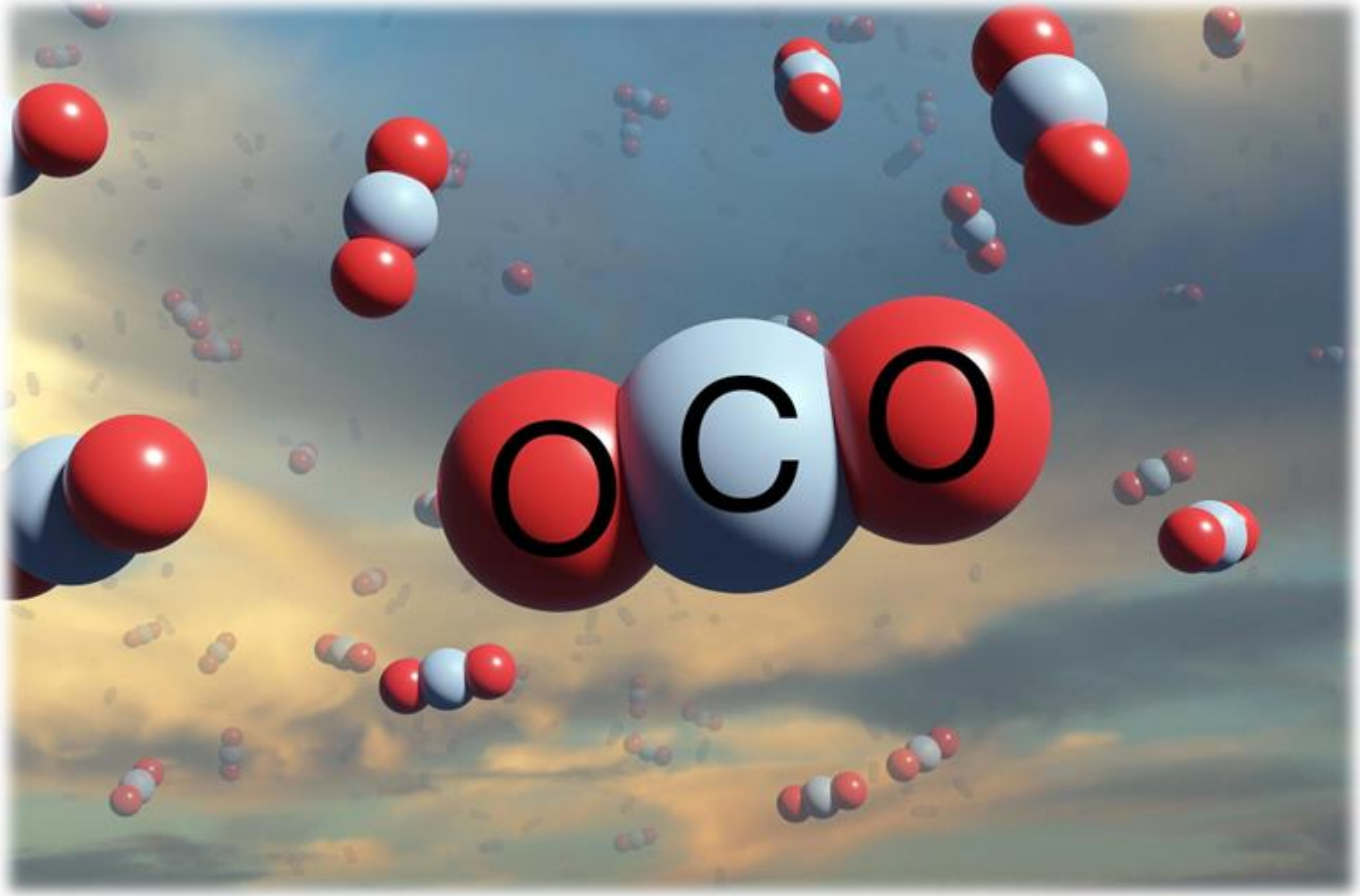
Predicting CO2 loadings

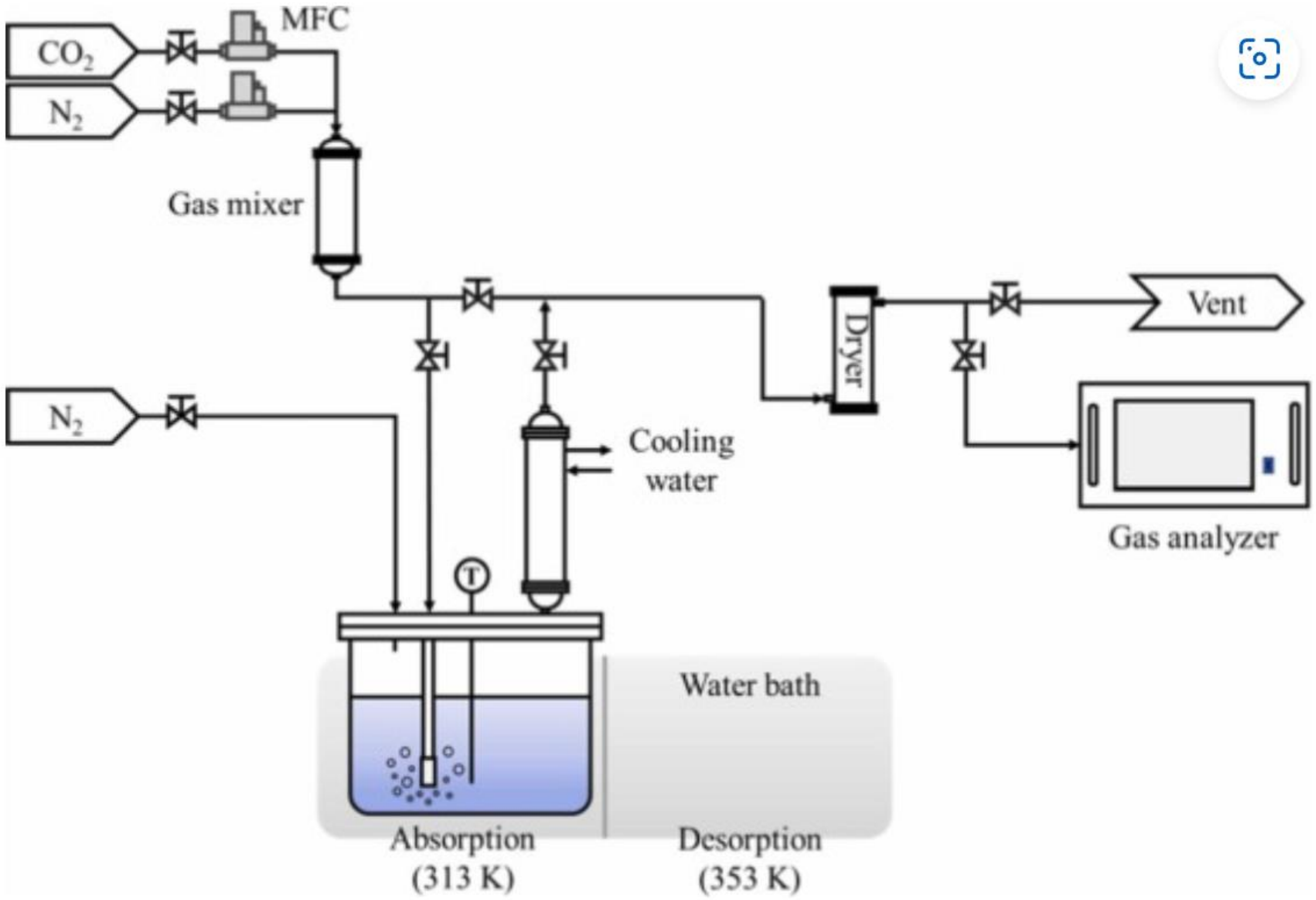


Team Vitamines!!

Marcos Cirne
Yihang fang
Youngku Lee
Tey Kim

Why CO2 loading?

- Separation of carbon dioxide from gas streams with respect to CO_2 negative environmental effects is one of the most significant parts of gas separation processes



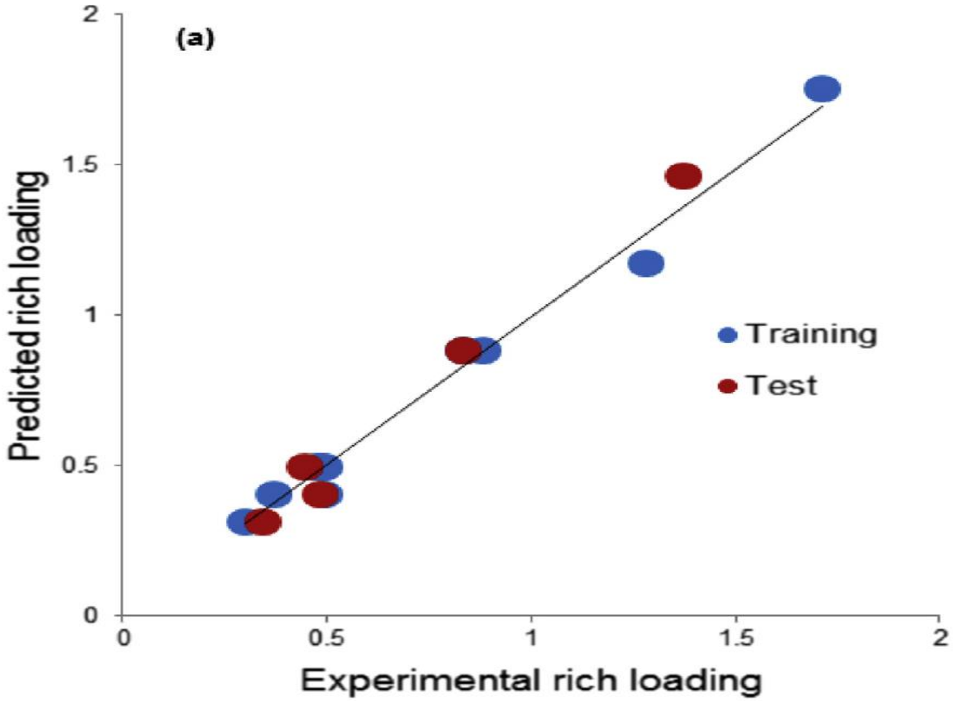
Previous Work

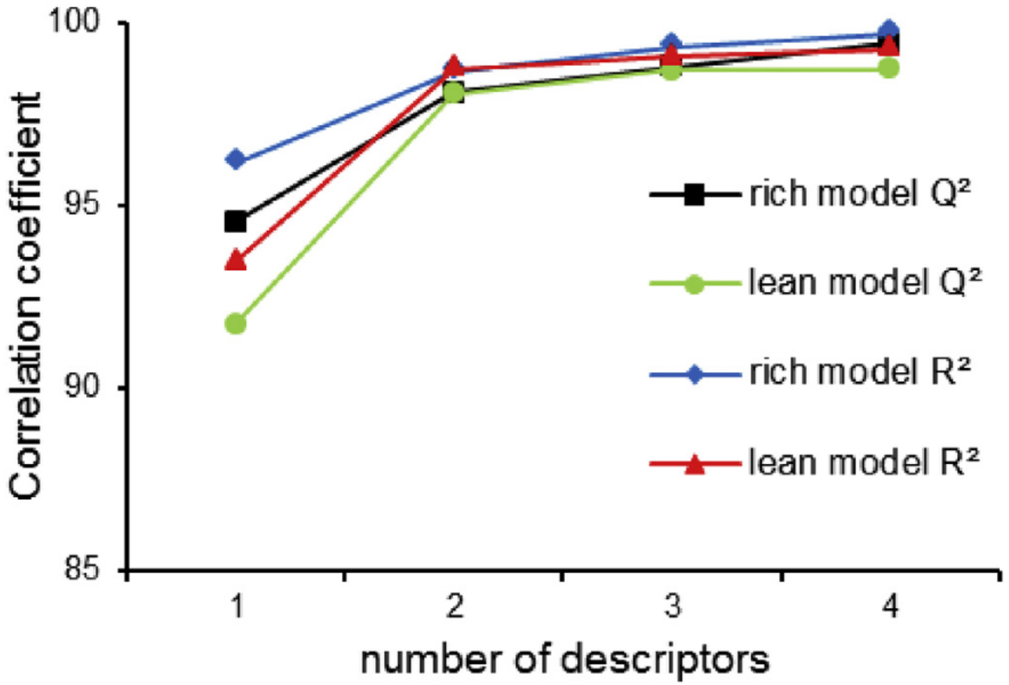
- pK_a has high linear correlation with carbon dioxide solubility in amine.

- CO_2 capture performances are governed by the molecular structure of amines

- Alkyl groups act as electron donors, increasing CO_2 loading, pK_a, *and cyclic capacity.*

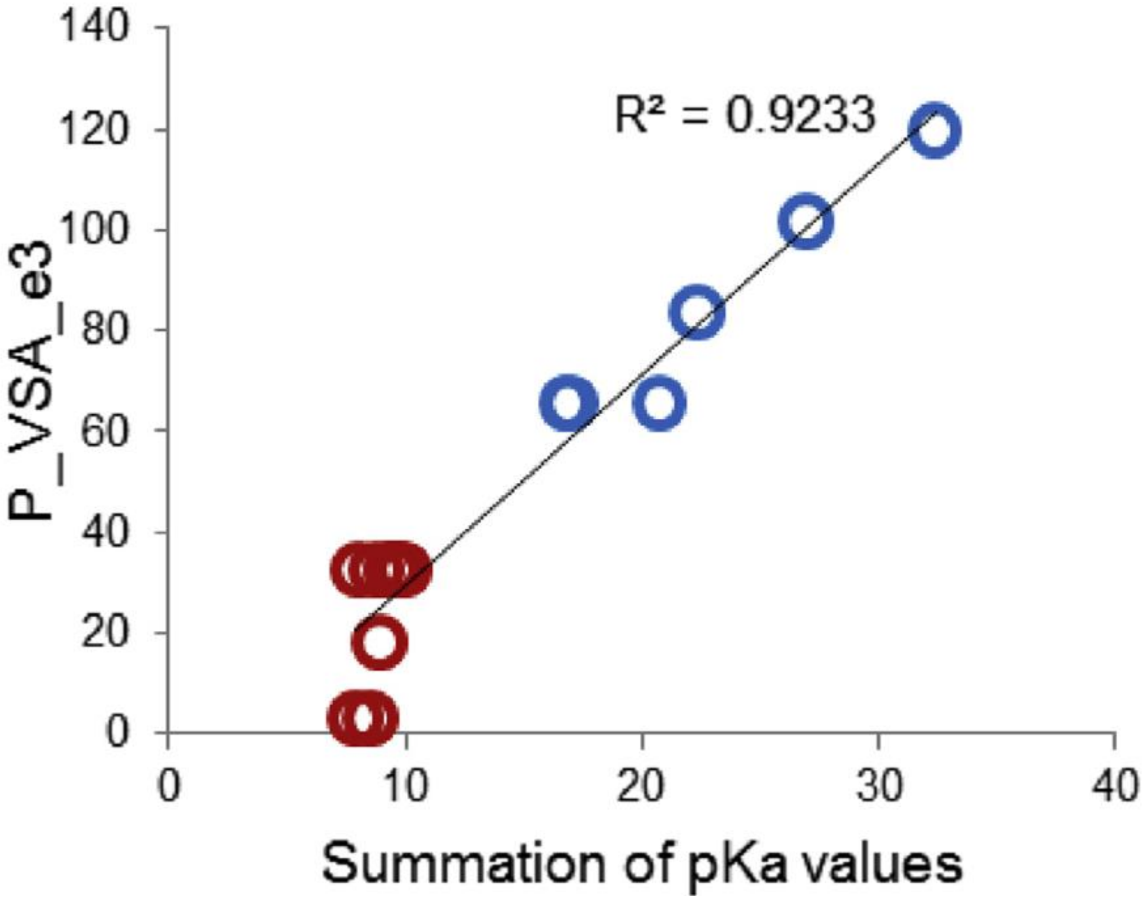
- Hydroxyl groups negatively affect the CO_2 loading, pK_a, *and cyclic capacity.*





Chemical Properties !

- ➢ pKa values has high linear correlation with the CO2 solubility in amines



🟥 Amines with only one nitrogen

🟦 Amines with more nitrogen

↓

Acidic group count & basic group count

Important Features

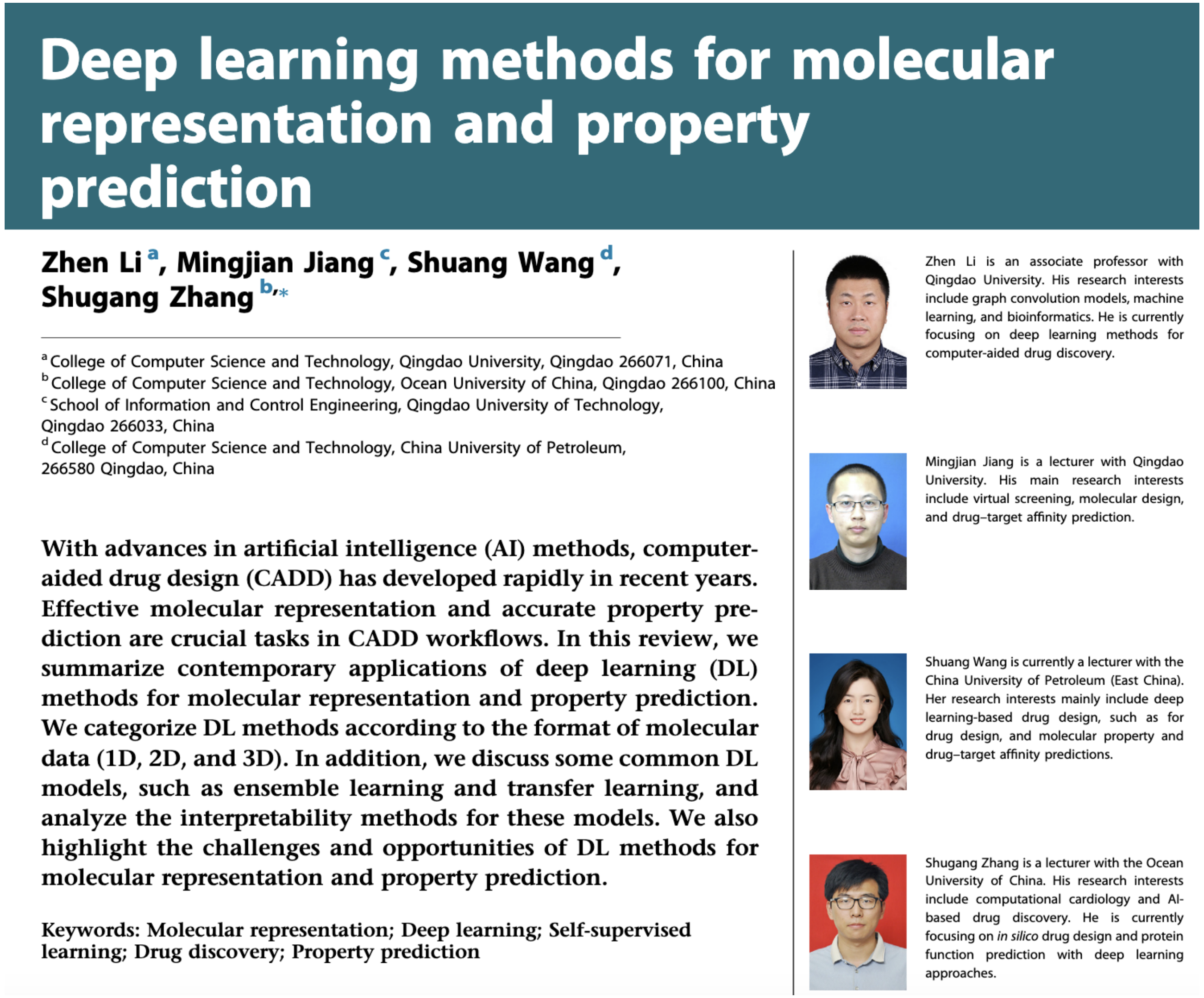
❑ OH count: Molecules with higher OH_count (e.g. polyhydroxy amines) are expected to show lower CO_2 capacities per amine, because the –OH groups withdraw electrons and can form intramolecular H-bonds that make the amine less reactive (33†L43-L51)

❑Alkyl chain count and length: These features reflect how many alkyl groups and how large the hydrocarbon backbone is. More alkyl chains (alkyl_chain_count) and longer chains (longest_alkyl_chain_length) generally indicate a more hydrophobic, electron-rich environment around the amine

❑N_substituent_count and max_N_substituent_length

❑Electron-donating environment (partial charge on N)

Literature Review



Deep learning methods for molecular representation and property prediction

Zhen Li [a], Mingjian Jiang [c], Shuang Wang [d], Shugang Zhang [b,*]

[a] College of Computer Science and Technology, Qingdao University, Qingdao 266071, China
[b] College of Computer Science and Technology, Ocean University of China, Qingdao 266100, China
[c] School of Information and Control Engineering, Qingdao University of Technology, Qingdao 266033, China
[d] College of Computer Science and Technology, China University of Petroleum, 266580 Qingdao, China

With advances in artificial intelligence (AI) methods, computer-aided drug design (CADD) has developed rapidly in recent years. Effective molecular representation and accurate property prediction are crucial tasks in CADD workflows. In this review, we summarize contemporary applications of deep learning (DL) methods for molecular representation and property prediction. We categorize DL methods according to the format of molecular data (1D, 2D, and 3D). In addition, we discuss some common DL models, such as ensemble learning and transfer learning, and analyze the interpretability methods for these models. We also highlight the challenges and opportunities of DL methods for molecular representation and property prediction.

Keywords: Molecular representation; Deep learning; Self-supervised learning; Drug discovery; Property prediction

Zhen Li is an associate professor with Qingdao University. His research interests include graph convolution models, machine learning, and bioinformatics. He is currently focusing on deep learning methods for computer-aided drug discovery.

Mingjian Jiang is a lecturer with Qingdao University. His main research interests include virtual screening, molecular design, and drug–target affinity prediction.

Shuang Wang is currently a lecturer with the China University of Petroleum (East China). Her research interests mainly include deep learning-based drug design, such as for drug design, and molecular property and drug–target affinity predictions.

Shugang Zhang is a lecturer with the Ocean University of China. His research interests include computational cardiology and AI-based drug discovery. He is currently focusing on *in silico* drug design and protein function prediction with deep learning approaches.

Approaches Considered **(1-D data)**:

1. Graph Neural Networks

2. Bidirectional LSTMs

3. Autoencoders

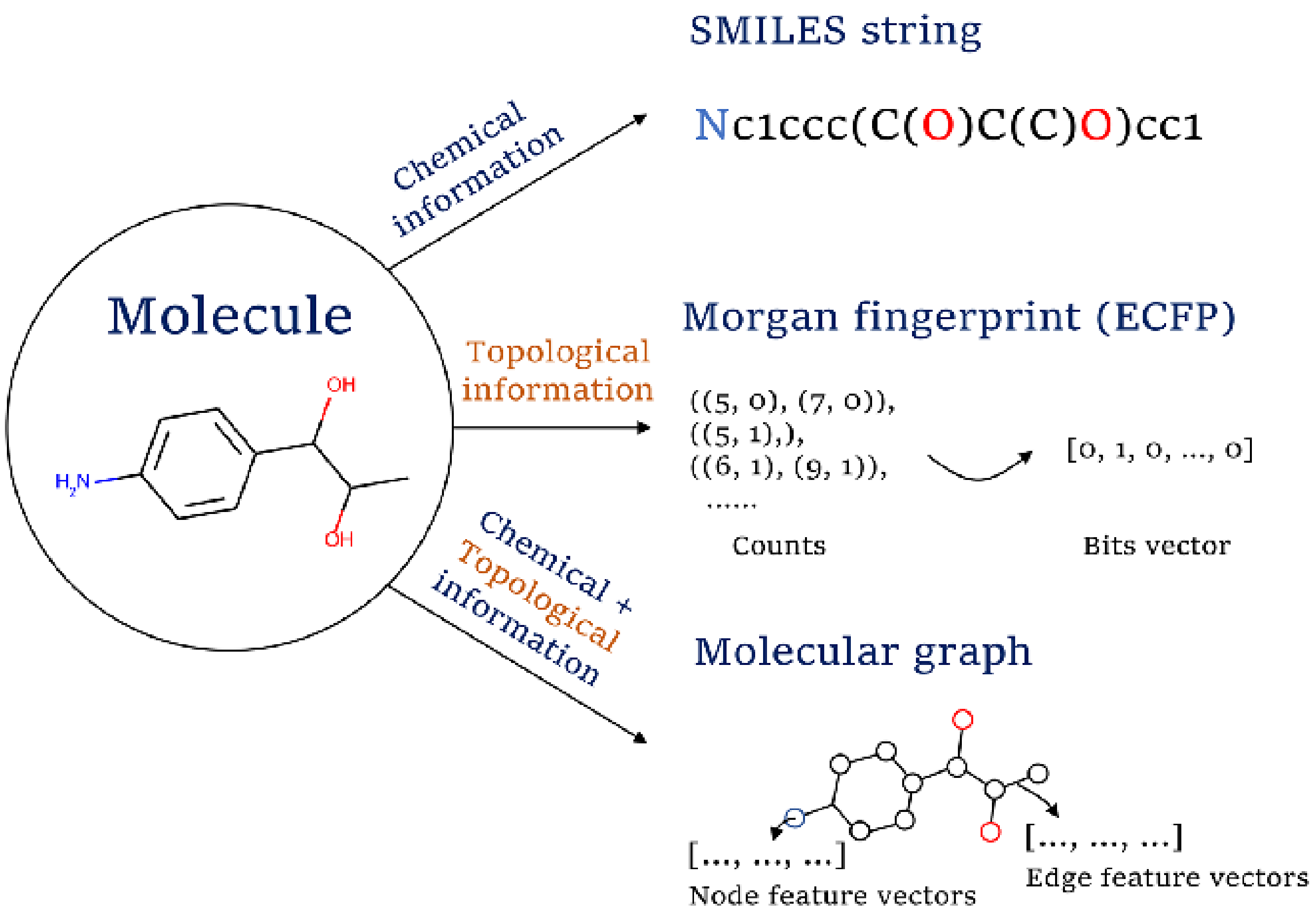
Feature Selection and Feature Engineering

Extra information included in the original set of features: **number of acids** and **number of bases**

Removal of potentially irrelevant features: **number of nitrogens** and **absorption capacity classes**

Principal Component Analysis (PCA): selects the **most meaningful components** that encompass the original information from the dataset

SMILES to Fingerprints



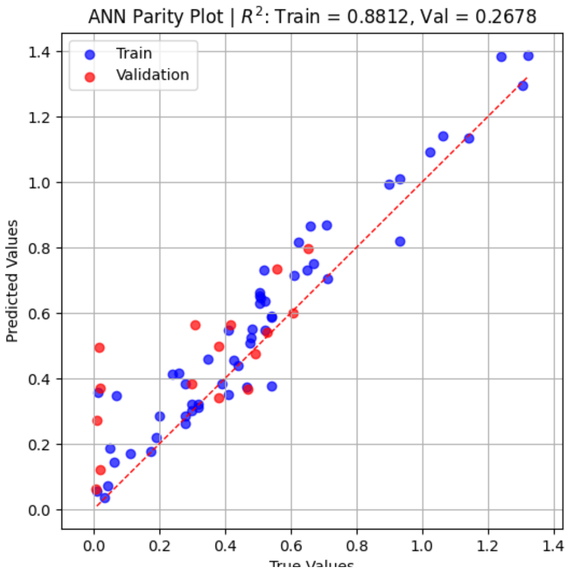
Source: https://www.researchgate.net/figure/Molecular-representations-SMILES-string-Morgan-fingerprint-Extendedconnectivity_fig1_369507722

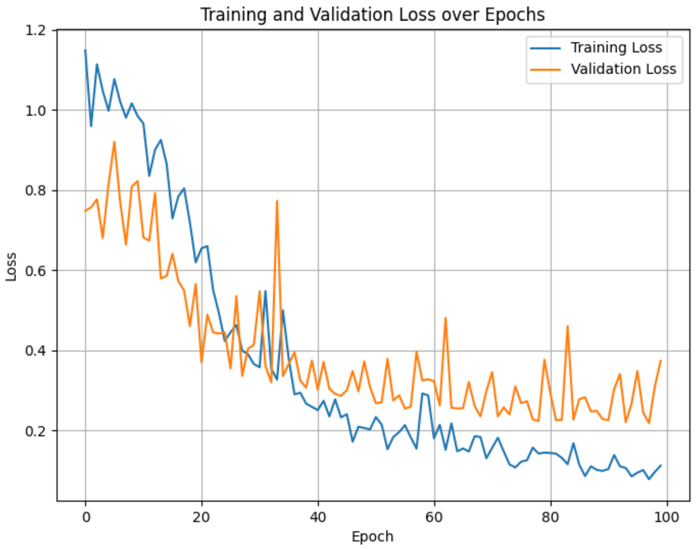
Shifts from variable-size to fixed-size representation (1024 bits)

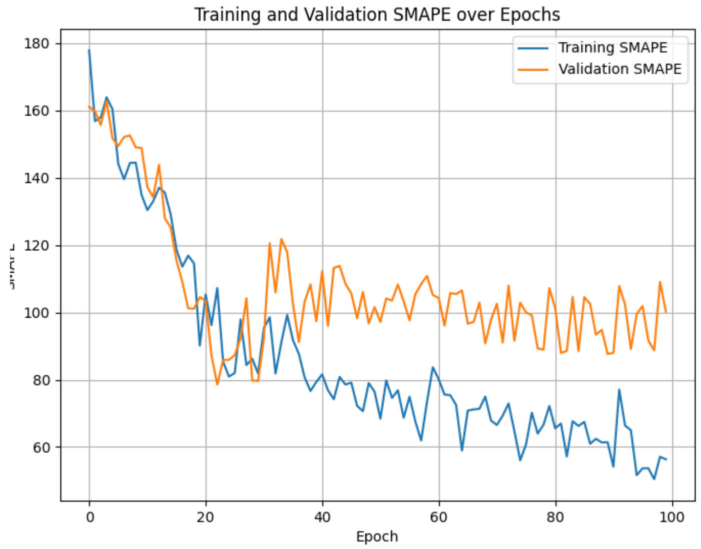
Conveys higher topological information about the molecule structures

Results (Simple ANN)

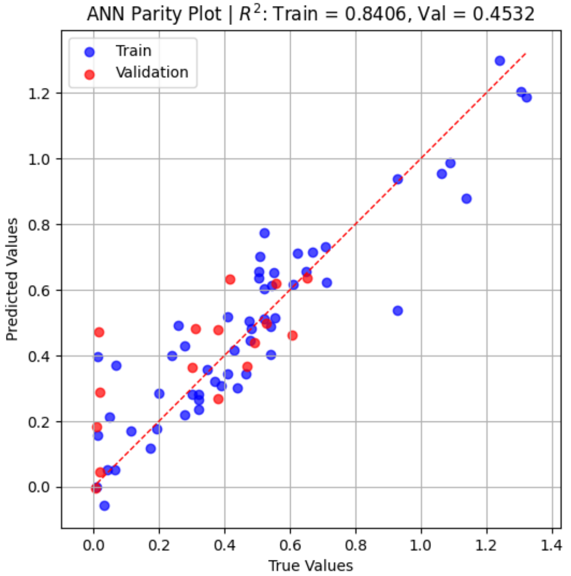
Baseline (Original Data)

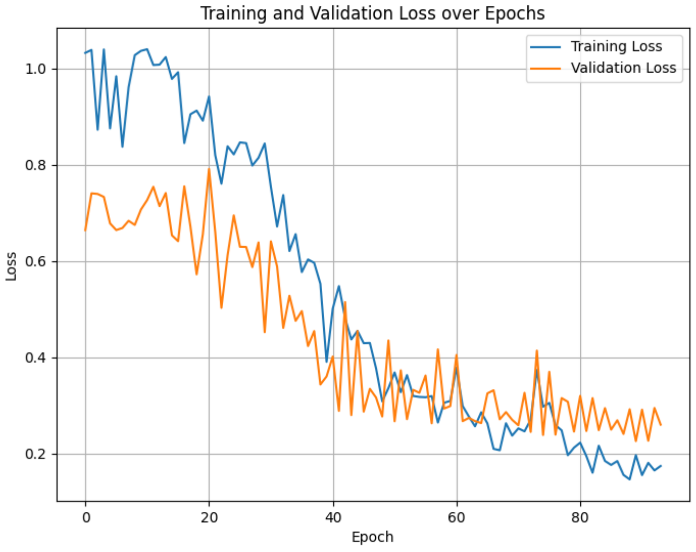


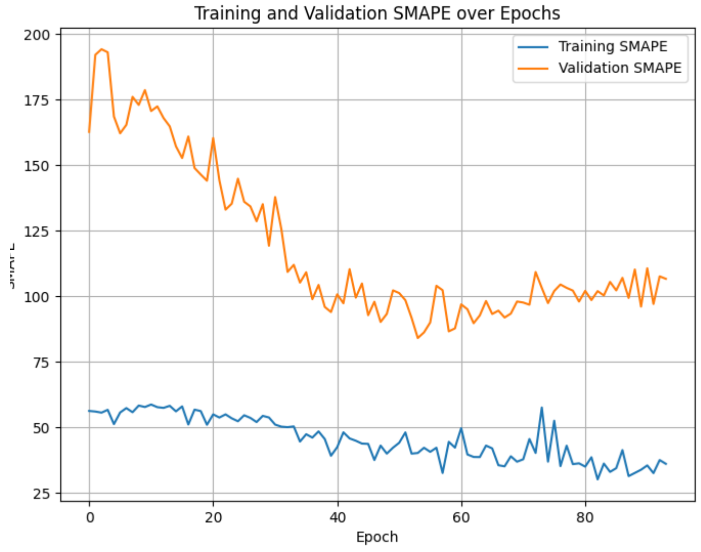




SMILES + Extra Features

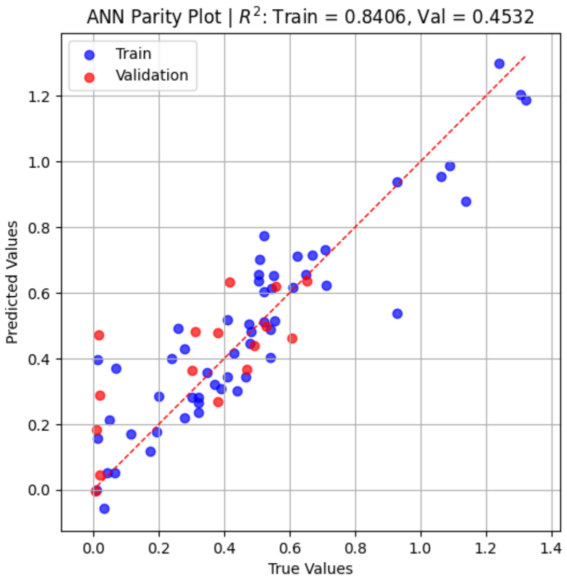


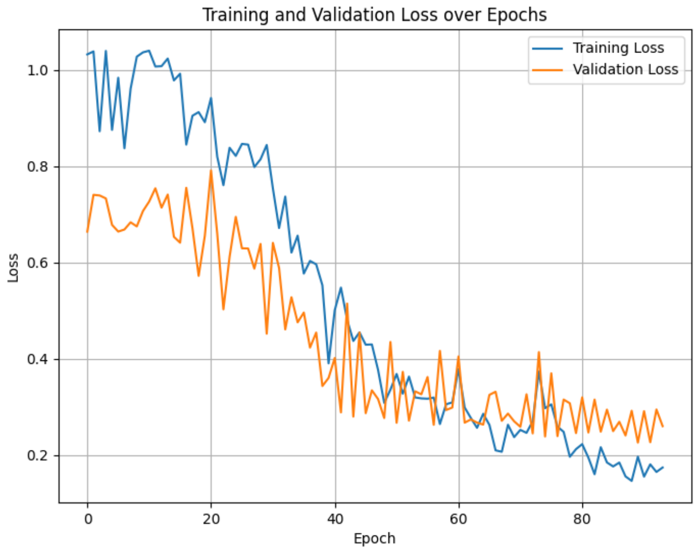


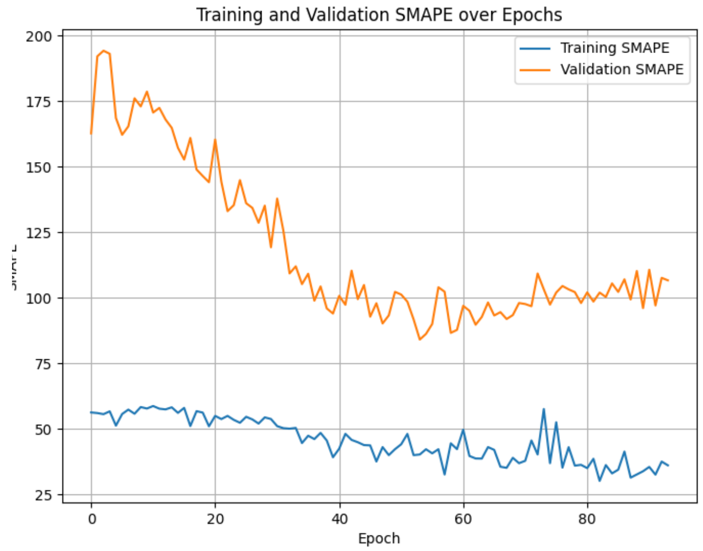


Results (Simple ANN)

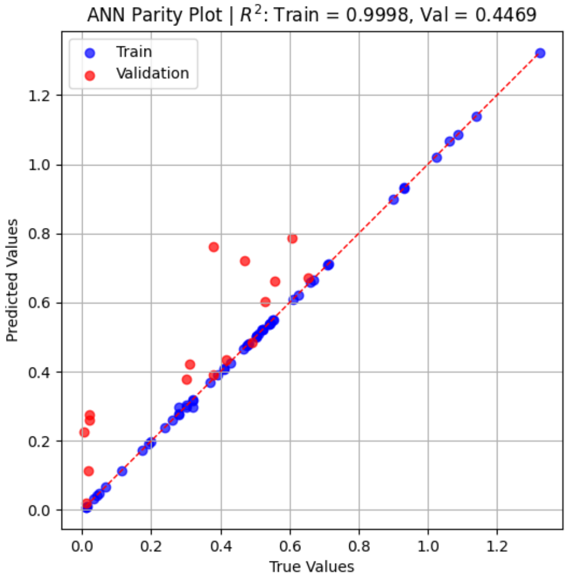
SMILES + Extra Features

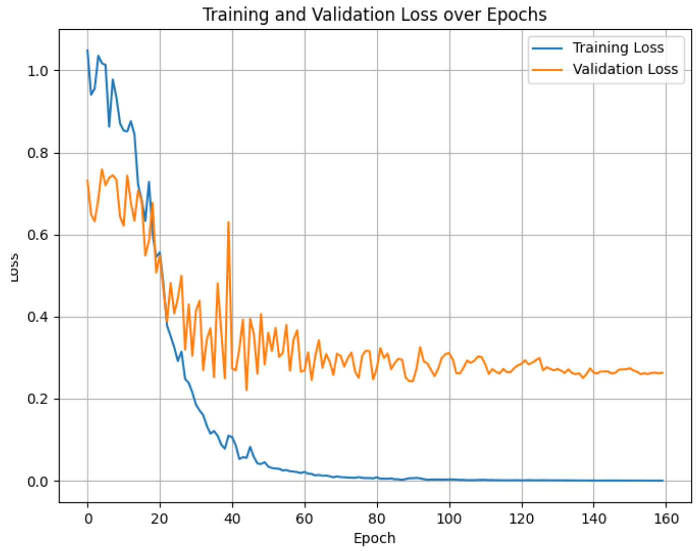

ANN Parity Plot | R^2: Train = 0.8406, Val = 0.4532

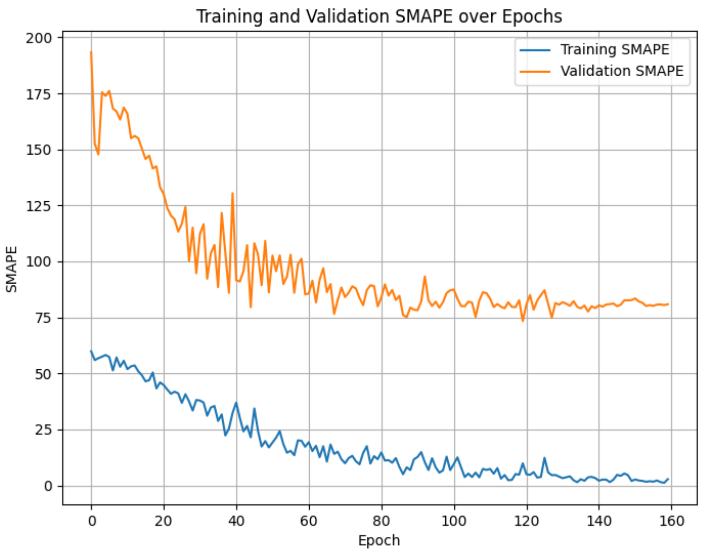

Training and Validation Loss over Epochs


Training and Validation SMAPE over Epochs

Fingerprints + Extra Features

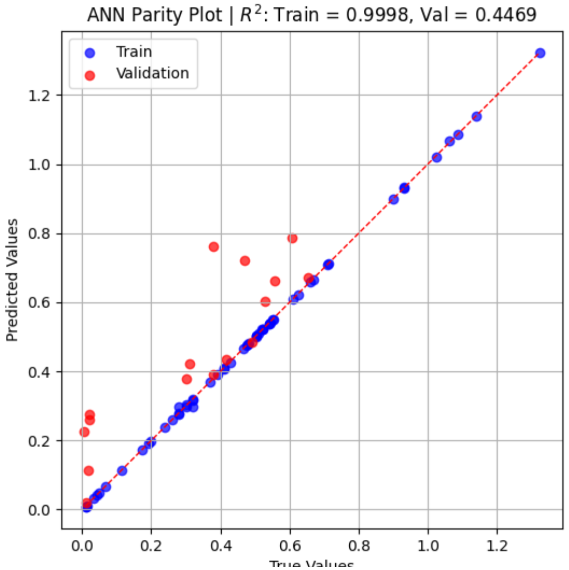

ANN Parity Plot | R^2: Train = 0.9998, Val = 0.4469

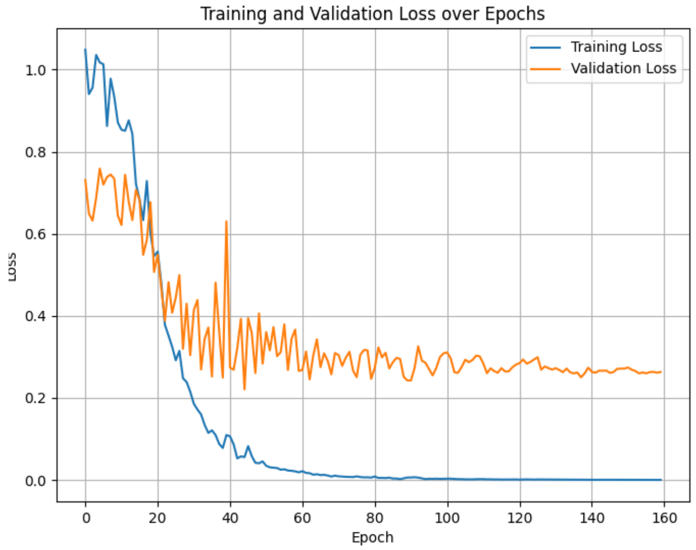

Training and Validation Loss over Epochs

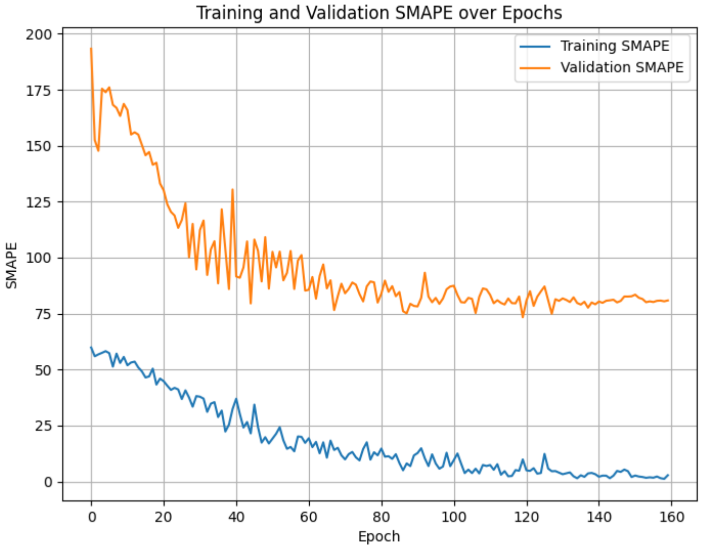

Training and Validation SMAPE over Epochs

Results (Simple ANN)

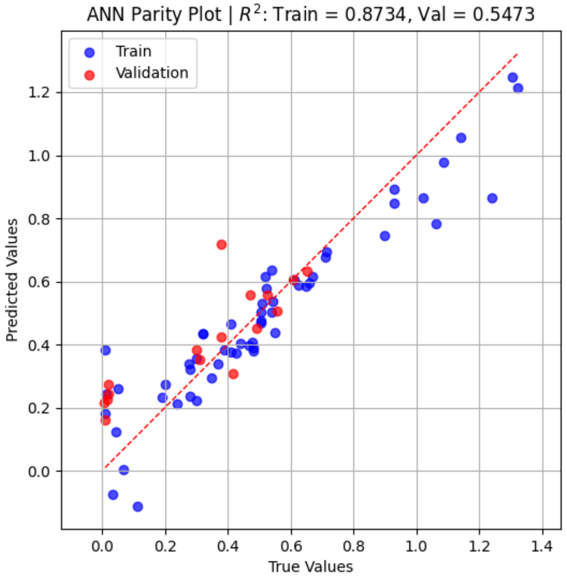
Fingerprints + Extra Features

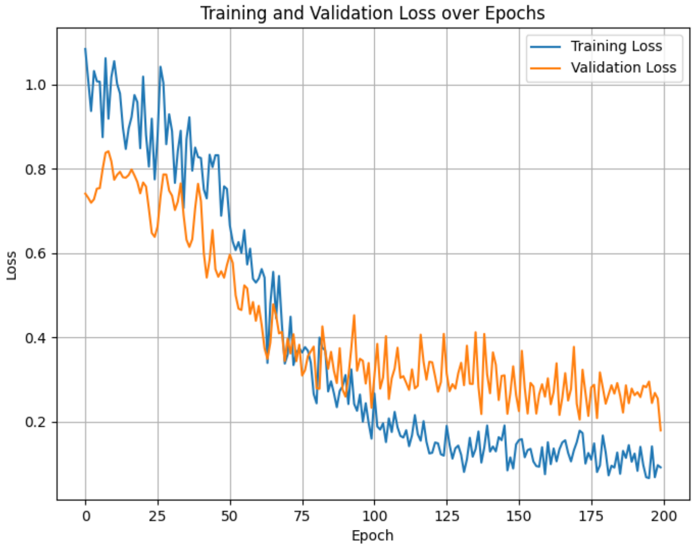

ANN Parity Plot | R^2: Train = 0.9998, Val = 0.4469

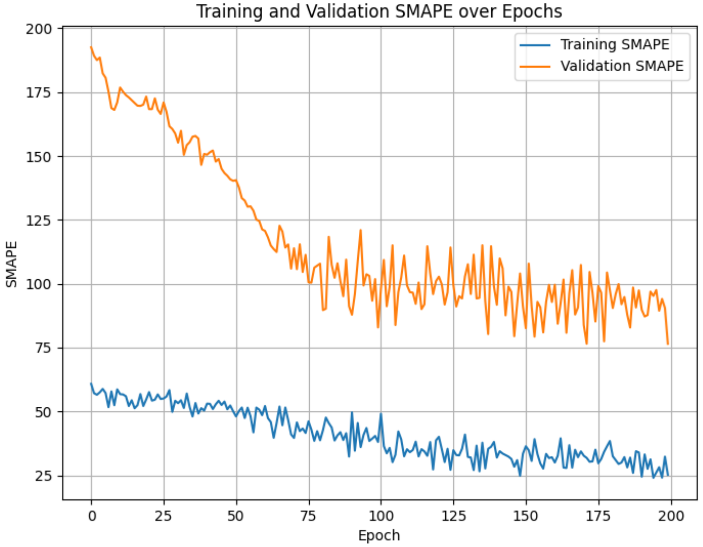

Training and Validation Loss over Epochs


Training and Validation SMAPE over Epochs

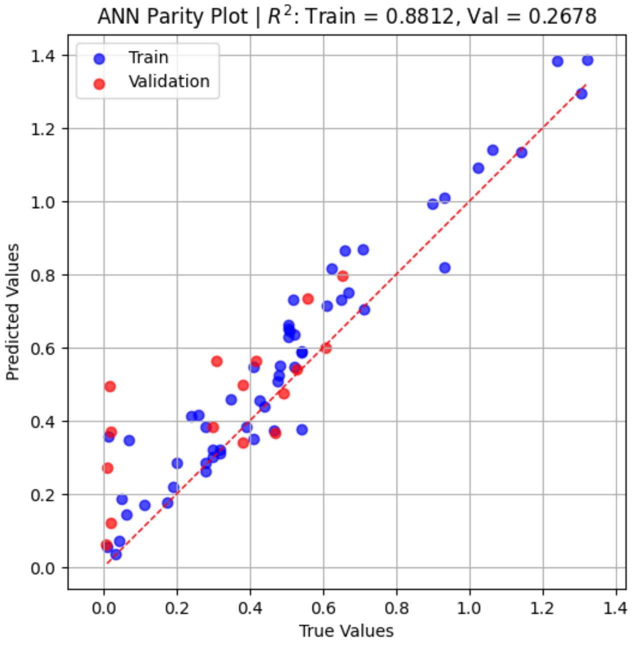
Adding Dropout (50%)

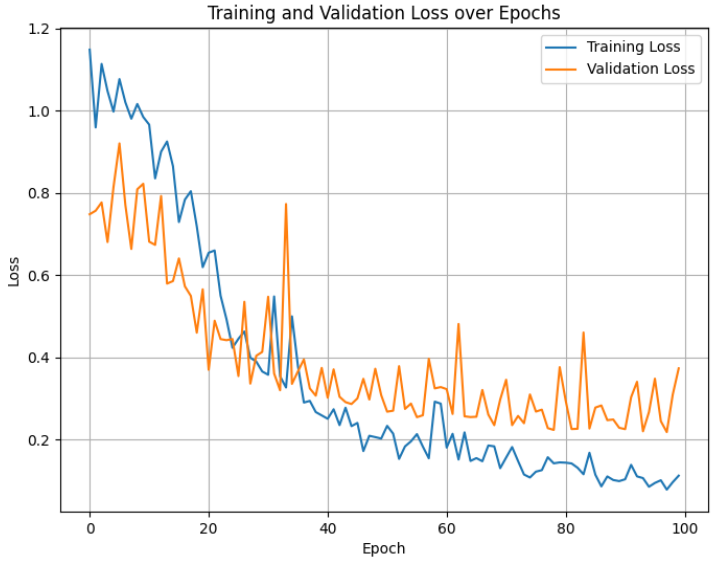

ANN Parity Plot | R^2: Train = 0.8734, Val = 0.5473


Training and Validation Loss over Epochs

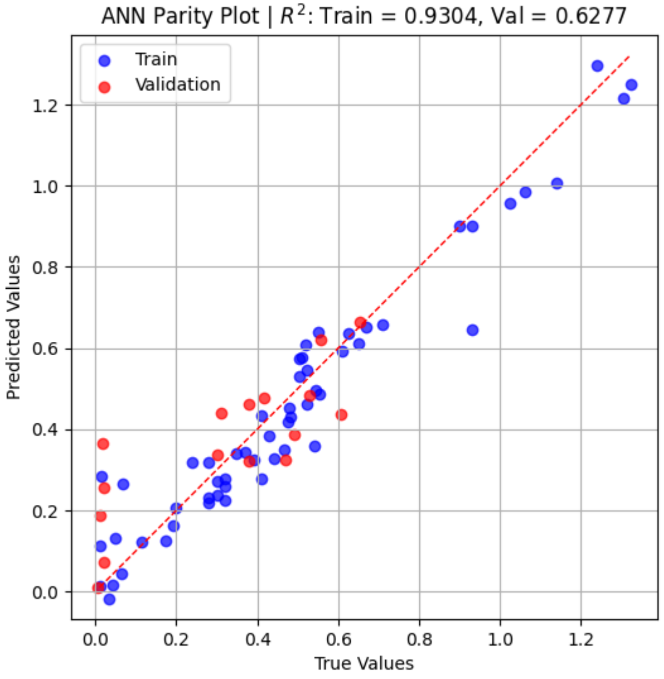

Training and Validation SMAPE over Epochs

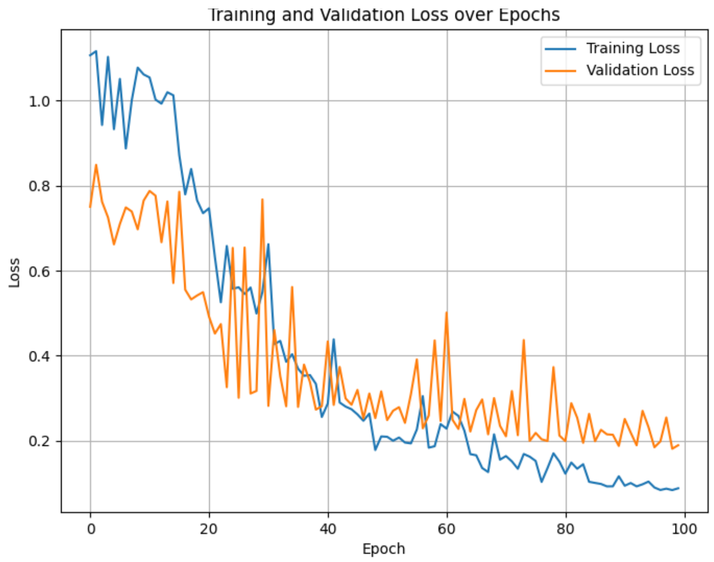
Baseline


ANN Parity Plot | R^2: Train = 0.8812, Val = 0.2678

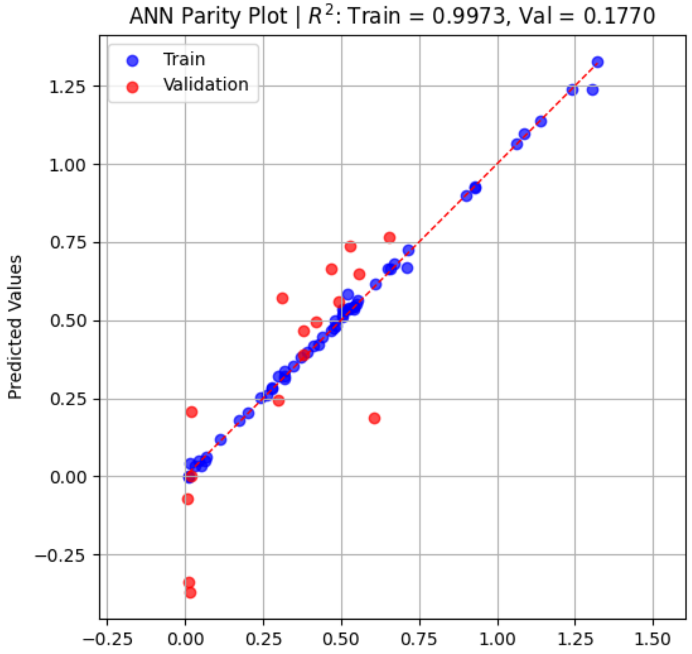

Training and Validation Loss over Epochs

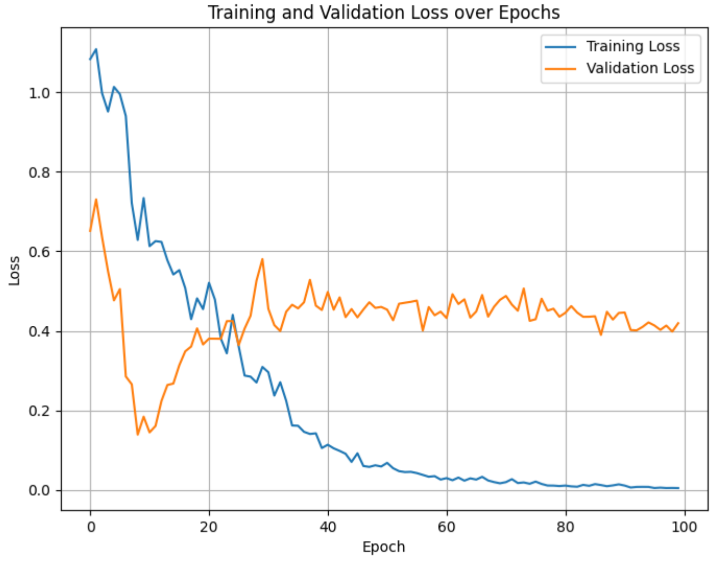
Extra Features (nAcid & nBase)


ANN Parity Plot | R^2: Train = 0.9304, Val = 0.6277


Training and Validation Loss over Epochs

Extra Features + PCA (10 components)

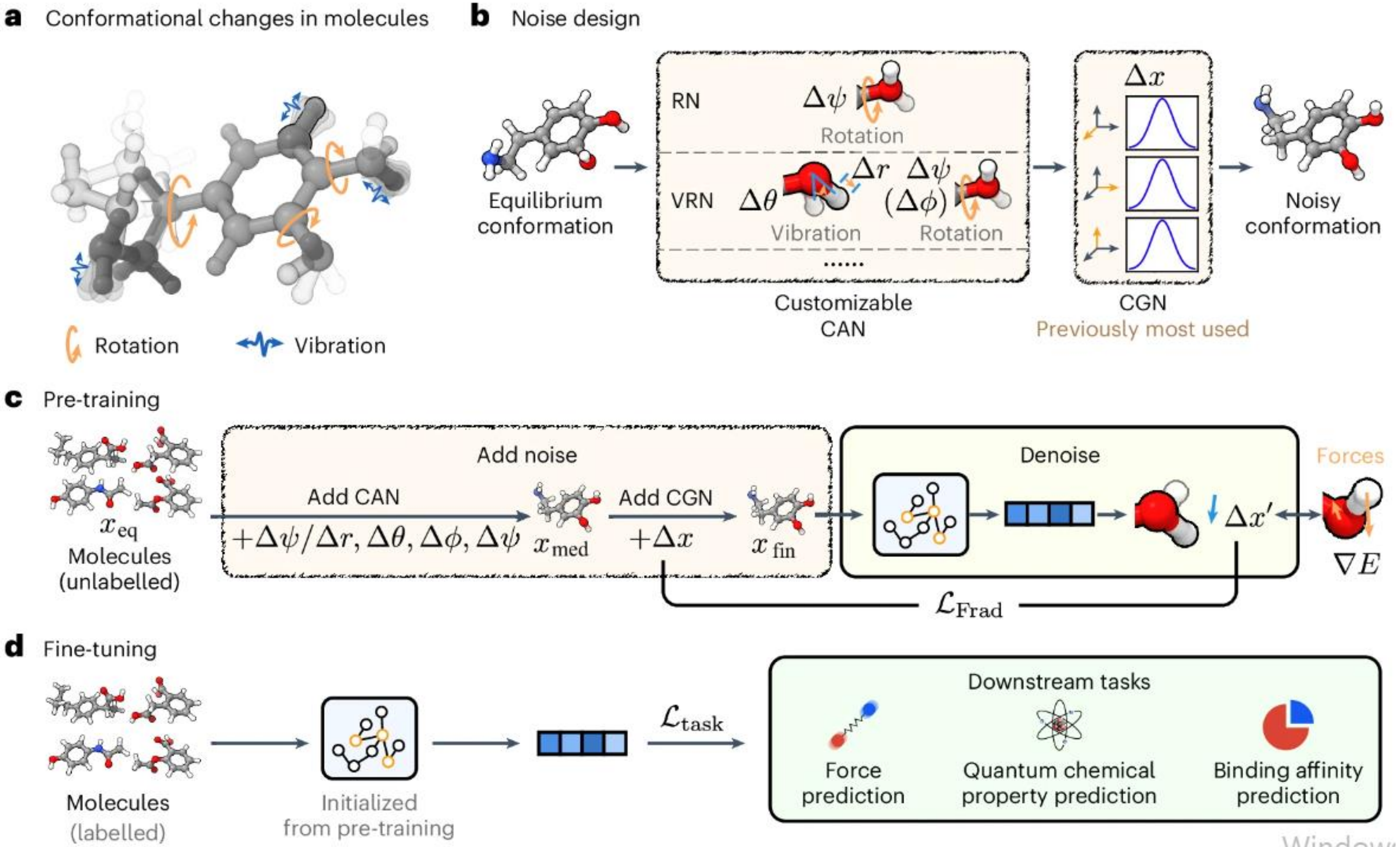

ANN Parity Plot | R^2: Train = 0.9973, Val = 0.1770


Training and Validation Loss over Epochs

Frad (Fractional Denoising) framework: base model

A novel molecular pre-training method

1. Uses a hybrid noise strategy to enhance the accuracy of molecular property predictions.
2. Captures the structural diversity of molecules while respecting chemical constraints

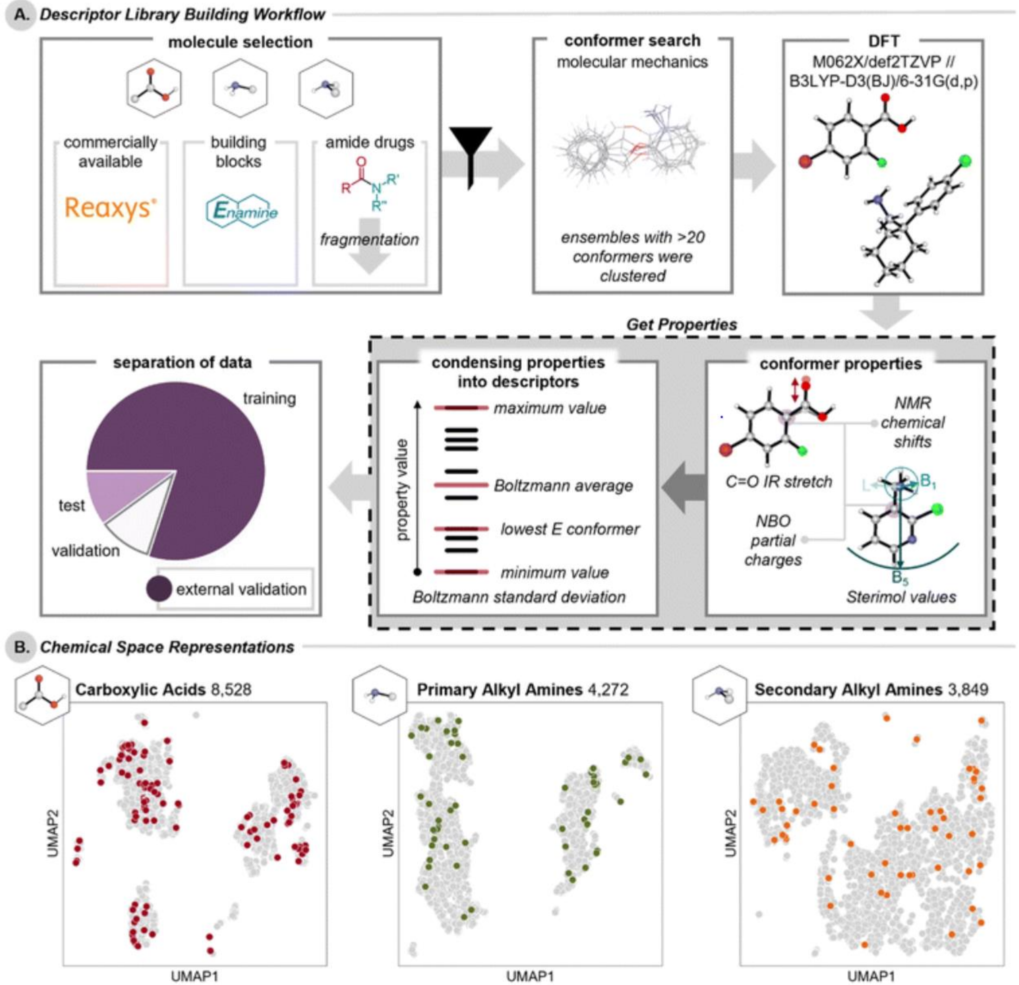


Nature Machine Intelligence volume 6, pages1169–1178 (2024)

2D amine dataset 1st Transfer

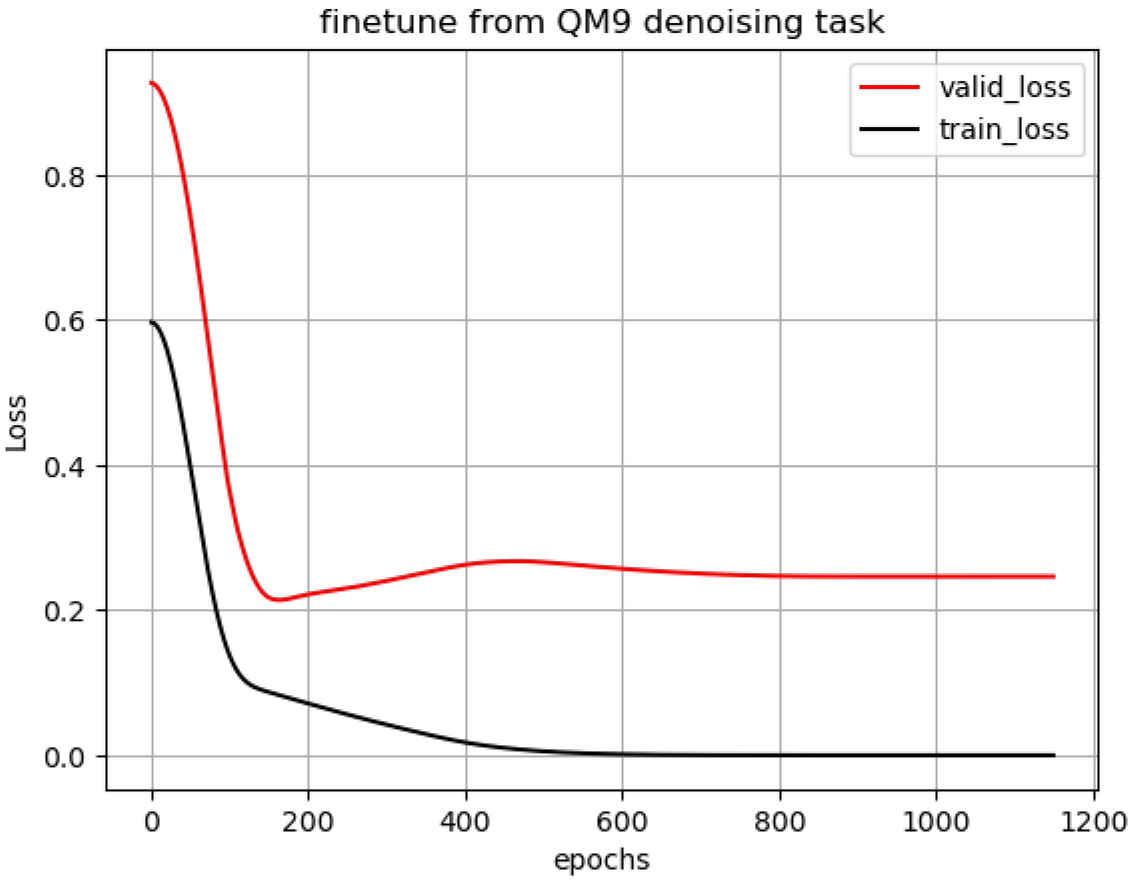
Prior transfer dataset :

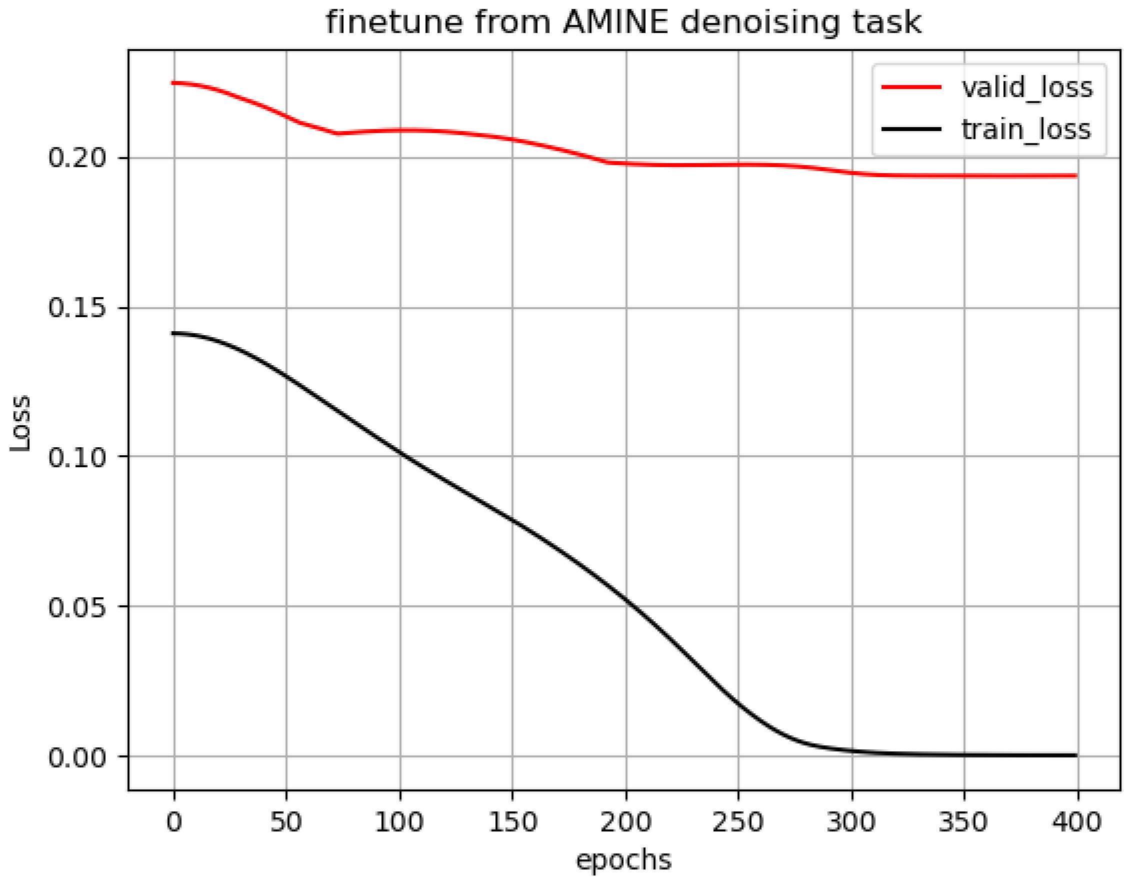
1. Because our target molecule for Co2-loading is amine, we select similar dataset.
2. Also, the basicity has essential impact on co2 loading, so our pretrain target label is HOMO value.



Haas, Brittany C., et al. Digital Discovery 4.1 (2025): 222-233.

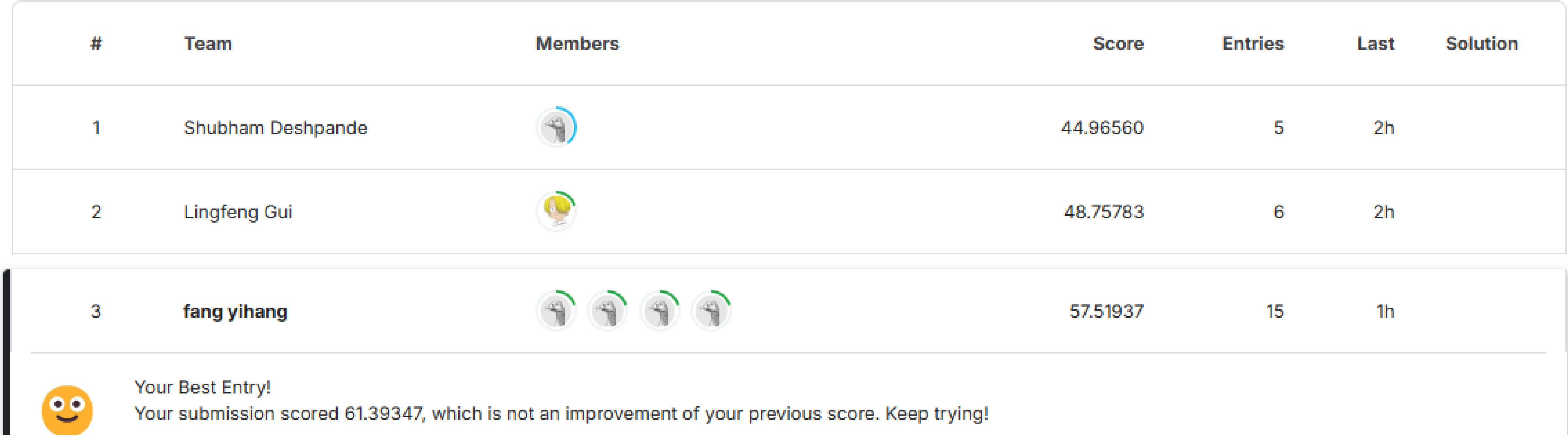
Results


finetune from QM9 denoising task


finetune from AMINE denoising task

If we start training from amine dataset, the valid loss showed little better performance

Results



#	Team	Members	Score	Entries	Last	Solution
1	Shubham Deshpande		44.96560	5	2h	
2	Lingfeng Gui		48.75783	6	2h	
3	**fang yihang**		57.51937	15	1h	

Your Best Entry!
Your submission scored 61.39347, which is not an improvement of your previous score. Keep trying!