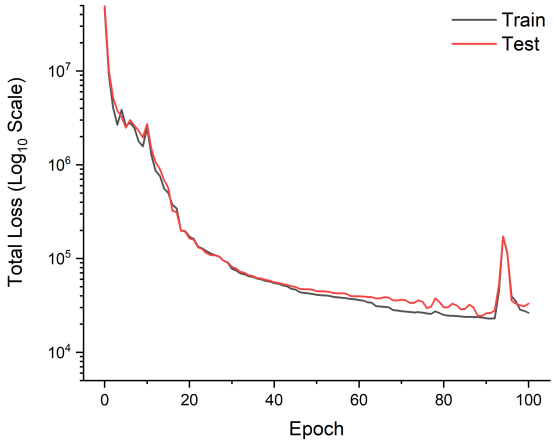
The SchNorb model (as implemented in SchNetPack) has been modified to enable the simultaneous prediction of molecules with different size and different number of basis functions. This requires the inclusion of masking and unmasking operation within the model to ensure that consistent array sizes are passed into the output layer.

7.3 Training

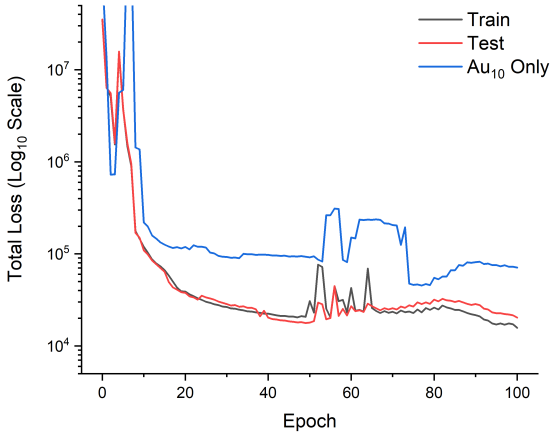
We have conducted and analyzed a number of training runs with the ML-DFTB model, which has substantial memory and GPU-load requirements. The corresponding learning curves can be found in figures 5-8. These training runs were carried out on a small subset of the total training database, they offer an initial insight into the model capabilities. Each run used the same sub-set of data comprised of 100 molecule-cluster systems with 50 training and 10 test examples per system. In these runs, only the total energy, dipole vector and charge deltas were included in the loss function due to time constraints. External charge updates were performed at every second step to minimise the disruptive effects associated with infrequent atomic charge updates. Initially, all two-center interaction splines were fitted irrespective of their parent elements, system-a (see figures 5-8a). However, in order to place a greater emphasis on improving the metal-organic interactions, which is the primary focus of this work, changes were implemented to restrict the training process to only fit Au-X splines, system-b (figures 5-8b).



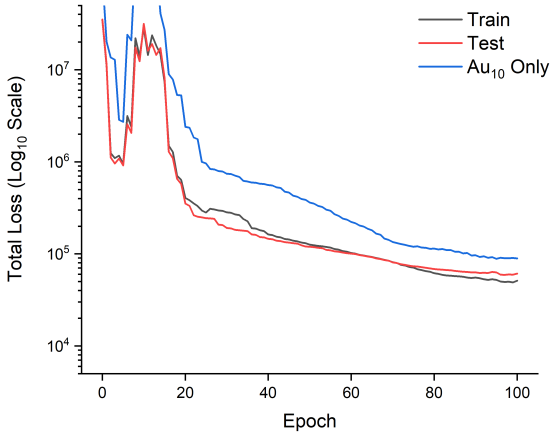
(A) All splines fitted.



(B) Only Au-X splines fitted.



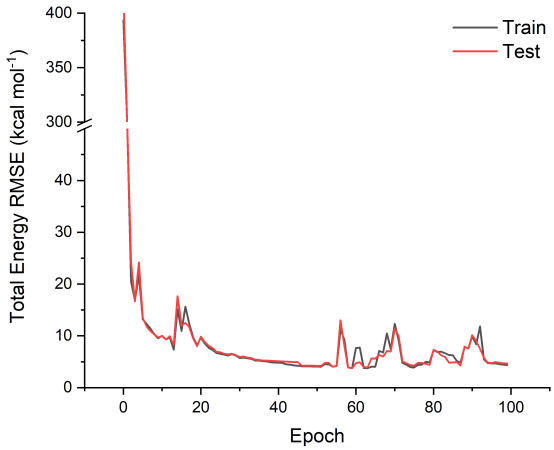
(C) Repeat of 5(a) with Au_{10} containing systems limited to the test set.



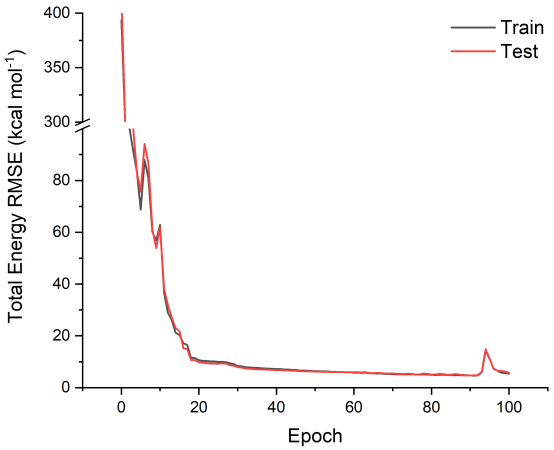
(D) Repeat of 5(b) with Au_{10} containing systems limited to the test set.

Figure 5. Learning curves showing "Total Loss" as a function of training epoch. Losses for just the Au_{10} containing systems are plotted in blue where appropriate.

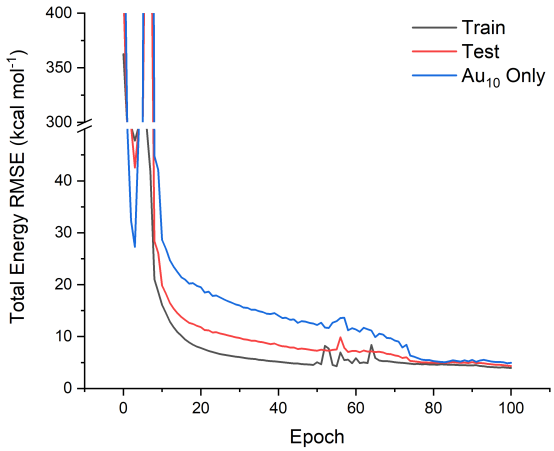
While initial results suggest a good fit, e.g. ΔE_{Total} rmse $= \sim 4-5$ kcal mol^{-1}, it can immediately be seen that there is a non-trivial degree of correlation between the test and train error; to the point that they are, at times, nearly indistinguishable from one another. Although such correlation can be indicative of a well generalised model, the lack of test-train divergence found upon continued training tends to contradict this. A more probable cause is a lack of distinction between the test and training data; i.e lack of data, lack of data diversity, or an insufficient form of data representation. This is an important feature of models that contain substantial prior physical information such as ML-DFTB: The structural configurations of the same molecules contained in the test and training datasets are not sufficiently different to define a well-balanced cost that leads to an optimisation of the interaction integral splines.



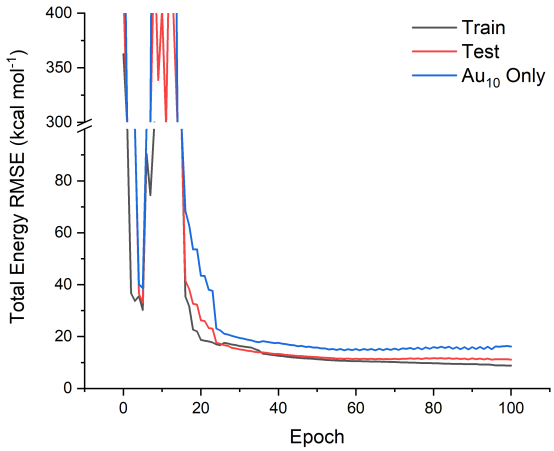
(A) All splines fitted.



(B) Only Au-X splines fitted.



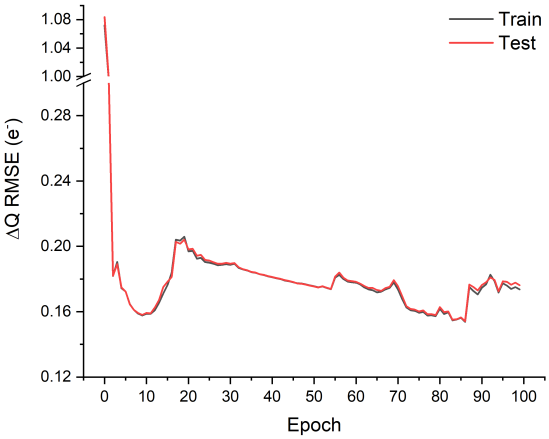
(C) Repeat of 6(a) with Au_{10} containing systems limited to the test set.



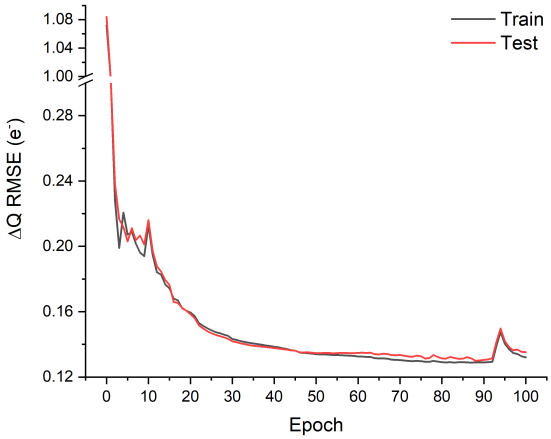
(D) Repeat of 6(b) with Au_{10} containing systems limited to the test set.

Figure 6. Learning curves for the rmse ΔE_{Total}. Losses for just the Au_{10} containing systems are plotted in blue.

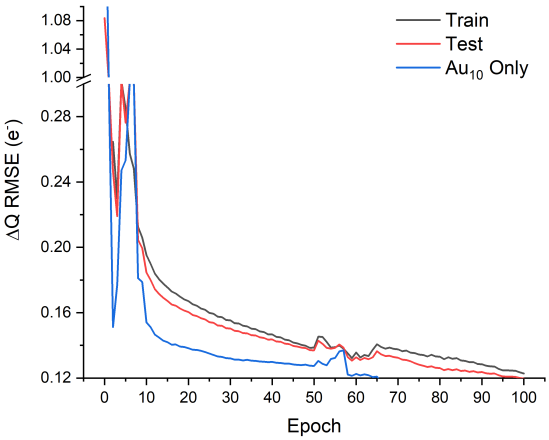
In an effort to combat this within the confines of the network's architecture, molecules containing Au_{10} clusters were confined to the test set only, systems-*c* & *d*. This specific selection was made, rather than an arbitrary distribution of molecule-cluster combinations, as it was i) feasible to implement within the code's current architecture, ii) as the network is only training Au-X interactions, and iii) as it allows training on smaller systems and testing on larger. The differences in molecular composition in test and training datasets which go beyond simple changes in atomic positions and are expected to improve the training. Indeed, upon partitioning Au_{10} containing systems to the test set, a decrease in train-test correlation is observed (figures 5-8c & d). Furthermore, this change introduces an instability to the test error, and, in some instances the train error. The training error instability of system-*c* is likely the same instability observed in system-*a*, only exacerbated by the reduction in the number of training points. This instability is most likely caused by the large number of free parameters compared to the size of the dataset and could be resolved by increasing the training set size to accommodate a larger portion of the database and by introducing a greater degree of regularisation by including the implemented PDOS cost functions into the training (work in progress).



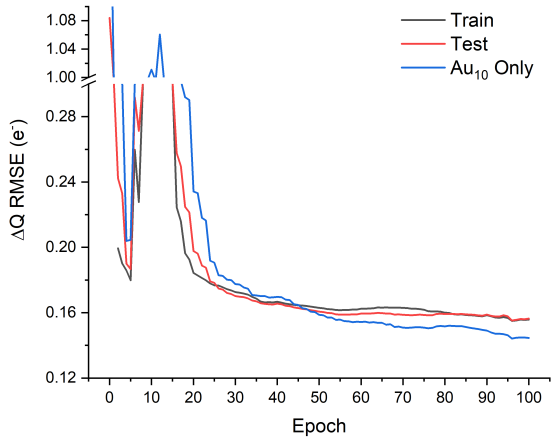
(A) All splines fitted.



(B) Only Au-X splines fitted.



(C) Repeat of 7(a) with Au_{10} containing systems limited to the test set.

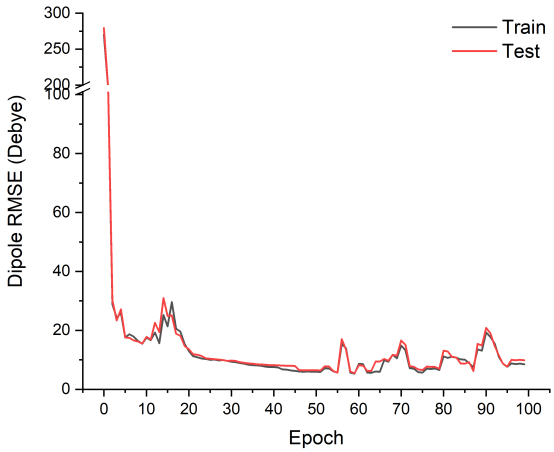


(D) Repeat of 7(b) with Au_{10} containing systems limited to the test set.

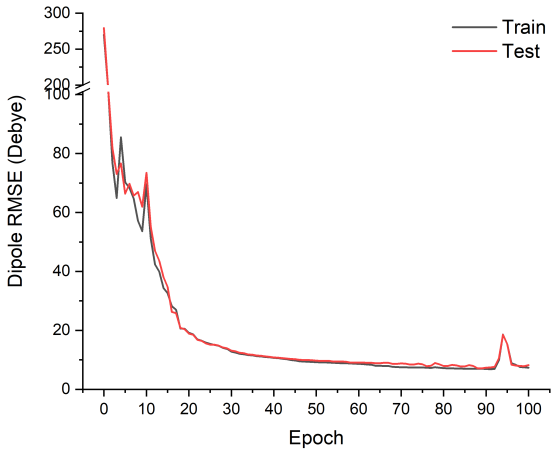
Figure 7. Learning curves for the rmse Δ-charge. Losses for just the Au_{10} containing systems are plotted in blue.

To resolve the instabilities in the test error, we will perform sensitivity correlation analysis on the network's properties and an optimization of the networks hyperparameters incl. the weights of the individual target costs. From this, changes will then be implemented to better generalise the model. Furthermore, an ensemble of training runs must be carried out on much larger portions of the data-set. The model hyperparameters incl. the weights of target costs need to be further optimized (currently underway). Successfully trained models can be used to extract splines and transfer them into the DFTB+ file format. The splines can then be used in a series of DFTB benchmark tests. From this it could be ascertained if the highly correlative test-train behaviour results from a well generalised model or originates from a more pathological origin such as the structure of the data / model. The necessary work to achieve this will be performed during the first half of 2020.

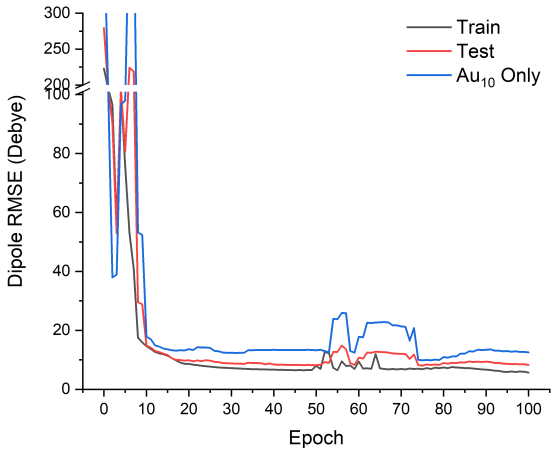
Simultaneously, we will train the transferable SchNorb model with the newly created QM9 dataset and compare its performance against ML-DFTB



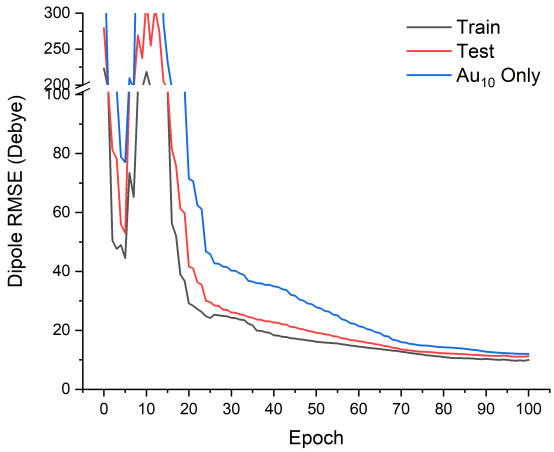
(A) All splines fitted.



(B) Only Au-X splines fitted.



(C) Repeat of 8(a) with Au_{10} containing systems limited to the test set.



(D) Repeat of 8(b) with Au_{10} containing systems limited to the test set.

Figure 8. Learning curves for the rmse Δ-dipole. Losses for just the Au_{10} containing systems are plotted in blue.

8 Outputs

Presentations

- Poster presentation: “Machine-Learning-based Parametrization of Density Functional Tight Binding for Metal-Organic Systems" at the CECAM-funded workshop “Thinking outside of the box: Beyond machine learning for quantum chemistry" in Bremen (07.10.2019-11.10.2019) by Dr. Adam McSloy
- Poster presentation: “Machine-Learning-based Parametrization of Density FUnctional Tight Binding for Metal-Organic Systems" at the annual AI3SD Network+ conference in Winchester (18.11.2019-19.11.2019) by Dr. Adam McSloy
- Invited talk: “Deep Learning Enhanced Quantum Chemistry: Pushing the Limits of Materials Discovery" at the annual AI3SD Network+ conference in Winchester by Dr. Reinhard Maurer

- (Upcoming) Contributed talk “Machine Learning Augmented Density Functional Tight Binding Theory" at the DPG Germany Physical Society spring meeting (15.-20.03.2020, Dresden, Germany) by Dr. Adam McSloy
- (Upcoming) Invited talk “Deep Learning Enhanced Quantum Chemistry: Pushing the Limits of Materials Discovery" at Workshop on Machine Learning Potentials and Long-Range Interactions in Lenggries, Germany (29.06.-02.07.2020) by Dr. Reinhard Maurer
- (Upcoming) Invited talk “Deep Learning Enhanced Quantum Chemistry: Pushing the Limits of Materials Discovery" at the ACS Fall meeting in San Francisco (16.-20.08.2020) by Dr. Reinhard Maurer

Publications

- (Upcoming) M. Gastegger, A. McSloy, M. Luya, K.-R. Mueller, R. J. Maurer, A transferable deep learning model of molecular wavefunctions, Invited article in Journal of Chemical Physics, in preparation
- (Upcoming) A. McSloy, D. Yaron, B. Hourahine, R. J. Maurer, Machine Learning-based DFTB+ Parametrization: An improved model for metal-molecule interaction, in preparation

9 Conclusions

We have improved the existing ML-DFTB network model to enable the study of metal-molecule interactions and more complex materials. We have created extensive training data for the construction of a new H-C-N-O-Au containing DFTB parameter set; The development of the ML-DFTB network has progressed at a much slower rate than intended, but the remaining optimization and training tasks can be concluded within the next few months.
The work conducted thus far has been invaluable to the progression of the project as it has enabled an in-depth exploration of each of the building blocks required to develop a viable approach. This includes the internal neural network, the DFTB output layer and the required cost function routines. The work of this project has recently been discussed with the DFTB+ main developers, which includes Dr. Ben Hourahine. All of the knowledge, tools, data and code developed throughout the course of this project will be carried over to form the foundation of a machine learning architecture, which will be incporporated into the DFTB+ software package. Furthermore, this project has enabled a diverse long-term international network of collaborators to be formed which will be a lasting outcome of this project.

10 Future Plans

Research into developing an ML-DFTB architecture will continue beyond the initial scope of this project. During this project, it became evident that the static and predefined nature of the existing ML-DFTB TensorFlow code hindered development and optimization tasks that require dynamic code adjustments. Firstly, a modular plug-and-play style codebase with a dynamically allocatable network graph is necessary to allow i) new features to be rapidly implemented, ii) other codebases to be used to incorporate inputs, outputs, losses, etc. and iii) changes to the graph during operation to be made without hindrance. Secondly, an alternative choice of representation (SchNet) is necessary to better express the differences in the input data’s feature space, which we expect to be a universal problem for ML representations with a lot of prior

built-in physics. In our future work, we will redevelop this model in a more flexible framework. We intend to build a basic, modular machine learning framework into which much of the code developed during this project can be placed. This code will subsequently be interfaced with SchNetPack, a powerful machine learning framework based on PyTorch.[6]

References

[1] H. Li, C. Collins, M. Tanha, G. J. Gordon and D. J. Yaron, *Journal of Chemical Theory and Computation*, 2018, **14**, 5764–5776.

[2] R. J. Maurer, V. G. Ruiz, J. Camarillo-Cisneros, W. Liu, N. Ferri, K. Reuter and A. Tkatchenko, *Progress in Surface Science*, 2016, **91**, 72 – 100.

[3] B. Aradi et al., *Journal of Chemical Physics*, 2020, under Review.

[4] P. Koskinen and V. Mäkinen, *Computational Materials Science*, 2009, **47**, 237–253.

[5] V. Mäkinen, P. Koskinen and H. Häkkinen, *European Physical Journal D*, 2013.

[6] K. T. Schütt, P. Kessel, M. Gastegger, K. A. Nicoli, A. Tkatchenko and K.-R. Müller, *Journal of Chemical Theory and Computation*, 2019, **15**, 448–455.

[7] K. T. Schütt, P.-J. Kindermans, H. E. Sauceda, S. Chmiela, A. Tkatchenko and K.-R. Müller, 2017.

[8] K. T. Schütt, M. Gastegger, A. Tkatchenko, K.-R. Müller and R. J. Maurer, *Nature Communications*, 2019, **10**, 5024.

[9] J. S. Smith, O. Isayev and A. E. Roitberg, *Scientific Data*, 2017, **4**, 170193.

[10] R. Ramakrishnan, P. O. Dral, M. Rupp and O. A. von Lilienfeld, *Scientific Data*, 2014, **1**, 140022.

[11] F. Neese, *WIREs Computational Molecular Science*, 2012, **2**, 73–78.

11 Data & Software Links

- All data will be uploaded to the NOMAD repository (http://nomad-repository.eu/). Datasets will receive a DOI, which will be referenced in the forthcoming publications.
- DFTB+: https://www.dftbplus.org/.