Benchmarking Compositional Generalisation for Machine Learning Interatomic Potentials

Amir Masoud Nourollah[1], Irtaza Khalid[1], Stefano Leoni[2], Steven Schockaert[1]

[1]Department of Computer Science, [2]Department of Chemistry, Cardiff University

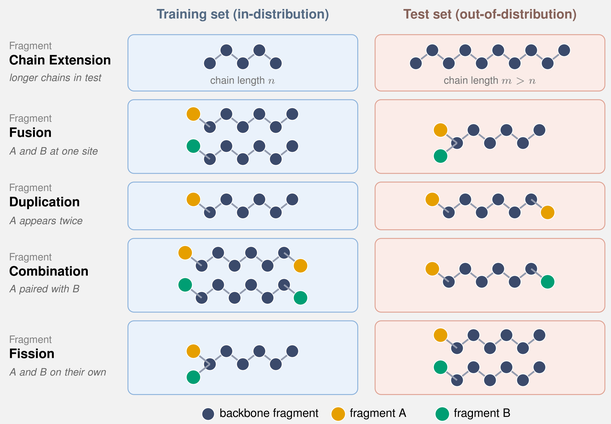
Why compositional generalisation for MLIPs?

Machine-learned interatomic potentials (MLIPs) promise near-quantum accuracy at a fraction of the cost, and are now widely used to simulate molecules.

But **chemical space is combinatorially infinite**: no training set can cover it, so a model cannot simply memorise the potential energy surface. It must **generalise** to molecules it has never seen.

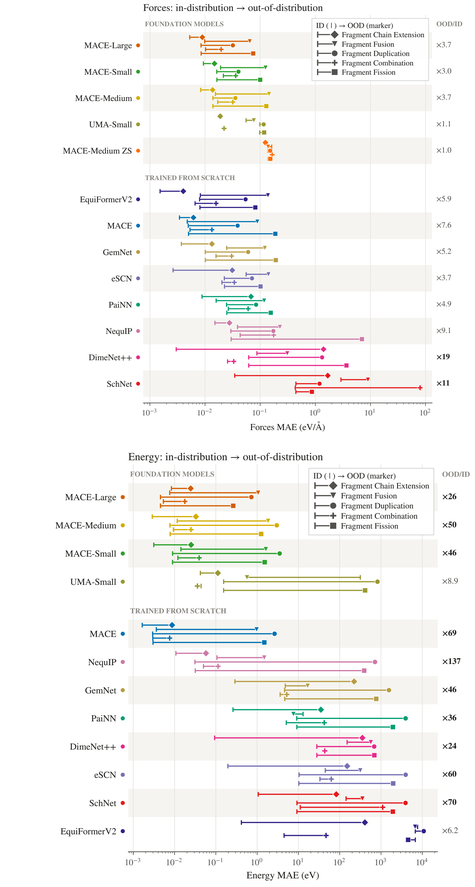
The challenge: If a model has learned every fragment of a molecule, can it predict the molecule itself?

The GMD benchmark



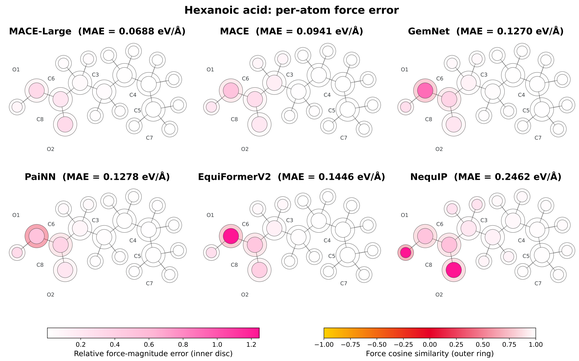
Task	Training set	OOD test set
Chain Extension	C_2–C_6 alkanes	C_7–C_{13} alkanes
Fusion	C_4–C_{10} alcohols and aldehydes; complex polyols and polycarbonyls	C_4–C_{10} carboxylic acids
Duplication	C_5–C_{10} carboxylic acids	C_5–C_{10} dicarboxylic acids
Combination	C_2–C_9 diamines; C_2–C_9 dicarboxylic acids	C_2–C_9 amino acids
Fission	C_5–C_{10} carboxylic acids	C_5–C_{10} alcohols and aldehydes

Performance of all models



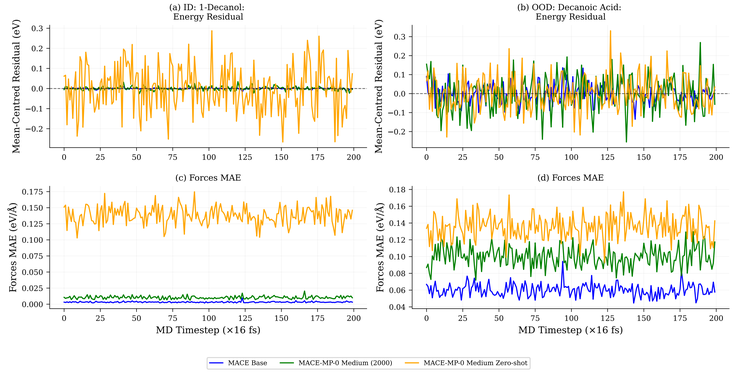
Each bar runs from a model's **ID error** to its **OOD error**: longer bars mean a larger generalisation gap. Fine-tuned foundation models (top group) do not close it.

Failure analysis



- Per-atom force error on hexanoic acid, an OOD molecule of the Fragment Fusion task, for six models chosen as representatives of their architecture families.
- Inner disc: error in force magnitude; outer ring: error in force direction.
- The error appears on the carboxyl group, where the two fragments are fused, while the carbon backbone is almost error-free: the models struggle to generalise to the new fragment.

Errors over time



Take-home

- **Seeing every fragment in training is not enough:** models generalise to longer chains, but on the other four tasks OOD force errors are often **an order of magnitude** above ID, and fine-tuned foundation models do not close the gap.
- **ID accuracy is a poor guide to OOD accuracy.** Test your model on GMD: the dataset, code and fine-tuned weights are all open.

Paper Code Dataset About Me