The four projects listed here will contribute to the Accelerate Programme for Scientific Discovery, an interdisciplinary research team that uses machine learning to solve scientific problems. Students working on these projects will help facilitate the use of machine learning techniques in Physics, Chemistry and Materials Science.

The QM9 challenge: learning quantum chemistry from small dataset

All molecules and materials on Earth can be described by the Schrödinger equation, but the computational cost of solving the equation makes this prohibitive. Machine learning (ML) has emerged as a promising tool to circumvent the high cost of quantum mechanical calculations, by performing statistical learning of relatively few examples, and then making speedy and accurate predictions about other materials and molecules. Such tool will vastly enhance our capability to explore the astronomically large chemical space, and accelerate chemical discovery.

Before any practical use, a ML for chemistry model needs to be validated. The most common benchmark dataset for this is called QM9, which consists of 134k smallest organic molecules containing up to 9 heavy atoms (C, O, N, or F; excluding H) along with their quantum properties. Several ML studies have already been published using QM9 (see Figure 1, Ref [1]). A popular challenge is to develop a next-generation ML model that can learn the quantum mechanical energy of the organic molecules with higher accuracy using less training data. This may be done by improving the representations of molecules, or by employing smarter training algorithms. It is quite likely that the next improvement will stem from a combination of supervised and unsupervised learning, e.g. first learn the representations using an unsupervised model, and then perform regression.

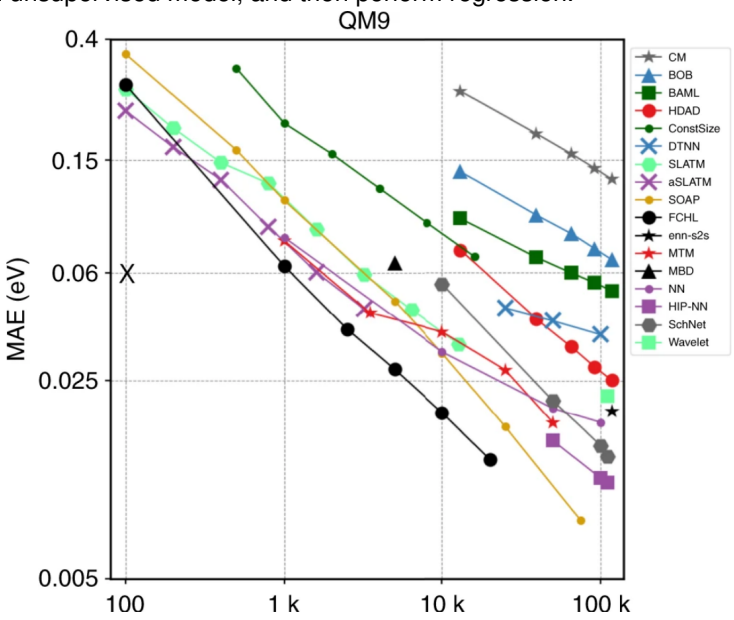

Figure 1: *Models shown differ by representation and architecture. The black X denotes the "QM9 challenge" of achieving 1* $\mathrm{kcal/mol}$ *accuracy on the QM9 dataset using only 100 molecules for training. Figure from Ref. [2].*

In this project, you will first learn about the state-of-the-art ML models for quantum chemistry, and then you will have the chance to tackle this "QM9 challenge". It is worth noting that the ML models and the skills are transferable to predicting other chemical or biological properties of materials and molecules, including drug-likeness, catalytic activities, and optical properties.

References

[1] Felix A Faber et al. "Prediction errors of molecular machine learning models lower than hybrid DFT error". In: *Journal of chemical theory and computation* 13.11 (2017), pp. 5255–5264.

[2] O. Anatole von Lilienfeld and Kieron Burke. "Retrospective on a decade of machine learning for chemical discovery". In: *Nature Communications* 11.1 (Sept. 2020). DOI: 10.1038/s41467-020-18556-9. URL: https://doi.org/10.1038/s41467-020-18556-9.

Measuring the mutual information content between different descriptors of materials and molecules

Machine learning (ML) of atomic-scale properties is revolutionizing computational physics and chemistry, by enabling accurate predictions without performing expensive quantum mechanical calculations. The accuracy, efficiency and reliability of these ML models, however, depends strongly on the choice of descriptors used as input for the ML method.

Many descriptors have been proposed in the past to represent molecular and material structures. Generally speaking, a good representation should be invariant to translation and rotation, as well as the permutation of atoms of the same species. Amongst these, the Atom Centered Symmetry Functions (ACSFs) [1] and the Smooth Overlap

of Atomic Positions (SOAP) [2] descriptors are probably the most popular ones. Many descriptors share a common theoretical foundation (e.g. most descriptors that are based on the atomic density differ only in the basis functions onto which the density is projected [3]), and often behave rather similarly. As an example, Figure 2 shows the principal component analysis (PCA) maps of the QM9 dataset (a popular benchmark set in chemistry that contains 134k small molecules of up to 9 heavy atoms) based on the SOAP and ACSF descriptors. Despite the distinct forms and the two orders of magnitude difference in the dimensionalities of the two types of descriptors, the commonalities in their PCA maps can be spotted.

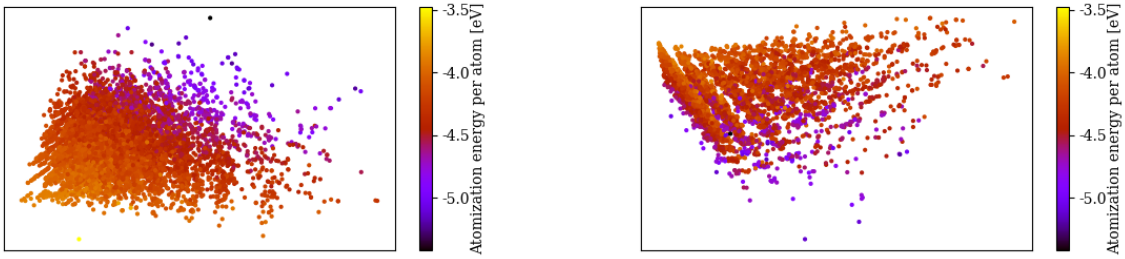

Figure 2: *PCA maps of the QM9 database using the SOAP descriptors (left) and the ACSFs (right). Each point indicates a small organic molecule in the dataset.*

This project focuses on investigating how these existing descriptors (different types of descriptors as well as descriptors of the same type but with different hyper parameters) are related in terms of their mutual information content. The methodology can be a theoretical analysis of the mathematical formulations of the descriptors. Alternatively, non-parametric measurements of the mutual information content between a pair of descriptors on benchmark datasets (e.g. QM7b, QM9) can be performed. The insights coming from this project will contribute to our understanding of how to best represent atomic systems when using ML models, help formulating next-generation descriptors, and ultimately make an impact on the ML for chemistry revolution.

References

[1] Jörg Behler. “Atom-centered symmetry functions for constructing high-dimensional neural network potentials”. In: *The Journal of chemical physics* 134.7 (2011), p. 074106.

[2] Albert P. Bartók, Risi Kondor, and Gábor Csányi. “On representing chemical environments”. In: *Phys. Rev. B* 87 (2013), p. 184115.

[3] Michael J Willatt, Félix Musil, and Michele Ceriotti. “Atom-density representations for machine learning”. In: *The Journal of chemical physics* 150.15 (2019), p. 154110.

Learning how atoms interact

The holy grail of computational physics and chemistry is to predict material properties by solving the fundamental equations of quantum mechanics (called “ab initio” methods). However, as illustrated in Figure 3, *ab initio* calculations are computationally demanding such that these calculations are restricted to small system sizes ($\sim$ 100 atoms) and short simulation time ($\sim 10^{-12}$ s), making them unpractical for modelling most systems. On the other hand, inexpensive empirical forcefields (i.e. using simple functions to approximate atomic interactions) may not be available for many systems or may lack quantitative accuracy.

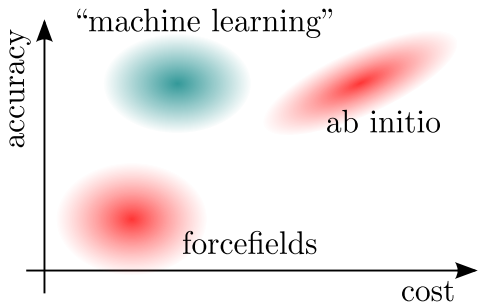

Figure 3: *A comparison between different ways of computing the potential energy surface.*

Machine learning (ML) has emerged as a way to sidestep the *ab initio* calculations, using a small number of reference evaluations to generate a data-driven model of the atomic interactions. The field of ML potentials is young but extremely dynamic. Thus far, ML potentials have been constructed for systems including small organic molecules, bulk condensed materials and interfaces [1]. As a recent example, we constructed a ML potential for high-pressure hydrogen, and revealed how hydrogen gradually turns into a metal in giant planets [2].

In this project, you will train a ML potential for a technologically important system. For example, titanium dioxide has unique optical and photocatalytic properties, and certain high-pressure hydrides are high temperature superconductors. In addition, you will learn a number of techniques for curating the training set, including sparsification (CUR and FPS), dimensionality reduction, and clustering. The outcome of the project will not only contribute to the understanding of the system of study, but also help building a general and automated workflow for constructing ML potentials.

References

[1] Volker L Deringer, Miguel A Caro, and Gábor Csányi. "Machine Learning Interatomic Potentials as Emerging Tools for Materials Science". In: *Advanced Materials* 31.46 (2019), p. 1902765. DOI: 10.1002/adma.201902765.

[2] Bingqing Cheng* et al. "Evidence for supercritical behaviour of high-pressure liquid hydrogen". In: *Nature* 585.7824 (Sept. 2020), pp. 217–220. DOI: 10.1038/s41586-020-2677-y.

Using ML to construct a coarse-grained model for water

Water, as the substrate in which life occurs, is perhaps the most important chemical compound around. Modelling water accurately and efficiently is of crucial importance in simulating pure water, solutions, biochemical systems, and many reactions happening in water.

A straightforward way of making a model for water is to treat each oxygen or hydrogen atom separately, and subsequently compute the interactions between atoms during simulations. To speed up the computation, one can employ a coarse-grained (CG) model in which each H_2O molecule is represented by a single particle. Coarse-graining is a way to enhance both the time and size of simulations that we can perform. Many approaches for coarse-graining have been developed [1], but designing the functional form for a CG potential often rely on human insights as well as trail-and-error processes.

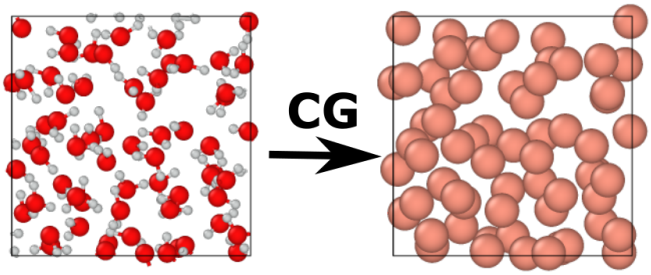

Figure 4: *A schematic of a fully atomistic and a coarse-grained (CG) representation of molecular water.*

In this project, you will construct a CG model of water, using a highly accurate atomisitic neural network potential of water [2] as the reference. Two recent works [3, 4] used machine-learning approaches for coarse graining, but they only work at a fixed temperature and pressure condition and do not account for nuclear quantum effects (NQEs) of light elements such as hydrogen. This project will expand the existing methodology, so the resulting CG model can be transferable to other thermodynamic conditions, and will include NQEs. Such a high quality CG model for water will be extremely useful for understanding the unique behaviors of water, and for modelling biological systems.

References

[1] Marissa G. Saunders and Gregory A. Voth. "Coarse-Graining Methods for Computational Biology". In: 42 (2013), pp. 73–93. ISSN: 1936-122X. DOI: 10.1146/annurev-biophys-083012-130348.

[2] Bingqing Cheng* et al. "Ab initio thermodynamics of liquid and solid water". In: *Proceedings of the National Academy of Sciences* 116.4 (2019), pp. 1110–1115.

[3] S. T. John and Gábor Csányi. "Many-Body Coarse-Grained Interactions Using Gaussian Approximation Potentials". In: *The Journal of Physical Chemistry B* 121.48 (Nov. 2017), pp. 10934–10949. DOI: 10.1021/acs.jpcb.7b09636.

[4] Linfeng Zhang et al. "DeePCG: Constructing coarse-grained models via deep neural networks". In: 149 (2018), p. 034101. ISSN: 0021-9606. DOI: 10.1063/1.5027645.