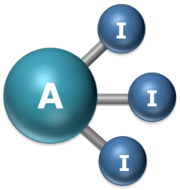

Artificial Intelligence and Augmented Intelligence for Automated Investigations for Scientific Discovery

Deep-Learning-Enhanced Quantum Chemistry: Pushing the Limits of Materials Discovery
Interim 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: 30/09/2019

Deep-Learning-Enhanced Quantum Chemistry: Pushing the Limits of Materials Discovery
AI3SD-Project-Series:Report-2_Maurer_Interim
Report Date: 30/09/2019
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	AI3SD-FundingCall1_027
Lead Institution	University of Warwick
Project Dates	01/07/2019 - 31/12/2019
Website	warwick.ac.uk/maurergroup
Keywords	DFTB, Hybrid Organic-Metallics, HOM

2 Project Team

2.1 Principal Investigator

Name and Title: Dr Reinhard Maurer
Association: University of Warwick, Chemistry Department
Work Email: r.maurer@warwick.ac.uk
Work Phone: 024 765 23228

2.2 Co-Investigators

Name and Title: Dr Benjamin Hourahine
Association: University of Strathclyde, Department of Physics
Work Email: benjamin.hourahine@strath.ac.uk
Work Phone: 0141 5482325

Name and Title: Prof. David Yaron
Association: Carnegie Mellon University, Department of Chemistry
Work Email: yaron@cmu.edu
Work Phone: +1-412-268-1351

2.3 Other 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 aim to 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.

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 provide proof-of-principle for the simulation of metal-organic hybrid materials and integrate the ML model 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. By building a highly efficient parametrised variant of DFT, namely DFTB, with the use of machine-learned representations of interaction integrals, we will retain the accuracy and prediction capability of DFT calculations at the computational efficiency of a numerically less demanding method. This goal can be broken down into the three tasks:

1. Build upon the DFTB-NN deep learning layer published by Li *et al.*[4] 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.

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.[3]

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

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.[7] 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.

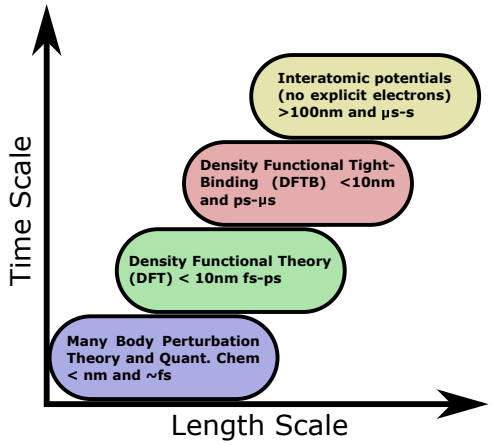

Figure 1: Typical time (fs/ps/µs) and length scale (nm) regimes of atomistic simulation methods.

Density-Functional Tight Binding (DFTB) is a semi-empirical method that bridges the divide between electronic structure methods and force fields (see figure 1).[6] This method provides 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

Building upon the recent work by Li *et al.*[4], a neural network with a purpose-built DFTB-layer will be 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.[2] Furthermore, we will move away from the standard spline model representation of the 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 characterisations cannot describe bond breaking.

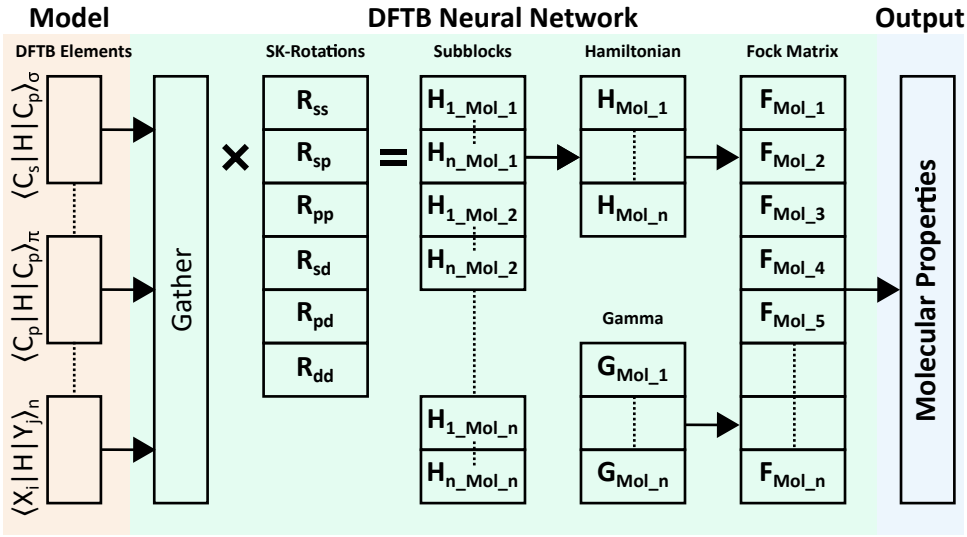

Figure 2: Schematic representation of the DFTB deep learning network's architecture.[4]

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.

7 Interim Results

7.1 Network Structure

Thus far, we have made notable upgrades to the density functional tight binding neural network's architecture implemented in TENSORFLOW. Such improvements include, but are not limited to, 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. These changes have made it possible for us to apply the neural network to systems containing metal atoms, such as gold. The only outstanding task relating to modification of the neural network is the introduction of a projected density of states (pDOS) cost function. Making use of in-house codes we can easily format datasets for training and extract the results in a format readable by DFTB+.

7.2 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 who's interatomic distance histograms *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[1] we have selected 97 molecules in non-equilibrium structures, with up to 4 heavy atoms, as 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 (CSIM). These systems, shown in figure 3, serve the purpose of ensuring coverage of all Au-X distances (up to 6 Å), where X=C,N,H,O. An additional 670464 systems, where n = 4, 6, 8 & 10, are being generated *via* a more traditional molecular dynamics method. These where intended to help the model better learn the behaviours associated with clusters and cluster-molecule interactions.* In addition to these systems, further isolated gold cluster

*Molecular dynamics based simulations are still being generated

and molecule systems have been selected for inclusion.

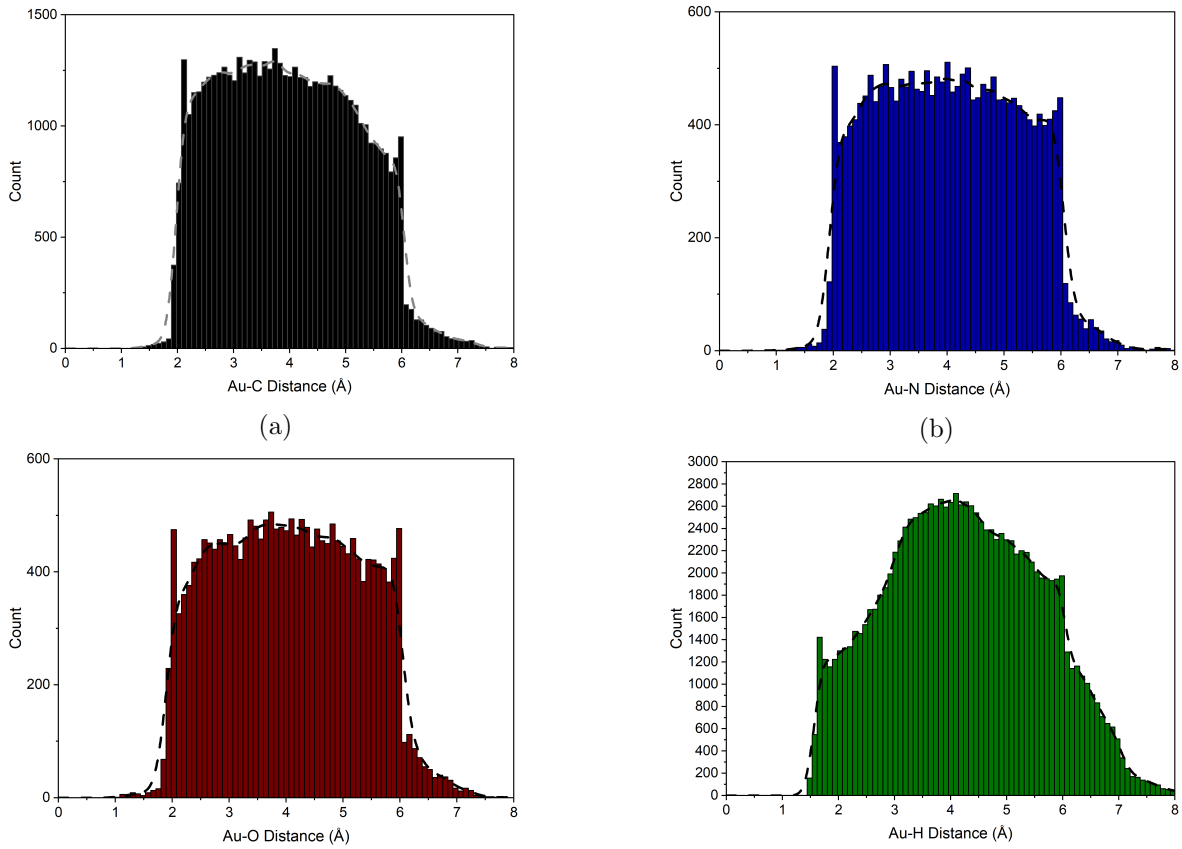

Figure 3: Au-X distance histograms for Au_1-molecule systems where, X = a) C, b) N, c) O and d) H. Generated by CSIM targeting C, N & O over a 6 Å range.

8 Outputs

The results of this project are scheduled to be presented in poster and presentation form at both the cecam-outbox (07.10.19 - 11.10.19) and the AI3SD conference (18.11.19 - 19.11.19), with more to follow as the project progresses. Furthermore, the dataset currently under construction will be made available upon completion.

9 Progress Summary

Since the start of this project we have successfully introduced the majority of the required changes to the neural network, as laid out in task 1 of section 5; with only one minor task still outstanding related to the integration of the projected density of states cost function. The dataset production task outlined in task 2, of section 5, is currently underway and is approaching its conclusion, with the training of the model with the resulting dataset to follow.

10 Next Steps

The next steps in the project are fivefold: ***i)*** the data collection stage will be wrapped up, and the results packed up into a readily accessible database. ***ii)*** a projected density of states

comparison method will be added to the cost function to improve the model's fitting capabilities. ***iii)***, the network will be trained on the final production level data to produce a set of new density functional tight binding theory parametrization sets. ***iv)*** the performance of these models will be evaluated, and the results published. ***v)*** Together with our coinvestigators and project partners, we will further pursue the permanent integration of the ML infrastructure into the DFTB+ codebase.

It is expected that this project will produce at least two publications in peer reviewed journals. The first publication will provide a detailed account of methods employed and will establish a protocol that can be followed which would allow others to more readily employ such methods in the future. The second publication will report on the newly derived hybrid organic-metallic DFTB parameter set, its validation and the type of systems to which it can be applied. Additional papers investigating various hybrid organic-metallic systems using the newly derived parameter set are also expected to follow. Finally, this work will be presented at a selection of topical conferences, e.g. CECAM OUTBOX and AI3SD conference.

11 References

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

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

[3] 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.

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

[5] V. Blum, R. Gehrke, F. Hanke, P. Havu, V. Havu, X. Ren, K. Reuter and M. Scheffler, *Computer Physics Communications*, 2009, **180**, 2175–2196.

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

[7] R. A. Hatcher, *The United States Pharmacopeia*, 27th edn, 2009.

[8] G. R. Schleder, A. C. M. Padilha, C. M. Acosta, M. Costa and A. Fazzio, *Journal of Physics: Materials*, 2019, **2**, 32001.

[9] G. B. Goh, N. O. Hodas and A. Vishnu, *Journal of Computational Chemistry*, 2017, **38**, 1291–1307.

12 Data & Software Links

N/A