Fitting the potential energy surface of CO_2 with neural networks

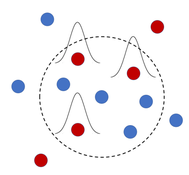
Alexandra BYBA. Suppervisors: Marco SAITTA, Michele CASULA, Michele CERIOTTI
PHYSIX, IMPMC, Sorbonne Université & COMSO, EPFL
Contact: alexandra.byba@epfl.ch
Acknowledgment: Julien Heu, PhD student at PHYSIX

Context & Motivation

The Deep Carbon Cycle plays an important part in the long term climatic trends on Earth and the carbon dioxide phase diagram at geological conditions is key in the understanding of this cycle. The PHYSIX team studied cinematic aspects of **phase transition between polymeric and molecular CO_2 phases** through ab initio molecular dynamics, using DFT-PBE to compute forces [1]. However, the **computational power** required to carry out such simulations motivated the development of **machine-learned potential energy surfaces** (ML-PES) that could be used as force fields for large-scale simulations.

1. Local Chemical Environment Representation

So as to efficiently train a ML model, one needs to represent a chemical environment in a **scalable, complete and unique** way using descriptors which are **differentiable and invariant** to basic symmetry operations and permutations. **SOAP** descriptors [2] were used, allowing an arbitrary degree of accuracy with physically inspired and easy-to-tune parameters.



2. Neural Network Architecture

a. State of the art architecture: High-dimensional NN (HD-NN) [3]
Within the HD-NN architecture, a **unique atomic-NN** is created for each chemical species. It predicts the atomic contribution of every atom of this species which are then summed to predict the total energy of the system.

b. Proposed architecture: Interaction NN
We proposed to further subdivide the atomic-NNs in **pair interaction contributions**, taking advantage of the SOAP vectors structure. This more constrained architecture yields better accuracy with and equivalent model complexity and scales well with the number of species.

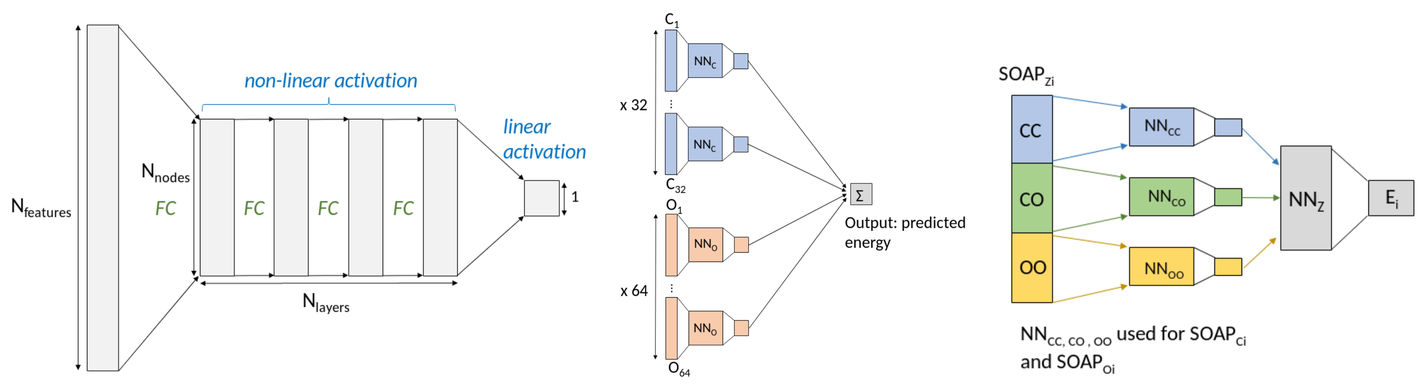


Figure 1: Schematic representation of the atomic-NN (left), HD-NN (middle) and interaction NN (right)

c. Algorithm Overview
A pre-processing of the SOAP vectors consisting in a **PCA**, **standardizations** before and after the PCA, and a **re-normalization of the energies to** $[-N_{\mathbf{part}}; N_{\mathbf{part}}]$ is defined for each species or interaction type. The Adam optimizer with early-stopping is used for training.

3. Energy predictions

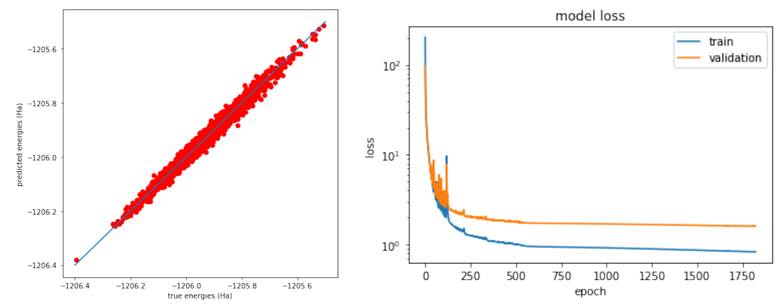


Figure 2: Predicted VS target energy (left) and learning curve (right) of HD-NN with full pre-processing (d=10 Å, T=3000 K).

The energy rescaling yields good prediction accuracy for all the energy range and over-fitting is avoided by using early-stopping as a regularizer (Figure 2).

Architecture	d (Å)	T (K)	DFT std (eV/atom)	RMSE (eV/atom)
HD-NN (no pre-process.)	10.0	3000	3.1e-02	7.9e-03 (± 1.4e-03)
HD-NN	10.0	3000	3.1e-02	3.8e-03 (± 2.1e-04)
HD-NN	10.0, 9.8, 9.6 & 9.35	3000	1.2e-01	2.5e-02 (± 1.8e-03)
Interaction NN	10.0	3000	3.1e-02	3.4e-03 (± 4.2e-04)

Table 1: Results of 5-folds cross-validation for different architectures and polymeric CO_2 conditions.

The proposed pre-processing scheme addresses the low compacity of SOAP descriptors while significantly increasing prediction accuracy. The developed interaction-NN yields slightly better prediction accuracy than the HD-NN (Table 1).

4. Validation of PES quality with Monte Carlo simulations

Monte Carlo (MC) simulations were performed using the HD-NN to further evaluate the quality of the PES by comparing the radial distribution functions generated to the ones of the original dataset (ab initio simulations). Figure 3 demonstrates that the MC simulations generate **energies and radial distributions very close to the original ones and manage to preserve the molecular structure of the system**. Artificial minima however appear in the CO radial pair distribution functions, evidencing instabilities of the fitted PES.

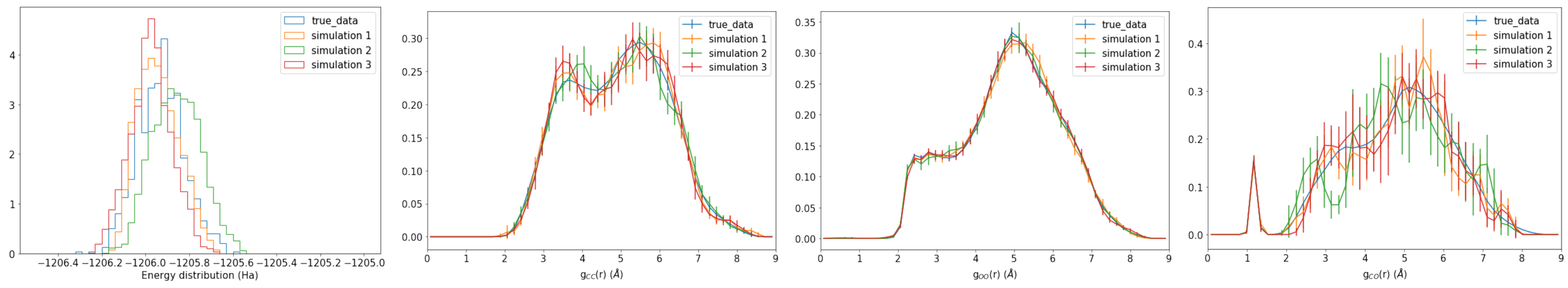


Figure 3: Energy (left) and radial distribution functions of MC simulations (from left to right: $g_{\mathrm{CC}}(\mathrm{r})$, $g_{\mathrm{OO}}(\mathrm{r})$ and $g_{\mathrm{CO}}(\mathrm{r})$). For each simulation, 100,000 "all-particles" MC steps were performed with a maximal 0.006 Å step-size. The HD-NN was used to compute the energies.

Conclusions and Future Work

The HD-NN architecture was compared to the Interaction NN, which appears to return slightly better prediction accuracies on specific datasets. Besides, MC simulations suggest that the quality of the prediction is very sensitive to the similarity between the studied configuration and the train set. An **iterative training** could be implemented, where the NN re-training of the model is performed on the fly, during MC or MD simulations, on generated configurations with low-fidelity NN predictions.

References

[1] M. Moog. Carbon dioxide at extreme conditions. *Thèse, Sorbonne Université*, 2018.
[2] Albert P. Bartók. On representing chemical environments. *Phys. Rev. B*, 2013.
[3] Jörg Behler. Neural network potential-energy surfaces in chemistry. *PCCP*, 2011.