The 2nd International Symposium on Machine Learning in Quantum Chemistry

Invited Talks
(abstracts)

In alphabetic order of presenting author

Integrating Explainability into Graph Neural Networks for the Prediction of X-ray Absorption Spectra

A. Kotobi[1], K. Singh[2,3], D. Höche[1], S. Bari[4,5], R. H. Meißner[1,6], A. Bande[2,7]

[1]Helmholtz-Zentrum Hereon, Institute of Surface Science, Geesthacht, DE 21502, Germany; [2]Helmholtz-Zentrum Berlin für Materialien und Energie GmbH, Berlin, DE 10409, Germany;
[3]Institute of Chemistry and Biochemistry, Freie Universität Berlin, Berlin, DE 14195, Germany; [4]Deutsches Elektronen-Synchrotron (DESY), Hamburg, DE 22607, Germany; [5]Zernike Institute for Advanced Materials, University of Groningen, Groningen 9712, Netherlands; [6]Hamburg University of Technology, Institute of Polymers and Composites, Hamburg, DE 21073, Germany; [7]Leibniz University Hannover, Institute of Inorganic Chemistry, Hannover, DE 30167, Germany

The use of sophisticated machine learning (ML) models, such as graph neural networks (GNNs), to predict complex molecular properties or all kinds of spectra has grown rapidly [1,2]. However, ensuring the interpretability of these models' predictions remains a challenge. For example, a rigorous understanding of the predicted X-ray absorption spectrum (XAS) generated by such ML models requires an in-depth investigation of the respective black-box ML model used. Here, this is done for different GNNs based on a comprehensive, custom-generated XAS data set for small organic molecules. We show that a thorough analysis of the different ML models with respect to the local and global environments considered in each ML model is essential for the selection of an appropriate ML model that allows a robust XAS prediction. Moreover, we employ feature attribution to determine the respective contributions of various atoms in the molecules to the peaks observed in the XAS spectrum. By comparing this peak assignment to the core and virtual orbitals from the quantum chemical calculations underlying our data set, we demonstrate that it is possible to relate the atomic contributions via these orbitals to the XAS spectrum [3].

[1] Singh, K., Münchmeyer, J., Weber, L., Leser, U., Bande, A., *J. Chem. Theory Comput.* **18** (2022) 4408.
[2] Kotobi, A., Schwob, L., Vonbun-Feldbauer, G. B., Rossi, M., Gasparotto, P., Feiler, C., Berden, G., Oomens, J., Oostenrijk, B., Scuderi, D., Bari, S., Meisner, R. H., *Commun. Chem*. **6** (2023), 6, 46.
[3] Kotobi, A., Singh, K., Höche, D., Bari, S., Meißner, R., Bande, A., *J. Am. Chem. Soc.* **145** (2023) 22584.

AniSOAP: An Expansion of Density-Correlation Machine Learning Representations for Anisotropic Coarse-Grained Particles

Rose K. Cersonsky[1], Arthur Lin[1], Kevin K. Hugeunin-Dumittan[2], Yong-Cheol Cho[1,3], Jigyasa Nigam[2]

1)Department of Chemical and Biological Engineering, University of Wisconsin, Madison, WI, USA
2)Laboratory of Computational Science and Modeling, 'Ecole Polytechnique F'ed'erale de Lausanne, Lausanne, Switzerland
3)Department of Computer Science and Engineering, University of Wisconsin, Madison, WI, USA

In understanding molecular interactions and modeling their resultant behavior, it is very often a worthwhile endeavor to group one or more atoms as a theoretical ``unit" or particle; coarse-grained and top-down simulation approaches are an important tool and a lens with which to study chemical systems. Similarly, machine learning (ML) methods have emerged as a powerful tool for scientific inquiry, with the ability to elucidate new patterns within or relationships between chemical spaces and observed properties, often in order to predict the properties of unseen systems. How do we incorporate the idea of atom grouping in the context of machine learning? Density-based frameworks present a compelling avenue for expansion, given that they can putatively be made flexible to any density expansion, even anisotropic density fields or hard particle volumes. Furthermore, by extending these frameworks to decouple molecule-level interactions from atom-atom site potentials, we gain the ability to combine or compare representations across multiple scales. We propose and demonstrate the first such anisotropic expansion of symmetrized density-based frameworks for ML representations by taking the popular SOAP (Smooth Overlap of Atomic Positions) formalism and demonstrating its expansion to simple anisotropic bodies'. While we here demonstrate the expansion for multivariate Gaussian densities, similar expansions can be made for arbitrary anisotropic density fields. From this, we can derive fundamental insights on how molecular shape influences mesoscale behavior, as well as understand and incorporate where individual atom-atom interactions remain important, as demonstrated via analysis of benzene interactions determined by first principles. Moving forward, we present AniSOAP as a powerful and flexible coarse-graining framework to systematically reduce molecular degrees of freedom in complex, multiscale simulation.

Machine Learning Electronic Excitations in Complex and Multichromophoric Systems

Lorenzo Cupellini, Edoardo Cignoni, Amanda Arcidiacono, Benedetta Mennucci

Department of Chemistry and Industrial Chemistry, Via Moruzzi 13, 56124 Pisa, Italy

Most photophysical processes of chemical or biological interest occur in condensed phase, within "complex" systems where chromophore-environment interactions and structural disorder play a primary role. Faithful quantum chemical modeling of these processes requires a multiscale strategy that accounts for all interactions,[1] as well as a thorough sampling of the configurational ensemble of the system. Complete modeling of these processes is therefore limited by the computational cost of numerous accurate multiscale calculations. Machine learning (ML) techniques hold great promise in alleviating the computational burden of quantum mechanical calculations, and have been successfully employed for the prediction of excited-state properties and dynamics.[2,3] However, only few ML models have been designed to account for the environment effects on excited states.[4]

Here I will outline potential ML strategies to model excited states in complex systems, showing that it is possible to incorporate the environment effects with a physically constrained Gaussian process regression model. I will focus on light-harvesting pigment-protin complexes as an example of complex biological system where excitations are delocalized over multiple chromophores and also tuned by the interactions with the protein.[5] The obtained ML models[6] not only reproduce small differences between identical pigments in different protein pockets, but can also extrapolate to completely unseen environments. Finally, these models can also help interpreting the solvatochromic effect of the protein.
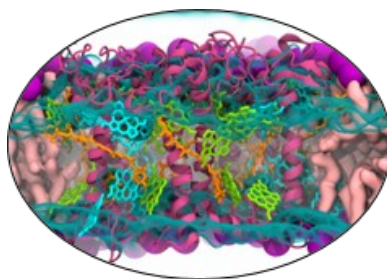

Figure 1: Structure of a light-harvesting complex with the embedded pigments, carotenoids (orange) and chlorophylls (blue and green)

1 M. Nottoli, *et al.,Annu. Rev. Phys. Chem.* **2021**, *72*, 489. doi:10.1146/annurev-physchem-090419-104031

2 J. Westermayr, P. Marquetand. *Chem. Rev.* **2021**, *121*, 9873. doi:10.1021/acs.chemrev.0c00749

3 P. Dral, M. Barbatti. *Nat. Rev. Chem.* **2021**, *5*, 388. doi:10.1038/s41570-021-00278-1

4 M. S. Chen *et al.*, *J. Phys. Chem. Lett.* **2020**, *11*, 7559. doi:10.1021/acs.jpclett.0c02168

5 L. Cupellini *et al.*, *BBA − Bioenerg.* **2020**, *1861*, 148049. doi:10.1016/j.bbabio.2019.07.004

6 E. Cignoni *et al.*, *J. Chem. Theory Comput.* **2023**, 19, 965. doi:10.1021/acs.jctc.2c01044

Towards practical AI-enhanced computational chemistry

Pavlo O. Dral

State Key Laboratory of Physical Chemistry of Solid Surfaces, College of Chemistry and Chemical Engineering, Fujian Provincial Key Laboratory of Theoretical and Computational Chemistry, and Innovation Laboratory for Sciences and Technologies of Energy Materials of Fujian Province (IKKEM), Xiamen University, Xiamen, Fujian 361005, China
E-mail: dral@xmu.edu.cn. Website: http://dr-dral.com

I will present our methods and software tools enabling practical AI-enhanced computational chemistry simulations and demonstrate their applications. The methods include the general-purpose, artificial intelligence-enhanced quantum mechanical method 1 (AIQM1),[1] which approaches the accuracy of the golden-standard, traditional CCSD(T)/CBS approach for many properties. Other methods focus on novel approaches for learning dynamics, such as our AI-quantum dynamics[2] and 4D-spacetime atomistic AI[3] approaches, which predict dynamics properties such as nuclear coordinates as the function of time and do not require iterative trajectory propagation as in classical MD. AIQM1 and AI-QD, along with many other methods such as a host of ML interatomic potentials, are implemented in our MLatom program package for user-friendly atomistic machine learning simulations, which can be run online using our MLatom@XACS (Xiamen atomistic computing suite) cloud-based service.[4]
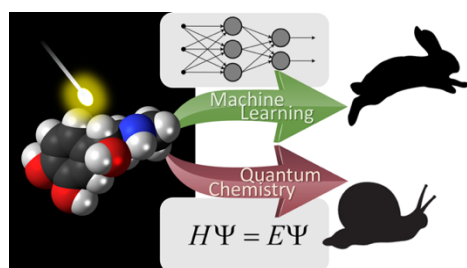

Figure 1: Machine learning greatly speeds up quantum chemistry simulations as discussed in our reviews[5,6] and an edited book[7].

[1] P. Zheng, R. Zubatyuk, W. Wu, O. Isayev, P. O. Dral. Artificial Intelligence-Enhanced Quantum Chemical Method with Broad Applicability. *Nat. Commun.* **2021,** *12*, 7022.

[2] A. Ullah, P. O. Dral. Predicting the future of excitation energy transfer in light-harvesting complex with artificial intelligence-based quantum dynamics. *Nat. Commun.* **2022,** *13*, 1930.

[3] F. Ge, L. Zhang, Y.-F. Hou, Y. Chen, A. Ullah, P. O. Dral. Four-Dimensional-Spacetime Atomistic Artificial Intelligence Models. *J. Phys. Chem. Lett.* **2023,** *14*, 7732–7743.

[4] P. O. Dral, F. Ge, B.-X. Xue, Y.-F. Hou, M. Pinheiro Jr, J. Huang, M. Barbatti. MLatom 2: An Integrative Platform for Atomistic Machine Learning. *Top. Curr. Chem.* **2021,** *379*, 27. See MLatom.com @ XACScloud.com.

[5] P. O. Dral, M. Barbatti. Molecular Excited States Through a Machine Learning Lens. *Nat. Rev. Chem.* **2021,** *5*, 388–405.

[6] P. O. Dral. Quantum Chemistry in the Age of Machine Learning. *J. Phys. Chem. Lett.* **2020,** *11*, 2336–2347.

[7] *Quantum Chemistry in the Age of Machine Learning*. Ed. P. O. Dral. Elsevier: Amsterdam, Netherlands, **2023**.

Restricted variance optimization: geometries, intersections and then some

Ignacio Fdez. Galván, Roland Lindh

Department of Chemistry – BMC, Uppsala University, P. O. Box 576, SE-75123 Uppsala, Sweden

Multidimensional optimization problems are ubiquitous and routine in computational chemistry. A paradigmatic example is molecular structure optimization, where a local minimum in a potential energy surface is to be found. The comparative simplicity of setup and interpretation of such calculations should not, however, hide the fact that it is a complex task: finding specific points in a high-dimensional space in as few iterations as possible, with only local information. Typical (and robust and successful) optimization methods are often based on a second-order Taylor expansion of the object function, that becomes more and more accurate as the optimization reaches convergence.[1]

Recently, a number of methods have been proposed that use machine learning techniques to construct a surrogate model that goes beyond the second-order expansion and are thus able to better utilize all the information generated during the optimization process, to find the target more efficiently. One such method is the restricted variance optimization (RVO),[2] based on a Gaussian process regression surrogate model aided by chemical heuristics.

The RVO method has been applied to stable structure optimizations, as well as transition states, reaction paths[3] and conical intersections,[4] showing an improved efficiency and robustness compared to established conventional second-order optimization methods. An extension to wave function (orbital) optimization is being developed, and results for single-determinant SCF are promising, demonstrating the versatility of the approach.

1 H. B. Schlegel. Geometry Optimization. *Wiley Interdiscip. Rev.: Comput. Mol. Sci.* **2011**, *1*, 790–809. doi:10.1002/wcms.34

2 G. Raggi, I. Fdez. Galván, C. L. Ritterhoff, M. Vacher, R. Lindh. Restricted-Variance Molecular Geometry Optimization Based on Gradient-Enhanced Kriging. *J. Chem. Theory Comput.* **2020**, *16*, 3989–4001. doi:10.1021/acs.jctc.0c00257

3 I. Fdez. Galván, G. Raggi, R. Lindh. Restricted-Variance Constrained, Reaction Path, and Transition State Molecular Optimizations Using Gradient-Enhanced Kriging. *J. Chem. Theory Comput.* **2021**, *17*, 571–582. doi:10.1021/acs.jctc.0c01163

4 I. Fdez. Galván, R. Lindh. Smooth Things Come in Threes: A Diabatic Surrogate Model for Conical Intersection Optimization. *J. Chem. Theory Comput.* **2023**, *19*, 3418–3427. doi:10.1021/acs.jctc.3c00389

Ab-Initio Potential Energy Surfaces via Graph Neural Networks

Nicholas Gao, Stephan Günnemann

Department of Computer Science & Munich Data Science Institute
Technical University of Munich, Germany

Recent advances in quantum chemistry lead to neural network-based electronic wave functions within the Variational Monte Carlo (VMC) framework. While offering unprecedented accuracy on various molecular systems, their training requires large computational resources. This problem is amplified in the context of potential energy surfaces (PES), where one must traditionally repeat the computation for different structures. In our works[1][2][3], we tackle this problem by exploiting the generalization capabilities of graph neural networks to accurately compress different wave functions within a single neural network. We accomplish this by developing a reparameterization scheme where the electronic wave function is conditioned on the molecular structure of the nuclei. Such adaptation allows the joint optimization of different wave functions. In our experimental evaluation, we find that such a joint optimization comes at no additional computational cost or loss of accuracy.

---

1 N. Gao & S. Günnemann, „Ab-Initio Potential Energy Surfaces by Pairing GNNs with Neural Wave Functions", International Conference on Representation Learning 2022

2 N. Gao & S. Günnemann, „Sampling-free Inference for Ab-Initio Potential Energy Surface Networks", International Conference on Representation Learning 2023

3 N. Gao & S. Günnemann, „Generalizing Neural Wave Functions", International Conference on Machine Learning 2023

Machine learning the Hohenberg-Kohn map for electronic excited states

Yuanming Bai,[1,2,3] Leslie Vogt-Maranto,[3] Mark E. Tuckerman,[2,3,4,5]
William J. Glover[1,2,3]

[1]*NYU Shanghai, 567 W. Yangsi Road, 200124, Shanghai, China;* [2]*NYU-ECNU Center for Computational Chemistry at NYU Shanghai, 3663 Zhongshan Road North, 200062, Shanghai, China;* [3]*Department of Chemistry, New York University, New York, NY, 10003, USA;* [4]*Simons Center for Computational Physical Chemistry at New York University, New York, NY, 10003, USA;* [5]*Courant Institute of Mathematical Science, New York University, New York, NY, 10012, USA*

Time-Dependent Density-Functional Theory (TDDFT) is the workhorse for computing electronic excitations in molecules and materials; however, the approximations inherent in practical TDDFT calculations, together with their computational expense, motivate finding a cheaper, more direct map for electronic excitations. The existence of such a map is provided by the Hohenberg-Kohn theorem of density-functional theory, which proves a bijection between the ground-state electron density and the external potential of a many-body system. This guarantees a one-to-one map from the electron density to all observables of interest, including electronic excited-state energies. Here, we show that multistate density and energy functionals can be constructed via machine learning.[1] The framework is used to perform the excited-state molecular dynamics simulations with a machine-learned functional on malonaldehyde. The simulations correctly capture the kinetics of an excited-state intramolecular proton transfer, allowing insight into how mechanical constraints can be used to control the proton transfer reaction in this molecule.

1 Y. Bai, L. Vogt-Maranto, M. E. Tuckerman, W. J. Glover. *Nat. Comm.*.**2022**, *13*, 7044.

Electron-density and long-range machine learning methods for the study of electrochemical interfaces

Andrea Grisafi*, Augustin Bussy**, Mathieu Salanne*, Rodolphe Vuilleumier***

Institute of Computing and Data Sciences, Sorbonne Université, CNRS, 75005 Paris, France
**Department of Chemistry, University of Zurich, Winterthurerstrasse 190, 8057 Zürich, Switzerland*
***PASTEUR, Department of Chemistry, Ecole Normale Supérieure, PSL University, Sorbonne Université, CNRS, 75005 Paris, France*

The computational study of energy storage and conversion processes calls for simulation techniques that can reproduce the interaction of a metallic surface with an electrolyte solution under an applied electric bias. I will first introduce an equivariant machine-learning model of the Kohn-Sham electron density of a system [1], together with a class of long-range structural representations able to incorporate non-local physical effects [2]. I will then show how the combination of the two methods can be used to reproduce the electronic charge transfer in model metal electrodes, and over arbitrarily large distances. Moreover, I will show how a finite-field extension of the method can be derived in order to predict the non-local polarization effect induced by an externally applied electric field. I will conclude by demonstrating the capability of the method to reproduce the charge-density response in a gold/electrolyte capacitor under an applied voltage, predicting the differential capacitance of the system with a greater accuracy than state-of-the-art classical atomic-charge models [3].

[1] A. Grisafi, A. M. Lewis, M. Rossi, M. Ceriotti, J. Chem. Theory Comput., 19, 4451 (2023)
[2] A. Grisafi, M. Ceriotti, J. Chem. Phys. **151**, 204105 (2019)
[3] A. Grisafi, A. Bussy, M. Salanne, R. Vuilleumier, arXiv preprint arXiv:2304.08966, (2023)

AIMNet2: A Neural Network Potential to Meet your Neutral, Charged, Organic, and Elemental-Organic Needs

Dylan M. Anstine, Roman Zubatyuk, Olexandr Isayev

Department of Chemistry, Mellon College of Science, Carnegie Mellon University, Pittsburgh, Pennsylvania 15213, USA

Machine learned interatomic potentials (MLIPs) are reshaping computational chemistry practices because of their ability to drastically exceed the accuracy-length/time scale tradeoff. Despite this attraction, the benefits of such efficiency are only impactful when an MLIP uniquely enables insight into a target system or is broadly transferable outside of the training dataset, where models achieving the latter are seldom reported. In this work, we present the 2nd generation of our atoms-in-molecules neural network potential (AIMNet2)[1], which is applicable to species composed of up to 14 chemical elements in both neutral and charged states, making it a valuable model for modeling the majority of non-metallic compounds. Using an exhaustive dataset of 20 million hybrid quantum chemical calculations, AIMNet2 combines ML-parameterized short-range and physics-based long-range terms to attain generalizability that reaches from simple organics to diverse molecules with "exotic" element-organic bonding. We show that AIMNet2 outperforms semi-empirical GFN-xTB and is on par with reference density functional theory for interaction energy contributions, conformer search tasks, torsion rotation profiles, and molecular-to-macromolecular geometry optimization. Overall, the demonstrated chemical coverage and computational efficiency of AIMNet2 is a significant step toward providing access to MLIPs that avoid the crucial limitation of curating additional quantum chemical data and retraining with each new application.
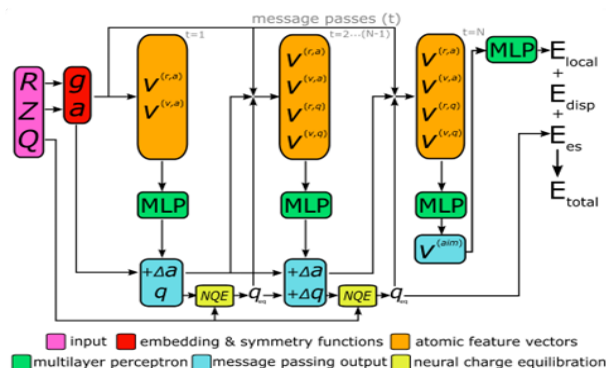
Figure 1: Architecture of the AIMNet2 model

1 Anstine D, Zubatyuk R, Isayev O. ChemRxiv Preprint. 2023; https://doi.org/10.26434/chemrxiv-2023-296ch

Universal machine learning for the response of atomistic systems to external fields

Bin Jiang

Department of Chemical Physics, University of Science and Technology of China, Hefei, Anhui 230026, China

Machine learning methods have been widely used in computational chemistry research in recent years, among which atomistic neural networks as the representation of potential energy surface and response properties have achieved great success. We have developed a physically inspired machine learning framework, namely embedded atom neural network (EANN), which uses the electron density of the embedded atom constructed from the linear combination of atomic orbitals as an atomic descriptor for the local environment[1, 2]. More recently, we propose a universal field-induced recursively embedded atom neural network (FIREANN) model, which integrates a pseudo field vector-dependent feature into atomic descriptors to represent system-field interactions with rigorous rotational equivariance. This "all-in-one" approach correlates various response properties like dipole moment and polarizability with the field-dependent potential energy in a single model, very suitable for spectroscopic and dynamics simulations in molecular and periodic systems in the presence of electric fields[3].
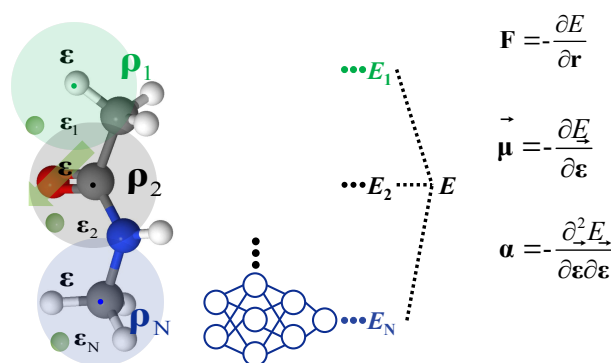

Figure 1: Schematic of FIREANN framework.

1. Y. Zhang, C. Hu and B. Jiang, *J. Phys. Chem. Lett.*, **2019**, 10, 4962-4967.
2. Y. Zhang, J. Xia and B. Jiang, *Phys. Rev. Lett.*, **2021**, 127, 156002.
3. Y. Zhang and B. Jiang, *Nat. Commun.*, **2023**, 14, 6424.

A data-driven Robotic AI-Chemist

Jun Jiang

School of Chemistry and Materials Science, University of Science and Technology of China, Hefei, Anhui 230026, China

The realization of automated chemical experiments by robots unveiled the prelude of artificial intelligent laboratory. Several AI-based systems or robots with specific chemical skills have been demonstrated, but conducting all-round scientific research remains challenging.

We have recently built a robotic AI-chemist system that is capable of proposing scientific hypothesis after reading/disgusting existing literature, executing a full set of experiments (synthesis, characterization, and performance testing) for multiple chemical tasks, and building predictive models utilizing theoretical calculations with experimental data feedback, allowing to propose new hypothesis for next iteration. With the help of computations, AI chemist has the ability to find the optimal result beyond the chemical space covered by the experiments. It means that we have created a robotic AI chemist that is capable of executing all-round chemical research with data driven intelligence. In the future, the more advanced all-round AI-Chemists equipped with scientific data intelligence may cause changes to chemical R&D.
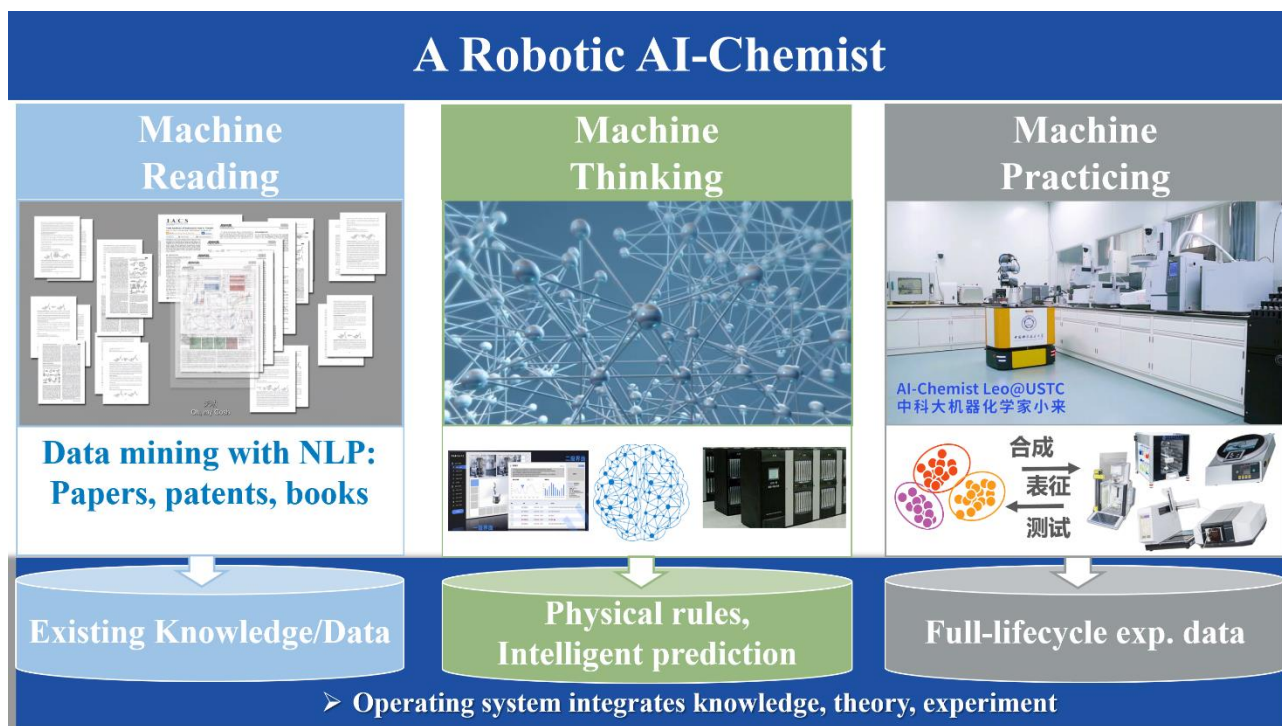

Figure 1: Three basic components of the robotic AI-Chemist system

Semi-empirical property predictors for inverse design

Kjell Jorner

ETH Zürich, Institute of Chemical and Bioengineering, Department of Chemistry and Applied Biosciences, Vladimir-Prelog-Weg 1, Zürich, Switzerland, CH-8093

Inverse design aims to deduce molecular structures directly from desired function.[1] Property predictors compute the function of the molecules suggested by generative models. Oversimplistic predictors or machine learning models are often used, and the generative models learn to "hack" these with rather meaningless molecules as a result.[2] Quantum-chemical calculations are better, but prohibitively expensive.

Here, we described the development of several property predictors based on semi-empirical (quantum chemistry) methods for applications to organic electronic materials and chemical reactivity. The methods are old and have been neglected over time in favor of more rigorous, but more expensive quantum chemical methods. For aromatic compounds, we have developed an automated version of the superaromatic stabilization energy of Aihara.[3] For OLED materials, we have developed a method based on perturbation-theory-augmented Pariser-Parr-Pople theory.[4] For chemical reactivity, we have developed a black-box implementation of the SEAM force field method[5] and applied it for a task in the TARTARUS benchmark suite for generative models. All models are open-sourced and available on GitHub.

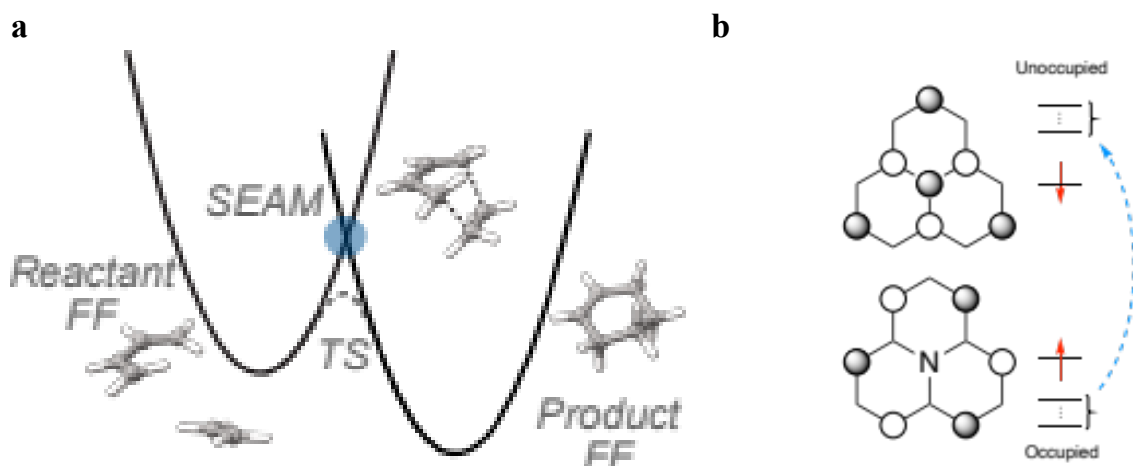a   b

Figure 1 (a) Fast reactivity estimation using the SEAM model. (b) Perturbation theory for description of OLED materials with inverted singlet-triplet gaps.

1 Sanchez-Lengeling, B.; Aspuru-Guzik, A. *Science* **2018**, *361* (6400), 360–365

2 Gendreau, P.; Turk, J.-A.; Drizard, N.; Ribeiro Da Silva, V. B.; Descamps, C.; Gaston-Mathé, Y. *J. Chem. Inf. Model.* **2023**, *63* (13), 3983–3998.

3 Aihara, J. *Bull. Chem. Soc. Jpn.* **2018**, *91* (2), 274–303.

4 Kollmar, H.; Staemmler, V. *Theor. Chim. Acta* **1978**, *48* (3), 223–239.

5 Jensen, F. *J. Am. Chem. Soc.* **1992**, *114* (5), 1596–1603.

Machine Learning Aided Quantum Chemistry Discovery in the Solution Phase

Fang Liu, Eugen Hruska, Ariel Gale, Xu Chen, Fangning Ren, Patrick Li, Rohit Gadde, Sangni Xun

Emory University, Atlanta, Georgia, USA

Although numerous critical chemical processes occur in the solution phase, datasets (computational or experimental) and machine-learning (ML) models for solution-phase molecular systems are still scarce. My research group's objective is to overcome these challenges for quantum chemistry discovery in the solution phase.

To enable the efficient generation of computational datasets of solvated molecular systems, we developed strategies to accelerate both the implicit and explicit solvent models for quantum chemistry calculations. For the implicit conductor-like polarization model (C-PCM), we developed algorithms on the graphical processing units (GPUs) to accelerate the calculation.[1][2][3] For the explicit solvent model, we developed AutoSolvate,[4] an open-source toolkit to streamline the QC calculation workflow of explicitly solvated molecules. To make these tools more accessible, we are developing a web-based platform to offer automated simulations of solvated molecules on cloud computing resources and publicly disseminate the datasets to the computational molecular science communities.

To improve the accuracy of the quantum chemistry generated solution-phase chemistry datasets, we develop ML models to reduce the discrepancy between experimental measurements and computationally predicted molecular properties in both implicit and explicit solvent models.[5] This ML correction technique has been applied to predict redox potential and absorption/fluorescence wavelength in the solution phase.[6] We are also developing data-driven recommendations of parameters for quantum chemistry calculations in the implicit solvent in order to reduce errors in the calculated excited state properties.

[1] F. Liu, N. Luehr, H. J. Kulik, and T. J. Martínez, *J. Chem. Theory Comput.* 11, 3131 (2015).

[2] F. Liu, D. M. Sanchez, H. J. Kulik, and T. J. Martínez, *Int. J. Quantum Chem* 119, e25760 (2019).

[3] A. Gale, E. Hruska, and F. Liu, *J Chem. Phys.* 154, 244103 (2021).

[4] E. Hruska, A. Gale, X. Huang, and F. Liu, *ibid.* 156, 124801 (2022).

[5] E. Hruska, A. Gale, and F. Liu, *J. Chem. Theory Comput.* 18, 1096 (2022).

[6] X. Chen, P. Li, E. Hruska, and F. Liu, *Phys. Chem. Chem. Phys.* 25, 13417 (2023).

LASP Software for Large-Scale Catalysis Simulations

Zhi-Pann Liu

Department of Chemistry, Fudan University, Shanghai, China

This lecture introduces our recent progress in LASP (Large-scale Atomic Simulation with neural network Potential) software (www.lasphub.com) and its application in catalysis. In 2018 we designed a "Global-to-Global" approach for material discovery by combining our SSW (stochastic surface walking) global optimization method with neural network (NN) techniques, which leaded to the SSW-NN method and the global neural network (G-NN) potential method[1]. The SSW-NN method is the key functionality of LASP software and has been utilized widely in solving different challenging problems. A recent progress in LASP is the implemtation of the many-body function corrected global neural network potential (G-MBNN), which shows a great advantage in improving the large-range interaction and reaction actiity prediction.[2] I will also introduce a number of LASP applications from our group, for example gas phase clusters, complex catalysts and interfaces in semiconducting devices, to demonstrate the automated global data set generation, the improved NN training procedure and the application in global structure determination. As a general tool for material simulation, the SSW-NN method provides an efficient and predictive platform for obtaining material properties.
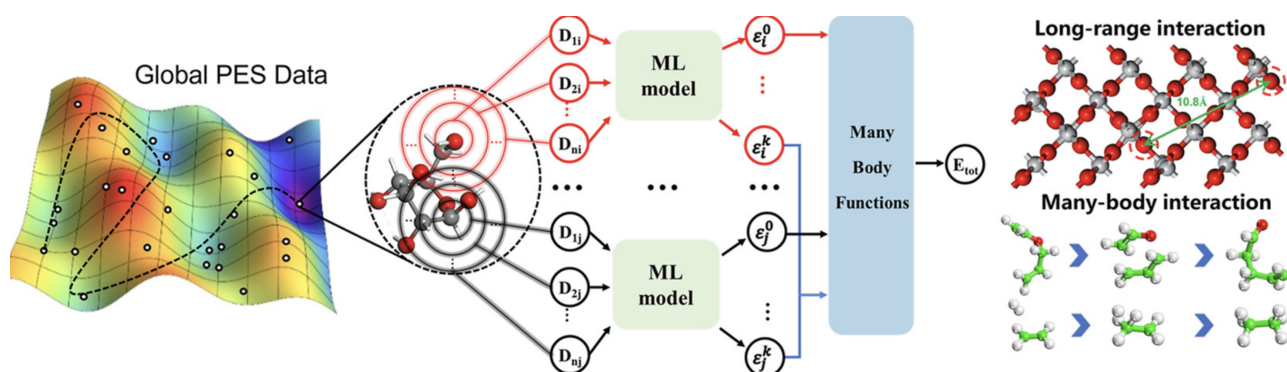

Figure 1: G-MBNN potential

1. Pei-Lin Kang, Cheng Shang, Zhi-Pan Liu, Large-Scale Atomic Simulation via Machine Learning Potentials Constructed by Global Potential Energy Surface Exploration, *Acc. Chem. Res* .2020, 53, 2119
2. Pei-Lin Kang, Zheng-Xin Yang, Cheng Shang, Zhi-Pan Liu, Global Neural Network Potential with Explicit Many-body Functions for Improved Descriptions of Complex Potential Energy Surface, J. Chem. Theory Comput. , 2023, https://doi.org/10.1021/acs.jctc.3c00873s

Machine learning of electronic structure for quantum dynamics and molecular design

Reinhard J. Maurer

Department of Chemistry and Department of Physics, University of Warwick, UK

Atomistic simulation based on quantum mechanics (QM) is currently being revolutionized by machine-learning (ML) methods. Many existing approaches use ML to predict molecular properties from quantum chemical calculations. This has enabled molecular property prediction within vast chemical compound spaces and high-dimensional parametrization of energy landscapes for the efficient simulation of measurable observables. However, as all properties derive from the QM wave function, an ML model that can predict the wave function also has the potential to predict other properties. In this talk, I will explore ML approaches that directly represent wave functions and QM Hamiltonians and their derivatives for developing methods that use ML and QM in synergy. Using example systems from heterogeneous catalysis and organic electronics, I will discuss the challenges associated with encoding physical symmetries and invariance properties into machine learning models of electronic structure. Upon overcoming these challenges, integrated ML-QM methods offer the combined benefits of data-driven parametrization and first-principles-based methods. I will discuss several opportunities associated with building ML-augmented quantum chemical methods, including Inverse Chemical Design based on ML-predicted wave functions and the development of efficient and accurate surrogate models to study materials chemistry.

Deep generative models for biomolecular engineering

Rocío Mercado Oropeza

Department of Computer Science and Engineering
Chalmers University of Technology

AI is transforming our approach to molecular engineering. Driven by the goal of accelerating drug development, our aim is to develop AI-driven molecular engineering methods which will enhance our approach to biomolecular discovery, such as drug discovery, drug repurposing, and chemical probe identification. This entails the development of generative and predictive tools that can learn from biochemical data, such as molecular structures, chemical reactions, and biomedical data. While AI can be applied to a range of molecular engineering tasks, one ideal area is de novo molecular design. De novo design is the concept of designing molecules with desired properties from scratch so as to minimize experimental screening, and is poised to allow scientists to more efficiently traverse chemical space in search of optimal molecules, and delegate error-prone decisions to computers via the use of predictive and generative models. In drug development, de novo design methods can aid medicinal chemists in the design and selection of drug candidates, with the added advantage that they can learn from datasets of billions of molecules in minutes and be constantly updated with new data. Deep molecular generative models are a particular approach to de novo design which uses deep neural networks to generate new molecules in silico, and works by proposing atom-by-atom (or fragment-by-fragment) modifications to an initial graph structure to generate compounds predicted to achieve a certain property profile. Such models can be applied to a range of therapeutic modalities.

In this talk, I will discuss the development of deep generative models for various molecular engineering tasks relevant to early-stage drug discovery. These include a model for synthesizability-constrained molecular generation, a reinforcement learning framework for molecular graph optimization, and recent applications from our group to the design of large modalities for targeted protein degradation.

Natural Quantum Monte Carlo Computation of Excited States

David Pfau[1,2], Simon Axelrod[1,3], Halvard Sutterud[2], Ingrid von Glehn[1], James S. Spencer[1]

[1]*Google DeepMind, London UK*
[2]*Imperial College London, London UK*
[3]*Harvard University, Cambridge, MA, USA*

In recent years, tools from machine learning have found useful application in computational quantum mechanics, especially in making variational quantum Monte Carlo (VMC) calculations far more accurate. These calculations mostly focus on the ground state, while excited state calculations remain more challenging. In this talk, I will present a VMC algorithm for estimating the lowest excited states of a quantum system which is a natural generalization of the estimation of ground states. The method has no free parameters and requires no explicit orthogonalization of the different states, instead transforming the problem of finding excited states of a given system into that of finding the ground state of an expanded system. Expected values of arbitrary observables can be calculated, including off-diagonal expectations between different states such as the transition dipole moment. Although the method is entirely general, it works particularly well in conjunction with recent work on using neural networks as variational Ansatze for many-electron systems, and we show that by combining this method with the FermiNet and Psiformer Ansatze we can accurately recover vertical excitation energies and oscillator strengths on molecules as large as benzene. Beyond the examples on molecules presented here, we expect this technique will be of great interest for applications of variational quantum Monte Carlo to atomic, nuclear and condensed matter physics.

Lifelong Machine Learning Potentials

Markus Reiher

ETH Zurich, Dep. of Chemistry and Applied Biosciences, Vladimir-Prelog-Weg 2, 8093 Zurich, Switzerland.

Machine learning potentials (MLPs) can retain the high accuracy of underlying quantum chemical training data and enable simulations of extended systems with little computational cost. On the downside, the training needs to be performed for each individual system. However, extensions of MLP representations by learning additional data usually require demanding retraining on all training data to avoid forgetting of previous knowledge. Moreover, common MLP descriptors cannot represent efficiently many different chemical elements. As a consequence, a large number of MLPs has been trained from scratch in recent years. We tackle these issues by the introduction of element-embracing atom-centered symmetry functions that are combined descriptors for structure and element information. In addition, we exploit uncertainty quantification to enable a continuously adapting lifelong machine learning potential (lMLP)[1] instead of a fixed, pre-trained MLP by ensuring a predefined accuracy level. To extend lMLP representations, we apply incremental learning strategies relying on data rehearsal, parameter regularization, and model architecture. Further, we propose the continual resilient (CoRe) optimizer[1,2] for fast convergence and high accuracy in lifelong learning of deep neural networks on a continuous stream of new data. Moreover, we show that the CoRe optimizer outperforms many other state-of-the-art optimizers in diverse machine learning tasks.

[1] M. Eckhoff, M. Reiher, Lifelong Machine Learning Potentials, *J. Chem. Theory Comput.* **2023**, *19*, 3509–3525.

[2] M. Eckhoff, M. Reiher, CoRe Optimizer: An All-in-One Solution for Machine Learning, *arXiv:2307.15663 [cs.LG]* **2023**.

Construction of orbital-free DFT scheme and its evaluation by ML

Junji Seino[1,2]

[1]*Department of Chemistry and Biochemistry, School of Advanced Science and Engineering, Waseda University, Tokyo, 169-8555, Japan*
[2]*Waseda Research Institute for Science and Engineering, Tokyo, 169-8555, Japan*

Density functional theory (DFT) is one of the most popular schemes for obtaining electronic states and their properties in molecules and materials. The total energy of electrons can be written as a functional of the electron density information. The kinetic and exchange-correlation energies, which are components in the total energy, have approximate formulae because the exact expressions are unknown. For the kinetic energy (KE), the majority use the Kohn−Sham (KS) expression, which introduces a set of one-electron KS orbitals instead of using the explicit functional in terms of electron density. However, the orbital-free (OF-)DFT with the practically accurate KEDF has possibilities to realize further efficient calculations of electronic states for large molecules. The present study attempts to predict the KEDF using machine learning (ML). The present scheme adopts electron density information and distances between grid points and centers of nuclei as descriptors, and the KE density of KS as the objective value. The ML KEDF provides closer KEs in KS than conventional KEDFs.[1] The scheme reproduces the potential energy curves in KS for single, double, and triple bonds with several elements.[2] Furthermore, the ML potential functional were constructed for the optimization of electron density in the OF-DFT framework with the Gaussian basis functions.[3]

For the correlation energy, we also proposed an ML correlation model that is built using the regression between density variables such as electron density and correlation energy density.[4] The correlation energy density of coupled cluster singles, doubles, and perturbative triples [CCSD(T)] is derived based on grid-based energy density analysis. In addition, the scheme was extended to the frozen core calculations.[5]

In DFT calculations based on ML, the applicability in terms of accuracy is considerably important. Thus, a scheme for evaluations of DFT calculations using a clustering method based on ML and structural/electronic descriptors was developed.[6] We generated 36 clusters consistent with human intuition using 30,436 carbon atoms from the QM9 dataset. The results in the clustering method were used to evaluate results in DFT.

[1]J. Seino, R. Kageyama, M. Fujinami, Y. Ikabata, and H. Nakai, *J. Chem. Phys.* **2018**, 148, 241705.
[2]J. Seino, R. Kageyama, M. Fujinami, Y. Ikabata, and H. Nakai, *Chem. Phys. Lett.* **2019**, 734, 136732.
[3]M. Fujinami, R. Kageyama, J. Seino, Y. Ikabata, and H. Nakai, *Chem. Phys. Lett.* **2020**, 748, 137358.
[4]T. Nudejima, Y. Ikabata, J. Seino, T. Yoshikawa, and H. Nakai, *J. Chem. Phys.* **2019**, 151, 024104.
[5]Y. Ikabata, R. Fujisawa, J. Seino, T. Yoshikawa, and H. Nakai, *J. Chem. Phys.* **2020**, 153, 184108.
[6]Y. Nakajima, T. Ohmura, and J. Seino, submitted.

On Electrons and Machine Learning Force Fields

Alexandre Tkatchenko

Department of Physics and Materials Science, University of Luxembourg, Luxembourg

Machine Learning Force Fields (MLFF) should be accurate, efficient, and applicable to molecules, materials, and interfaces thereof. The first step toward ensuring broad applicability and reliability of MLFFs requires a robust conceptual understanding of how to map interacting electrons to interacting "atoms". Here I discuss two aspects: (1) how electronic interactions are mapped to atoms with a critique of the "electronic nearsightedness" principle, and (2) our developments of symmetry-adapted gradient-domain machine learning (sGDML) framework for MLFFs generally applicable for modeling of molecules, materials, and their interfaces. I highlight the key importance of bridging fundamental physical priors and conservation laws with the flexibility of non-linear ML regressors to achieve the challenging goal of constructing chemically-accurate force fields for a broad set of systems. Applications of sGDML will be presented for small and large (bio/DNA) molecules, pristine and realistic solids, and interfaces between molecules and 2D materials. In addition, I will briefly report on the results of the recent "Crash Testing MLFF" workshop held in Luxembourg that convened many of the leading developers of MLFFs and MLPs.

[Refs] Nature Commun. 13, 3733 (2022); Sci. Adv. 3, e1603015 (2017); Nature Commun. 9, 3887 (2018); Comp. Phys. Comm. 240, 38 (2019); J. Chem. Phys. 150, 114102 (2019); Sci. Adv. 5, eaax0024 (2019).

Active learning for data-efficient optimisation of functional materials

Milica Todorović

Dept. of Mechanical and Materials Engineering, University of Turku, 20014 Finland

The arrival of materials science data infrastructures in the past decade has ushered in the era of data-driven materials science based on artificial intelligence (AI) algorithms, which has facilitated breakthroughs in materials optimisation and design. Of particular interest are active learning algorithms, where datasets are collected on-the-fly in the search for optimal solutions. We encoded such a probabilistic algorithm into the Bayesian Optimization Structure Search (BOSS) Python tool for materials optimisation[1].

BOSS builds N-dimensional surrogate models for materials' energy or property landscapes to infer global optima, allowing us to conduct targeted materials engineering. The models are iteratively refined by sequentially sampling materials data with high information content. This creates compact and informative datasets. We utilised this approach for computational density functional theory studies of molecular surface adsorbates[2], thin film growth[3], solid-solid interfaces[4] and molecular conformers[5]. With experimental colleagues, we applied BOSS to accelerate the development of novel materials with targeted properties[6], and to optimise materials processing[7]. With recent multi-objective and multi-fidelity implementations for active learning, BOSS can make use of different information sources to help us discover optimal solutions faster.

1 M. Todorović, *npj Comput. Mater.*, 5, 35 (2019)
2 J. Järvi, *Beilstein J. Nanotechnol.* 11, 1577-1589 (2020), *Adv. Func. Mater.*, 31, 2010853 (2021)
3 A. Egger, et al., *Adv. Sci.* 7, 2000992 (2020)
4 A. Fangnon, et al.,*ACS Appl. Mater. Interfaces* 14 (10), 12758-12765 (2022)
5 L. Fang, et al., *J. Chem. Theory Comput.* 17, 1955 (2020)
6 S-A Jin, et al., *MRS Bulletin* 47, 29-37 (2022)
7 J. Löfgren, et al., *ACS Sustainable Chem. Eng.* 10, 9469 (2022)

Exploration of the Two-electron Correlation Space with Data-driven Quantum Chemistry

Konstantinos D. Vogiatzis

Department of Chemistry, University of Tennessee, Knoxville, TN, USA

The data-driven computational methodology developed by our group combines quantum chemistry with machine learning (ML) to surpass size limitations of accurate but computationally demanding methods such as coupled-cluster (CC). We have previously demonstrated the speedup and transferability that the data-driven CCSD (DDCCSD) model can achieve.[1,2,3] One major limitation of the DDCC models is the size of the training sets that increases exponentially with the system size. We have recently introduced effective sampling of the amplitude space as a solution to this issue. Five different amplitude selection techniques that reduce the amount of data used for training were evaluated, an approach that also prevents model overfitting and increases DDCCSD portability to more complex molecules or larger basis sets. Extension to perturbative triples (T) and alternative architectures based on graph neural networks will be also discussed. Finally, we have extended this approach to other quantum chemical methods, such as the variational 2-RDM (v2RDM).[4]

1 Townsend, K. D. Vogiatzis, *J. Phys. Chem. Lett.*, **2019**, *10*, 4129.
2 J. Townsend, K. D. Vogiatzis, *J. Chem. Theory Comput.*, **2020**, *16*, 7453.
3 G. M. Jones, P. D. V. S. Pathirage, K. D. Vogiatzis, Data-driven Acceleration of Coupled-Cluster Theory and Perturbation Theory Methods, in: "Quantum Chemistry in the Age of Machine Learning", 2022, Editor: Pavlo Dral, Elsevier, pp. 509-529.
4 G. M. Jones, R. R. Li, A. E. DePrince III, K. D. Vogiatzis *J. Phys. Chem. Lett.*, **2023**, *14*, 6377.

Quantum chemical Hamiltonians as flexible and interpretable model forms for machine learning

David Yaron, and Frank Hu

Department of Chemistry, Carnegie Mellon University, Pittsburgh PA 15213

The high computational cost of quantum chemistry is currently a primary imitation on the size and type of systems that can be studied. Machine learning (ML) provides a means to dramatically lower cost while maintaining high accuracy. However, ML models often sacrifice interpretability by using components, such as the artificial neural networks of deep learning, that function as black boxes. These components impart the flexibility needed to learn from large volumes of data but make it difficult to gain insight into the physical or chemical basis for the predictions. Here, we demonstrate that semiempirical quantum chemical (SEQC) models can learn from large volumes of data without sacrificing interpretability. The SEQC model is that of Density Functional based Tight Binding (DFTB) with fixed atomic orbital energies and interactions that are one-dimensional functions of interatomic distance. This model is trained to ab initio data in a manner that is analogous to that used to train deep learning models. Using benchmarks that reflect the accuracy of the training data, we show that the resulting model maintains a physically reasonable functional form while achieving an accuracy, relative to coupled cluster energies with a complete basis set extrapolation (CCSD(T)*/CBS), that is comparable to that of density functional theory (DFT). This suggests that trained SEQC models can achieve low computational cost and high accuracy without sacrificing interpretability. Use of a physically-motivated model form also substantially reduces the amount of ab initio data needed to train the model compared to that required for deep learning models.

Finite-field molecular dynamics simulation goes machine-learned

Chao Zhang

Department of Chemistry - Ångström Laboratory, Uppsala University
Lägerhyddsvägen 1,
Box 538, 75121 Uppsala, Sweden
chao.zhang@kemi.uu.se

Physical chemistry of electrified interfaces and ionic solutions plays a fundamental role in energy-related applications such as electrocatalysis, supercapacitors, fuel cells, and batteries. A realistic representation of these electrochemical systems requires treating electronic, structural, and dynamic properties on an equal footing. Density functional theory-based molecular dynamics (DFTMD) is perhaps the only approach that can provide a consistent atomistic description.

Here, I will first discuss our recent progress in modelling the protonic double layer at metal-oxide/electrolyte interfaces using finite-field DFTMD simulations [1]. Then, this is followed by our new effort on simulating electrified metal/electrolyte interfaces and deploying its machine-learning version with the PiNN code [2].
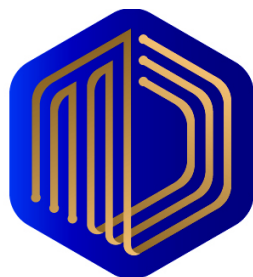
Reference:
[1] Knijff, L., Jia, M. and Zhang, C. *Encyclopedia of Solid-Liquid Interfaces*, 2024, 2: 567
[2] Dufils, T., Knjiff, L., Shao, Y. and Zhang, C. *J. Chem. Theory Comput.*, 2023, 19: 5199

Sponsored by