



ECO-AI Hackathon challenge: Predicting properties of amine molecules for carbon capture using AI Part I

Predicting viscosities of aqueous amine solutions

Team: Rice Cake

Team member: Eman, Yuhui, Dennis

- **Viscosity** is a critical property in the search for novel amine molecules, as it affects various aspects of the CO₂ absorption/desorption process, including mass transfer, equipment sizing, and energy consumption, among others.
- **Challenge:** To predict viscosity
- **Goal:** develop AI or machine learning models to predict the viscosities of amine solutions at two temperatures with the highest possible accuracy.

Outline

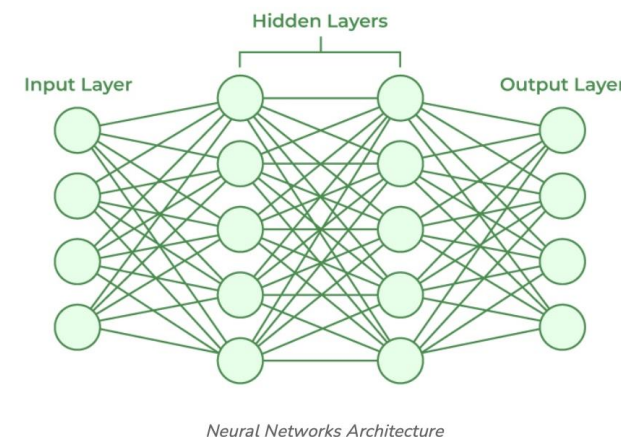
- Literature
- Methods
- Molecular representations
- Results
- Conclusion

Machine Learning in Viscosity Prediction: Insights from Recent Literature

Study	ML Model Used	Dataset Size	Key Features	Reference
Study 1	Random Forest, ANN, XGBoost	559 experimental data from the literature	XGBoost, outperforms the other models	(Shukla et al., 2024)
Study 2	Random Forest	352 samples, based on molecular dynamics and literature data	primary amine (viscosity prediction with $R^2 = 0.90$), secondary amine (viscosity prediction with $R^2 = 0.93$), and tertiary amine (viscosity prediction with $R^2 = 0.87$)	(Keer et al., 2025)
Study 3	Artificial Neural Networks (customized ANN)	12 alkanolamine and diamine systems	MAE: R-ANN (0.42), C-ANN (0.53)	(Tang et al., 2025)
Study 4	Cascade-Forward Neural Network (CFNN)	1682 training data + 220 testing data	Viscosity prediction for CO ₂ -loaded and CO ₂ -free aqueous amines; outperforms semi-empirical models.	(Aminian & ZareNezhad, 2020)
Study 5	Graph-Based Neural Network	Large dataset collected by National Institute of Standards and Technology (NIST)	Predicting viscosity of binary liquid mixtures; MAE = 0.043, RMSE = 0.080.	(Bilodeau et al., 2023)

Methods:

- 1645 amines for training & 705 amines for predicting .
- Neural Network Models:
 - Artificial Neural Network (ANN)
 - Convolutional Neural Network (CNN)
 - Simple ResNet
- ML Random Forest (RF) and LightGBM
- molecular representations: One-hot encoder, fingerprints and descriptors
- Performance of all models for predicting viscosity was evaluated using Symmetric mean absolute percentage error (sMAPE).



Molecular representations for ML

One hot Encoding

CNCCC(C)(N)CO
(String)



[[1., 0., 0., 0., 0., 0.], [0., 0., 0., 1., 0., 0.], ...]
(Vector)

It's easy and simple, but it lacks molecular structure!

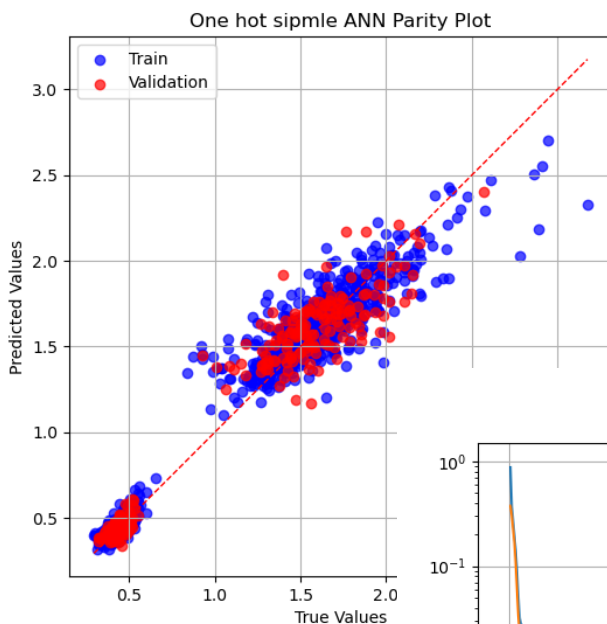
Inputs that contain information about molecular structures and physical properties are needed

Molecular representations for ML

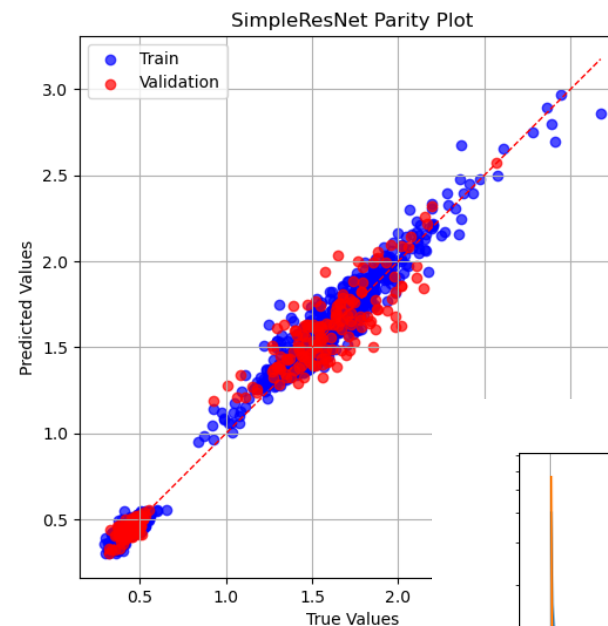
4 Methods are combined to complement each other

	Method	Strengths	Limitations	Complemented by
Structure {	Morgan Fingerprint	Captures local structural features	Hash collisions, lacks full connectivity	RDKit FP, 2D Descriptor
	RDKit Fingerprint	Captures bond connectivity and long-range patterns	May overlook functional group importance	MCAA Key, 2D Descriptor
Property {	MCAA Key	Highlights functional groups explicitly	Lacks overall structural info	Morgan FP, RDKit FP
	2D Descriptor	Directly encodes physicochemical properties	Lacks structural representation	Morgan FP, RDKit FP

One-hot with different NN model

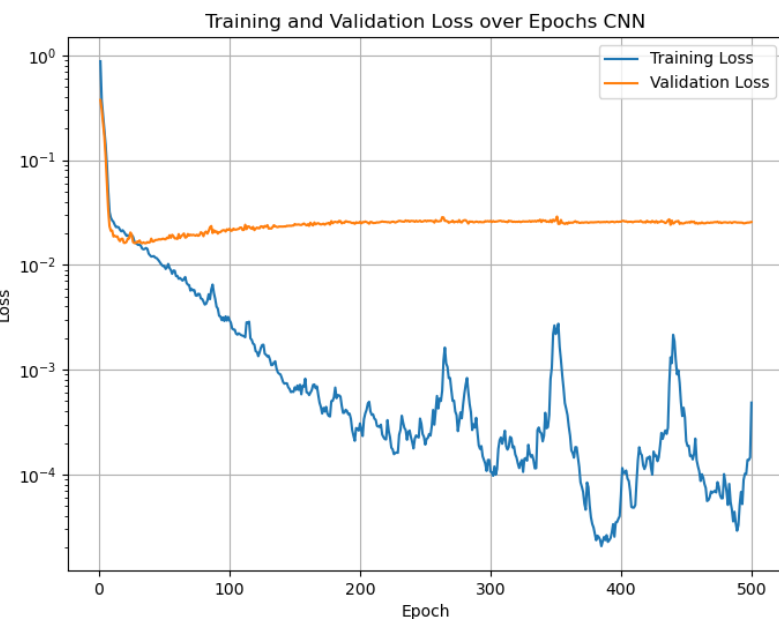


Training sMAPE: 6.7496
Validation sMAPE: 7.8435



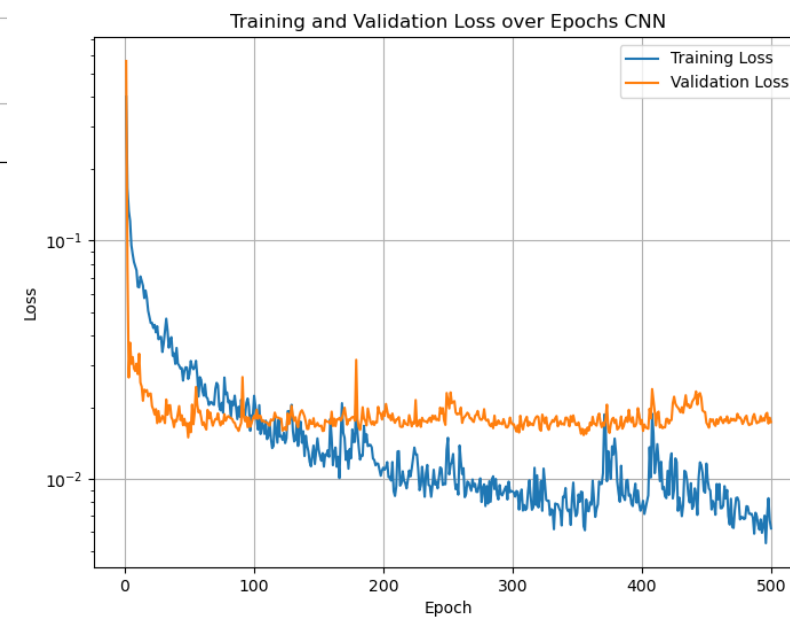
Training sMAPE: 5.8334
Validation sMAPE: 7.4436

Log y-axis



10s

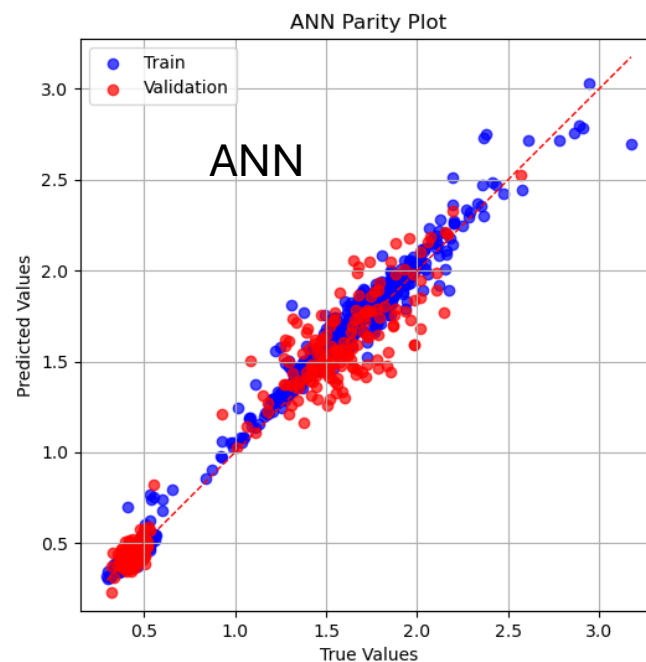
Overfitting!



30speed

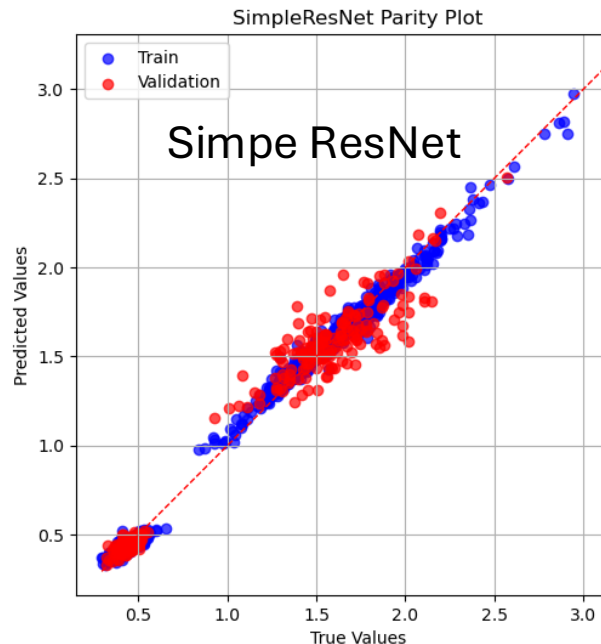
RDkit with different NN model

Leaky relu for dying ReLU problem



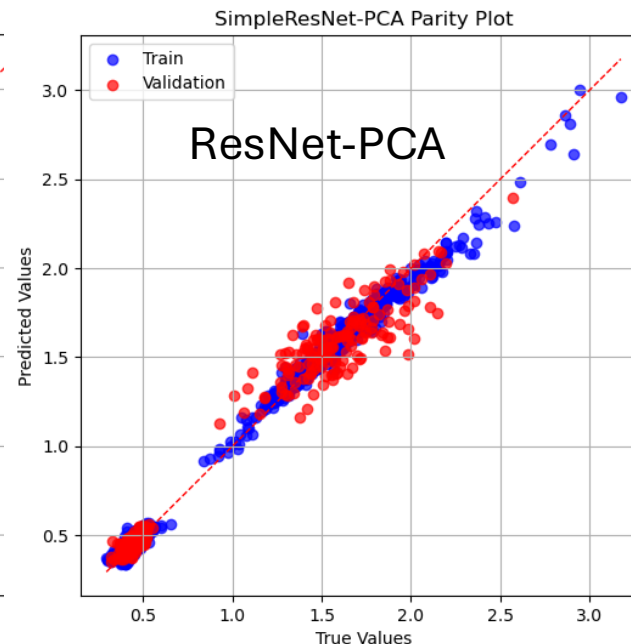
Training sMAPE: 4.2020
Validation sMAPE: 8.5971

ANN still have some overfitting



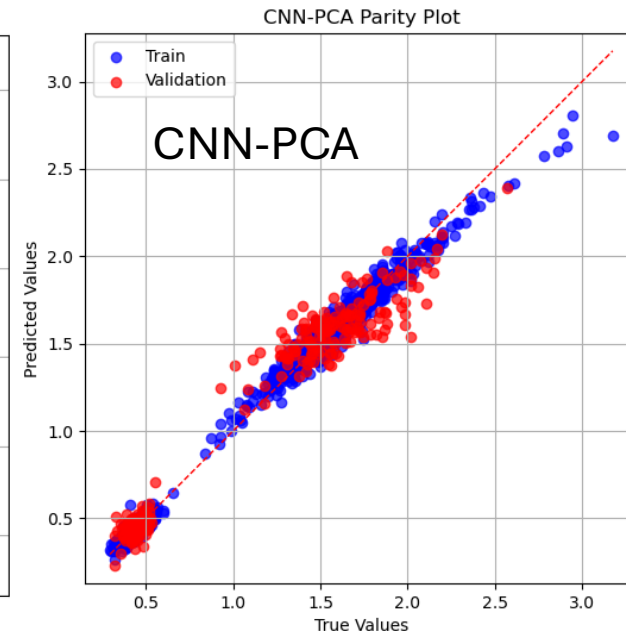
Training sMAPE: 3.4141
Validation sMAPE: 6.2788

ResNet have good results
(slow training speed)



Training sMAPE: 4.0457
Validation sMAPE: 6.8399

Faster



Training sMAPE: 4.0244
Validation sMAPE: 8.2570

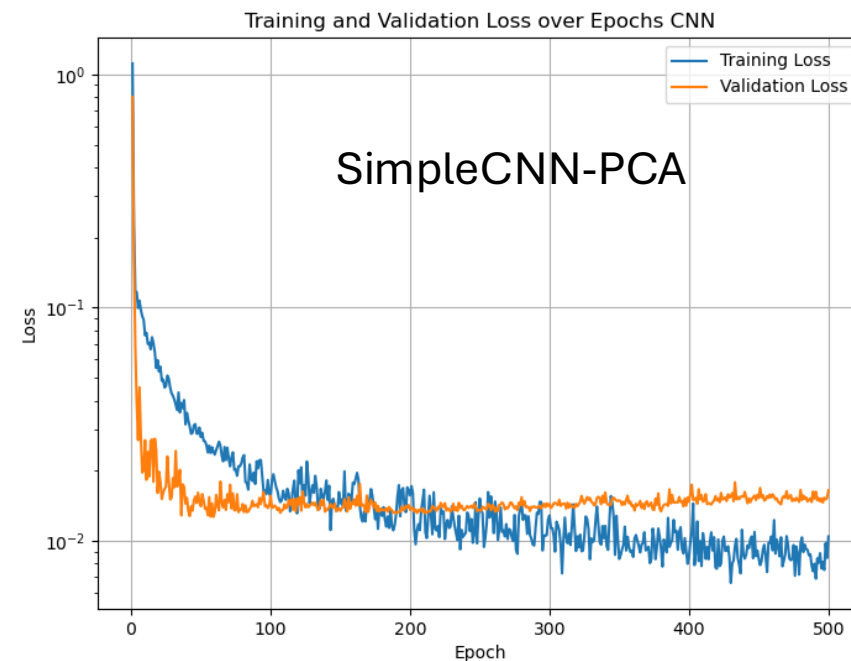
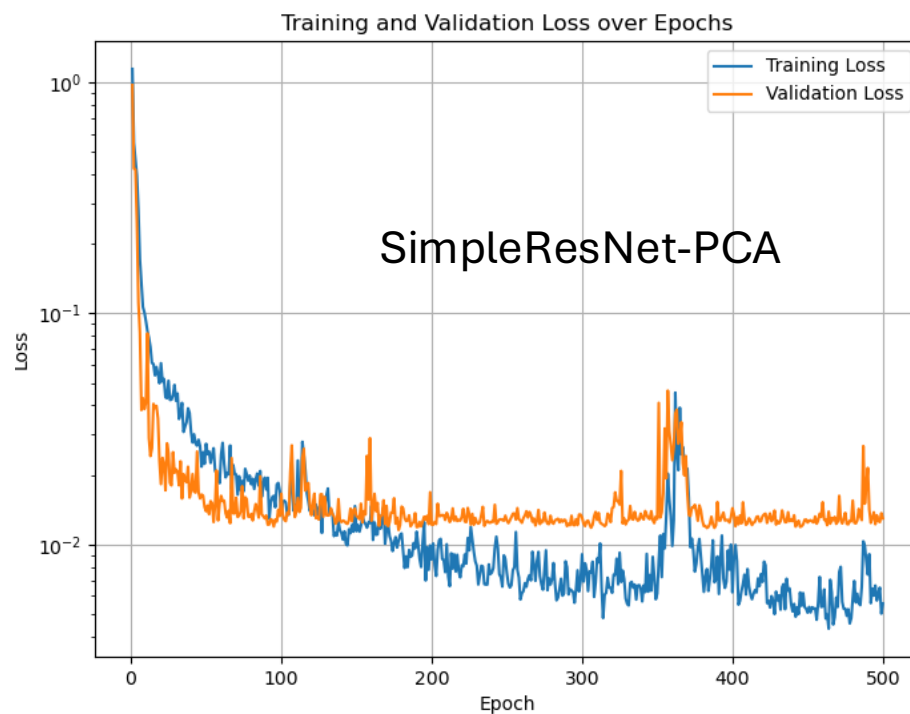
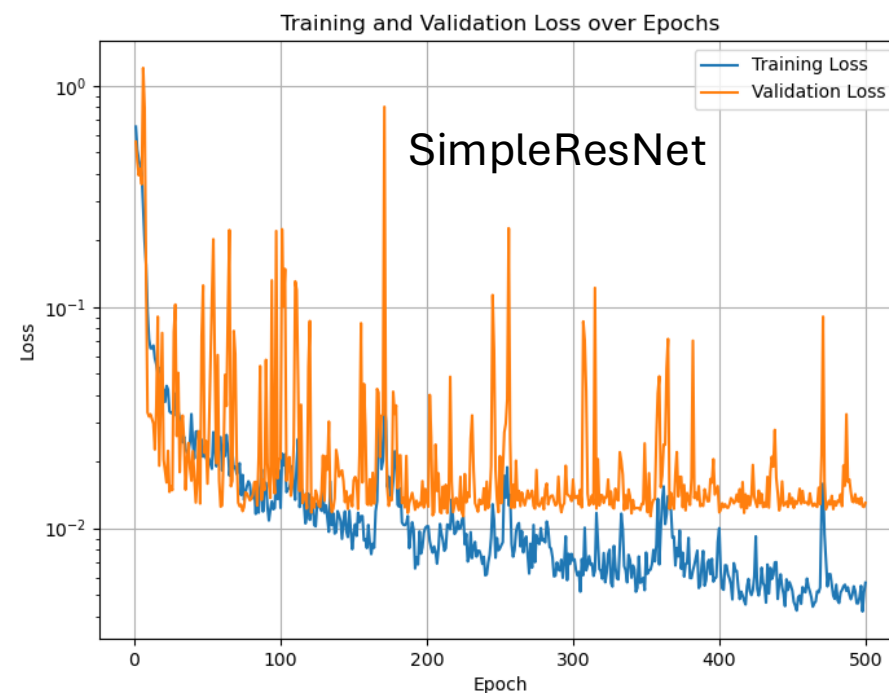
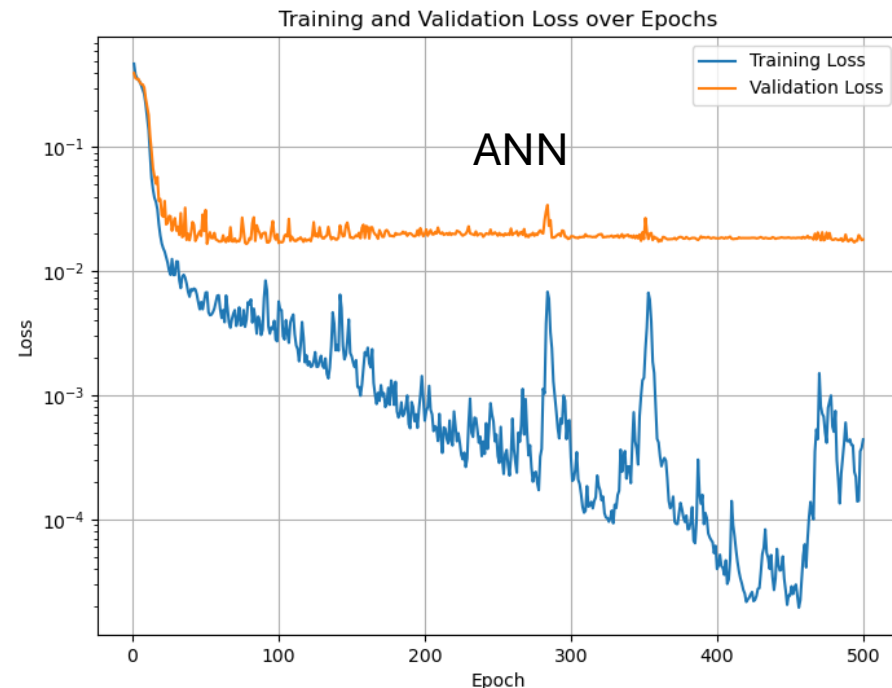
Not very well
and very slow

Convolution for fingerprinter

All of these model were chosen by the lowest val loss during training process

RDkit with different NN model

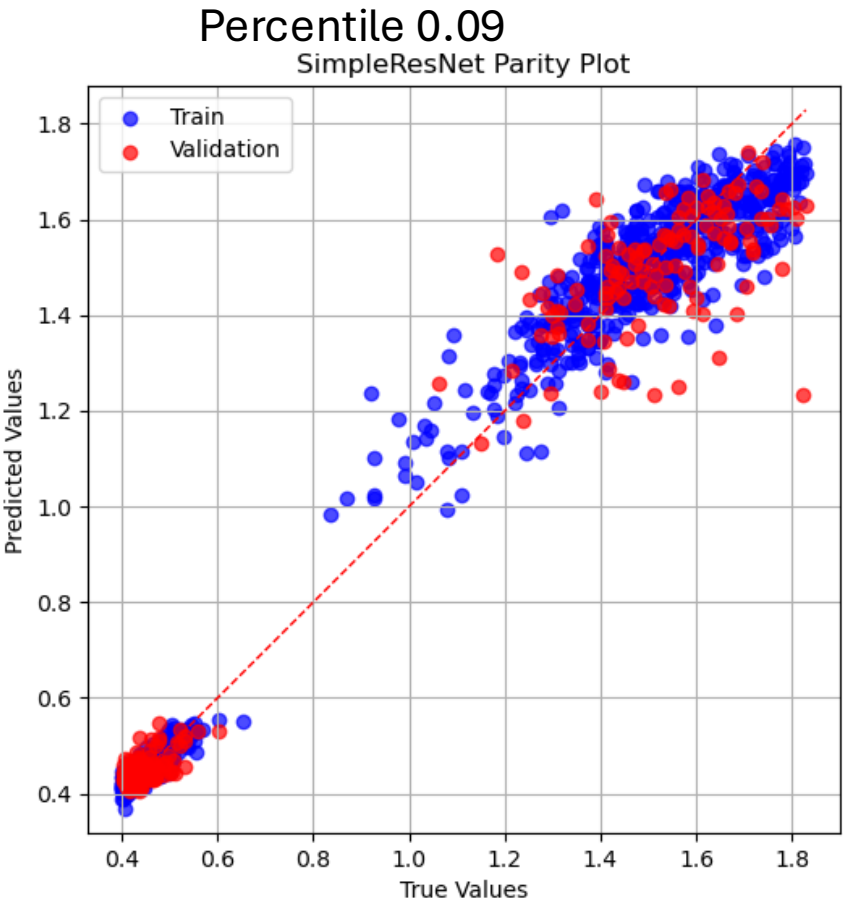
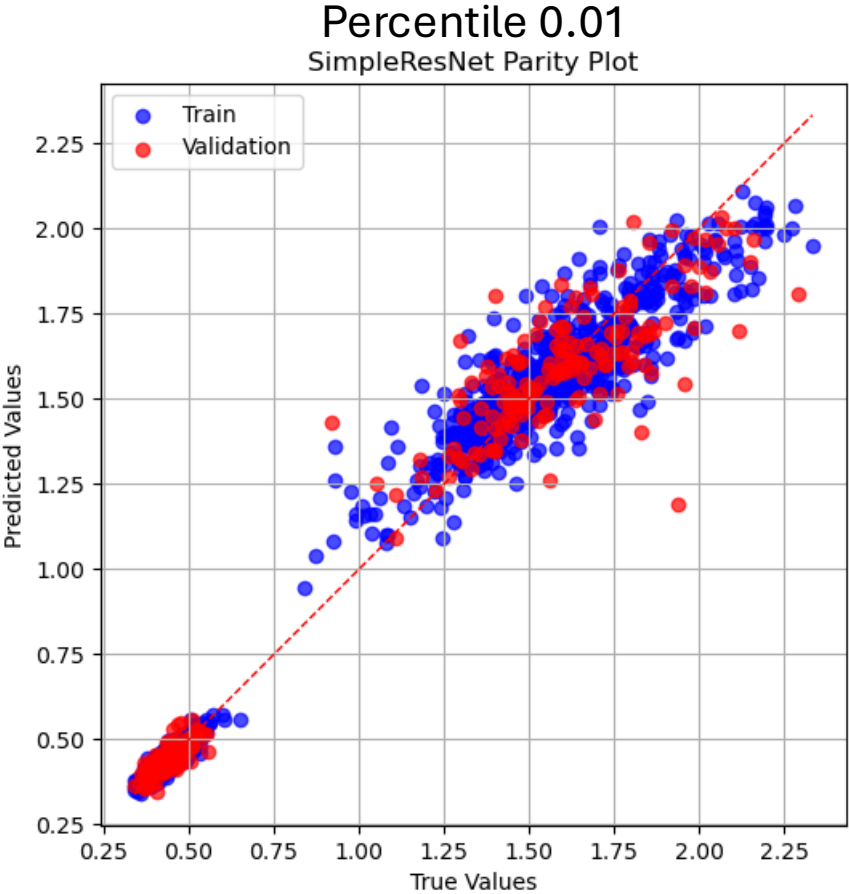
Loss figures



Random forest and LightGBM






10 folds CV

GridSearch (Parameter):
'max_depth': [5,6,7,8,9]
'learning_rate':
[0.05,0.06,0.07,0.08]



percent ile	0.01	0.02	0.03	0.04	0.05	0.06	0.07	0.08	0.09	0.10
Train smape	4.7175	4.2519	3.9811	4.5627	4.502	4.2312	4.2506	4.3254	4.037	4.0797
Val smape	5.8439	6.2165	5.8043	5.2277	5.4493	6.1381	5.6499	5.4111	5.6418	5.7937
	Score: 5.67				Score: 5.8				Score: 6.15	

Final Score:

#	△	Team	Members	Score	Entries	Last	Solution
1	▲ 1	Yuhui Yin	  	6.43293	14	1h	
2	▼ 1	Shubham Deshpande		6.55540	5	32m	
3	—	Florian Baakes		6.76178	3	1h	

Thank You!