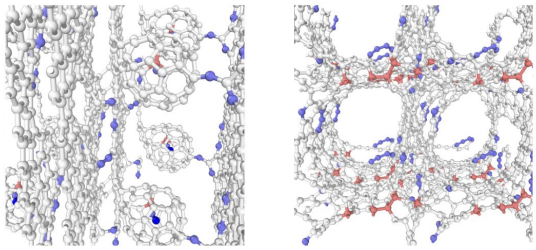
Synthetic Data Enable Experiments in Atomistic Machine Learning

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

University of Oxford

@jla_gardner

Synthetic Data

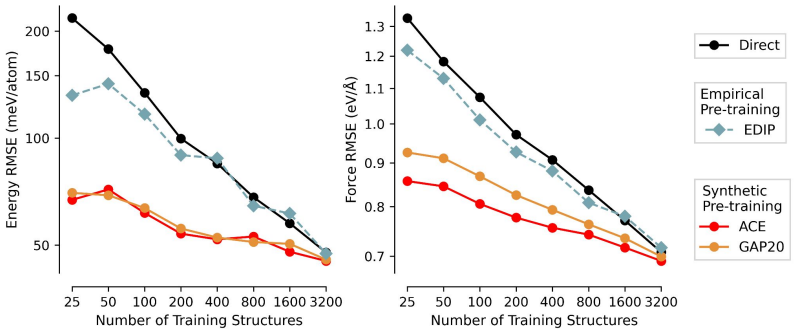


We show that "synthetic" data, generated using a fast machine-learned model rather than the quantum-mechanical ground truth, are useful in their own right.

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

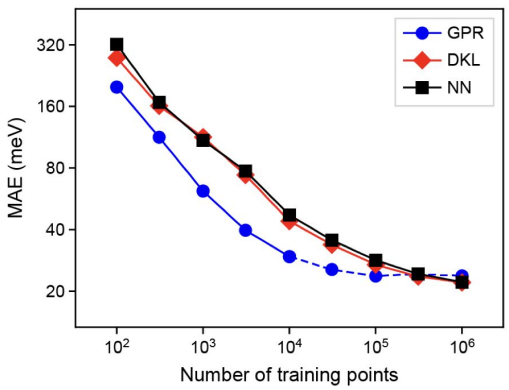
Fine-tuning



We pre-train a series of NequIP[3] 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 ACE and GAP models for carbon), lead to **strong positive transfer** upon fine-tuning. This technique is particularly useful in the low data regime, leading to data efficiency improvements of **up to 32×** for a given accuracy.

Regression Experiments



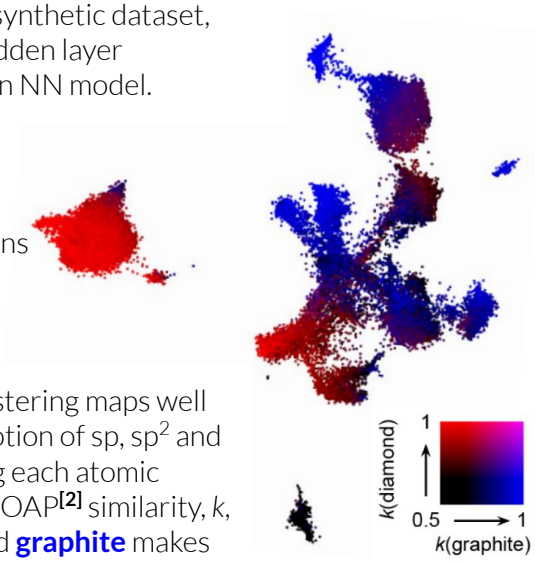
We train Gaussian Process Regression (GPR), Deep Kernel Learning (DKL) and Neural Network (NN) models to predict per-atom local energy labels from SOAP[2] descriptors.

While GPR models excel in the low-data limit due to their strong regularisation, we find that network (NN and DKL) models are, in the limit of large data, the most accurate, with NNs significantly faster than DKL models.

Chemical Maps

We make a "chemical map" of each atomic environment in our synthetic dataset, based on the final hidden layer representations of an NN model.

We find that this learned space is rich with chemical information, and aligns well with human chemical intuition.



In particular, the clustering maps well onto the chemical notion of sp, sp^2 and sp^3 atoms. Colouring each atomic environment by its SOAP[2] similarity, *k*, to both **diamond** and **graphite** makes this very clear!

Such well defined clustering is not present in the original SOAP[2] space, showing that these synthetic labels are allowing the NN to learn a meaningful and compressed representation of chemical structure.

Paper at:
Digital Discovery, 2023, 2, 651 - 662

References:
[1] *Phys. Rev. B 95*, 094203 (2017)
[2] *Phys. Rev. B 87*, 184115 (2013)
[3] *Nat Commun. 13*, 2453 (2022)

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 Research Fund.