Electronic Supplementary Information

Enhancing Predictive Models for Solubility in Multicomponent Solvent Systems using Semi-Supervised Graph Neural Networks

Hojin Jung[a,‡], Christopher D. Stubbs[a,‡], Sabari Kumar[a], Raúl Pérez-Soto[a], Su-min Song[a], Yeonjoon Kim[b,*], and Seonah Kim[a,*]

[a]Department of Chemistry, Colorado State University, Fort Collins, CO, 80523, USA.

[b]Department of Chemistry, Pukyong National University, Busan 48513, Republic of Korea.

[‡]Equal contribution.

*Corresponding author: Yeonjoon Kim (Email: yeonjoonkim@pknu.ac.kr), and Seonah Kim (Email: seonah.kim@colostate.edu)

S1. Functional Group analysis

In this section, we present the counts of functional groups on the solute molecules in MixSolDB dataset.

Table S1. Distribution of Functional Groups of solute molecules in MixSolDB.

Functional Groups	Counts	Functional Groups	Counts	Functional Groups	Counts
Alkane	41,983	Alkene	5,074	Thiol	452
Arene	40,531	Nitro	4,735	Nitrate	447
Ether	16,219	Sulfone	4,268	Alkyne	376
Alcohol	15,713	Sulfide	3,576	Phosphoric Acid	135
Cycloalkanes	15,295	Ketone	3,325	Phosphoric Ester	134
Halide	14,745	Phenol	3,142	Imine	98
Fused Ring Aromatics	14,438	Fused Ring Cycloalkanes	2,653	Nitroso	94
Amine	13,274	Nitrile	2,067	Sulfoxide	54
Carboxylic Acid	8,364	Enamine	1,013	Azide	21
Amide	7,961	Aldehyde	640	Acyl Halide	16
Ester	5,842	Peroxide	456	Sulfinate	8

S2. Cross-Validation Results for Machine Learning Models

In this section, we present the results of a cross-validation procedure aimed to evaluate the predictive performance of two machine learning models: a concatenation model and a subgraph model. This procedure offers a robust means of quantifying model accuracy. Consequently, it helps to mitigate overfitting and provides a more reliable assessment of the models' capabilities.

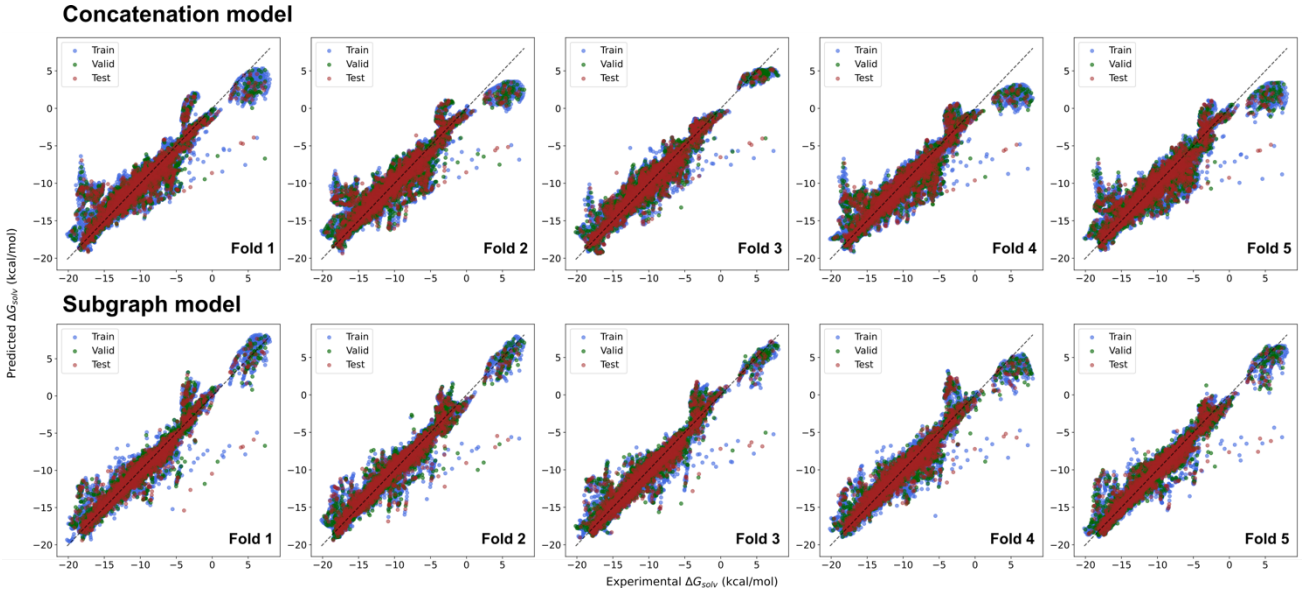

Figure S1. Parity plots for the concatenation and subgraph models applied to binary solvent systems.

Table S2. Model metrics by fold for both the concatenation and subgraph models in binary solvent systems.

Model	Metric		Fold 1	Fold 2	Fold 3	Fold 4	Fold 5
Concatenation Model	MAE	Train	0.77	0.81	0.64	0.81	0.83
		Valid	0.78	0.85	0.67	0.83	0.84
		Test	0.84	0.88	0.68	0.87	0.87
	RMSE	Train	1.34	1.36	0.98	1.37	1.35
		Valid	1.34	1.4	1.02	1.39	1.38
		Test	1.46	1.48	1.08	1.48	1.45
Subgraph Model	MAE	Train	0.52	0.57	0.59	0.65	0.55
		Valid	0.57	0.63	0.66	0.69	0.61
		Test	0.63	0.67	0.7	0.76	0.65
	RMSE	Train	0.88	0.9	0.93	1.04	0.88
		Valid	0.97	0.98	1.04	1.07	0.94
		Test	1.13	1.12	1.17	1.25	1.1

Binary Solvent Systems: Figure S1 displays parity plots for both the concatenation and subgraph models in binary solvent systems. These plots compare the predicted values from each model to the corresponding reference values for each fold. Table S2 presents the mean absolute error (MAE) and root

mean squared error (RMSE) metrics for each fold. Collectively, Figure S1 and Table S1 offer a comprehensive depiction of how both models perform across multiple folds in binary solvent systems.

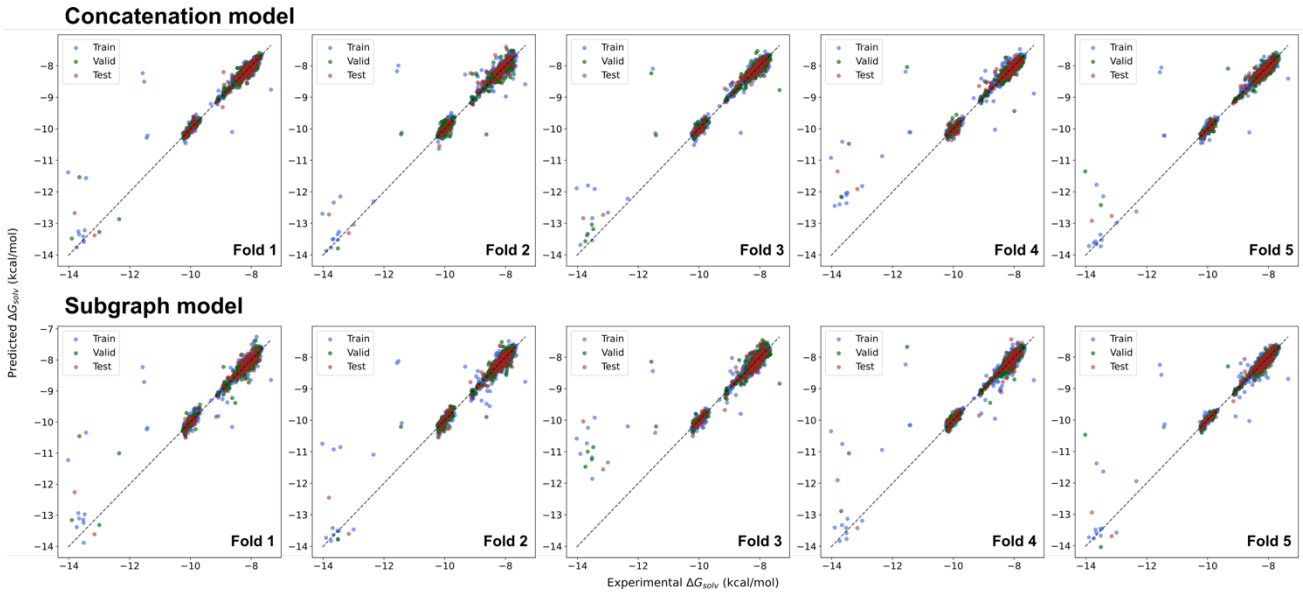

Figure S2. Parity plots for the concatenation and subgraph models applied to ternary solvent systems.

Table S3. Model metrics by fold for both the concatenation and subgraph models in ternary solvent systems.

Model	Metric		Fold 1	Fold 2	Fold 3	Fold 4	Fold 5
Concatenation Model	MAE	Train	0.08	0.09	0.08	0.09	0.07
		Valid	0.09	0.11	0.11	0.11	0.1
		Test	0.09	0.11	0.09	0.12	0.09
	RMSE	Train	0.19	0.2	0.17	0.24	0.19
		Valid	0.16	0.19	0.25	0.31	0.21
		Test	0.15	0.18	0.14	0.25	0.14
Subgraph Model	MAE	Train	0.09	0.09	0.1	0.08	0.07
		Valid	0.12	0.11	0.15	0.11	0.1
		Test	0.12	0.12	0.14	0.11	0.09
	RMSE	Train	0.22	0.24	0.28	0.21	0.19
		Valid	0.25	0.17	0.38	0.29	0.24
		Test	0.19	0.18	0.35	0.21	0.14

Ternary Solvent Systems: Figure S2 shows analogous parity plots for ternary solvent systems, again for both the concatenation and subgraph approaches. As with the binary solvent data, predicted values are plotted by fold. Table S3 contains the MAE and RMSE for each fold in ternary solvent systems. These results underscore the models' performances in more complex solvent environments and highlight their consistency across different data partitions.

S3. Solvent Performance Analysis

We examined the performance of concatenation and subgraph architectures for the five most common binary solvent systems in our multicomponent solubility database. We choose to consider only binary solvent performance due to the larger number of unique solvents (>100) within the binary solvent database compared to the ternary solvent database. We find that the trends in model performance for the subgraph and concatenation architectures remain consistent in our solvent system analysis. In particular, we see that the subgraph architecture has a lower average error (MAE) with fewer and less severe outliers (RMSE) when compared to the concatenation architecture. Additionally, the MAE and RMSE of each solvent system remains close to the overall model performance (test set MAE/RMSE) over all molecules. Therefore, we believe that the developed GNNs remain relevant for a wide variety of solvents and applications.

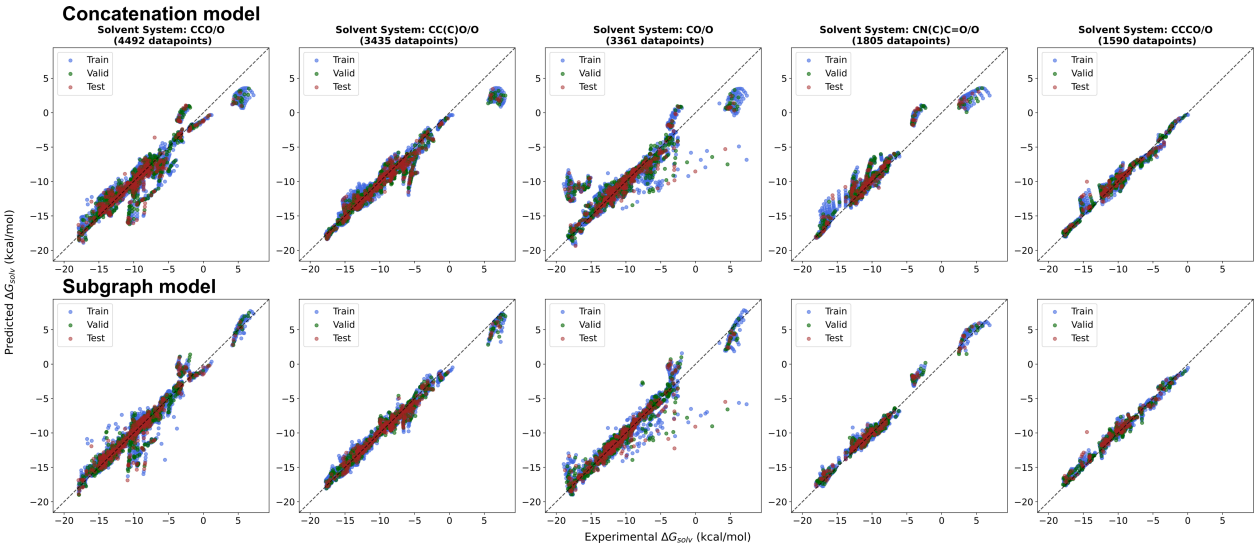

Figure S3. Parity plots for the 5 most common binary solvent systems for both the concatenation and subgraph models (for cross-validation fold 2).

Table S4. Model metrics for the 5 most common binary solvent systems for both the concatenation and subgraph models (for cross-validation fold 2).

Model	Metric		Solvent System 1	Solvent System 2	Solvent System 3	Solvent System 4	Solvent System 5
Concatenation Model	MAE	Train	0.81	0.77	1.07	0.92	0.53
		Valid	0.88	0.85	1.17	0.91	0.49
		Test	0.95	0.82	1.13	0.96	0.55
	RMSE	Train	1.3	1.23	1.93	1.32	0.73
		Valid	1.39	1.37	2.04	1.32	0.66
		Test	1.49	1.29	2.06	1.36	0.85
Subgraph Model	MAE	Train	0.6	0.51	0.61	0.51	0.45
		Valid	0.67	0.54	0.71	0.54	0.47
		Test	0.76	0.55	0.75	0.54	0.54
	RMSE	Train	1.01	0.69	1.07	0.69	0.58
		Valid	1.1	0.75	1.28	0.73	0.6
		Test	1.29	0.74	1.41	0.74	0.75

S4. Model Baseline

To provide a baseline for our GNN performance, we trained a random forest model on our binary solubility database, using the temperature/stoichiometry/Morgan fingerprint of each solute and solvent system pair as the model input. Our morgan fingerprint used a bit length of 4096 and a radius of 3 following benchmarking over the entire binary solubility database. To allow for more direct comparison to our GNN models, we used the same train/validation/test assignments (80/10/10) as the binary models shown in **Figure 4,** while omitting the validation set due to incompatibility with the Scikit-Learn API. We find that the random forest model has poor train/test performance (MAE of 0.02/2.80) compared to the GNN models trained on the same training set and evaluated on the same test set. We believe this model baseline highlights the utility of GNN-based approaches over more traditional models.

S5. Model Hyperparameter Evaluation

In order to identify optimal hyperparameters for our models, we benchmarked hyperparameter performance against the test set MAE/RMSE for the binary solvent prediction task. We found that the GNN model using the hyperparameters outlined in the paper (in **bold**) was the most effective in achieving high test set performance while reducing overfitting.

Table S5. Hyperparameter Benchmark

Hyperparameters	Values Tested
Number of epochs	[10, 100, **1000**]
Number of message blocks	[1, 3, **5**]
Number of neurons	[32, 64, **128**]