Multi-Fidelity Deep Learning for Data-Efficient Molecular Property Models from Experimental and Computational Data

Kevin P. Greenman, Temujin Orkhon, William H. Green, and Rafael Gómez-Bombarelli

kpg@mit.edu

Introduction

Chemical and materials design frequently involve data of different fidelities to the true property to be optimized, for instance experimental measurements and computational approximations, or simulations at different levels of theory. The different fidelity levels typically involve a cost-accuracy trade-off, where high-fidelity (HF) results are relatively expensive or slow to acquire, while low-fidelity (LF) results are cheaper and faster but include more bias and/or noise in the data. Multi-fidelity methods show some advantages to improve the generalizability of high-fidelity predictions. However, many questions regarding when multi-fidelity methods should be expected to perform well remain unanswered, such as the relationship between dataset needs at both fidelities, and how the accuracy mismatch influences the results.

Models

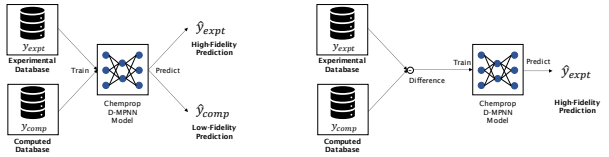
To address these questions, we are creating a comprehensive benchmark of multi-level methods. We begin by comparing:

Multi-target:

Predict both low- and high-fidelity targets at the end of the network using the same embedding

Delta-ML:

Learn *difference* between low- and high-fidelity targets (requires exact overlap of datasets)

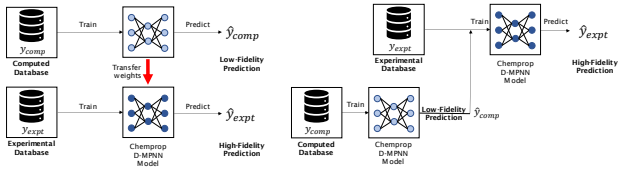


Transfer learning:

Pre-train on low-fidelity data, then transfer weights to train and predict on high-fidelity data

Multi-fidelity:

Auxiliary network learns low-fidelity targets, provides a "guess" to help with high-fidelity prediction

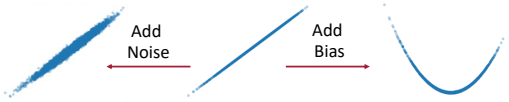


Methods

Bias and Noise

We systematically add noise and bias to a large synthetic dataset of *enthalpies of formation*:

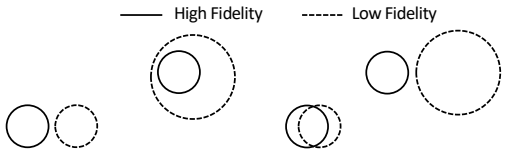
- **Bias as a function of property:** constant and polynomial bias.
- **Bias as a function of molecular features:** polynomial bias as a function of a set of molecular descriptors from RDKit.
- **Random Gaussian noise:** with several levels of variance.



Overlap, Size Ratios, and Splitting

We split high- and low-fidelity data in ways that mimic realistic use cases:

- **Overlap:** low-fidelity set of molecules overlapping with high-fidelity set versus not overlapping.
- **Size Ratios:** Different ratios of low- to high-fidelity dataset sizes.



- **Splitting:** random, scaffold, by property, by molecular weight, and by atom types.

Non-Synthetic, Realistic Datasets

- **Optical properties:** from experiments and time-dependent density functional theory calculations
- **Solubility:** from experiments and COSMO-RS calculations
- **Drug efficacy vs. potency:** from single point and concentration response curves

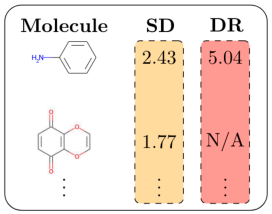
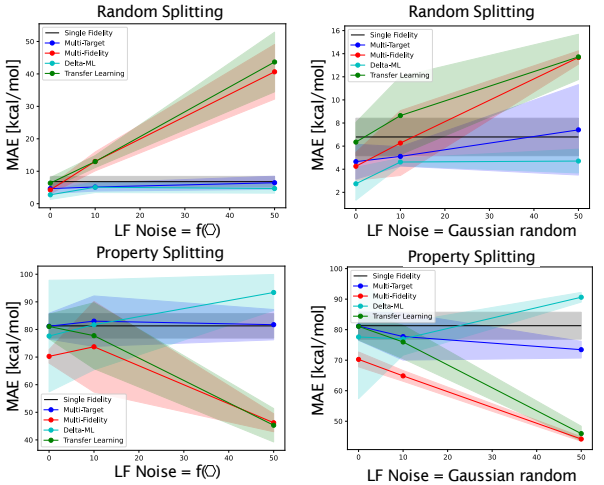


Figure: Buterez et al. *chemRxiv*. (2022).

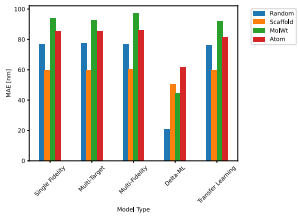
Results

Synthetic Dataset

- Multi-target and delta-ML models are more robust to added noise and bias than multi-fidelity and transfer learning models (for random splits).
- Noise as a function of molecular structure is more challenging to learn from than random noise.
- For more challenging split types, multi-fidelity and transfer learning high-fidelity performance improves with noisier low-fidelity data.



Optical Properties



- Delta-ML performance is best across all split types
- Multi-fidelity and transfer learning perform similarly to multi-target and single-fidelity
- Suggests this task is in a low-noise / low-bias regime

Acknowledgements

K. P. G. was supported by the National Science Foundation Graduate Research Fellowship Program under Grant No. 1745302. This work was also supported by the DARPA Accelerated Molecular Discovery (AMD) program under contract HR00111920025. We acknowledge the MIT Engaging cluster and MIT Lincoln Laboratory Supercloud cluster at the Massachusetts Green High Performance Computing Center (MGHPCC) for providing high-performance computing resources to train our deep learning models.