nature computational science

Supplementary information

https://doi.org/10.1038/s43588-025-00768-y

Harnessing large language models for data-scarce learning of polymer properties

In the format provided by the authors and unedited

Table of Contents

Supplementary Algorithm 1 The key steps of the two-phase training strategy.

LLM encoder pretraining:
Use a large dataset of unlabeled SMILES representations of polymers to pretrain an LLM encoder $\tilde{\mathcal{M}}_{encode}$.
Phase-1 supervised pretraining:
Use physics-based hypothetical polymer generation methods, such as group contribution (GC), to generate a large dataset of physically meaningful synthetic polymer structures $\{\boldsymbol{X}_i\}_{i=1}^{S_{GC}}$ with the correlation of fundamental thermophysical properties;
Build a physics-based model $\mathcal{M}_{physics}$ of the real-world physical process, by leveraging on the fundamental properties calculated from physically meaningful synthetic polymers;
Construct a physically meaningful synthetic dataset $\mathcal{D}_{GC} := \{(\boldsymbol{X}_i, \mathcal{M}_{physics}(\boldsymbol{X}_i))\}_{i=1}^{S_{GC}}$;
Apply supervised pretraining to the LLM decoder/predictor using the synthetic dataset $\mathcal{D}_{GC}$, to obtain an LLM decoder $\tilde{\mathcal{M}}_{decode}$ with physically consistent initial state;
Phase-2 finetuning:
Collect a (usually small) set of high-fidelity measurements from experiments, denoted as $\mathcal{D}_{HF} := \{(\boldsymbol{X}_i^{HF}, \boldsymbol{Y}_i^{HF})\}_{i=1}^{S_{HF}}$;
Split the high-fidelity experimental dataset $\mathcal{D}_{HF}$ as a training set $\mathcal{D}_{HF}^{train}$ and a test set $\mathcal{D}_{HF}^{test}$, and finetune the phase-1 LLM $\tilde{\mathcal{M}}_{decode}$ using $\mathcal{D}_{HF}^{train}$;
Obtain the final physics-guided LLM, and report the prediction accuracy on the test dataset $\mathcal{D}_{HF}^{test}$.