Supporting Information

Accurate prediction of aqueous solvation free energies using 3D atomic feature-based graph neural network with transfer learning

Dongdong Zhang[1], Song Xia[1] and Yingkai Zhang[1,2]*

[1]Department of Chemistry, New York University, New York, New York 10003, United States

[2]NYU-ECNU Center for Computational Chemistry at NYU Shanghai, Shanghai 200062, China

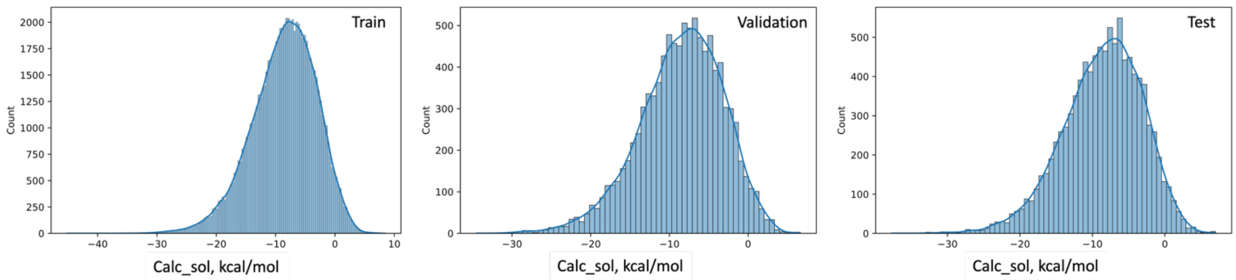

Figure S1. Data distributions for Frag20-Aqsol-100K with fixed data split. Sizes for train/validation/test are 80K, 10K and 10K, respectively.

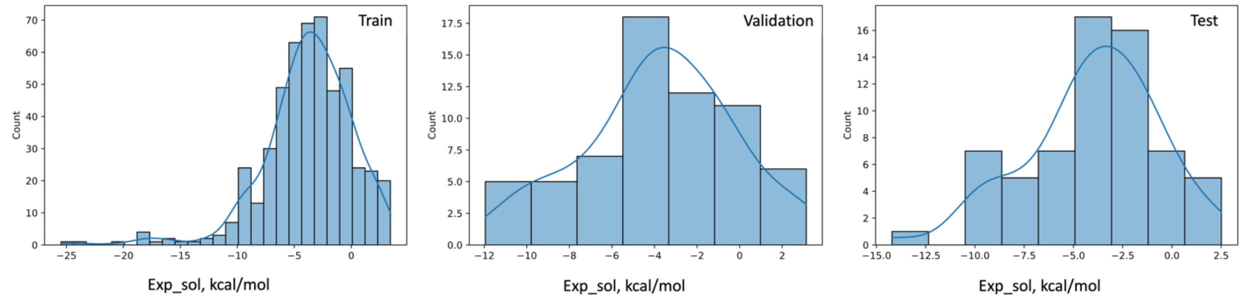

Figure S2. Data distributions for FreeSolv with fixed data split. Sizes for train/validation/test are 502, 63 and 65, respectively.

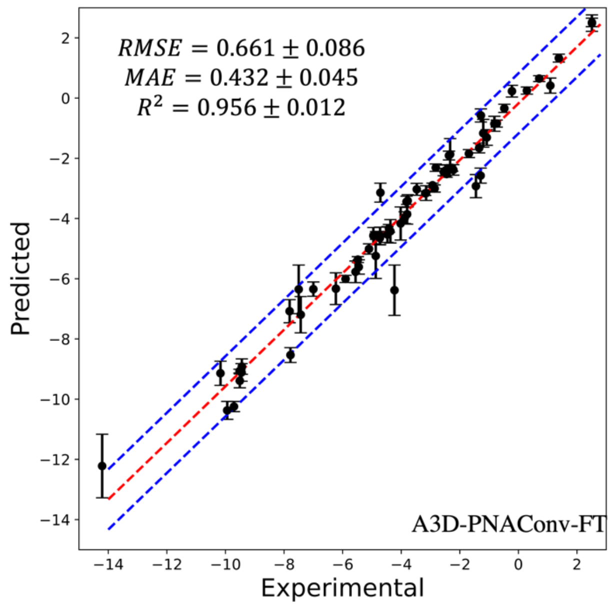

Figure S3. Scatter plots between predicted values and experimental values on the test set (65 compounds) of FreeSolv for A3D-PNAConv-FT. The dashed blue lines correspond to ± 1 kcal/mol.

Table S1. 2D bond attributes and the corresponding descriptions and encoding methods

Attributes	Descriptions	Size	Encoding
Type	Single, double, triple, aromatic	4	One-hot
conjugated	whether the bond is conjugated	1	-
In ring	whether the bond is part of an aromatic system	1	-

Table S2. The related functions in MPNN for selected GNN modules

GNN[a]	Message function	Updating function	Aggregation method	Use bond information or not	Total # of trainable parameters[b]
GIN[1]	$\sum_{w\in N(v)} h_w^t$	$h_v^t + m_v^{t+1}$	SUM	NO	(79K, 100K)
GINEConv[2]	$\sum_{w\in N(v)} \mathrm{ReLU}(h_w^t + e_{vw})$	$h_v^t + m_v^{t+1}$	SUM	YES	(80K, 110K)
NNConv[3]	$\sum_{w\in N(v)} h_w^t \cdot e_{vw}$	$concat(h_v^t, m_v^{t+1})$	SUM	YES	(5.31M, 5.33M)
PNAConv[4]	$\bigoplus (h_v^t + h_w^t + e_{vw})$	$concat(h_v^t, m_v^{t+1})$	PNA	YES	(1.21M, 1.23M)
SuperGAT[5]	$\sum_{w\in N(v)} \alpha_{vw} h_w^t$	$h_v^t + m_v^{t+1}$	Weighted summation	NO	(81K, 110K)

[a]Besides the listed GNNs, we also applied D-MPNN for comparison. D-MPNN adopted more complex way to construct the message and updating functions which are based on Loopy Belief Propagation. Edge information was used. Check the methods section 'Directed MPNN' in the original paper[6] for more details.

[b]Total number of trainable parameters for 2D feature based and 3D feature-based models under each GNN. In the parenthesis, the first number is for 2D and the second is for 3D, respectively. For D-MPNN, it has 71K for 2D and 120K for 3D.

Table S3. 2D atomic attributes and the corresponding descriptions and encoding methods.

Attributes	**Descriptions**	**Size**	**Encoding**
Types	Element symbols	11	One-hot
Hybridizations	sp, sp2, sp3, sp3d, or sp3d2	5	One-hot
Degree	0, 1, 2, 3, 4, 5	6	One-hot
Explicit valence	0, 1, 2, 3, 4, 5, 6	7	One-hot
Implicit valence	0, 1, 2, 3, 4, 5	6	One-hot
Aromaticity	whether the atom is part of an aromatic system	1	-
In ring of different sizes*	Whether the atom is in a ring of size 3 or larger	11	One-hot

* Ring size ranges from 3 to 20. This atom attribute takes up 11 positions in the final atom feature vector because we assign the first 10 positions for whether one atom is in a ring of size within [3,12], and the last 1 position for whether one atom is in a ring of size within [13,20].

Table S4. Key hyper parameters for the model building and training. All internal hyper parameters in each GNN modules are set as default.

Hyper parameters	**Values**
# of encoder layers	3
# hidden size	120
# read out layers	3
Weights initialization method	Xavier_norm
Optimizer	Adam
Batch size	100
Learning rate	0.001
Learning rate scheduler	Constant
Loss function	L1

Table S5. 95% confidence interval (CI) on test set (10,000 samples) of Frag20-Aqsol-100K by each GNN under 2D and A3D featurization from bootstrapping (500,000 iterations)

Models	**Median**	**CI**
2D-PNAConv	1.168	(1.133, 1.205)
$A3D_{MM}$-PNAConv	0.681	(0.659, 0.704)
$A3D_{QM}$-PNAConv	0.432	(0.416, 0.448)
2D-DMPNN	1.181	(1.148, 1.216)
A3D-DMPNN	0.757	(0.733, 0.785)
$A3D_{QM}$- DMPNN	0.544	(0.519, 0.572)
2D-GINConv	1.209	(1.179, 1.239)
$A3D_{MM}$-GINConv	0.799	(0.773, 0.825)
$A3D_{QM}$-GINConv	0.557	(0.528, 0.577)
2D-GINEConv	1.175	(1.141, 1.207)
$A3D_{MM}$-GINEConv	0.742	(0.718, 0.767)
$A3D_{QM}$-GINEConv	0.487	(0.470, 0.507)
2D-NNConv	1.175	(1.145, 1.206)
$A3D_{MM}$-NNConv	0.715	(0.693, 0.738)
$A3D_{QM}$-NNConv	0.462	(0.443, 0.485)
2D-superGAT	1.198	(1.166, 1.232)
$A3D_{MM}$-superGAT	0.786	(0.758, 0.817)
$A3D_{QM}$-superGAT	0.531	(0.516, 0.547)
2D-DNN	2.805	(2.752, 2.858)
A3D-DNN	0.933	(0.906, 0.962)
$A3D_{QM}$-DNN	0.707	(0.684, 0.731)

Table S6. Statistical analysis between 2D-DMPNN-TS (Baseline) with other variants. The underscore bold components indicate where the difference between variant models and baseline locates.

Models	General performance (RMSE)	Median	90% CI	90% CI of difference between baseline	Probability of difference >0 between baseline
2D-DMPNN-TS (**Baseline**)	0.968±0.132	0.967	(0.636, 1.482)	-	-
2D-**PNAConv**-TS	0.954±0.084	0.930	(0.648, 1.255)	(-0.411, 0.573)	58.03%
A3D-DMPNN-TS	1.054±0.101	1.035	(0.780, 1.347)	(-0.485, 0.511)	55.60%
2D-DMPNN-**FT**	0.821±0.113	0.789	(0.537, 1.144)	(-0.357, 0.789)	71.02%
A3D-**PNAConv**-TS	0.987±0.131	0.939	(0.605, 1.420)	(-0.550, 0.611)	53.81%
2D-**PNAConv**-**FT**	0.780±0.099	0.743	(0.537, 1.090)	(-0.241, 0.790)	76.84%
A3D-DMPNN-**FT**	0.698±0.039	0.677	(0.476, 0.933)	(-0.143, 0.831)	85.31%
A3D-**PNAConv**-**FT**	0.661±0.086	0.647	(0.446, 0.884)	(-0.062, 0.837)	**89.76%**

References

1. Xu, K.; Hu, W.; Leskovec, J.; Jegelka, S. How Powerful Are Graph Neural Networks **2018**? arXiv: 1810.00826. arXiv.org e-Print archive. https://arxiv.org/abs/1810.00826

2. Hu, W.; Liu, B.; Gomes, J.; Zitnik, M.; Liang, P.; Pande, V.; Leskovec, J. Strategies for Pre-Training Graph Neural Networks **2019**. arXiv: 1905.12265. arXiv.org e-Print archive. https://arxiv.org/abs/1905.12265

3. Gilmer, J.; Schoenholz, S. S.; Riley, P. F.; Vinyals, O.; Dahl, G. E. Neural Message Passing for Quantum Chemistry **2017**. arXiv: 1704.01212. arXiv.org e-Print archive. https://arxiv.org/abs/1704.01212

4. Corso, G.; Cavalleri, L.; Beaini, D.; Liò, P.; Veličković, P. Principal Neighbourhood Aggregation for Graph Nets **2020**. arXiv: 2004.05718. arXiv.org e-Print archive. https://arxiv.org/abs/2004.05718

5. Kim, D.; Oh, A. How to find your friendly neighborhood: graph attention design with self-supervision. *ICLR **2021***.

6. Yang, K.; Swanson, K.; Jin, W.; Coley, C.; Eiden, P.; Gao, H.; Guzman-Perez, A.; Hopper, T.; Kelley, B.; Mathea, M.; Palmer, A.; Settels, V.; Jaakkola, T.; Jensen, K.; Barzilay, R. Analyzing Learned Molecular Representations for Property Prediction. *J. Chem. Inf. Model.* **2019**, *59*, 3370–3388.