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

Dibyendu Maity
S.N. Bose National Centre for Basic Sciences
Suman Chakrabarty
sumanc@bose.res.in

S.N. Bose National Centre for Basic Sciences

Research Article

Keywords:

Posted Date: July 23rd, 2025

DOI: https://doi.org/10.21203/rs.3.rs-6727155/v1

Additional Declarations: No competing interests reported.

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

Abstract

In-silico prediction of physicochemical properties such as solvation free energy is crucial for efficient drug discovery. However, accurate prediction remains challenging due to complexities inherent in molecular representations and model transferability. This study systematically evaluates the influence of different molecular representations, namely descriptor-based, fingerprint-based and graph-based, on the predictive performance and transferability of supervised machine learning (ML) models. Using three diverse datasets (MNSol, FreeSolv, and CombiSolv), we compared classical regression techniques (XGBoost, Random Forest, Support Vector Regression, Kernel Ridge Regression) against deep learning models, specifically the Chemically Interpretable Graph Interaction Network (CIGIN). Our findings indicate that while traditional models with interpretable descriptors provide insights into the important features, their transferability is limited by dataset size and chemical diversity. Molecular fingerprints show

improved performance, and a Multilayer Perceptron (MLP) Regressor demonstrates better regularization with high-dimensional fingerprints compared to traditional models. The graph-based CIGIN model exhibits strong performance and chemical interpretability but faces challenges in generalizing to novel chemical entities absent in the training data, showing increased errors for molecules with long hydrocarbon chains or polyol moieties. This research highlights the critical interplay between data quality, molecular representation, and model choice in achieving accurate and transferable predictions of molecular properties, underscoring the need for further refinement in handling novel chemical space and incorporating physics-informed features.

Scientific Contribution

This study provides a systematic comparative analysis of diverse molecular representations, highlighting their impact on predictive accuracy and model transferability in supervised learning of solvation free energies. By rigorously evaluating classical and deep learning models using multiple chemically diverse datasets, we uncover specific strengths and critical limitations in current methodologies, particularly emphasizing issues with transferability and dataset overlap. These insights will guide the future developments of robust, interpretable, and generalizable ML models for molecular property prediction and computational drug discovery.

Introduction

Efficient computational prediction of molecular properties, such as solvation free energy, is a cornerstone of modern drug discovery, offering significant reduction in time and resources. However, accurately predicting these properties remains a formidable challenge due to the intricate interplay between molecular features, computational complexity, and dataset limitations.[1–3] A fundamental aspect of predictive modeling in cheminformatics involves the

appropriate selection and representation of molecular data, directly influencing the effectiveness and generalizability of machine learning (ML) models.

In-silico, there are broadly two approaches to compute/predict physicochemical properties of chemical compounds: first principle methods, e.g. density functional theory (DFT) or other advanced electronic structure methods and molecular dynamics (MD) simulations based on classical Newtonian and statistical mechanics.[4–9] While these methods have solid theoretical foundations and can accurately calculate various molecular properties, they are computationally expensive and time-intensive due to the extensive numerical calculations involved.[7,10] As a result, there is a growing need for data-driven approaches to address these challenges efficiently.

In the context of computer-aided drug discovery, accurate prediction of the pharmacophore and pharmacokinetic properties to refine candidates is crucial.[11–22] Pharmacophore properties describe the spatial arrangement of chemical features, such as size, shape, and charge distribution, that are essential for binding to a biological target and eliciting a desired response. Wermuth et al. (1998) likened pharmacophores to a "fingerprint" responsible for a drug's biological effect.[23–28] Pharmacokinetics, on the other hand, involves a drug's Absorption, Distribution, Metabolism, and Excretion (ADME) kinetics. While binding affinity is crucial, a drug's pharmacokinetic profile determines whether it reaches the target site at sufficient concentrations with minimal safety concerns, ultimately influencing its clinical viability.[29–32]

Recent advancements in machine learning (ML) have revolutionized cheminformatics, enabling large-scale property predictions. Quantitative structure–activity relationship (QSAR) and quantitative structure–property relationship (QSPR) models predict molecular properties based on empirical or structural features.[33–38] These models use two main mathematical functions: an encoding function, which encodes the chemical structure into a molecular descriptor, and a mapping function, which predicts the target property or activity using the encoded descriptor.[39] Numerous databases of experimentally measured pharma-

cophore and pharmacokinetic properties have been compiled—many of which are publicly available—leading to the development of QSAR models trained on these datasets.[40–48]

Advancements in machine learning (ML) modeling, coupled with the availability of extensive chemical and biological datasets, have created a critical need to translate raw data into machine-readable formats before model training. Chemical data can be represented in various forms, including empirical, molecular, and structural formats such as molecular graphs, fingerprints, and descriptors.[49,50]

Traditionally, molecular descriptors, which are easily interpretable physicochemical parameters, have provided insights into structure-property relationships. Nevertheless, they often fall short in capturing the complex chemical environments influencing properties like solvation energy. Alternatively, molecular fingerprints encode substructural information into fixed-length vectors, striking a balance between interpretability and flexibility but occasionally struggling with capturing novel chemical motifs.

Traditional regressors and simple artificial neural networks (ANNs), when trained on these representations, have demonstrated success in predicting a range of molecular properties, often providing computationally efficient alternatives to first-principle methods. However, these models typically depend on recognizing specific substructural patterns, which restricts their ability to generalize beyond the training set. As a result, their performance often deteriorates when applied to new datasets, posing challenges for transferability.

Deep learning has significantly improved molecular property prediction due to its ability to model complex, nonlinear relationships between molecular features and properties. Recent methodological advancements, such as delta-learning and transfer learning, have further enhanced prediction accuracy by integrating domain knowledge and physics-based constraints into the models.[51–56] Among the most impactful developments is the message passing neural network (MPNN), which has transformed supervised learning for chemical property prediction.[57] This has led to a surge in deep learning models that represent molecules as graphs, capturing atomic interactions more effectively.

Several cutting-edge models have emerged in this space, including: Delfos, which predicts solvation free energy with high precision,[58] MLSolvA, an improved architecture incorporating pairwise atomistic interactions[59] etc. Recently, graph-based representations, coupled with deep learning models such as Chemically Interpretable Graph Interaction Network (CIGIN),[60] have demonstrated substantial potential, effectively encoding atomic-level interactions while maintaining chemical interpretability.

Deep learning models excel at detecting patterns, but their performance relies heavily on data quality and quantity. Biases in training data can lead to overfitting, where a model performs well on familiar data but fails to generalize to new molecules. Interpretability is another major challenge. Despite their predictive power, many ML models operate as black boxes, making their decision-making unclear. To address this, researchers are incorporating domain knowledge into model architectures and developing explainable AI (XAI) techniques to enhance transparency.

Despite these advances, critical challenges persist in transferring learned models across chemically diverse datasets. Performance often diminishes significantly when models encounter novel chemical structures outside their training data distributions. Thus, systematically evaluating different molecular representations and modeling strategies to enhance predictive accuracy and transferability is imperative. For example, Llompart et al. have raised the question "Will We Ever Be Able to Accurately Predict Solubility?", stressing the importance of external validation to the trained models.[61] They highlight how models that perform exceptionally well in-house often fail on newly compiled datasets, raising concerns about data leakage in complicated models and whether deep learning methods are truly learning or merely memorizing.

This study addresses these challenges by comparing descriptor-based, fingerprint-based, and graph-based representations across classical ML algorithms and state-of-the-art deep learning approaches, providing insights to guide the development of robust, generalizable computational models in drug discovery.

Computational details

Datasets

Numerous well-curated datasets are publicly available for developing accurate predictive models. In this work, we utilize three such datasets, dedicated for solvation free energy: the Minnesota Solvation Database (MNSol),[62] the FreeSolv Database,[63] and the CombiSolv-Exp-8780 (CombiSolv).[64,65] The MNSol dataset originally contained 3,037 experimental solvation free energies across 790 solutes in 92 solvents. To specifically focus on water's solvation effects, we filtered it down to 390 water-based entries.

The FreeSolv dataset comprises 641 unique solute entries, all in water. CombiSolv, the largest of the three, initially included solvation free energies for 10,145 solvent–solute pairs, compiled by Vermeire et al.[64,65] The publicly available version contains 8,780 entries, including 1,153 water-based samples. To investigate the impact of training data size, we selected datasets spanning a range of scales—from the smallest (MNSol) to the largest (CombiSolv). Despite variations in size and chemical entity overlap (Figure 1A), all datasets exhibit a broad functional group distribution (Figure 1B), making them well-suited for exploring the influence of dataset scale on model performance. After rigorous curation and merging of the water-based entries from all three datasets, we constructed a unified benchmark dataset comprising 1,333 entries. Dataset quality was thoroughly validated, and the models were subjected to robust intra- and inter-dataset cross-validation to ensure generalizability and performance consistency.

Molecular representations

In all the datasets used in this study, chemical compounds are represented using SMILES (Simplified Molecular Input Line Entry System) strings, alongside the target property—free energy of solvation(in kcal/mol). However, most traditional machine learning models cannot directly interpret SMILES strings, necessitating their conversion into a numerical format.

While significant progress has been made in molecular representation, a universally optimal encoding remains elusive. In this study, we employ three distinct approaches tailored to different model requirements: descriptors, fingerprints, and graph-based representations.

Descriptor-based modeling represents chemical compounds as feature vectors derived from measurable physicochemical properties calculated from their SMILES strings. This approach enhances interpretability by leveraging domain knowledge while improving predictive efficiency.

In this study, we adopted a comprehensive approach by computing 217 molecular descriptors using RDKit.[66] These descriptors include commonly used features such as: Topological polar surface area (TPSA), hydrogen bond donors/acceptors, molecular weight, fraction of different functional groups, Quantitative Estimation of Drug-likeness (QED), logP (octanol-water partition coefficient), etc. A full list is provided in Table S1 (Supporting Information).

To ensure feature quality and reduce redundancy, we carried out the following preprocessing steps. Descriptors with zero standard deviation (no variability) were discarded. Highly correlated descriptors were identified and removed using Pearson correlation analysis. The cutoff threshold was varied to evaluate its influence on model performance. We found (Figure S1) that a correlation threshold of 0.75 yields optimal performance, particularly when using ensemble models like XGBoost Regressor (XGBR)[67] and Random Forest Regressor (RFR).[68]

Unlike descriptors, molecular fingerprints offer a more localized, pattern-based encoding of structural features into fixed-length binary vectors. Fingerprints are particularly useful for encoding substructures or fragment-level information and fall into two broad categories: structural keys and hashed fingerprints.

Structural keys represent a molecule using a predefined set of fragments or substructures, where each bit corresponds to the presence or absence of a specific feature. MACCS (Molecular ACCess System) keys are one of the most commonly used structural keys.[69] There are two sets—one with 960 keys and a publicly available subset of 166 keys (which we used

in this study). PubChem fingerprints are 881-bit-long structural keys used by PubChem for similarity searches and structure neighboring.[70] While interpretable, structural keys are inherently limited by the predefined fragment library and may miss novel or uncommon features. Hashed fingerprints, on the other hand, generate molecular fragments dynamically (typically up to a specified radius or path length) and use hash functions to map them to a fixed-length binary vector. These offer greater coverage and flexibility compared to structural keys.

In this study we have compared the efficiency of four popular hashed fingerprints. They are Atom-Pair (AP) Fingerprints,[71] RDKit(RDK) Fingerprints,[66] Morgan Fingerprints,[72] Extended Connectivity Fingerprints (ECFPs),[73] a widely used variant of Morgan fingerprints. To determine an appropriate fingerprint size, we varied fingerprint lengths and evaluated model performance. Our results (Supporting Information, Figure S5) indicate that a fingerprint size of 2048 bits offers the best trade-off between information richness and model accuracy.

Graph-based representations of molecular structures have gained significant popularity due to their ability to efficiently encode complex chemical information in a format well-suited for deep learning models. A graph is a mathematical structure composed of nodes (representing atoms) and edges (representing chemical bonds), making it a natural fit for modeling molecular structures. In a graph, each node is assigned a feature vector encoding atomic properties, typically using a one-hot encoding scheme. These vectors collectively form a node feature matrix. The edges, which define the relationships between atoms, are represented in an adjacency matrix($\mathbf{a}$), where each element a_{ij} is 1 if atoms i and j are bonded, and 0 otherwise.

For this study, we followed the graph construction methodology prescribed in the CIGIN model,[74] converting SMILES strings into molecular graphs. Each atom (node) is annotated with a predefined set of features (Table S4), while bonds (edges) are assigned feature vectors tailored to capture bond-specific information (Table S5). This structured representation

enables the model to learn both atomic and bonding interactions effectively, enhancing predictive performance.

Dimensionality reduction methods

While high-dimensional or complex molecular representations provide a rich foundation for machine learning (ML) and deep learning (DL) models, they pose challenges for human interpretation. Our cognitive limitations make it difficult to intuitively grasp patterns in high-dimensional spaces. To break the curse of dimensionality, we apply dimensionality reduction techniques that project the data into a lower-dimensional, coarse-grained space where patterns and clusters become more visually and analytically accessible. In this study, we evaluated and compared three widely used dimensionality reduction methods for fingerprints and descriptors:

- Principal Component Analysis (PCA) — a linear projection technique,[75]
- t-Distributed Stochastic Neighbor Embedding (t-SNE) — a non-linear method focused on preserving local structure,[76–78]
- Uniform Manifold Approximation and Projection (UMAP) — a non-linear technique grounded in manifold learning and topological data analysis.[79]

Using the merged dataset, we applied these methods to project molecular representations into 2D space. This enabled us to visualize the spread, clustering, and overlap among chemical entities across different datasets(Figure 1C and D). Among the three techniques, t-SNE provided the clearest separation of molecular clusters, especially in larger datasets, making it more suitable for qualitative assessment in this context.

However, when working with graph-based molecular representations, direct application of traditional dimensionality reduction is not feasible due to their non-Euclidean structure. To address this, we employed a Graph Convolutional Neural Network–based Variational Autoencoder (GCN-VAE) to learn meaningful low-dimensional embeddings from molecular

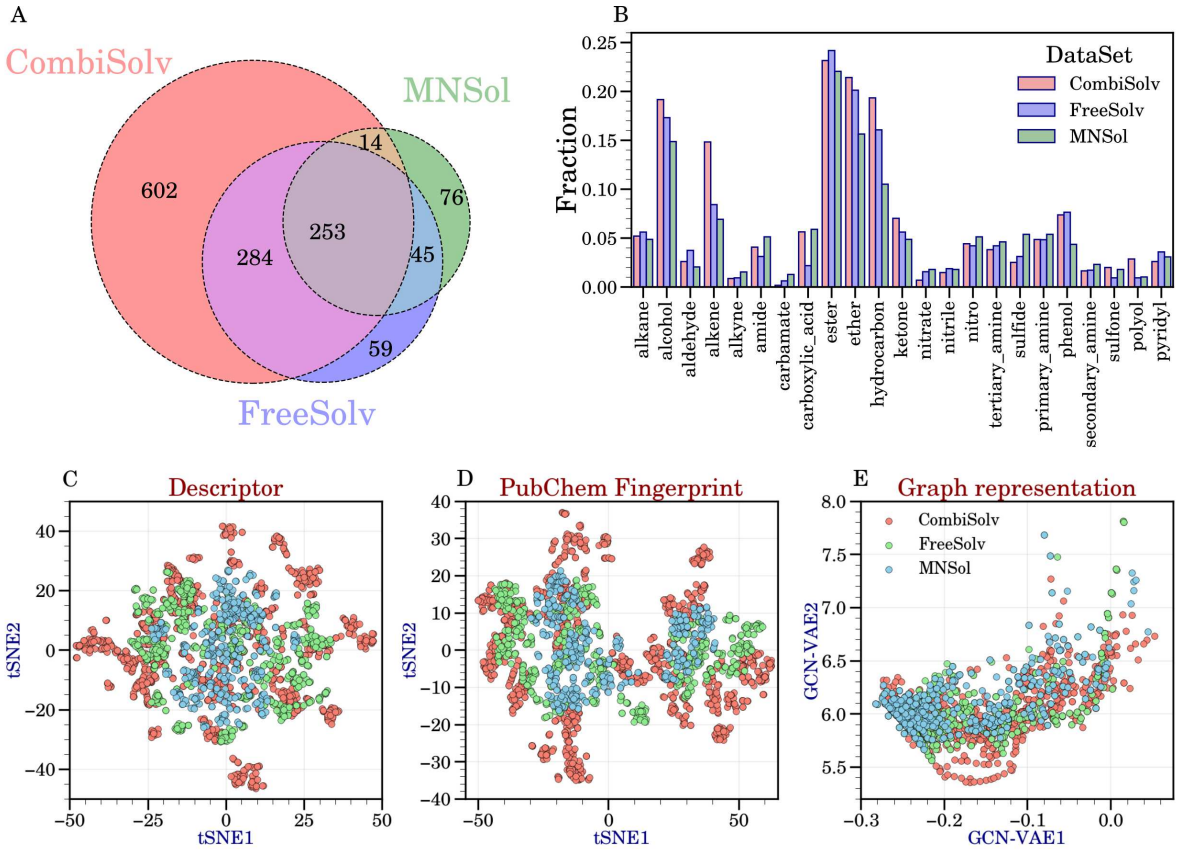


Figure 1: (A) A Venn diagram illustrating the overlap of chemical entities among the three datasets used. (B) A bar plot comparing the populations of different functional groups within each dataset. (C) Spread of molecules in tSNE space of descriptor and (D) 881 bit PubChem fingerprint. (E) Molecules in 2D latent space of GCN-VAE model. The color code for datasets remains same for each sub figure.

graphs. The GCN-VAE architecture operates on graph data (implemented using DGL[80]) derived from SMILES strings and consists of two main components:

- Encoder: Composed of multiple Relational Graph Convolutional Network (R-GCN)[81] layers followed by dense layers. The R-GCNs capture atom-bond relationships by assigning distinct weight matrices to different bond types. A global average pooling layer aggregates the node-level features into a graph-level representation, which is then transformed into latent space parameters—mean (z_{mean}) and log-variance (log_{var}).

- Decoder: Responsible for reconstructing the original graph structure and molecular features from the latent space.

The GCN-VAE is trained by minimizing a composite loss function, which includes:

- Reconstruction losses for both node features and adjacency matrix (measured via cross-entropy),
- Kullback-Leibler (KL) divergence loss[82] to regularize the latent space toward a standard normal distribution,
- Binary cross-entropy loss for auxiliary property prediction,
- An optional gradient penalty to improve training stability and prevent overfitting.

In the latent space of trained GCN-VAE model the richness of CombiSolv is prominent(Figure 1E). Comparison of datasets in every reduction tequnique applied to each representation is provided in Supporting information(Figure S6 - Figure S12)

ML and DL Models

We systematically compared several classical and widely adopted regression models for predicting solvation free energies based on abovementioned molecular feature representations. Preliminary analyses revealed that simple linear regression was insufficient, likely due to the inherently non-linear relationships between molecular descriptors and solvation properties. This observation motivated the adoption of more advanced machine learning algorithms capable of capturing these complex, nonlinear structure–property interactions. To model solubility behavior more effectively, we employed two powerful ensemble-based learning techniques, each with distinct advantages:

- XGBoost, a cutting-edge gradient boosting framework, has consistently demonstrated superior performance in recent solubility prediction tasks.[67,83–85] It operates by sequentially refining an ensemble of decision trees, allowing for effective handling of missing values, outliers, and large datasets with high dimensionality. One of its key strengths

lies in its built-in feature importance analysis, which offers mechanistic insight into the relationship between molecular structure and solvation behavior.

- Random Forest Regression (RFR), in contrast, builds an ensemble of uncorrelated decision trees in parallel, then averages their outputs to make predictions.[68] This approach is particularly robust when dealing with high-dimensional feature spaces and mixed data types, and it shows strong resistance to overfitting and noise in experimental data.

In parallel, we evaluated two kernel-based models: Kernel Ridge Regression (KRR)[86,87] and Support Vector Regression (SVR).[88] KRR blends the regularization benefits of ridge regression with the flexibility of kernel methods, making it well-suited for moderately sized datasets and smooth, non-linear patterns—even in the presence of noise. SVR, on the other hand, transforms input features into higher-dimensional spaces using kernel functions, allowing it to model complex, non-linear relationships. Its maximal margin approach helps it generalize well, especially in high-dimensional settings and with limited data. Thanks to its mathematical rigor and stability, SVR is effective at capturing subtle patterns in structure–property relationships. Among all models tested, both XGBoost and SVR consistently delivered strong results, showing high predictive accuracy and adaptability across different molecular representations. This comparison highlights the complementary strengths of ensemble and kernel-based methods for solvation energy prediction.

To test the performance of a deep learning approach against traditional models, we evaluated the CIGIN model (Chemically Interpretable Graph Interaction Network), which processes molecular structures, in terms of grpah, through three intuitive stages. First, the message-passing phase treats molecules as interconnected networks(graphs) where atoms(nodes) continuously update their representations by exchanging information with bonded(edge) neighbors. The interaction phase then examines how atoms of solute and solvent molecules influence each other. The model creates a detailed map of pairwise relationships, assessing how strongly each solute atom interacts with each solvent atom. Rather than treating

molecules as single entities, this approach captures the atomic-level interactions that may influence solubility. Finally, the prediction phase synthesizes all this information. The model combines the refined atomic representations(graph) with the interaction patterns, processing them through a series of neural network layers to estimate the solvation energy. What makes CIGIN particularly valuable is its ability to maintain chemical interpretability - while complex, the model's architecture preserves meaningful relationships between molecular features and the final prediction, unlike traditional black-box approaches. This three-stage design allows the model to capture both local atomic environments and global molecular interactions that govern solvation behavior. We have adapted the original idea of CIGIN as described in the paper[37] and used the code provided at https://github.com/devalab/CIGIN.

Hyperparameter tuning of the models

The performance of machine learning models critically depends on selecting appropriate hyperparameters. We conducted systematic optimization for each algorithm using exhaustive grid search (GridSearchCV) and randomized parameter sampling (RandomSearchCV) techniques. These methods evaluate multiple parameter combinations through cross-validation to identify optimal configurations that maximize predictive accuracy while preventing overfitting.

For traditional machine learning models including XGBoost, Random Forest, and kernel-based methods, we tuned key parameters such as tree depth, learning rates, regularization terms, and kernel coefficients. The CIGIN model demonstrated robust performance with its default configuration, as the original authors' parameter choices already represented an optimized balance between model complexity and generalization capability.

Metric to evaluate model performance

For each model tested, statistical metrics are computed: the squared correlation coefficient (R^2) and the Mean-Squared Error (MSE). These metrics are defined as:

$$R^2 = 1 - \frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2} \tag{1}$$

$$MSE = \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2 \tag{2}$$

where y_i represents the experimental data, $\hat{y}_i$ represents the model-predicted values, and $\bar{y}$ is the mean of the experimental data. Model performance was evaluated by training each model on a training set and then testing it on a test set, created by splitting the dataset at an 8:2 ratio. To ensure the robustness of the models, k-fold cross-validation was performed. We have used scikit-learn library to construct the classical regressor and to analyze the efficiency of the models.[89]

Results and discussion

Having established the molecular representations and committed to our set of machine learning models, we systematically evaluated their performance across all the datasets. We begin by discussing the descriptor-based modeling of chemical compounds.

Interpretable descriptors has limited information

Following rigorous preprocessing that involves eliminating descriptors with zero variance and removing highly correlated or redundant features, we reduced the feature space to a refined set of 116 descriptors. This final feature set was used to train and validate classical machine learning models using k-fold cross-validation, ensuring robust and unbiased performance estimates.

When trained individually on each dataset, the models achieved reasonably strong predictive performance, with Support Vector Regression (SVR) and XGBoost Regressor (XGBR) consistently emerging as top performers(Figure 2(a) and (b)). Interestingly, the FreeSolv dataset produced the best results, despite not being the largest. While the MNSol dataset underperformed—likely due to its limited size—the most surprising outcome came from CombiSolv, the largest dataset, which also lagged behind FreeSolv in predictive accuracy. This observation highlights a crucial point: data quality often outweighs quantity, and there may exist a threshold beyond which additional noisy data contributes little to model improvement.

One of the advantages of descriptor-based modeling is interpretability. By analyzing the relative feature importance from XGBR, we found that topological polar surface area (TPSA) alone contributed to over 50% of the model's predictive power. The number of hydrogen bond donors followed as the second most important feature, contributing approximately 10%. A detailed bar chart in the Supporting Information(Figure S2) illustrates the ranked importance of features. These two properties—both physically relevant to solvation—show strong correlations with the target variable (free energy of solvation), underscoring the value of such physically interpretable models.

Next, we evaluated the transferability of the models across datasets—that is, training a model on one dataset and testing it on a different one. This serves as a form of external validation, providing insight into the model's ability to generalize beyond its training distribution. As illustrated in Figure 2(c)-(e), when the XGBoost Regressor (XGBR) was trained on the smallest dataset, MNSol, it completely overfit the training data, achieving an MSE of 0 and an R^2 score of 1.0. However, this model failed to generalize, performing poorly on other datasets except for the overlapping entries. This outcome highlights a key limitation in model transferability, especially when trained on small, potentially narrow datasets. This pattern was consistent across other cross-dataset evaluations (see Supporting Information).

Why do these models fail to generalize, despite the datasets having similar functional

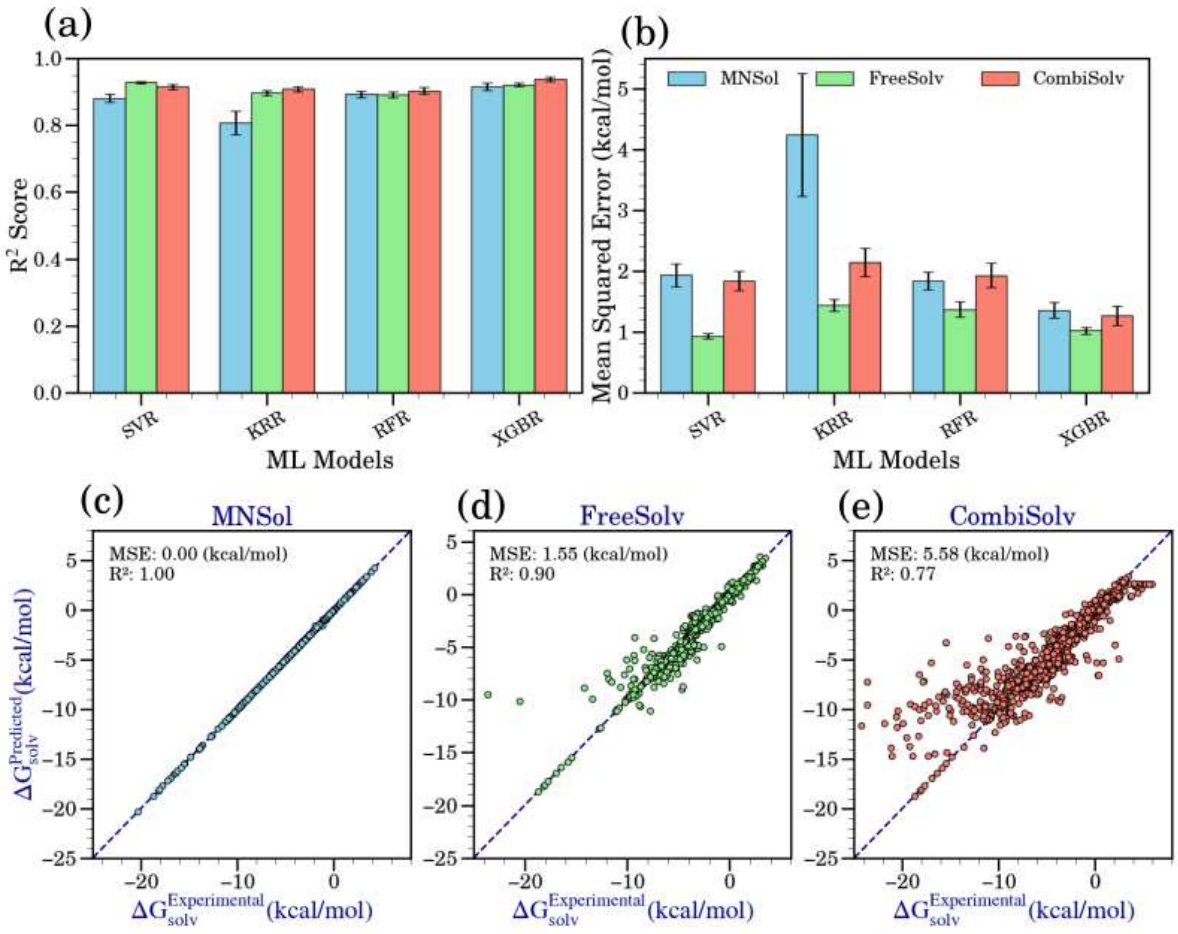


Figure 2: Performance of machine learning models using feature-based descriptions. (a) R^2 scores of the models and (b) MSE values (in kcal/mol) presented as bar plots with standard error bars for each dataset. (c) to (e) Predicted vs. experimental ΔG_{solv} plots, highlighting the transferability issue of classical models. In this case, an XGBoost model was trained on the MNSol dataset and tested on other datasets to assess generalization performance.

group distributions?

To investigate this, we conducted a comparative analysis of feature distributions across datasets. Although the most influential features (TPSA, hydrogen bond donors, etc.) remained consistent, we observed that descriptors from the smaller MNSol dataset did span the broader range seen in the larger datasets (see Figure S3 A–C and Figure S3 E–F). This implies that the current feature set lacks sufficient discriminatory power to capture the subtler distinctions present in larger or more diverse datasets.

In this context, a recent study by Yadav et al.[90] showed that a minimal but physics-

informed feature set—including polar surface area, hydrogen bond donors and acceptors, number of rotatable bonds, logP, and a simplified polar energy term derived from Generalized Born (GB) theory—can significantly improve both model interpretability and predictive accuracy. Notably, the GB energy term should directly correlate with the solvation free energy. However, computing the GB energy term can be expensive for larger molecules and it is not strictly a data-based description of the molecules.

Molecular fingerprints encodes local structure

Descriptors offer a coarse-grained representation of molecules and often fail to capture critical local neighborhood information, which can be especially important for properties like solvation free energy. To address this limitation, molecular fingerprints have been widely adopted due to their ability to encode local substructural patterns. Fingerprints represent molecules as binary vectors, where each bit corresponds to the presence or absence of specific substructures. Depending on the fingerprinting method, the bit size can vary and is often user-defined.

To identify the optimal fingerprint length, we followed a protocol similar to that used for descriptor selection—systematically varying the bit size and evaluating model performance. Our results indicate that a fingerprint length of 2048 bits provides the best balance between informativeness and model accuracy(Figure S5). Consequently, all subsequent tests were conducted using 2048-bit fingerprints.

From our analysis, structure-based fingerprints—specifically MACCS keys and PubChem fingerprints—consistently outperformed hashed fingerprints in terms of predictive accuracy. This was evident from the mean squared error (MSE) and R^2 scores (as provided in Figure 3). For example, using CombiSolv as the training dataset:

- XGBoost Regressor (XGBR) with MACCS keys achieved an MSE of 1.7 kcal/mol
- Random Forest Regressor (RFR) achieved an MSE of 1.8 kcal/mol

A possible reason for the inferior performance of hashed fingerprints lies in the nature of hash functions, which map arbitrary-sized data into fixed-length values. Enumerating all possible molecular fragments can generate a vast number of fragments, but hashing them into a fixed range introduces bit collisions, where distinct fragments are assigned the same numeric value and bit position.[73] Unlike structural keys, hashed fingerprints lack a one-to-one correspondence between fragments and fingerprint bits, which may challenge standard machine learning models in capturing meaningful relationships.

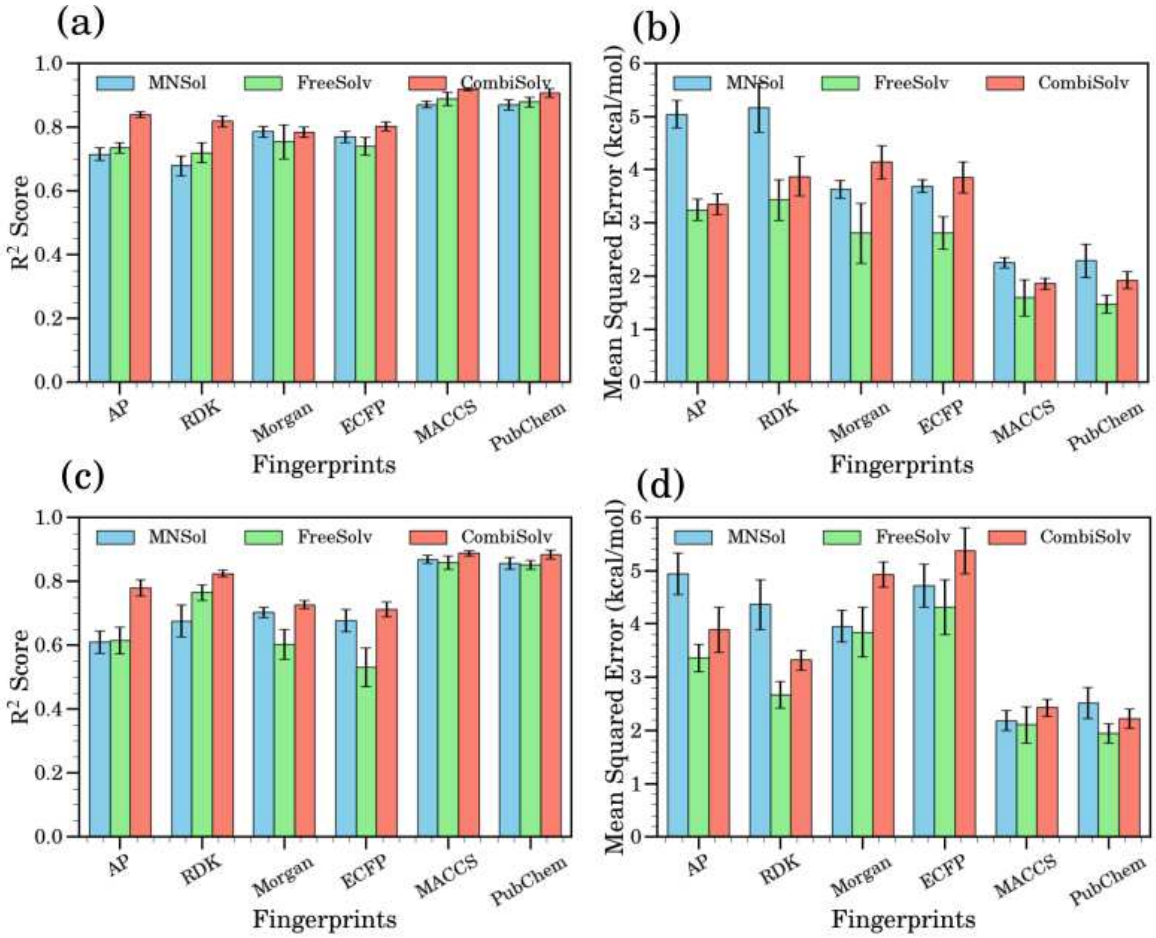


Figure 3: Performance comparison of different fingerprints using two popular machine learning models: XGBoost(upper panel) and Random Forest Regressor(lower panel). (a) and (c) display bar plots of the R^2 scores for all fingerprints across each dataset, while (b) and (d) show the corresponding MSE values (in kcal/mol). Five-fold cross-validation was employed for robustness, with standard error bars indicating the confidence in metric measurements.

To overcome the limitations of classical models with hashed fingerprints, we hypothesized

that artificial neural networks (ANNs)—known for their ability to model complex, non-linear relationships in high-dimensional data—might offer improved performance. To evaluate this, we implemented a Multilayer Perceptron Regressor (MLPR) with the following architecture: three hidden layers comprising 2048, 1024, and 1024 neurons, ReLU activation functions, and an Adam optimizer with a learning rate of 0.001.[91]

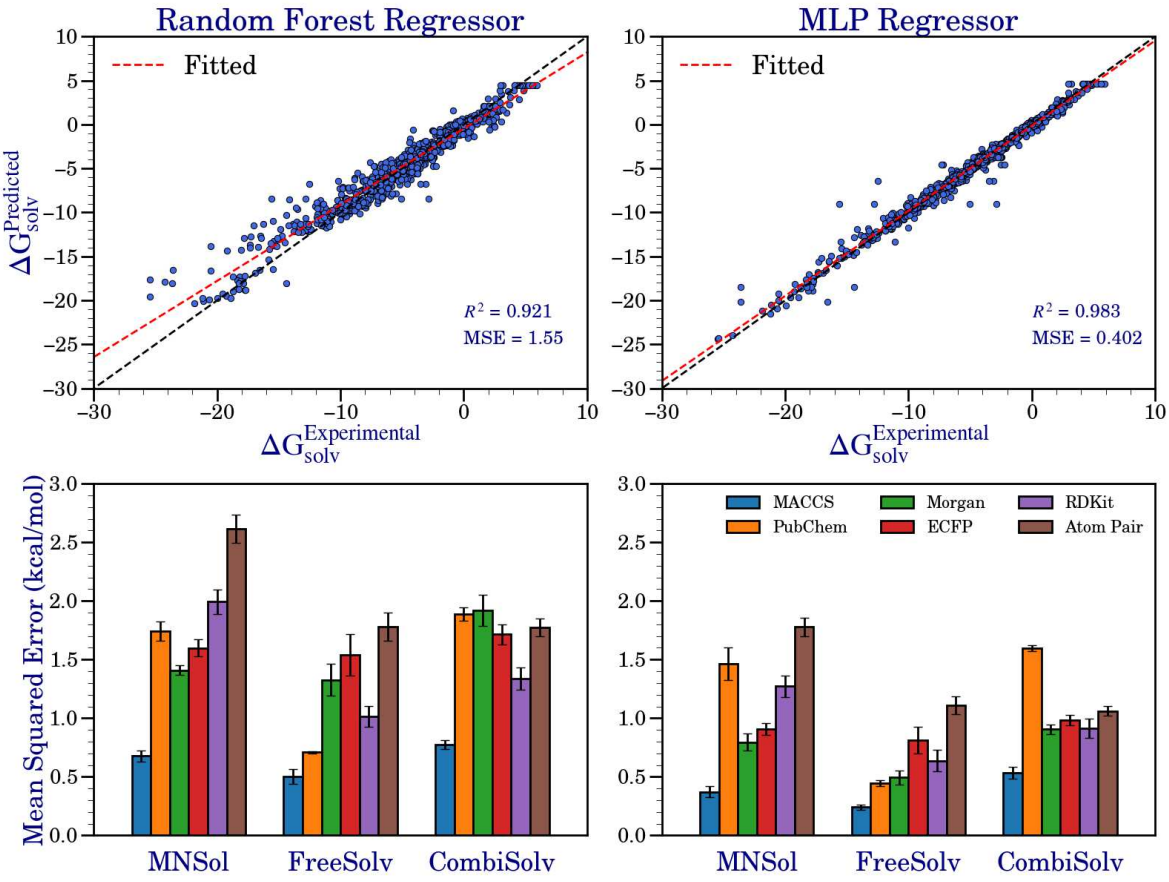


Figure 4: Multilayer Perceptron (MLP) Regressor outperforms traditional machine learning models. (a) and (b) compare the performance of the Random Forest Regressor (RFR) and MLP Regressor (MLPR) on the CombiSolv dataset using Morgan fingerprints, where RFR exhibits a skewed learning curve, while MLPR demonstrates better generalization. (c) and (d) show a bar plot of MSE scores for each dataset, comparing the efficiency of RFR and MLPR with different fingerprints.

The MLPR showed improved predictive performance over both XGBR and RFR models when trained on hashed fingerprints. This is reflected in lower mean squared error (MSE) and higher R^2 scores (Tables S2 and S3, Supporting Information). Unlike classical models,

which tend to skew toward underfitting in such high-dimensional spaces, the MLP was better able to regularize learning(Figure 3). However, despite this improvement, the model still yielded large prediction errors—exceeding 10 kcal/mol for certain molecules—which limits its practical reliability.

Among the fingerprint types, MACCS and PubChem again outperformed other representations, achieving MSE values of 1.01 and 1.10 kcal/mol, respectively for MLPR model. These results are within an acceptable error margin, but when evaluating cross-dataset transferability, the model exhibited the same shortcomings observed in descriptor-based approaches. As shown in Figure S4, the trained models failed to generalize well to external datasets.

So we conclude that while neural networks offer incremental gains, traditional machine learning models—regardless of feature representation—face a bottleneck when trained on small or narrowly distributed datasets. This underscores the need for more physics-informed feature engineering and/or models that incorporate physical priors to better capture the underlying structure–property relationships.

Transferability test for CIGIN

In recent years, the adoption of deep learning has become nearly inevitable due to advancements in computational resources and the development of sophisticated algorithms. In this study, we evaluated the performance of CIGIN (Chemically Interpretable Graph Interaction Neural Network), a state-of-the-art deep learning model proposed by Devapriyakumar and colleagues.[60,74] CIGIN leverages graph convolutional neural networks (GCNNs) and message-passing neural networks (MPNNs) to model solute–solvent interactions, offering a chemically interpretable and robust framework for molecular property prediction.

Following the original protocol established by the authors, we adapted the model specifically for solute–water interactions, in contrast to the broader solvent–solute scope of the original study. When trained on the MNSol dataset, the model generalized well to FreeSolv,

replicating the performance trends reported previously (Figure 5A–B). However, performance declined on the larger and more chemically diverse CombiSolv dataset (Figure 5C). This discrepancy can be attributed to the substantial overlap between MNSol and FreeSolv, while CombiSolv introduces numerous novel chemical entities absent from the training set.

Table 1: Cross-dataset transferability of models. Mean Squared Error (MSE) and R^2 are reported for each combination.

Trained on	**MNSol**		**FreeSolv**		**CombiSolv**	
	MSE	R^2	**MSE**	R^2	**MSE**	R^2
MNSol	0.786	0.960	1.804	0.878	5.479	0.778
FreeSolv	1.330	0.932	0.366	0.975	3.008	0.878
CombiSolv	2.225	0.886	0.481	0.968	0.366	0.985

A notable trend emerged in the prediction errors within CombiSolv, especially for alkane molecules (highlighted in Figure 4C), where errors scaled approximately linearly with molecule size (Figure 4D). Despite the use of undirected, heterogeneous graph representations that capture general chemical features, the model struggled to learn size-dependent behavior, particularly in nonpolar molecules. This suggests that the current graph features may lack key structural descriptors required to model such molecules effectively. A detailed table (Table 1) summarizes the cross-dataset performance and highlights this transferability limitation. Interestingly, the model consistently performed better on FreeSolv than on MNSol or CombiSolv, underscoring the critical role of dataset quality and overlap in deep learning-based predictions.

To further address this issue, we trained the CIGIN model on the merged dataset, employing cross-validation to ensure robustness. This broader training set significantly reduced overall prediction errors, and the previously observed alkane-related errors were largely mitigated (Figure 4E). However, certain molecules still exhibited errors exceeding 2 kcal/mol—a threshold that can be problematic in drug design and solubility prediction. A closer analysis of these outliers revealed consistent structural patterns, such as long hydrocarbon chains or polyol moieties, which appear challenging for the current model to accurately capture (Figure 6).

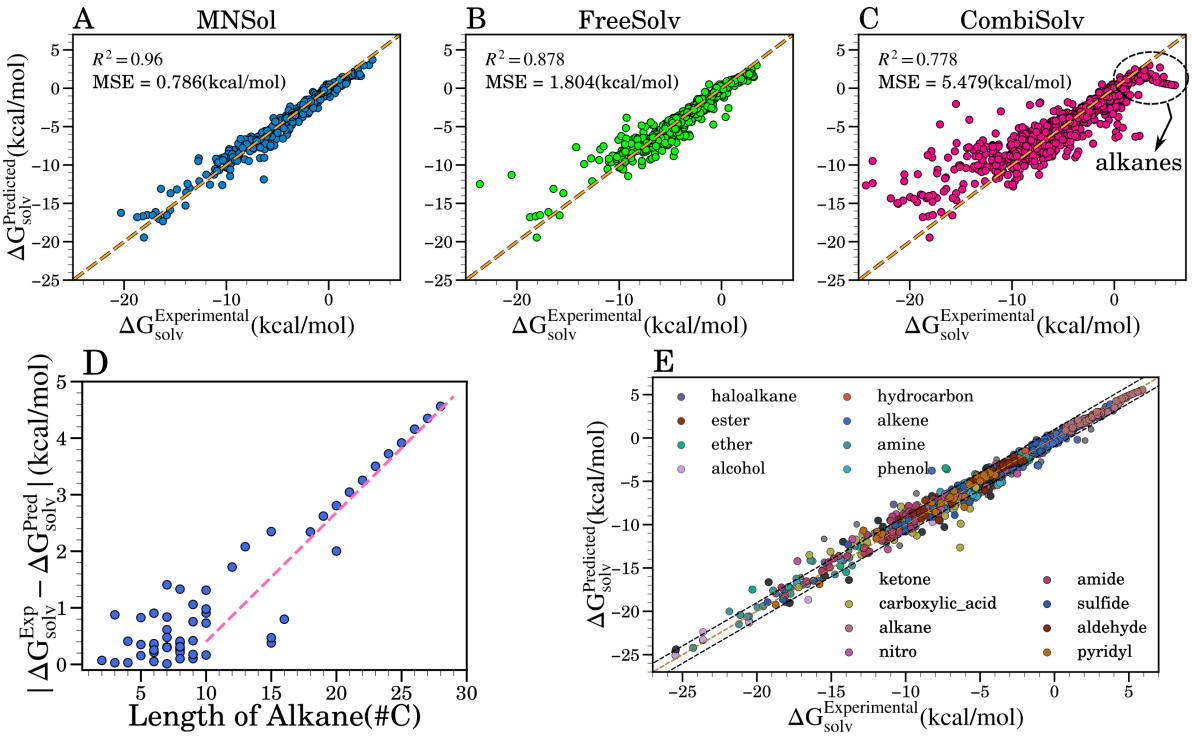


Figure 5: Transferability Test for CIGIN. (A–C) Predicted vs actual solvation energy plots, when the model is trained on the MNSol dataset and tested on other datasets, with performance shown in terms of R^2 scores and MSE values. A decline in performance is observed as the model encounters novel solutes.(D) presents a scatter plot showing the length of alkane molecules from the CombiSolv dataset against their absolute prediction errors, revealing higher errors for longer alkanes. These molecules are marked by circle in (C). (G) shows the overall model performance with a predicted vs. experimental ΔG_{solv} plot for the merged dataset using k-fold cross-validation, where data points are colored according to functional group classification.

These findings indicate that while data augmentation and diversity improve generalization, some structural motifs still evade accurate modeling, highlighting the need for improved feature engineering or architecture refinement in future work.

Dealing with broader datasets poses its own challenges. When we tested the performance of our well-trained model on the CombiSolv-QM dataset, which includes many molecules unfamiliar to the model, its performance degraded significantly. This indicates the model's limitations in generalizing to new, unseen data.

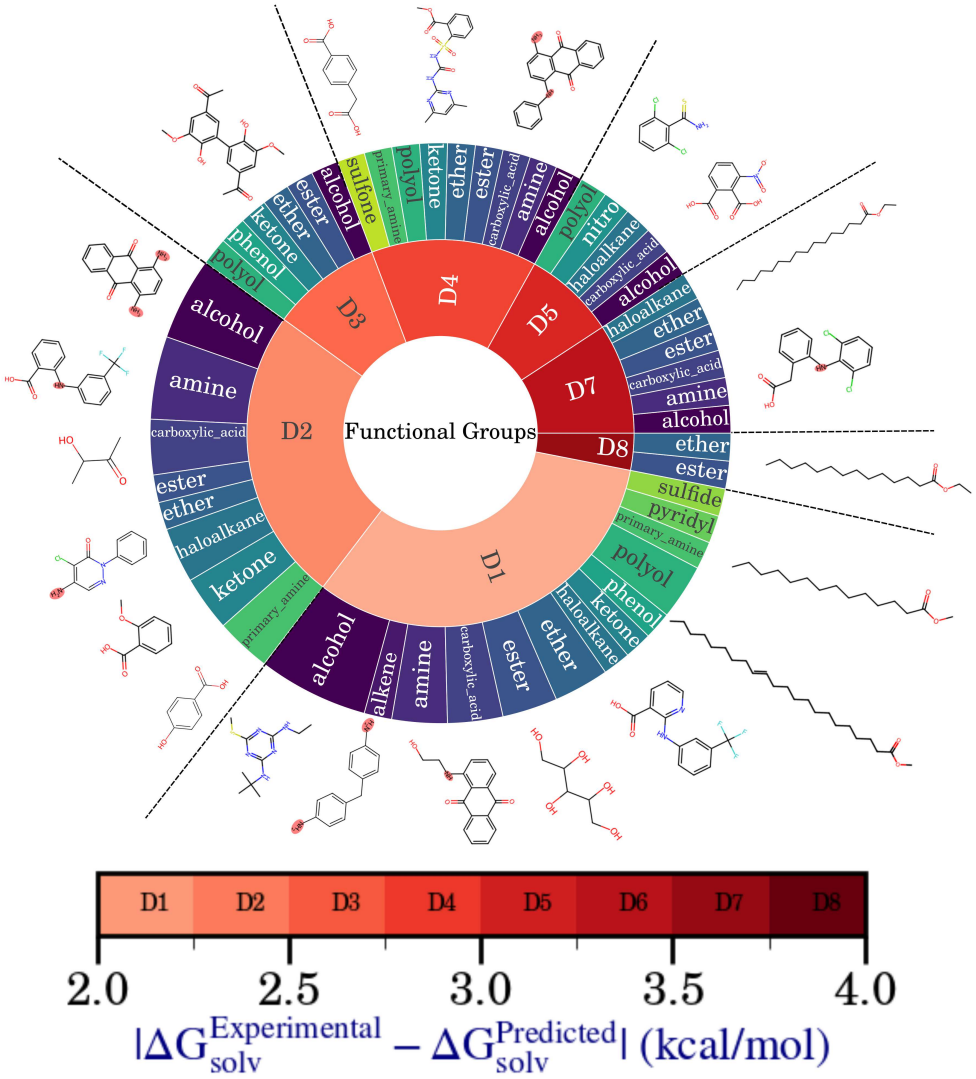


Figure 6: Outliers of CIGIN. Molecules that have mean error of prediction more than 2.0 kcal/mol are analyzed in the sun-burst plot. A few representative molecules are shown beside their respective functional group category.

Conclusion

In this study, we conducted a systematic investigation into the challenges of predicting molecular properties, specifically solvation free energy, using supervised learning. By comparing various molecular representations and machine learning models, we elucidated their strengths and limitations. Our findings underscore the critical impact of data quality, size, and chemical diversity on predictive performance and model transferability.

We demonstrated that traditional machine learning models, while offering interpretability through descriptor analysis and achieving reasonable accuracy on datasets with high overlap, struggle to generalize effectively to novel chemical space. The transferability issue is particularly pronounced when trained on smaller, potentially biased datasets, highlighting that data quantity alone does not guarantee robust model performance; data quality and representativeness are paramount.

Molecular fingerprints, especially structural keys, proved more effective in capturing local structural information compared to simple descriptors. The use of a Multilayer Perceptron Regressor showed promise in handling the complexity of high-dimensional fingerprint data, offering improved regularization over traditional methods. However, even with these advancements, achieving consistently low prediction errors across diverse datasets remains a challenge for classical ML approaches.

The application of the graph-based deep learning model, CIGIN, showcased the potential of learning complex atomic and bonding interactions for molecular property prediction. While CIGIN exhibited strong performance and maintained a degree of chemical interpretability, its predictive accuracy diminished when faced with chemical entities significantly different from its training distribution. Our analysis of outliers revealed that certain structural motifs, such as long hydrocarbon chains and polyol moieties, pose particular difficulties for the current model architecture and feature set.

This work reinforces that while sophisticated machine learning and deep learning models, coupled with rich molecular representations, have significantly advanced in-silico property prediction, challenges related to data transferability and the accurate modeling of diverse chemical structures persist. Future efforts should focus on developing more physics-informed molecular representations, enhancing model architectures to better capture complex interactions and novel chemical space, and employing strategies to mitigate dataset biases to improve the generalizability and reliability of supervised learning models for molecular property prediction. The identification and analysis of erroneous predictions, as demonstrated

in this study, can provide valuable feedback for iterative model refinement and feature engineering.

Acknowledgement

We thankfully acknowledge the high-performance computing facilities under the Technical Research Centre (TRC) at SNBNCBS, Kolkata. We also acknowledge the National Supercomputing Mission (NSM) for providing computing resources of 'Param Rudra' at SNBNCBS, Kolkata, which is implemented by C-DAC and supported by the Ministry of Electronics and Information Technology (MeitY) and Department of Science and Technology (DST), Government of India. D.M. thanks SNBNCBS, Kolkata for the fellowship.

Funding

This research has been funded by Science and Engineering Research Board (SERB), Govt. of India (project number: MTR/2021/000859).

Data Availability Statement

All data files and Python scripts required to reproduce the results reported in this manuscript are available from the GitHub repository at github.com/TeamSuman/Publication_Data.

Supporting Information Available

Table S1 summarizes the key molecular descriptors used to construct the feature vector. Tables S2 and S3 compare the performance of MLP Regressor (MLPR) with XGBoost Regressor (XGBR) and Random Forest Regressor (RFR) across different fingerprint (FPs) representations. Figure S1 illustrates correlation-based feature selection applied to the initial

brute-force feature set. Figure S2 highlights the relative importance of top features for predictive accuracy in the trained XGBR model. Figure S3 addresses the transferability challenge in descriptor-based modeling by comparing XGBoost performance across multiple datasets. Figure S4 evaluates the trade-off between fingerprint bit size and model effectiveness. Figure S5 visualizes molecular graphs projected into the 2D latent space of the trained Graph Convolutional Network-Variational Autoencoder (GCN-VAE). Figure S6 demonstrates the dataset-dependent performance of the Contextual Integration of Graph and Interaction Networks (CIGIN) model. Finally, Figure S7 confirms the stable convergence of the GCN-VAE training process.

References

(1) Shoichet, B. K. Virtual screening of chemical libraries. *Nature* **2004**, *432*, 862–865.

(2) Patrick, G. L. *An introduction to medicinal chemistry*, seventh edition ed.; Oxford University Press: New York, NY, 2023.

(3) Keserü, G. M.; Makara, G. M. The influence of lead discovery strategies on the properties of drug candidates. *Nature Reviews Drug Discovery* **2009**, *8*, 203–212.

(4) Kang, J.; Hagiwara, Y.; Tateno, M. Biological Applications of Hybrid Quantum Mechanics/Molecular Mechanics Calculation. *Journal of Biomedicine and Biotechnology* **2012**, *2012*, 1–11.

(5) Cavalli, A.; Carloni, P.; Recanatini, M. Target-Related Applications of First Principles Quantum Chemical Methods in Drug Design. *Chemical Reviews* **2006**, *106*, 3497–3519.

(6) Spiegel, K.; Magistrato, A. Modeling anticancer drug–DNA interactions via mixed QM/MM molecular dynamics simulations. *Org. Biomol. Chem.* **2006**, *4*, 2507–2517.

(7) Senn, H. M.; Thiel, W. QM/MM Methods for Biomolecular Systems. *Angewandte Chemie International Edition* **2009**, *48*, 1198–1229.

(8) Cardoso, F. J. B.; de Figueiredo, A. F.; da Silva Lobato, M.; de Miranda, R. M.; de Almeida, R. C. O.; Pinheiro, J. C. A study on antimalarial artemisinin derivatives using MEP maps and multivariate QSAR. *Journal of Molecular Modeling* **2007**, *14*, 39–48.

(9) Barua, H.; Gunnam, A.; Yadav, B.; Nangia, A.; Shastri, N. R. A n ab initio molecular dynamics method for cocrystal prediction: validation of the approach. *CrystEngComm* **2019**, *21*, 7233–7248.

(10) Senn, H. M.; Thiel, W. QM/MM studies of enzymes. *Current Opinion in Chemical Biology* **2007**, *11*, 182–187.

(11) Kalyaanamoorthy, S.; Chen, Y.-P. P. Structure-based drug design to augment hit discovery. *Drug Discovery Today* **2011**, *16*, 831–839.

(12) Meng, X.-Y.; Zhang, H.-X.; Mezei, M.; Cui, M. Molecular Docking: A Powerful Approach for Structure-Based Drug Discovery. *Current Computer Aided-Drug Design* **2011**, *7*, 146–157.

(13) Pantaleão, S. Q.; Fernandes, P. O.; Gonçalves, J. E.; Maltarollo, V. G.; Honorio, K. M. Recent Advances in the Prediction of Pharmacokinetics Properties in Drug Design Studies: A Review. *ChemMedChem* **2022**, *17*, e202100542.

(14) Sanachai, K.; Mahalapbutr, P.; Hengphasatporn, K.; Shigeta, Y.; Seetaha, S.; Tabtimmai, L.; Langer, T.; Wolschann, P.; Kittikool, T.; Yotphan, S.; Choowongkomon, K.; Rungrotmongkol, T. Pharmacophore-Based Virtual Screening and Experimental Validation of Pyrazolone-Derived Inhibitors toward Janus Kinases. *ACS Omega* **2022**, *7*, 33548–33559.

(15) Lin, T. E.; HuangFu, W.-C.; Chao, M.-W.; Sung, T.-Y.; Chang, C.-D.; Chen, Y.-Y.; Hsieh, J.-H.; Tu, H.-J.; Huang, H.-L.; Pan, S.-L.; Hsu, K.-C. A Novel Selective JAK2 Inhibitor Identified Using Pharmacological Interactions. *Frontiers in Pharmacology* **2018**, *9*.

(16) Jasuja, H.; Chadha, N.; Kaur, M.; Silakari, O. Dual inhibitors of Janus kinase 2 and 3 (JAK2/3): designing by pharmacophore- and docking-based virtual screening approach. *Molecular Diversity* **2014**, *18*, 253–267.

(17) Mafethe, O.; Ntseane, T.; Dongola, T. H.; Shonhai, A.; Gumede, N. J.; Mokoena, F. Pharmacophore Model-Based Virtual Screening Workflow for Discovery of Inhibitors Targeting Plasmodium falciparum Hsp90. *ACS Omega* **2023**, *8*, 38220–38232.

(18) Giordano, D.; Biancaniello, C.; Argenio, M. A.; Facchiano, A. Drug Design by Pharmacophore and Virtual Screening Approach. *Pharmaceuticals* **2022**, *15*, 646.

(19) Voet, A.; Banwell, E. F.; Sahu, K. K.; Heddle, J. G.; Zhang, K. Y. J. Protein Interface Pharmacophore Mapping Tools for Small Molecule Protein: Protein Interaction Inhibitor Discovery. *Current Topics in Medicinal Chemistry* **2013**, *13*, 989–1001.

(20) Seidel, T.; Wieder, O.; Garon, A.; Langer, T. Applications of the Pharmacophore Concept in Natural Product inspired Drug Design. *Molecular Informatics* **2020**, *39*.

(21) Kaserer, T.; Beck, K.; Akram, M.; Odermatt, A.; Schuster, D. Pharmacophore Models and Pharmacophore-Based Virtual Screening: Concepts and Applications Exemplified on Hydroxysteroid Dehydrogenases. *Molecules* **2015**, *20*, 22799–22832.

(22) Muhammed, M. T.; Aki-Yalcin, E. Homology modeling in drug discovery: Overview, current applications, and future perspectives. *Chem Biol Drug Des* **2018**, *93*, 12–20.

(23) Ehrlich, P. Über den jetzigen Stand der Chemotherapie. *Berichte der deutschen chemischen Gesellschaft* **1909**, *42*, 17–47.

(24) Wermuth, C. G. Pharmacophores: Historical Perspective and Viewpoint from a Medicinal Chemist. 2006; http://dx.doi.org/10.1002/3527609164.ch1.

(25) Opo, F. A. D. M.; Rahman, M. M.; Ahammad, F.; Ahmed, I.; Bhuiyan, M. A.; Asiri, A. M. Structure based pharmacophore modeling, virtual screening, molecular docking and ADMET approaches for identification of natural anti-cancer agents targeting XIAP protein. *Scientific Reports* **2021**, *11*.

(26) Gao, Q.; Yang, L.; Zhu, Y. Pharmacophore Based Drug Design Approach as a Practical Process in Drug Discovery. *Current Computer Aided-Drug Design* **2010**, *6*, 37–49.

(27) Dror, O.; Shulman-Peleg, A.; Nussinov, R.; Wolfson, H. Predicting Molecular Interactions in silico: I. A Guide to Pharmacophore Identification and its Applications to Drug Design. *Current Medicinal Chemistry* **2004**, *11*, 71–90.

(28) Zhu, H.; Zhou, R.; Cao, D.; Tang, J.; Li, M. A pharmacophore-guided deep learning approach for bioactive molecular generation. *Nature Communications* **2023**, *14*.

(29) Reichel, A.; Lienau, P. *New Approaches to Drug Discovery*; Springer International Publishing, 2015; p 235–260.

(30) Jang, G. R.; Harris, R. Z.; Lau, D. T. Pharmacokinetics and its role in small molecule drug discovery research. *Medicinal Research Reviews* **2001**, *21*, 382–396.

(31) Sharma, P.; Patel, N.; Prasad, B.; Varma, M. V. S. *Drug Discovery and Development*; Springer Singapore, 2021; p 297–355.

(32) Lai, Y.; Chu, X.; Di, L.; Gao, W.; Guo, Y.; Liu, X.; Lu, C.; Mao, J.; Shen, H.; Tang, H.; Xia, C. Q.; Zhang, L.; Ding, X. Recent advances in the translation of drug metabolism and pharmacokinetics science for drug discovery and development. *Acta Pharmaceutica Sinica B* **2022**, *12*, 2751–2777.

(33) Niazi, S. K.; Mariam, Z. Recent Advances in Machine-Learning-Based Chemoinformatics: A Comprehensive Review. *International Journal of Molecular Sciences* **2023**, *24*, 11488.

(34) Lo, Y.-C.; Rensi, S. E.; Torng, W.; Altman, R. B. Machine learning in chemoinformatics and drug discovery. *Drug Discovery Today* **2018**, *23*, 1538–1546.

(35) Jiménez-Luna, J.; Grisoni, F.; Weskamp, N.; Schneider, G. Artificial intelligence in drug discovery: recent advances and future perspectives. *Expert Opinion on Drug Discovery* **2021**, *16*, 949–959.

(36) Priya, S.; Tripathi, G.; Singh, D. B.; Jain, P.; Kumar, A. Machine learning approaches and their applications in drug discovery and design. *Chem Biol Drug Des* **2022**, *100*, 136–153.

(37) Paul, D.; Sanap, G.; Shenoy, S.; Kalyane, D.; Kalia, K.; Tekade, R. K. Artificial intelligence in drug discovery and development. *Drug Discovery Today* **2021**, *26*, 80–93.

(38) Goel, M.; Aggarwal, R.; Sridharan, B.; Pal, P. K.; Priyakumar, U. D. Efficient and enhanced sampling of drug-like chemical space for virtual screening and molecular design using modern machine learning methods. *WIREs Computational Molecular Science* **2023**, *13*, e1637.

(39) Mitchell, J. B. O. Machine learning methods in chemoinformatics. *WIREs Computational Molecular Science* **2014**, *4*, 468–481.

(40) Rutz, A.; Sorokina, M.; Galgonek, J.; Mietchen, D.; Willighagen, E.; Gaudry, A.; Graham, J. G.; Stephan, R.; Page, R.; Vondrášek, J.; Steinbeck, C.; Pauli, G. F.; Wolfender, J.-L.; Bisson, J.; Allard, P.-M. The LOTUS Initiative for Open Natural Products Research: Knowledge Management through Wikidata. **2021**,

(41) Sorokina, M.; Steinbeck, C. Review on natural products databases: where to find data in 2020. *Journal of Cheminformatics* **2020**, *12*.

(42) Banerjee, P.; Erehman, J.; Gohlke, B.-O.; Wilhelm, T.; Preissner, R.; Dunkel, M. Super Natural II—a database of natural products. *Nucleic Acids Research* **2014**, *43*, D935–D939.

(43) Zeng, X.; Zhang, P.; He, W.; Qin, C.; Chen, S.; Tao, L.; Wang, Y.; Tan, Y.; Gao, D.; Wang, B.; Chen, Z.; Chen, W.; Jiang, Y. Y.; Chen, Y. Z. NPASS: natural product activity and species source database for natural product research, discovery and tool development. *Nucleic Acids Research* **2017**, *46*, D1217–D1222.

(44) Xue, R.; Fang, Z.; Zhang, M.; Yi, Z.; Wen, C.; Shi, T. TCMID: traditional Chinese medicine integrative database for herb molecular mechanism analysis. *Nucleic Acids Research* **2012**, *41*, D1089–D1095.

(45) Mendez, D. et al. ChEMBL: towards direct deposition of bioassay data. *Nucleic Acids Research* **2018**, *47*, D930–D940.

(46) Gilson, M. K.; Liu, T.; Baitaluk, M.; Nicola, G.; Hwang, L.; Chong, J. BindingDB in 2015: A public database for medicinal chemistry, computational chemistry and systems pharmacology. *Nucleic Acids Research* **2015**, *44*, D1045–D1053.

(47) Wishart, D. S. et al. DrugBank 5.0: a major update to the DrugBank database for 2018. *Nucleic Acids Research* **2017**, *46*, D1074–D1082.

(48) Sussman, J. L.; Lin, D.; Jiang, J.; Manning, N. O.; Prilusky, J.; Ritter, O.; Abola, E. E. Protein Data Bank (PDB): Database of Three-Dimensional Structural Information of Biological Macromolecules. *Acta Crystallographica Section D Biological Crystallography* **1998**, *54*, 1078–1084.

(49) Haghighatlari, M.; Li, J.; Heidar-Zadeh, F.; Liu, Y.; Guan, X.; Head-Gordon, T. Learning to Make Chemical Predictions: The Interplay of Feature Representation, Data, and Machine Learning Methods. *Chem* **2020**, *6*, 1527–1542.

(50) David, L.; Thakkar, A.; Mercado, R.; Engkvist, O. Molecular representations in AI-driven drug discovery: a review and practical guide. *Journal of Cheminformatics* **2020**, *12*.

(51) Grumet, M.; von Scarpatetti, C.; Bučko, T.; Egger, D. A. Delta Machine Learning for Predicting Dielectric Properties and Raman Spectra. *The Journal of Physical Chemistry C* **2024**, *128*, 6464–6470.

(52) Zhao, Q.; Anstine, D. M.; Isayev, O.; Savoie, B. M. $\Delta 2$ machine learning for reaction property prediction. *Chemical Science* **2023**, *14*, 13392–13401.

(53) Fralish, Z.; Chen, A.; Skaluba, P.; Reker, D. DeepDelta: predicting ADMET improvements of molecular derivatives with deep learning. *Journal of Cheminformatics* **2023**, *15*.

(54) Chen, X.; Li, P.; Hruska, E.; Liu, F. $\Delta-$Machine learning for quantum chemistry prediction of solution-phase molecular properties at the ground and excited states. *Physical Chemistry Chemical Physics* **2023**, *25*, 13417–13428.

(55) Buterez, D.; Janet, J. P.; Kiddle, S. J.; Oglic, D.; Lió, P. Transfer learning with graph neural networks for improved molecular property prediction in the multi-fidelity setting. *Nature Communications* **2024**, *15*.

(56) Hoffmann, N.; Schmidt, J.; Botti, S.; Marques, M. A. L. Transfer learning on large datasets for the accurate prediction of material properties. *Digital Discovery* **2023**, *2*, 1368–1379.

(57) Gilmer, J.; Schoenholz, S. S.; Riley, P. F.; Vinyals, O.; Dahl, G. E. Neural Message Passing for Quantum Chemistry. 2017; https://arxiv.org/abs/1704.01212.

(58) Lim, H.; Jung, Y. Delfos: deep learning model for prediction of solvation free energies in generic organic solvents. *Chemical Science* **2019**, *10*, 8306–8315.

(59) Lim, H.; Jung, Y. MLSolvA: solvation free energy prediction from pairwise atomistic interactions by machine learning. *Journal of Cheminformatics* **2021**, *13*.

(60) Pathak, Y.; Laghuvarapu, S.; Mehta, S.; Priyakumar, U. D. Chemically Interpretable Graph Interaction Network for Prediction of Pharmacokinetic Properties of Drug-Like Molecules. *Proceedings of the AAAI Conference on Artificial Intelligence* **2020**, *34*, 873–880.

(61) Llompart, P.; Minoletti, C.; Baybekov, S.; Horvath, D.; Marcou, G.; Varnek, A. Will we ever be able to accurately predict solubility? *Scientific Data* **2024**, *11*.

(62) Marenich, A. V.; Kelly, C. P.; Thompson, J. D.; Hawkins, G. D.; Chambers, C. C.; Giesen, D. J.; Winget, P.; Cramer, C. J.; Truhlar, D. G. Minnesota Solvation Database (MNSOL) version 2012. 2020; http://hdl.handle.net/11299/213300.

(63) Mobley, D. L.; Guthrie, J. P. FreeSolv: a database of experimental and calculated hydration free energies, with input files. *Journal of Computer-Aided Molecular Design* **2014**, *28*, 711–720.

(64) Vermeire, F. H.; Green, W. H. Transfer learning for solvation free energies: From quantum chemistry to experiments. *Chemical Engineering Journal* **2021**, *418*, 129307.

(65) Yu, J.; Zhang, C.; Cheng, Y.; Yang, Y.-F.; She, Y.-B.; Liu, F.; Su, W.; Su, A. SolvBERT for solvation free energy and solubility prediction: a demonstration of an NLP model for predicting the properties of molecular complexes. *Digital Discovery* **2023**, *2*, 409–421.

(66) Landrum, G. rdkitrdkit: 2024_03_4 (Q1 2024) Release. 2024; https://zenodo.org/doi/10.5281/zenodo.591637.

(67) Chen, T.; Guestrin, C. XGBoost: A Scalable Tree Boosting System. Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. 2016.

(68) Svetnik, V.; Liaw, A.; Tong, C.; Culberson, J. C.; Sheridan, R. P.; Feuston, B. P. Random Forest: A Classification and Regression Tool for Compound Classification and QSAR Modeling. *Journal of Chemical Information and Computer Sciences* **2003**, *43*, 1947–1958.

(69) Durant, J. L.; Leland, B. A.; Henry, D. R.; Nourse, J. G. Reoptimization of MDL Keys for Use in Drug Discovery. *Journal of Chemical Information and Computer Sciences* **2002**, *42*, 1273–1280.

(70) Fernández-de Gortari, E.; García-Jacas, C. R.; Martinez-Mayorga, K.; Medina-Franco, J. L. Database fingerprint (DFP): an approach to represent molecular databases. *Journal of Cheminformatics* **2017**, *9*.

(71) Carhart, R. E.; Smith, D. H.; Venkataraghavan, R. Atom pairs as molecular features in structure-activity studies: definition and applications. *Journal of Chemical Information and Computer Sciences* **1985**, *25*, 64–73.

(72) Morgan, H. L. The Generation of a Unique Machine Description for Chemical Structures-A Technique Developed at Chemical Abstracts Service. *Journal of Chemical Documentation* **1965**, *5*, 107–113.

(73) Rogers, D.; Hahn, M. Extended-Connectivity Fingerprints. *Journal of Chemical Information and Modeling* **2010**, *50*, 742–754.

(74) Pathak, Y.; Mehta, S.; Priyakumar, U. D. Learning Atomic Interactions through Solvation Free Energy Prediction Using Graph Neural Networks. *Journal of Chemical Information and Modeling* **2021**, *61*, 689–698.

(75) Pearson, K. LIII. On lines and planes of closest fit to systems of points in space. *The London, Edinburgh, and Dublin Philosophical Magazine and Journal of Science* **1901**, *2*, 559–572.

(76) van der Maaten, L.; Hinton, G. Visualizing Data using t-SNE. *Journal of Machine Learning Research* **2008**, *9*, 2579–2605.

(77) van der Maaten, L. Accelerating t-SNE using Tree-Based Algorithms. *Journal of Machine Learning Research* **2014**, *15*, 3221–3245.

(78) Belkina, A. C.; Ciccolella, C. O.; Anno, R.; Halpert, R.; Spidlen, J.; Snyder-Cappione, J. E. Automated optimized parameters for T-distributed stochastic neighbor embedding improve visualization and analysis of large datasets. *Nature Communications* **2019**, *10*.

(79) Healy, J.; McInnes, L. Uniform manifold approximation and projection. *Nature Reviews Methods Primers* **2024**, *4*.

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

(81) 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/1703.06103`.

(82) Belov, D. I.; Armstrong, R. D. Distributions of the Kullback–Leibler divergence with applications. *British Journal of Mathematical and Statistical Psychology* **2011**, *64*, 291–309.

(83) Ghanavati, M. A.; Ahmadi, S.; Rohani, S. A machine learning approach for the prediction of aqueous solubility of pharmaceuticals: a comparative model and dataset analysis. *Digital Discovery* **2024**,

(84) Mohammadi, M.-R.; Hadavimoghaddam, F.; Atashrouz, S.; Abedi, A.; Hemmati-Sarapardeh, A.; Mohaddespour, A. Modeling hydrogen solubility in alcohols using machine learning models and equations of state. *Journal of Molecular Liquids* **2022**, *346*, 117807.

(85) Yang, A.; Sun, S.; Mi, H.; Wang, W.; Liu, J.; Kong, Z. Y. Interpretable Feedforward Neural Network and XGBoost-Based Algorithms to Predict CO2 Solubility in Ionic Liquids. *Industrial & Engineering Chemistry Research* **2024**, *63*, 8293–8305.

(86) Vovk, V. *Empirical Inference*; Springer Berlin Heidelberg, 2013; p 105–116.

(87) Exterkate, P. Model selection in kernel ridge regression. *Computational Statistics & Data Analysis* **2013**, *68*, 1–16.

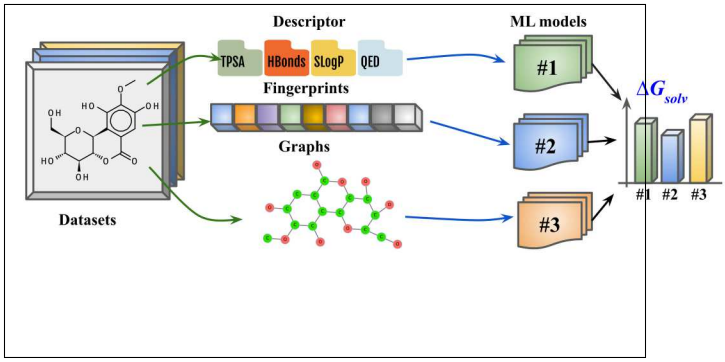
(88) Jingqing, J.; Chuyi, S.; Chunguo, W.; Maurizio, M.; Yangchun, L. *Lecture Notes in Computer Science*; Springer Berlin Heidelberg, 2006; p 547–554.

(89) Pedregosa, F. et al. Scikit-learn: Machine Learning in Python. *Journal of Machine Learning Research* **2011**, *12*, 2825–2830.

(90) Yadav, A. K.; Prakash, M. V.; Bandyopadhyay, P. Physics-Based Machine Learning to Predict Hydration Free Energies for Small Molecules with a Minimal Number of Descriptors: Interpretable and Accurate. *The Journal of Physical Chemistry B* **2025**, acs.jpcb.4c07090.

(91) Kingma, D. P.; Ba, J. Adam: A Method for Stochastic Optimization. 2014; https://arxiv.org/abs/1412.6980.

TOC Graphic


Descriptor
ML models
TPSA
HBonds
SLogP
QED
#1
Fingerprints
ΔG_{solv}
#2
Graphs
#1 #2 #3
Datasets
#3

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- SI.pdf