Analysis of Atom-level pretraining with Quantum Mechanics data for GNN Molecular property models

Jose A. Arjona-Medina & Ramil Nugmanov

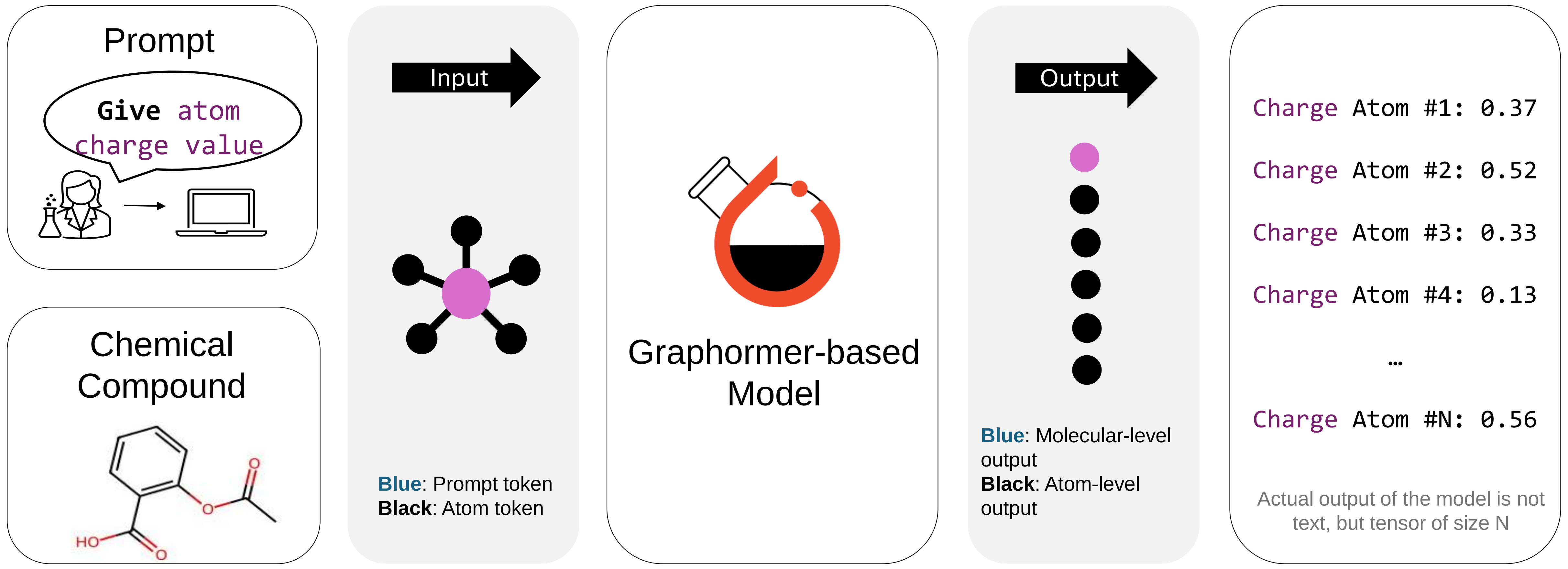
jarjonam@its.jnj.com , rnugmano@its.jnj.com

Abstract

The challenge of learning robust molecular representations that effectively generalize in real-world scenarios to novel compounds remains an elusive and unresolved task.

This work examines how atom-level pretraining with quantum mechanics (QM) data can mitigate violations of assumptions regarding the distributional similarity between training and test data and therefore improve performance and generalization in downstream tasks.

Atom-level Pretraining


Prompt

Give atom charge value

Chemical Compound

Input

Blue: Prompt token
Black: Atom token

Graphormer-based Model

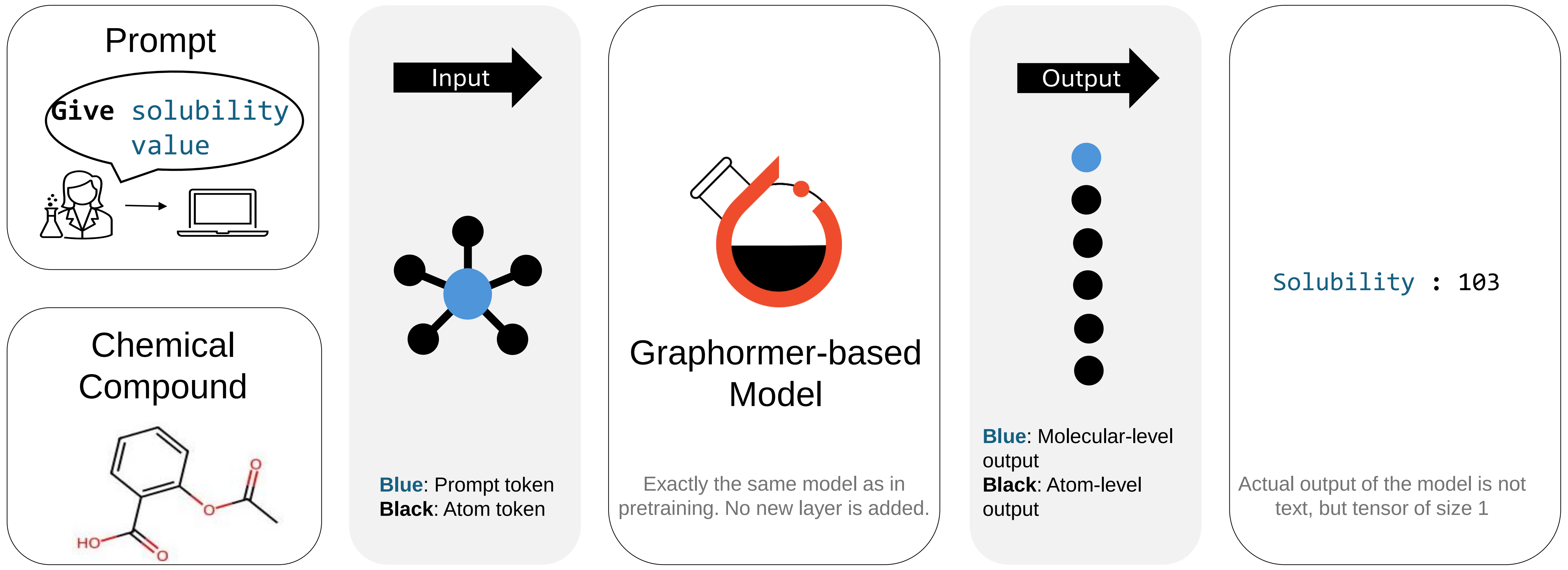
Output

Blue: Molecular-level output
Black: Atom-level output

```
Charge Atom #1: 0.37

Charge Atom #2: 0.52

Charge Atom #3: 0.33

Charge Atom #4: 0.13

…

Charge Atom #N: 0.56
```

Actual output of the model is not text, but tensor of size N

Molecular-level Finetuning


Prompt

Give solubility value

Chemical Compound

Input

Blue: Prompt token
Black: Atom token

Graphormer-based Model

Exactly the same model as in pretraining. No new layer is added.

Output

Blue: Molecular-level output
Black: Atom-level output

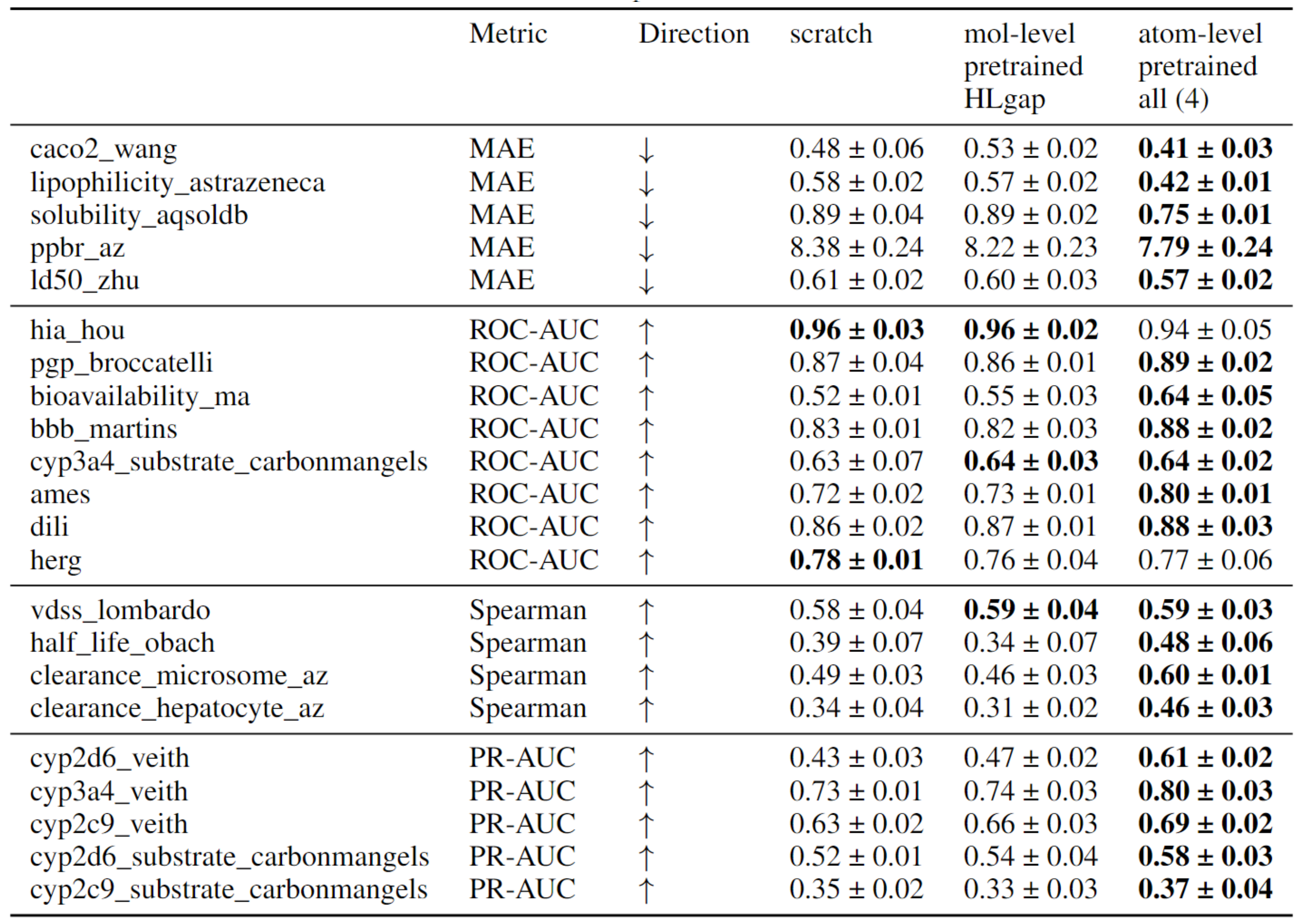
```
Solubility : 103
```

Actual output of the model is not text, but tensor of size 1

Results

Table 1

Performance on the TDC benchmark
(https://tdcommons.ai/)

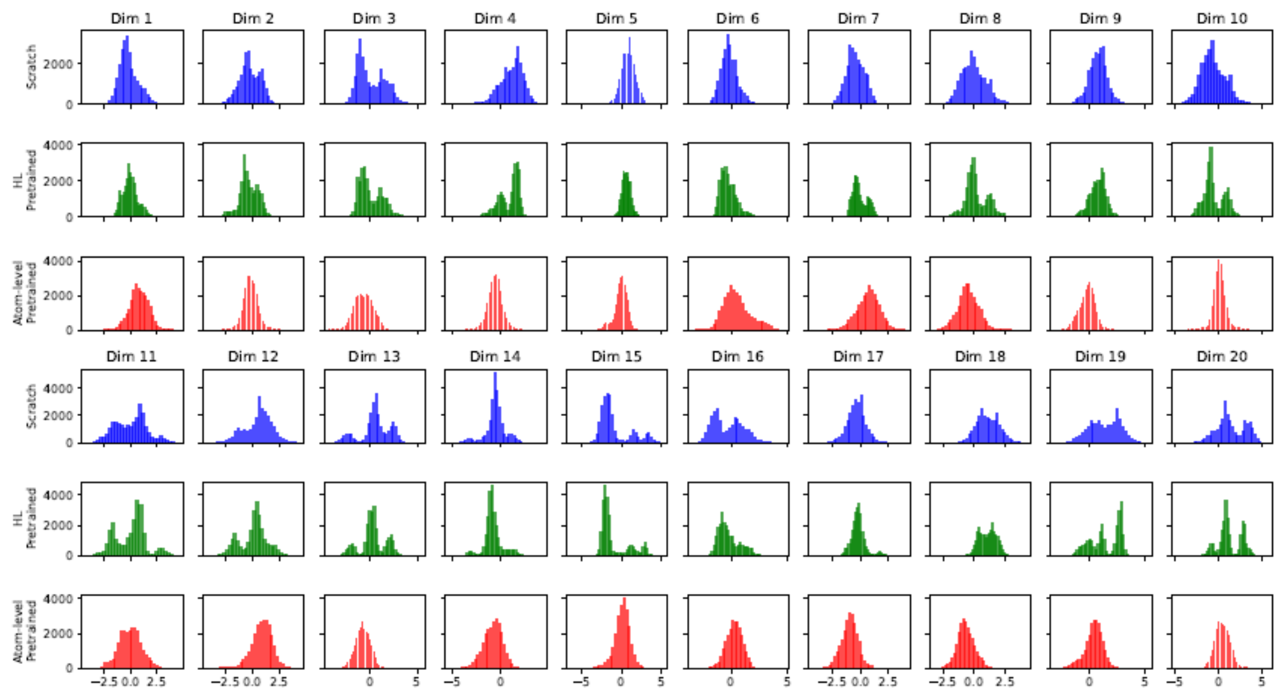


	Metric	Direction	scratch	mol-level pretrained HLgap	atom-level pretrained all (4)
caco2_wang	MAE	↓	0.48 ± 0.06	0.53 ± 0.02	**0.41 ± 0.03**
lipophilicity_astrazeneca	MAE	↓	0.58 ± 0.02	0.57 ± 0.02	**0.42 ± 0.01**
solubility_aqsoldb	MAE	↓	0.89 ± 0.04	0.89 ± 0.02	**0.75 ± 0.01**
ppbr_az	MAE	↓	8.38 ± 0.24	8.22 ± 0.23	**7.79 ± 0.24**
ld50_zhu	MAE	↓	0.61 ± 0.02	0.60 ± 0.03	**0.57 ± 0.02**
hia_hou	ROC-AUC	↑	**0.96 ± 0.03**	**0.96 ± 0.02**	0.94 ± 0.05
pgp_broccatelli	ROC-AUC	↑	0.87 ± 0.04	0.86 ± 0.01	**0.89 ± 0.02**
bioavailability_ma	ROC-AUC	↑	0.52 ± 0.01	0.55 ± 0.03	**0.64 ± 0.05**
bbb_martins	ROC-AUC	↑	0.83 ± 0.01	0.82 ± 0.03	**0.88 ± 0.02**
cyp3a4_substrate_carbonmangels	ROC-AUC	↑	0.63 ± 0.07	**0.64 ± 0.03**	**0.64 ± 0.02**
ames	ROC-AUC	↑	0.72 ± 0.02	0.73 ± 0.01	**0.80 ± 0.01**
dili	ROC-AUC	↑	0.86 ± 0.02	0.87 ± 0.01	**0.88 ± 0.03**
herg	ROC-AUC	↑	**0.78 ± 0.01**	0.76 ± 0.04	0.77 ± 0.06
vdss_lombardo	Spearman	↑	0.58 ± 0.04	**0.59 ± 0.04**	**0.59 ± 0.03**
half_life_obach	Spearman	↑	0.39 ± 0.07	0.34 ± 0.07	**0.48 ± 0.06**
clearance_microsome_az	Spearman	↑	0.49 ± 0.03	0.46 ± 0.03	**0.60 ± 0.01**
clearance_hepatocyte_az	Spearman	↑	0.34 ± 0.04	0.31 ± 0.02	**0.46 ± 0.03**
cyp2d6_veith	PR-AUC	↑	0.43 ± 0.03	0.47 ± 0.02	**0.61 ± 0.02**
cyp3a4_veith	PR-AUC	↑	0.73 ± 0.01	0.74 ± 0.03	**0.80 ± 0.03**
cyp2c9_veith	PR-AUC	↑	0.63 ± 0.02	0.66 ± 0.03	**0.69 ± 0.02**
cyp2d6_substrate_carbonmangels	PR-AUC	↑	0.52 ± 0.01	0.54 ± 0.04	**0.58 ± 0.03**
cyp2c9_substrate_carbonmangels	PR-AUC	↑	0.35 ± 0.02	0.33 ± 0.03	**0.37 ± 0.04**

In **Table 1**, we present the outcomes from benchmarking three distinct training approaches:scratch, molecule-level QM pretrained, and atom-level QM pretrainedwith all properties for 5 different seeds, as described in the guidelines provided by the TDC dataset. We have excluded the results for atom-level pretraining on individual QM properties from this table. These results show that atom-level pretraining notably enhances model performance compared to training from scratch for 21 of the 22 datasets.

Figure 1

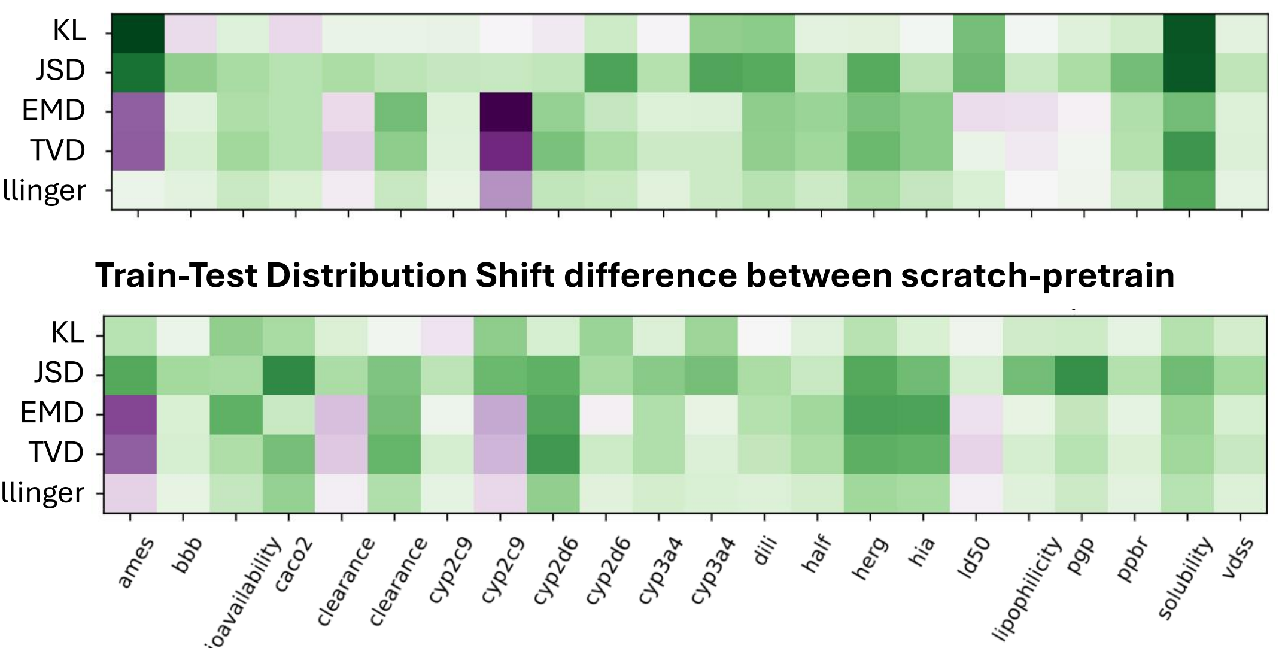
Distribution of Activations for first 20 features
(out of 256, first model layer)



Distribution of first 20 features from the first layer of the Graphormer network for three different training approaches —scratch, HOMO-LUMO pretrained and atom-level pretrained— across test split of lipophilicity dataset.

Figure 2

Train-Val Distribution Shift difference between scratch-pretrain



Train-Test Distribution Shift difference between scratch-pretrain

Conclusions

In this study, we have demonstrated that pretraining of graph-based neural networks with atom-level quantum mechanics (QM) data significantly enhances performance on downstream tasks related to ADMET properties within the TDC dataset, as illustrated in **Table 1**.

- We showed the change in the distributions of activations of the internal model's features due to specific pretraining. After atom-level pretraining with QM data, these distributions become more Gaussian-like, which is known to be conducive to better learning dynamics and thus improved performance (**Figure 1**).
- Moreover, our findings indicate that pretrained models exhibit smaller distribution shifts from training to testing datasets, further supporting the efficacy of QM data pretraining in enhancing model robustness (**Figure 2**).

To our knowledge, this is the first study that elucidates how atom-level pretraining can optimize molecular representations by analyzing the model's internal representation and robustness to distribution shifts.