Transfer Learning for Molecular Activity Prediction

Xinhao Li & Denis Fourches*

Department of Chemistry, Bioinformatics Research Center, North Carolina State University, Raleigh, NC 27695, United States

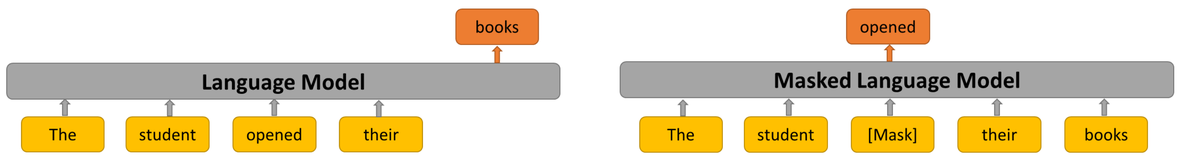
Background

Quantitative Structure–Activity Relationships (QSAR) are statistical, data-driven models that establish quantitative links between an experimental activity (*e.g.*, binding affinity, inhibition potency) and chemical structures. QSAR models are typically developed using supervised machine learning algorithms and further validated using a variety of statistical procedures and metrics. Good model performance usually requires a decent amount of labeled data, but collecting labels is expensive and hard to be scaled up. Thus, it would be highly relevant to utilize the tremendous unlabeled compounds from publicly-available datasets. Self-supervised learning opens up a huge opportunity for better utilizing unlabeled data. In this study, we propose the **Mol**ecular **P**rediction **Mo**del **Fi**ne-**T**uning (**MolPMoFiT**) approach, an effective transfer learning method based on **self-supervised pre-training + task-specific fine-tuning** for QSAR modeling.

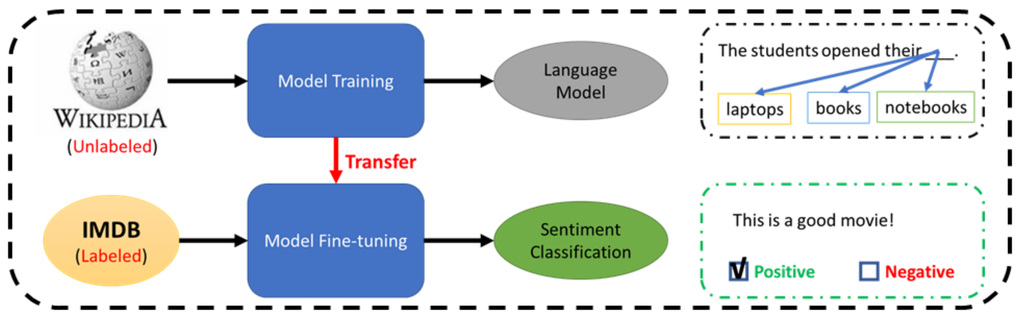
Self-Supervised Learning and Transfer Learning

Self-Supervised Learning: Filling in the Bl_nk.

The self-supervised learning creates labels for unlabeled data and trains unsupervised dataset in a supervised manner. It achieves this by framing a supervised learning task in a special form to predict only a subset of information using the rest.



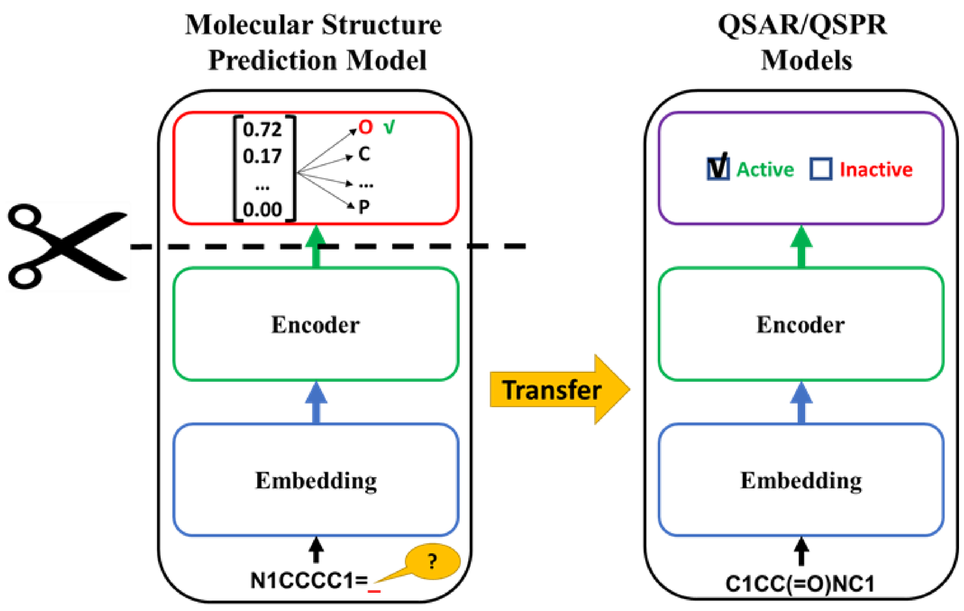
Transfer learning is a machine learning method where a model developed for a task is reused as the starting point for a model on a second task.



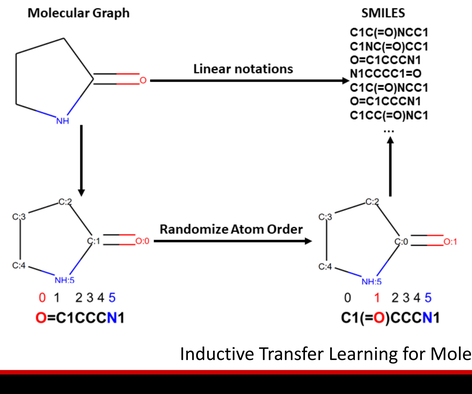
Universal language model fine-tuning for text classification. *ACL*, **2018**, 328–339.

Molecular Prediction Model Fine-Tuning (MolPMoFiT)

MolPMoFiT is a transfer learning method that can be applied to any QSAR problems.



A molecular structure prediction model (MSPM) is pre-trained on one million bioactive molecules from ChEMBL in a self-supervised manner, and then fine-tuned on various QSPR/QSAR tasks. The QSPR/QSAR models are initialized using the embedding layer and encoder from the pre-train MSPM.



- **SMILES** (tokenized at atom level) are used as input for both the MSPM and QSPR/QSAR models.
- A single molecular structure can be represented by multiple SMILES. Randomized SMILES can be generated by randomizing the atom ordering in the molecular graph.
- The **SMILES augmentation** technique was applied to train both the MSPM and QSAR/QSPR models.
- **Test Time SMILES Augmentation** (TTA) was applied for further improving model performances.

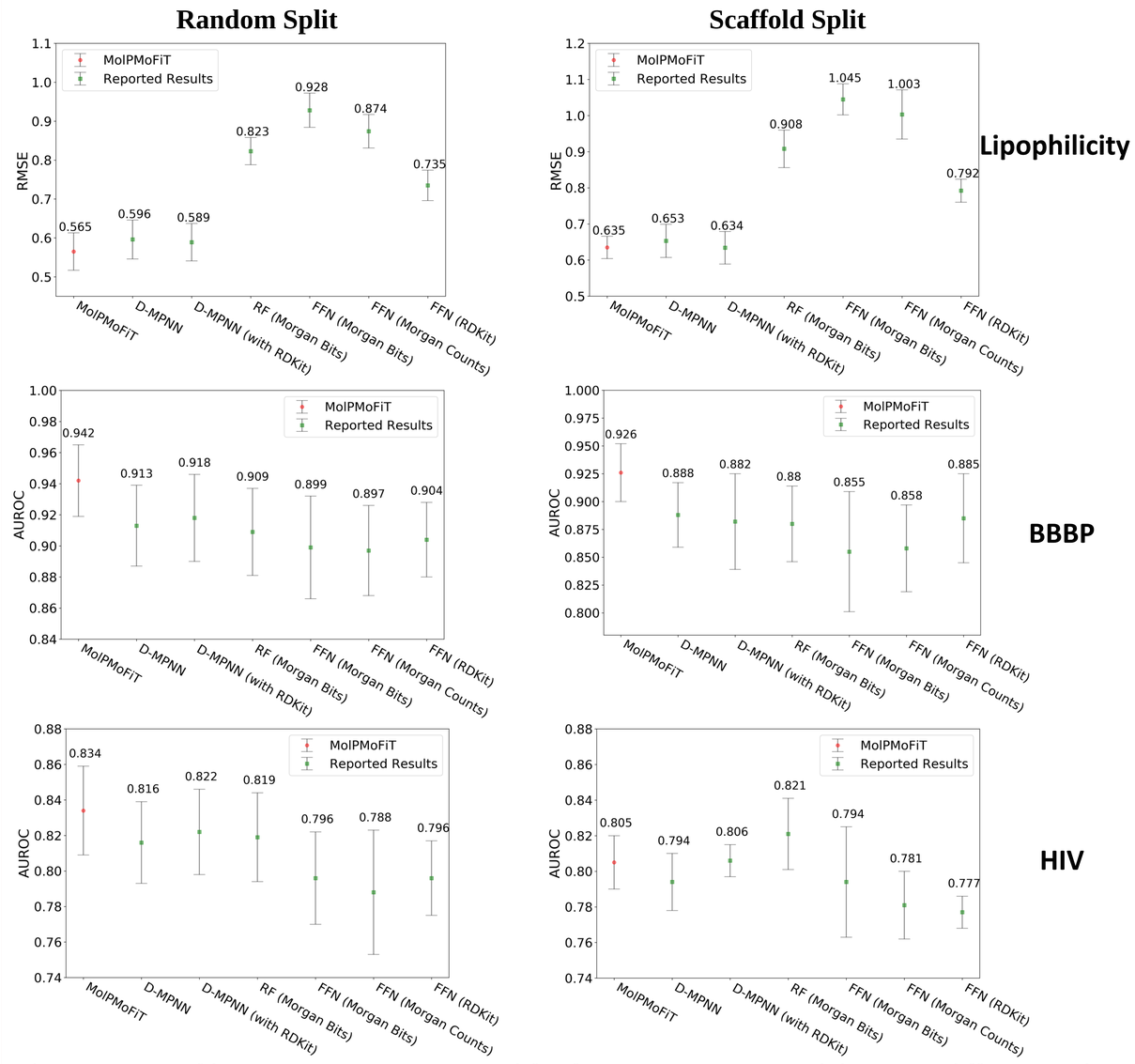
Inductive Transfer Learning for Molecular Activity Prediction: Next-Gen QSAR Models with MolPMoFiT. *ChemRxiv*, **2019**.

Benchmark Data Sets

Data Set	Description	Size	# of Active Compound	Task
Lipophilicity	Octanol/water distribution coefficient	4,200		*Regression*
HIV	Inhibition of HIV replication	41,127	1,443	*Classification*
BBBP	Ability to penetrate the blood-brain barrier	2,039	1,560	*Classification*

MoleculeNet: A Benchmark for Molecular Machine Learning. *arXiv*, **2017**.

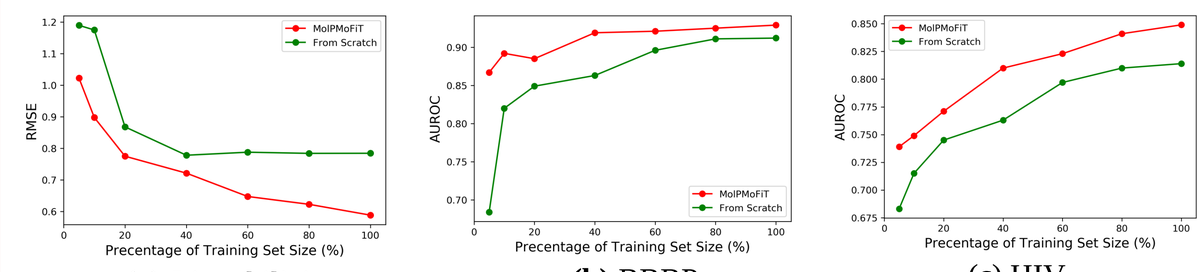
Model Comparison



The same set of hyperparameters was used for fine-tuning QSAR models across different tasks. The MolPMoFiT can provide strong baselines *out-of-box*. Compared Models are from Yang et al*.

*Analyzing Learned Molecular Representations for Property Prediction. *J. Chem. Inf. Model.*, **2019**. *59*, 3370-3388.

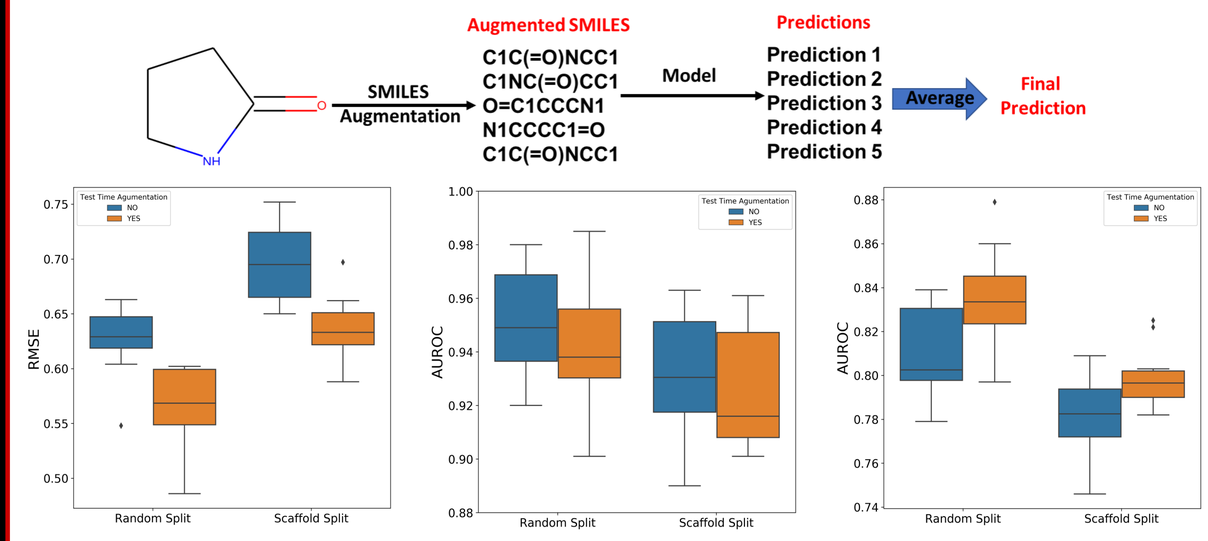
Impact of Transfer Learning



(a) Lipophilicity **(b)** BBBP **(c)** HIV

MolPMoFiT models (Red lines) always outperforms the models trained from scratch (Green lines). The models were trained on different number of training data and tested on the same test set on a single 80:10:10 random split. The hyperparameters and training procedures were kept fixed.

Impact of Test Time Augmentation (TTA)



Comparison of model performances on predictions from Canonical SMILES and Test time augmentation. All data sets were evaluated on ten 80:10:10 random/scaffold splits.

Conclusions

- **MolPMoFiT** is a novel and effective transfer learning method for QSPR/QSAR tasks
- A **molecular structure prediction model** was trained using one million bioactive molecules from ChEMBL and fine-tuned for three (***Lipophilicity, HIV and BBBP***) QSAR tasks. **MolPMoFiT** showed strong performance compared to the current *state-of-the-art* techniques.
- Transfer learning techniques such as MolPMoFiT could significantly contribute in boosting the reliability of next-generation QSAR models.

Acknowledgement DARPA and Chancellor's Faculty Excellence Program