Challenges in Transferable Prediction of Solvation Free Energy: A Comparative Analysis of Molecular Representations and Machine Learning Methods

Dibyendu Maity and Suman Chakrabarty*

Department of Chemical and Biological Sciences, S.N. Bose National Centre for Basic Sciences, Kolkata

E-mail: sumanc@bose.res.in

Descriptor Based Models

Table S1: Molecular descriptors with definitions

Descriptor	**Description**
Basic Molecular Properties	
Molecular Weight	Exact mass of the molecule (in daltons)
TPSA	Topological Polar Surface Area (in Å^2)
CrippenClogP	Wildman-Crippen octanol-water partition coefficient
Fraction SP3	Ratio of sp^3-hybridized carbon atoms to total carbons
Bond and Ring Characteristics	

Continued on next page

Table S1 – continued from previous page

Descriptor	**Description**
Number of Rotatable Bonds	Count of non-terminal single bonds excluding amides
Number of Rings	Total count of all ring systems
Number of Aromatic Rings	Count of rings with aromatic character
Number of Aliphatic Rings	Count of non-aromatic rings
Number of Saturated Rings	Count of fully saturated rings
Number of Bridgehead Atoms	Atoms shared between rings with ≥ 2 bonds
Heteroatom and Functional Groups	
Number of Heteroatoms	Total non-carbon, non-hydrogen atoms
#O_atoms	Total oxygen atoms
#N_atoms	Total nitrogen atoms
#F_atoms	Total fluorine atoms
#Cl_atoms	Total chlorine atoms
NumAmideBonds	Count of CONH groups
fr_bicyclic	Number of bicyclic rings
fr_ketone	Number of ketones
fr_para_hydroxylation	Number of para-hydroxylation sites
fr_sulfone	Number of sulfone groups
Hydrogen Bonding	
Number of H-Bond Donors	Count of NH or OH groups
lipinskiHBD	Lipinski rule-compliant H-bond donors
Stereochemistry	
Number of Atom Stereo Centers	Total stereocenters (specified + unspecified)

Continued on next page

Table S1 – continued from previous page

Descriptor	Description
NumUnspecifiedAtomStereoCenters	Stereocenters without defined configuration
Other Indices	
hallKierAlpha	Hall-Kier alpha value (Rev. Comp. Chem. 2, 367–422, 1991)
kappa1–3	Hall-Kier shape indices κ_1–κ_3
MaxEStateIndex	Maximum EState index
MinEStateIndex	Minimum EState index
MinAbsEStateIndex	Minimum absolute EState index
qed	Quantitative estimate of drug-likeness
MaxAbsPartialCharge	Maximum absolute Gasteiger atomic charge
FpDensityMorgan1	Morgan fingerprint density, radius 1
BalabanJ	Chemical distance-based topological index
Chi4v	Valence molecular connectivity index
Chi4n	Variant of Chi4v using nVal
Kappa3	Third-order shape/connectivity index
SlogP_VSAN, N = 4, 5, 8, 11	MOE-type descriptors using LogP and surface area contributions

Model Description: Graph Convolutional Network - Variational Autoencoder (GCN-VAE)

Model Architecture:

The GCN-VAE model handles molecular graphs(dgl[1]) from SMILES strings through two key components: the Encoder and the Decoder.

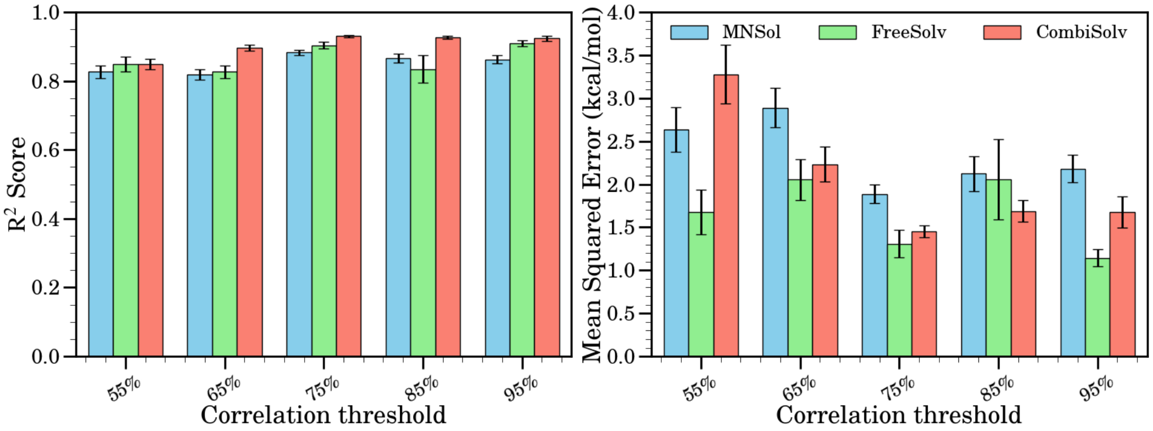

Figure S1: By varying the threshold value of correlation criteria we found optimal number of features to form feature vector. A XGBoost model was used to evaluate the efficiency of the representations. 10-fold cross validation was employed for reliable statistics. The standard error in measurement is provided.

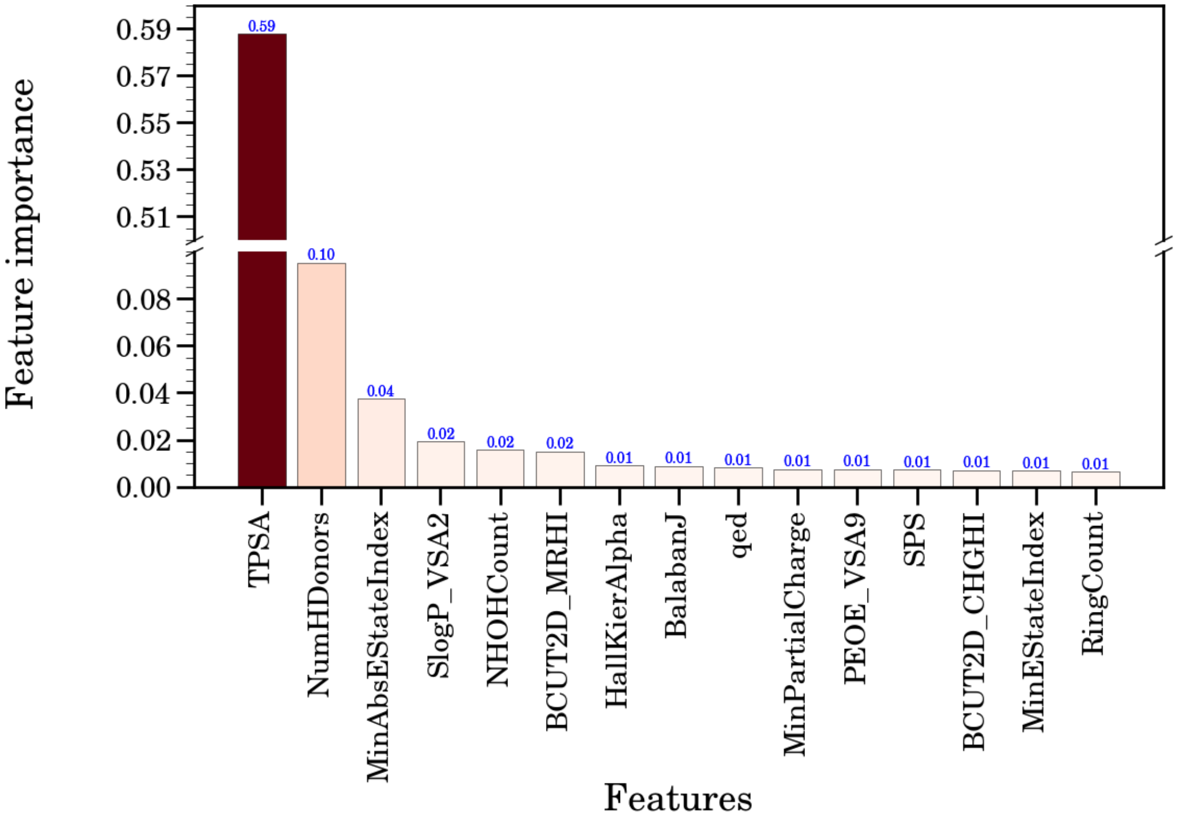

Figure S2: Relative importance of the features in terms of Gini score. TPSA has the most contribution towards the predictability of the model, followed by number of hydrogen bond donors.

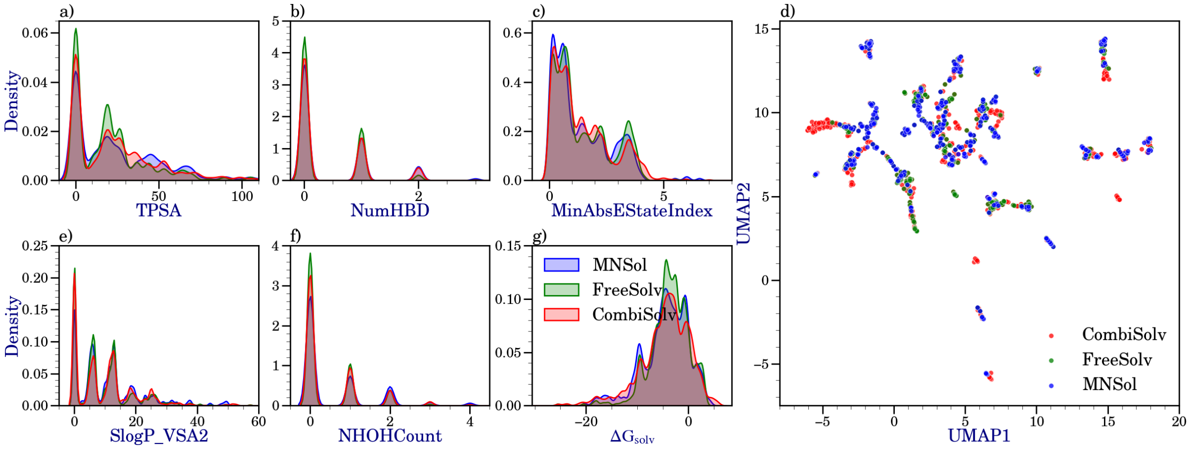

Figure S3: Distribution plots of key features contributing to the feature vector for analyzing transferability issues. (a)–(c) and (e)–(f) show the distributions of important features, while (g) presents the distribution of the target property, free energy of solvation (ΔG_{solv}) for each dataset considered. (d) displays a lower-dimensional projection of high-dimensional feature vectors, generated using UMAP and plotted along two UMAP embedding axes to visualize the data spread and clustering patterns.

The Encoder utilizes multiple Relational Graph Convolutional Layers (R-GCN)[2] followed by dense layers. The R-GCN captures atom and bond relationships, assigning distinct weight matrices for different bond types. A global average pooling layer aggregates node features into a single vector, which is then passed through dense layers to produce the latent space parameters—mean (z_mean) and log variance (log_var).

The Decoder reconstructs the adjacency matrix and node features from the latent vector. Dense layers process the latent representation, and two output layers generate the adjacency matrix and node features. Softmax activation ensures valid probability distributions for these outputs.

Loss Function:

The total loss comprises multiple components: adjacency and feature reconstruction losses (measured using cross-entropy), KL divergence loss (to enforce a standard normal distribution in the latent space), and binary cross-entropy loss for molecular property prediction. An optional gradient penalty may be included for regularization and training stability.

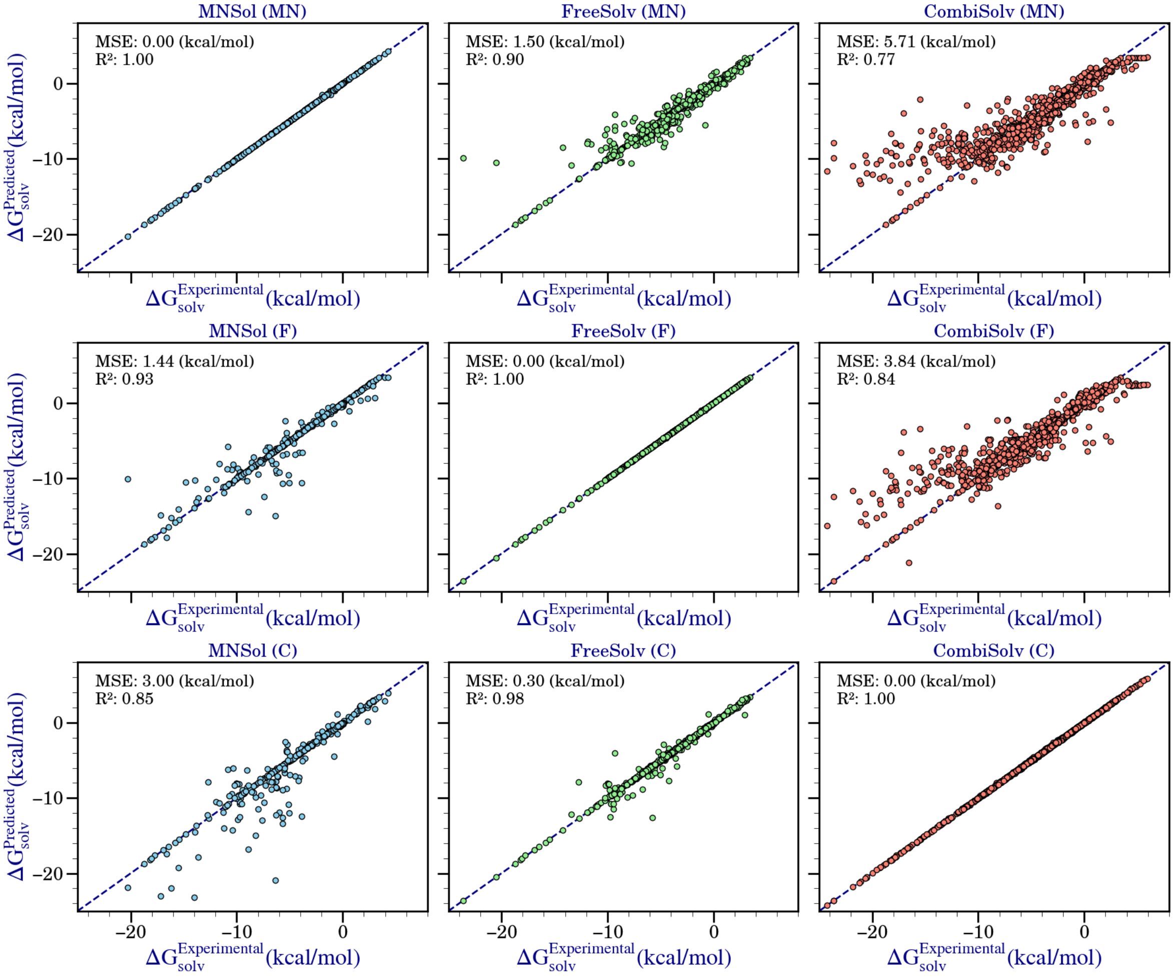

Figure S4: Popular ML models often overfit to training data, leading to reduced performance on evaluation datasets. Here, the XGBoost model is trained on each dataset (rows) and tested on the other two datasets (columns) to assess its generalization ability.

Hyperparameters:

The model is defined by several hyperparameters, including the maximum number of atoms (NUM_ATOMS), bond types (BOND_DIM), atom feature dimensions (ATOM_DIM), graph convolution output sizes (gconv_units), dense layer dimensions (dense _units), latent space dimension (latent_dim), dropout rate (dropout_rate), training epochs (epochs), and optimizer learning rate (learning_rate).

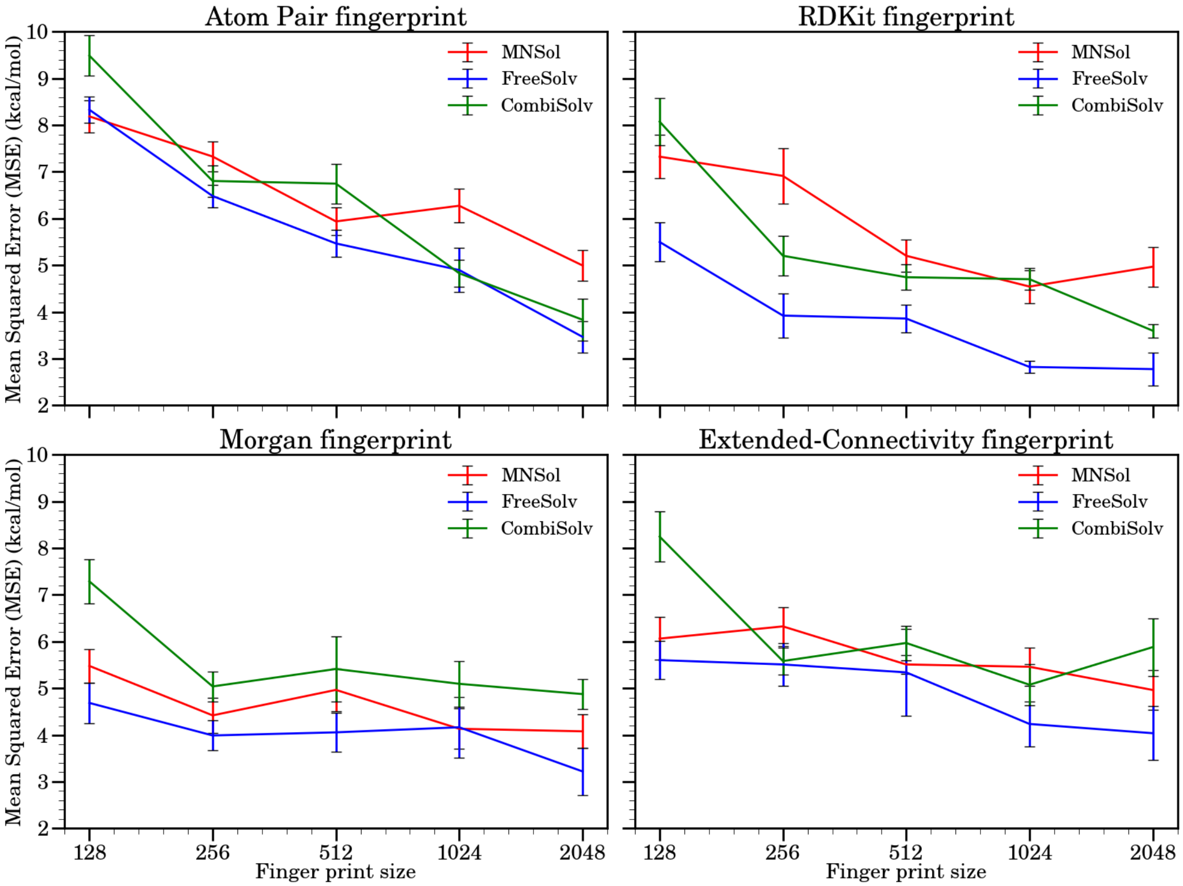

Figure S5: By varying the bit size of the fingerprints we found optimal size of fingerprints. A XGBoost model was used to evaluate the efficiency of the FPs. 10-fold cross validation was employed for reliable statistics. The standard error in measurement is provided in terms of error bar.

Training Loop

The training process begins with a forward pass to generate the latent vector, adjacency matrix, and node features. Loss components are computed and combined, followed by a backward pass to calculate gradients. Model parameters are updated through an optimization step. Validation is conducted regularly to assess generalization on unseen data.

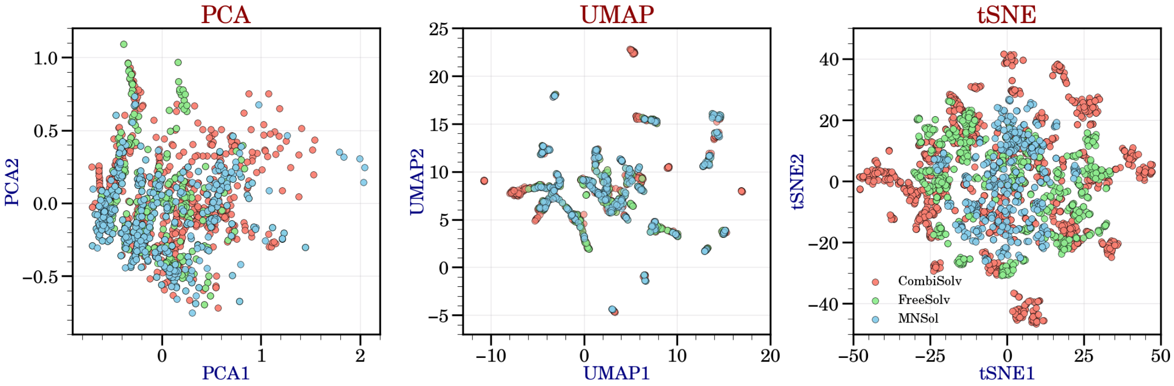

Figure S6: Molecular entries(represented in terms of descriptor) in the 2D space of PCA, UMAP and tSNE.

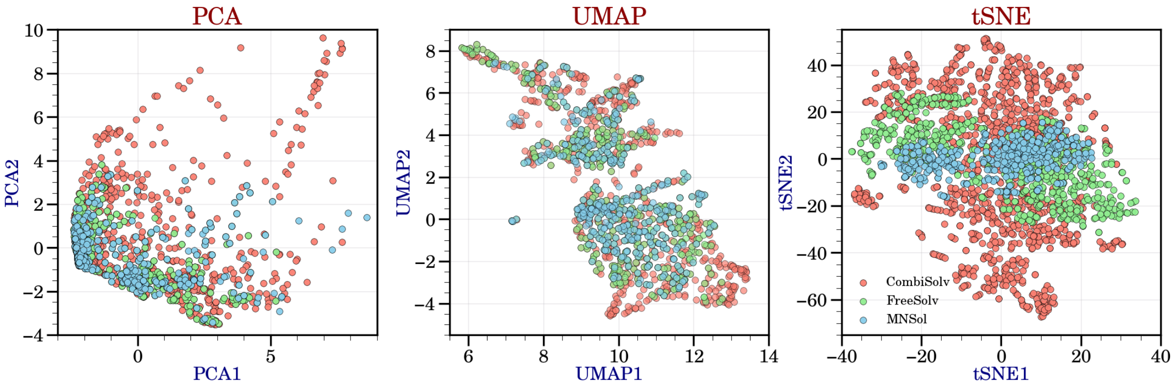

Figure S7: Molecular entries(represented in terms of atom-pair fingerprints) in the 2D space of PCA, UMAP and tSNE.

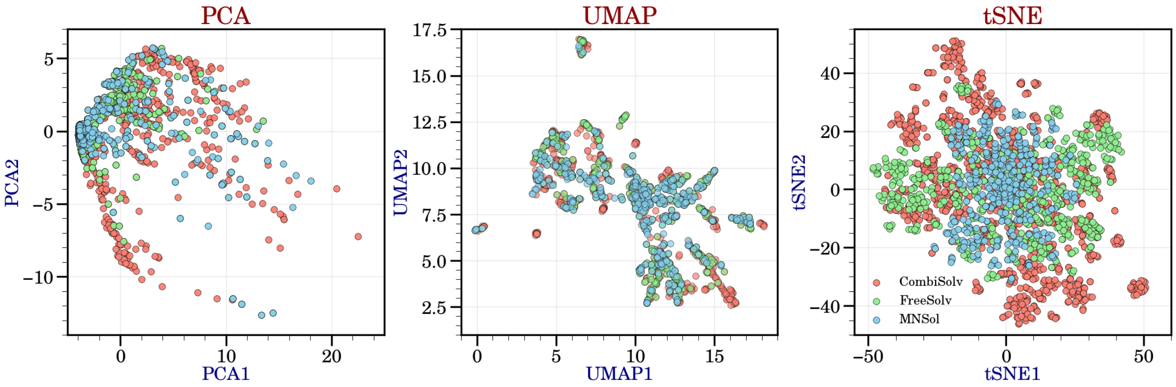

Figure S8: Molecular entries(represented in terms of RDKit fingerprints) in the 2D space of PCA, UMAP and tSNE.

Dimensionality reduction of feature space

References

(1) Wang, M.; Zheng, D.; Ye, Z.; Gan, Q.; Li, M.; Song, X.; Zhou, J.; Ma, C.; Yu, L.; Gai, Y.; Xiao, T.; He, T.; Karypis, G.; Li, J.; Zhang, Z. Deep Graph Library: A

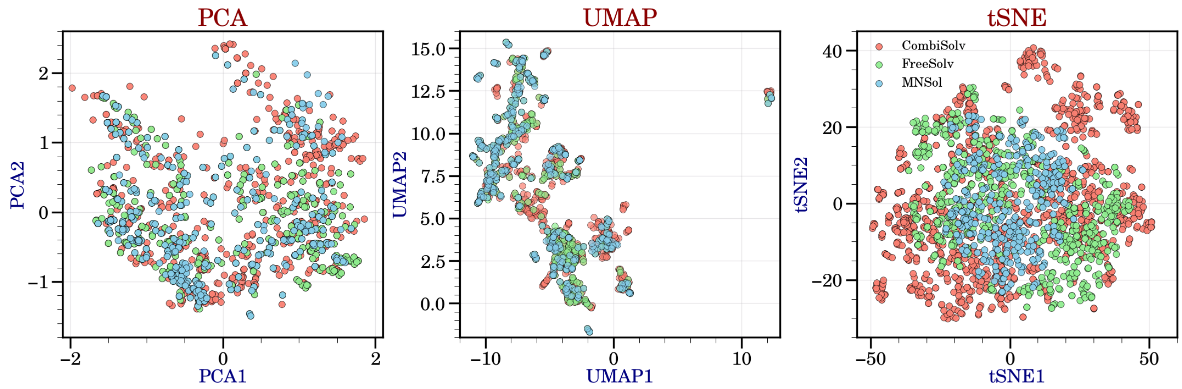

Figure S9: Molecular entries(represented in terms of Morgan fingerprints) in the 2D space of PCA, UMAP and tSNE.

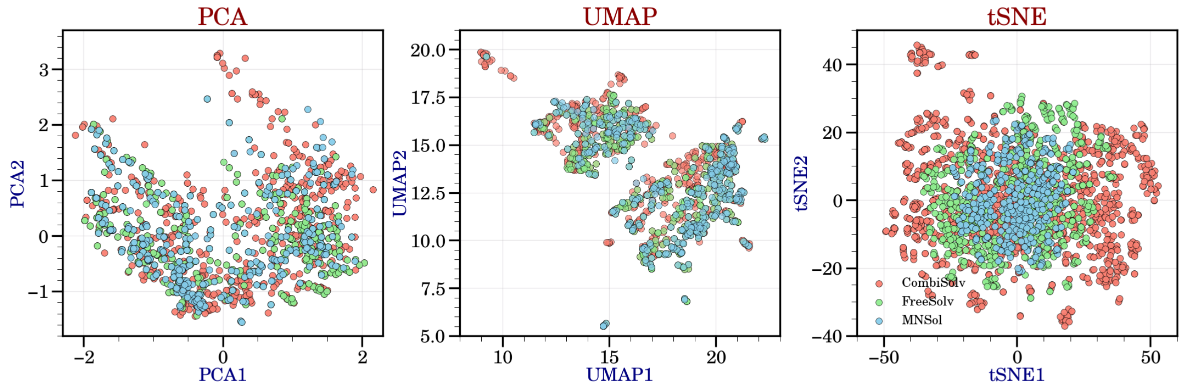

Figure S10: Molecular entries(represented in terms of extended connectivity fingerprints(ECFPs)) in the 2D space of PCA, UMAP and tSNE.

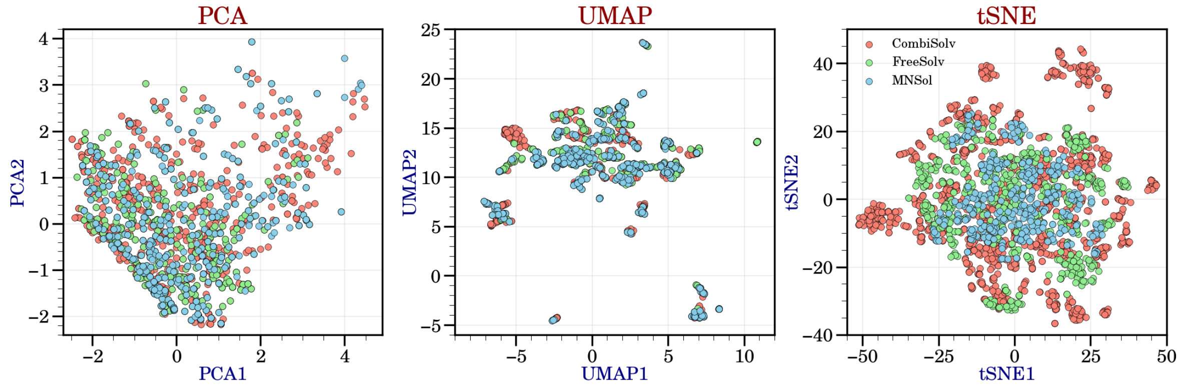

Figure S11: Molecular entries(represented in terms of MACCS keys) in the 2D space of PCA, UMAP and tSNE.

Graph-Centric, Highly-Performant Package for Graph Neural Networks. *arXiv preprint arXiv:1909.01315* **2019**,

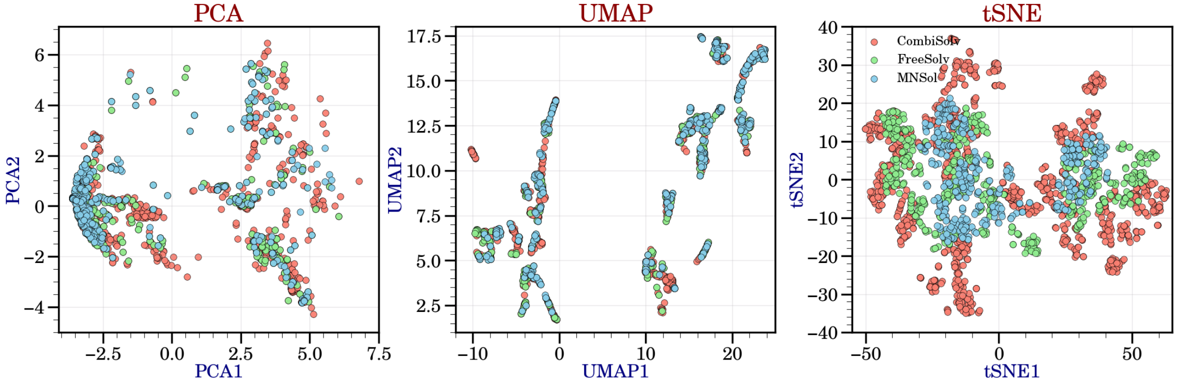

Figure S12: Molecular entries(represented in terms of PubChem keys) in the 2D space of PCA, UMAP and tSNE(color codes follow the legend of last subplot).

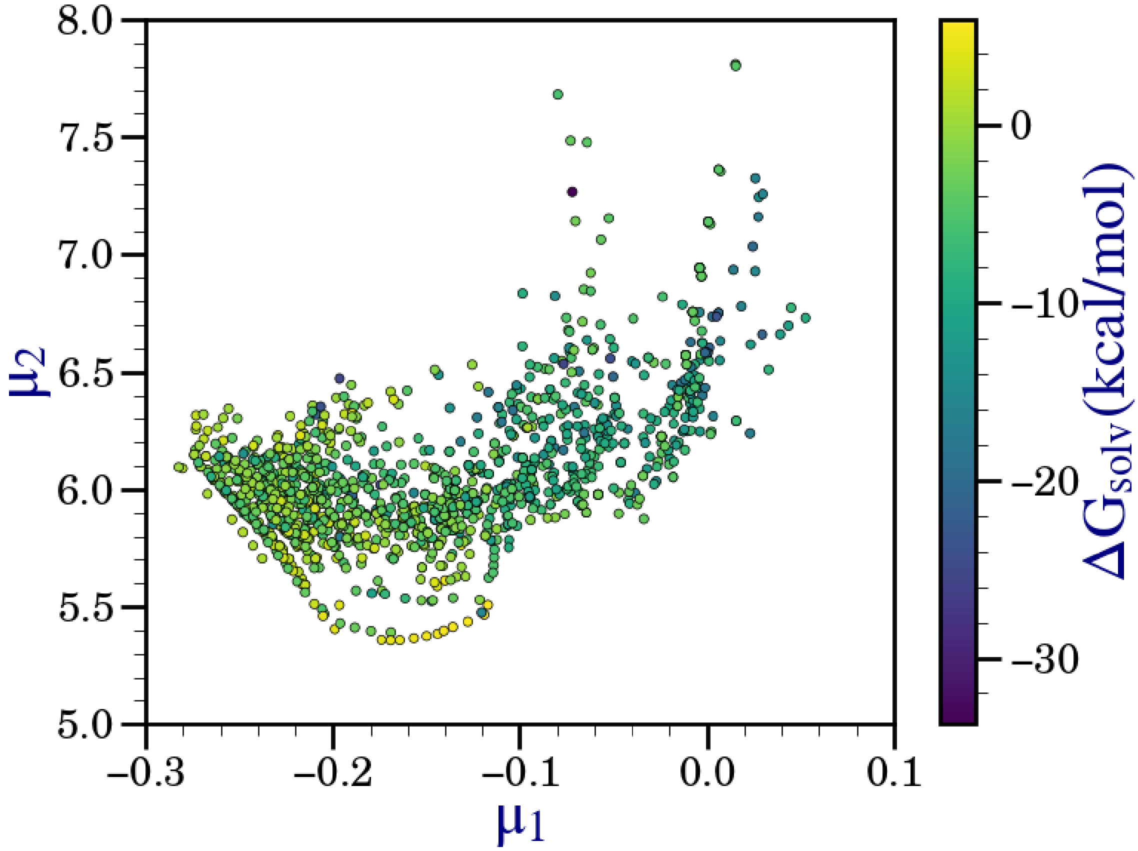

Figure S13: Molecular entries in the 2D latent space of trained GCN-VAE(merged dataset). Molecules are colored by the value of their free energy of solvation.

(2) Schlichtkrull, M.; Kipf, T. N.; Bloem, P.; Berg, R. v. d.; Titov, I.; Welling, M. Modeling Relational Data with Graph Convolutional Networks. 2017; `https://arxiv.org/abs/`

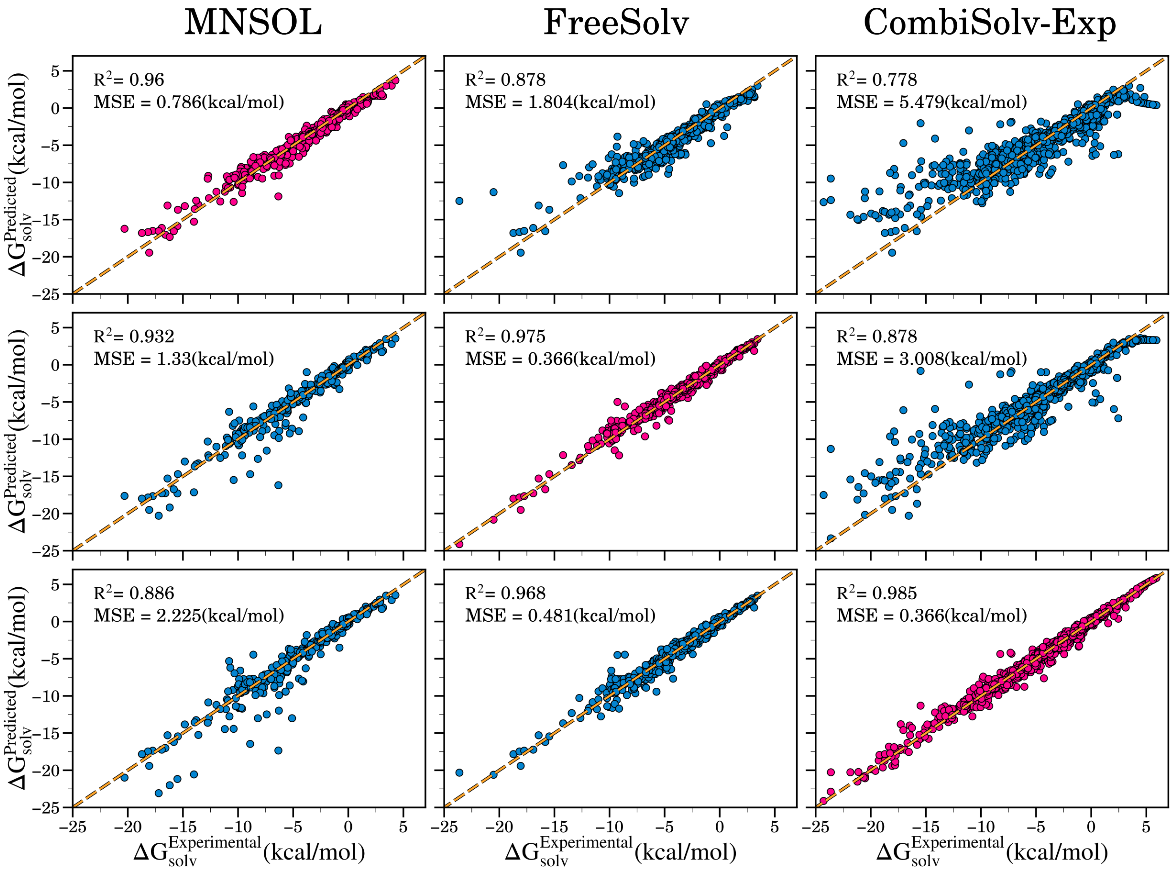

Figure S14: Dataset dependency of CIGIN model. Each column represents a dataset, while each row corresponds to a model trained on the same dataset. In the top panel, the CIGIN model is trained on the MNSol dataset and tested on the other two datasets. The middle panel shows the model trained on FreeSolv, and in the lower panel, the model is trained on CombiSolv before being tested across datasets.

1703.06103.

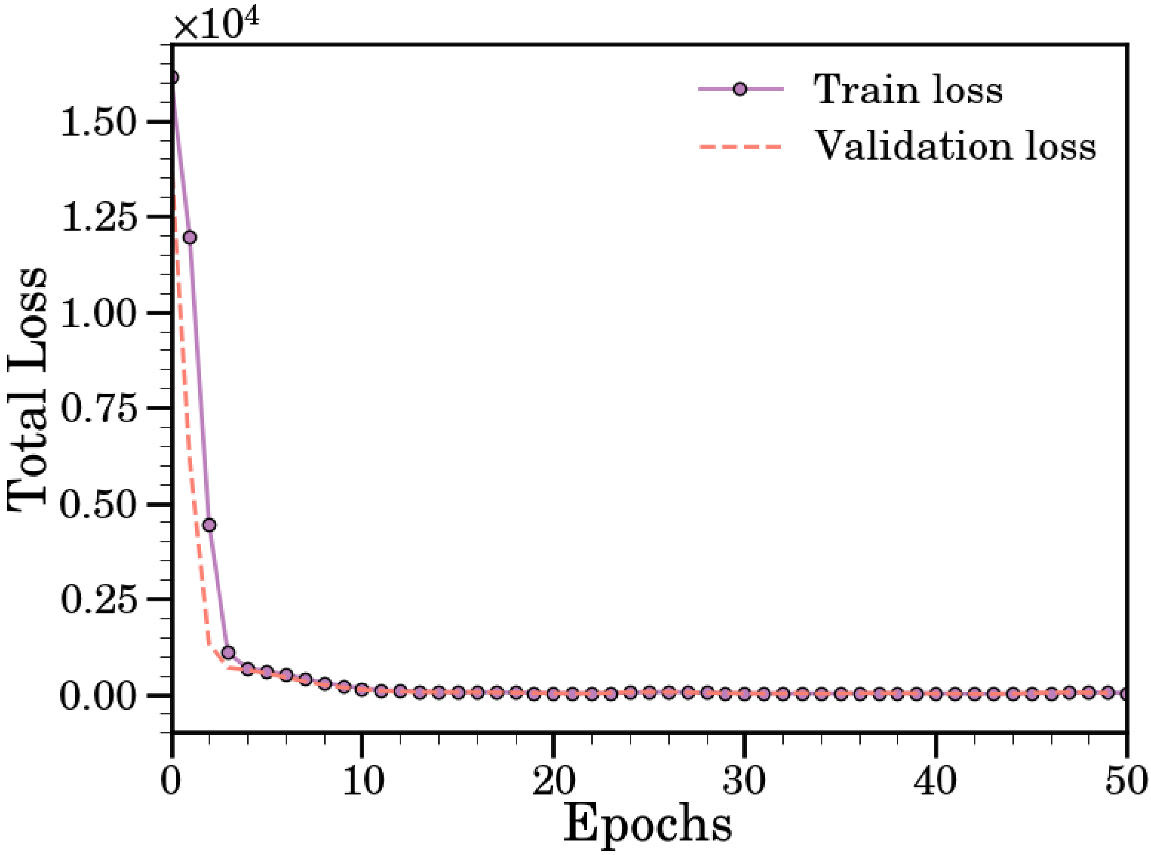

Figure S15: Loss function vs epoch for GCN-VAE showing steady convergence.

Table S2: Trained MLP regressor(MLPR), XGBoost regressor(XGBR) and Random forest regressor(RFR) on FreeSolv dataset.

Fingerprint	MLPR		RFR		XGBR	
	R^2	MSE	R^2	MSE	R^2	MSE
AP	0.751	2.93	0.633	3.41	0.719	3.19
RDK	0.741	2.70	0.743	3.01	0.762	2.62
Morgan	0.771	2.73	0.541	4.43	0.786	2.56
ECFP	0.573	4.21	0.344	5.86	0.666	3.33
MACCS	0.920	1.20	0.868	1.63	0.910	1.14
PubChem	0.910	1.22	0.857	1.82	0.910	1.14

Table S3: Trained MLP regressor(MLPR), XGBoost regressor(XGBR) and Random forest regressor(RFR) on CombiSolv dataset.

Fingerprint	MLP		RFR		XGB	
	R^2	MSE	R^2	MSE	R^2	MSE
AP	0.854	2.77	0.763	4.36	0.847	3.19
RDK	0.822	3.36	0.824	3.25	0.837	3.71
Morgan	0.847	3.04	0.707	5.45	0.778	4.29
ECFP	0.766	4.14	0.648	6.28	0.765	4.69
MACCS	0.921	1.78	0.893	2.22	0.914	1.83
PubChem	0.911	1.87	0.884	2.28	0.914	1.83

Table S4: The atom (node) features used for molecular graph representation

Atom Feature	**Description**
Atom Type	Element identity (H, C, N, O, F, etc.) represented using one-hot encoding
Implicit Valence	Presence of implicit valence electrons (Binary)
Radical Electrons	Presence of radical electrons (Binary)
Chirality	Chirality configuration: R, S, or None (one-hot)
Number of Hydrogens	Number of neighboring hydrogen atoms (one-hot)
Hybridization	Hybridization state: sp, sp^2, sp^3, sp^{3}d (one-hot)
Acidic	Atom is acidic in nature (Binary)
Basic	Atom is basic in nature (Binary)
Aromatic	Atom is part of an aromatic group (Binary)
Donor	Donates electrons (Binary)
Acceptor	Accepts electrons (Binary)

Table S5: The bond (edge) features used for molecular representation

Bond Feature	**Description**
Bond Type	Bond order: single, double, triple, or aromatic (one-hot)
Bond is in Conjugation	Indicates if the bond is part of a conjugated system (Binary)
Bond is in Ring	Indicates if the bond is part of a ring structure (Binary)
Bond Chirality	Stereochemistry of bond: E or Z (one-hot)