UNIVERSITY OF OXFORD

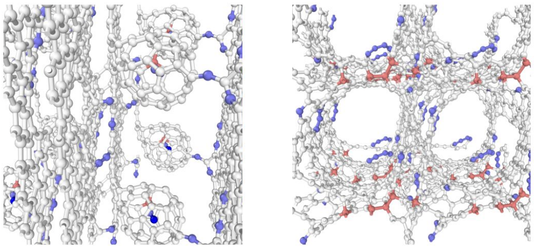
Synthetic pre-training for neural-network interatomic potentials

John L. A. Gardner, Zoé Faure Beaulieu, Kathryn Baker, and Volker L. Deringer

Department of Chemistry, University of Oxford, Oxford, UK

@jla_gardner

Synthetic data

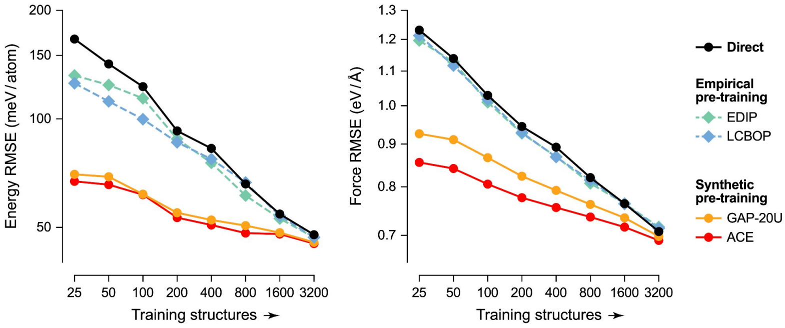


We show that "synthetic" data, generated using a fast machine-learned model rather than the quantum-mechanical ground truth, are useful in their own right.

We used the existing, machine-learned C-GAP-17[1] potential to create a 22.9 million atom carbon dataset. As a synthetic regression target, each atom has been labelled with a local energy from this potential.

You can find and download this dataset on GitHub: github.com/jla-gardner/carbon-data

Fine-tuning



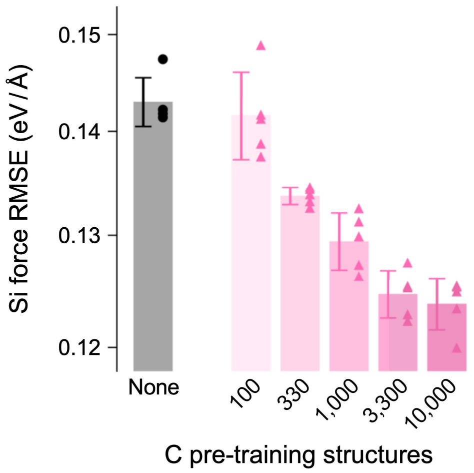
We pre-train a series of NequIP[2] models to mimic the energy and force labels generated from several existing synthetic sources on our synthetic dataset.

We find that empirical potentials provide very small improvements in both force and energy errors. In contrast, ML based pre-training sources (in this case, previously published GAP and ACE models for carbon), lead to **strong positive transfer** upon fine-tuning. This technique is particularly useful in the low data regime, improving data efficiency by **up to 32×** for a given accuracy.

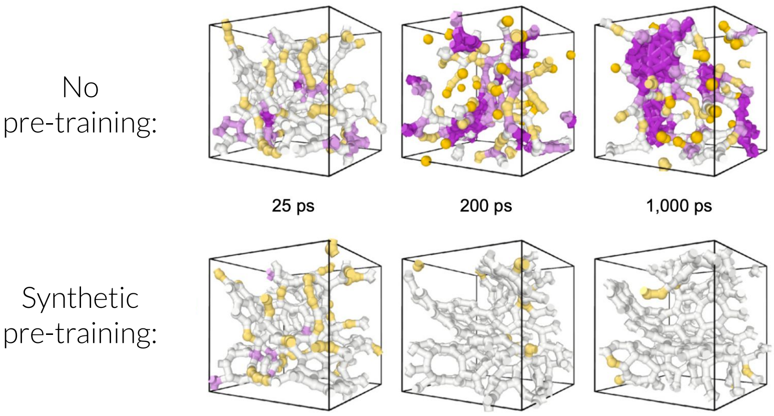
Alchemical transfer

We show a proof-of-concept for alchemical transfer learning: we pre-train on (linearly scaled) **carbon** structures and synthetic labels, before fine-tuning on **silicon** structures with ground-truth DFT labels.

We find significant positive transfer when pre-training on modest amounts of Carbon data.



Improved robustness



Synthetic pre-training leads to increased robustness in MD.

We directly train a NequIP on a small dataset of carbon nanotubes. This model has little general physical knowledge, and quickly breaks down in MD under mild conditions.

In contrast, a generally pre-trained NequIP fine-tuned on the same dataset gives sensible predictions, and is robust to loss of atoms and other strange behaviour. All weights and code can be found at: github.com/jla-gardner/nnp-pre-training

References:

[1] *Phys. Rev. B* 95, 094203 (2017)

[2] *Nat Commun.* 13, 2453 (2022)

Digital Discovery (2023) →

arXiv (2023) ←

J.L.A.G. acknowledges a UKRI Linacre - The EPA Cephalosporin Scholarship, support from an EPSRC DTP award (EP/T517811/1), and from the Department of Chemistry, University of Oxford. V.L.D. acknowledges a UK Research and Innovation Frontier Research grant [EP/X016188/1] and support from the John Fell OUP