Machine Learning interatomic potentials from accurate quantum Monte Carlo calculations: application to water solvation.

In recent years, the modelling of physical systems and the numerical prediction of their properties have seen an unprecedented boost driven by artificial intelligence (AI) technologies applied to fundamental science. In the realm of condensed matter and computational chemistry, machine learning interatomic potentials (MLIP) trained on datasets generated by density functional theory (DFT) calculations or by more advanced coupled cluster techniques have proven their reliability and shown breakthrough speedup in molecular dynamics simulations. They allowed one to reach systems sizes and time scales previously inaccessible, bridging more tightly a microscopic description with the macroscopic world, by retaining at the same time the quantum nature of matter and by revealing its implications at the macroscale. In this regard, paradigmatic examples are quantum criticality and phase transitions appearing in pristine hydrogen[1], as well as in liquid water and water ice[2], particularly challenging because they need an accurate quantum description of both electrons and nuclei.

While an impressive progress has been made in the development of new neural network (NN) architectures to improve the flexibility and representability of their target models[3], the dependence of the NN predictability on the quality of the training set is emerging more and more as a severe weakness in the framework of atomistic simulations. This is a major bottleneck for next-generation AI schemes applied to the condensed phase, because they badly need accurate data training to improve their self-sustained reliability. With the aim at overcoming these flaws, seminal works explored two main directions, on one side by creating more sophisticated workflows that guarantee data provenance and curation obeying FAIR-data principles[4], on the other side by improving the reference theories used to generate the training sets, with the replacement of DFT by more advanced quantum chemistry methods, such as coupled cluster[5]. In this spirit, only preliminary attempts have been presented so far to derive MLIP based on training sets generated by quantum Monte Carlo (QMC), a powerful family of methods that retains good scalability and high precision in electronic structure calculations. Two main factors have been detrimental for a wider deployment of this promising ML+QMC combination: the stochastic noise inherent in the QMC technique and the difficulty in generating MLIP based on energy-only datasets, without the information of nuclear forces, harder to access in QMC. In this project, we want to exploit the so-called Δ-machine learning approach and the information of nuclear forces readily available in the TurboRVB QMC package[6] to construct an accurate MLIP for water. Preliminary applications to liquid hydrogen have shown how Δ-learning is able to harness the QMC correction with respect to DFT forces and energies despite the stochastic noise[7].

During this thesis, a systematic analysis will be carried out to develop a ML+QMC for water by studying in particular:

- new strategies to generate improved training sets based either on sparsification or on active learning (or on both), obtained from configurations produced by path integral molecular dynamics, thus natively including quantum nuclear effects;
- advanced workflows to optimize the QMC wavefunction for the target system to generate low-variance nuclear forces and highly-accurate energies in a high-throughput mode deployed across the training set configurations. This will naturally include non-trivial long-range interactions and polarization effects at the QMC level of accuracy;

- optimal machine learning schemes to combine baseline and Δ models, efficiently and accurately, by relying on state-of-the-art machine learning interatomic potentials with higher order equivariant message passing, such as MACE[8].

The developed ML+QMC framework will be applied to water solvation. Water is a very active area of research; it is a playground where quantum nuclear and many-body effects lead to exotic phases and a rich variety of ice structures. Water solvation depends strongly on thermodynamic conditions[9] and it is a crucial aspect to understand many properties, ranging from ion conductivity to protein folding. Accurate potentials for water already exist, such as q-TIP4P and MB-pol, but they are not easily used in combination with conventional force fields. In our approach, both water and the solute will be treated at the QMC level, guaranteeing a seamless integration of different components in the final MLIPs, and their transferability. As a solute, we will start from the simple proton, a charge defect that is relevant for benchmarking the anomalous ion conductivity in water. As a second step, we will study H_2, which will allow us to analyze H_2O-H_2 interactions in host-guest structures, important for hydrogen storage applications of high environmental impact.

The content of the present project is aligned with PostGenAI@Paris project's scientific missions as disruptive technology. Indeed, the deployment of the ML+QMC approach can lead to the generation of a new family of MLIP potentials retaining the same quality as the first-principles QMC simulations. This new family of MLIP will have a breakthrough character as it will overcome deficiencies related to the training sets generation, affecting a substantial fraction of modern MLIPs, and will rely on a more robust integration between machine learning schemes and unbiased QMC data. The digital infrastructure for the data, training, and MLIP repository will be supported through the DIAMOND initiative of the PEPR DIADEM. The IMPMC partner has indeed hired, within this initiative, a CDD research engineer specialized in the automatization of MLIP training and generation.

We are looking for a student with a solid background in theoretical physics, statistical physics and physical chemistry of materials. Familiarity with mathematical formalism and numerical programming is an important prerequisite. He/she will be supervised by Michele Casula, expert in quantum Monte Carlo methods and condensed matter theorist, by Rodolphe Vuilleumier, expert in quantum chemistry and quantum nuclear dynamics, and by Marco Saitta, in the CoPil of DIAMOND and expert in both machine learning and atomistic simulations.

[1] *Quantum phase diagram of high-pressure hydrogen,* L. Monacelli, M. Casula, K. Nakano, S. Sorella and F. Mauri, Nature Physics **19**, 845 (2023).
[2] *Water Is Cool: Advanced Phonon Dynamics in Ice Ih and Ice XI via Machine Learning Potentials and Quantum Nuclear Vibrations*, A. Živković, U. Terranova and N. H. de Leeuw, Journal of Chemical Theory and Computation **21**, 1978 (2025).
[3] *Neural network potentials: A concise overview of methods*, E. Kocer, T.W. Ko, J. Behler, Annual Review of Physical Chemistry **73**, 163 (2022).
[4] *Shared metadata for data-centric materials science*, L. M. Ghiringhelli et al., Scientific Data **10**, 626 (2023).
[5] *Approaching coupled cluster accuracy with a general-purpose neural network potential through transfer learning*, J. S. Smith et al., Nature Communications **10**, 2903 (2019).
[6] *TurboRVB: A many-body toolkit for ab initio electronic simulations by quantum Monte Carlo*, K. Nakano et al., The Journal of Chemical Physics **152**, 204121 (2020).
[7] *Principal deuterium Hugoniot via quantum Monte Carlo and Δ-learning,* G. Tenti, K. Nakano, A. Tirelli, S. Sorella, and M. Casula, Physical Review B **110**, L041107 (2024).
[8] *Hydrogen liquid-liquid transition from first principles and machine learning*, G. Tenti, B. Jäckl, K. Nakano, M. Rupp, and M. Casula, arXiv:2502.02447 (2025).
[9] *Equilibrium Magnesium Isotope Fractionation between Aqueous Mg2+ and Carbonate Minerals: Insights from Path Integral Molecular Dynamics*, C. Pinilla, M. Blanchard, E. Balan, S. K.Natarajan, R. Vuilleumier and F. Mauri, Geochimica et Cosmochimica Acta **163**, 126 (2015).