# Chemically Interpretable Graph Interaction Network for Prediction of Pharmacokinetic Properties of Drug-like Molecules

by

Yashaswi Pathak, Siddhartha Laghuvarapu, Sarvesh Mehta, U Deva Priyakumar

Report No: IIIT/TR/2020/-1

Centre for Computational Natural Sciences and Bioinformatics
International Institute of Information Technology
Hyderabad - 500 032, INDIA
February 2020

# Chemically Interpretable Graph Interaction Network for Prediction of Pharmacokinetic Properties of Drug-like Molecules

**Yashaswi Pathak*, Siddhartha Laghuvarapu*, Sarvesh Mehta*, U Deva Priyakumar†**
Center for Computational Natural Sciences and Bioinformatics
IIIT Hyderabad

**Abstract**

Solubility of drug molecules is related to pharmacokinetic properties such as absorption and distribution, which affects the amount of drug that is available in the body for its action. Computational or experimental evaluation of solvation free energies of drug-like molecules/solute that quantify solubilities is an arduous task and hence development of reliable computationally tractable models is sought after in drug discovery tasks in pharmaceutical industry. Here, we report a novel method based on graph neural network to predict solvation free energies. Previous studies considered only the solute for solvation free energy prediction and ignored the nature of the solvent, limiting their practical applicability. The proposed model is an end-to-end framework comprising three phases namely, message passing, interaction and prediction phases. In the first phase, message passing neural network was used to compute inter-atomic interaction within both solute and solvent molecules represented as molecular graphs. In the interaction phase, features from the preceding step is used to calculate a solute-solvent interaction map, since the solvation free energy depends on how (un)favorable the solute and solvent molecules interact with each other. The calculated interaction map that captures the solute-solvent interactions along with the features from the message passing phase is used to predict the solvation free energies in the final phase. The model predicts solvation free energies involving a large number of solvents with high accuracy. We also show that the interaction map captures the electronic and steric factors that govern the solubility of drug-like molecules and hence is chemically interpretable.

## 1 Introduction

One of the most convenient, least painful and safest ways of administering drugs is oral administration, which also aids high patient compliance. However, the major challenge involved in oral drug administration is achieving high bioavailability, the fraction of the orally administered drug that is available for its action. Absorption and distribution are two key pharmacokinetic properties that determine the bioavailability of a given drug. It is reported that about 40% of the new chemical entities developed by the pharmaceutical industry suffer from low bioavailability due to which further development of a significant number of these compounds is discontinued. Low bioavailability is primarily due to poor solubility and permeability. Drug molecules that can be orally administered have to be hydrophilic enough for them to be soluble in aqueous biological fluids and hydrophobic enough for them to be able to permeate across the hydrophobic lipid bilayer environment, which are related to how well they dissolve in polar and nonpolar solvents respectively. Hence, solubility of drug-like molecules is a crucial property that determines the viability of a drug to be marketable among other pharmacokinetic properties such as metabolism, excretion and toxicity. In addition to the role of solubility in pharmaceutical industry, it is significant in wide ranging areas such as chemistry, biochemistry, food industry, biotechnology industry, etc. Solvation free energy is also related to a number of physicochemical properties that are relevant to wide ranging areas of science and technology.

Solvation free energy is the change in free energy for a molecule to be transferred from gas phase to a given solvent. A large negative value of the solvation free energy indicates high solubility, while lower magnitudes/positive values indicate poor solubility. Experimental measurement of solvation free energies is a difficult task especially given the requirement that the molecule has to be synthesized and purified. It is desirable to have computational models in the drug discovery toolkit that are able to predict highly accurate solvation free energies within reasonable time and effort so that experimental realization of only those with desired solvation properties are realized in the laboratory thus minimizing the time and expenditure. Using these calculated solvation free energies, the solubility and drug permeation properties can be assessed with ease. Alchemical free energy methods such as free energy of perturbation and thermodynamic integration methods have been shown to be efficient methods for the calculation of solvation free energies. Philosophically, the molecule in its gas phase, in its solvated state, and a number of non-physical intermediate states (hence the name alchemical) are simulated using molecular dynamics or Monte Carlo simulations with respect to a parameter from

*Authors contributed equally
†Corresponding Author Email: deva@iiit.ac.in; Phone: +91 40 6653 1161

which the difference in the free energies of the molecule in the gas phase and solvated state is calculated. However, reliable force field is not available for all drug-like molecules and needs to be derived for each molecule, which in turn involves a large number of quantum mechanical calculations and molecular dynamics simulations. In short, experimental and computational evaluation of accurate solvation free energies is a difficult task and hence, fast and accurate computational models are necessary for high throughput evaluation of a large number of molecules, usually in the order of several millions in case of drug discovery projects.

Recent advances in deep learning methods have found varied applications in chemistry, biochemistry and drug design. Some recent problems that have been tackled with ingenious use of modern machine learning methods include protein structure prediction, generation of focused molecular libraries for drug development, prediction of biological activity, toxicity, and protein-protein binding, etc (Liu et al. 2018; You et al. 2018; Zhou et al. 2019). These techniques have not only been proven to be accurate but also brought about significant reduction in computational costs(Schütt et al. 2017; Behler and Parrinello 2007; Butler et al. 2018; Goh, Hodas, and Vishnu 2017; Rupp et al. 2012; Laghuvarapu, Pathak, and Priyakumar 2019). The field though still in its infancy has the potential to revolutionize the way computational chemistry/biology is traditionally practiced.

Traditional molecular representations (Cartesian and internal coordinates) are molecular size dependent and are not invariant to rotations/translations/permutation of similar atoms. Hence, initial efforts in the area focused on development of feature vectors for representing molecules for regression/classification tasks based on chemical intuition. Several approaches have been proposed to hand engineer machine learning friendly representation for molecules: Coulomb matrix proposed by Rupp et al., bag of bonds approach by Hansen et al. and symmetry functions by Behler and Parrinello. Recently, the focus is on developing new methods to machine learn the molecular features resulting in robust representations. Primarily, these approaches consider molecules as graphs, where atoms denote the nodes and bonds the edges (Duvenaud et al. 2015; Li et al. 2015; Kearnes et al. 2016; Gilmer et al. 2017).

In this paper, we propose a method based on graph neural networks using molecular graph representation of molecules to predict solvation free energies that is generalized for all drug-like molecules and all generic organic solvents. More importantly, we demonstrate that the interaction map calculated for any solute-solvent pair can be used to derive chemically meaningful insights on the solvation free energies, and the electronic/steric factors that determine the free energies.

## 2 Related Work

Molecular dynamics simulations employing free energy of perturbation and thermodynamic integration methods is the current state of the art strategy for accurate solvation free energy calculations(Duarte Ramos Matos et al. 2017). These calculations employ force fields that consist of several empirical parameters, which are obtained based on further high level quantum mechanical calculations and experimental data. The best of the methods available today employing molecular dynamics simulations gives rise to an RMSE of about 1.5 kcal/mol for the hydration (in water) free energies(Duarte Ramos Matos et al. 2017). Calculation of solvation free energies of other organic solvents are more complicated due to lack of good solvent models and hence in general are not done computationally. On the other hand, other computationally expensive quantum mechanical methods involving dielectric continuum models yield RMSEs of 0.75 to 4.8 kcal/mol compared to experimental solvation free energies (Marenich, Cramer, and Truhlar 2009; Klamt and Diedenhofen 2010).

During the last two years, few machine learning models to predict the hydration free energies have been reported based on the FreeSolv dataset (Mobley and Guthrie 2014) with root mean square errors ranging from 0.82 to 2.13 kcal/mol (Goh, Hodas, and Vishnu 2017; Wu et al. 2018; Goh et al. 2018; Cho, Choi, and others 2018; Hutchinson and Kobayashi 2019). Prediction of solvation free energies for variable solute with a constant solvent (water in this case) is a relatively straightforward exercise. However, when multiple solvents of diverse polarities (wide ranging dielectric constant, $\epsilon$) are considered, the nature of intermolecular interactions vary significantly due to different solutes and solvents. Hutchinson and Kobayashi proposed multiple models for different solvents using XGBoost along with functional class fingerprint featurizer using the MNSol dataset. The only unifying model that has been developed for organic solvents in general was proposed by Lim and Jung recently. They used Recurrent Neural Networks and attention mechanism with SMILES sequences as the input. Features for SMILES were first obtained using Mol2Vec(Jaeger, Fulle, and Turk 2018), an embedding technique inspired from Word2Vec, which involves unsupervised training over a huge corpus of SMILES ($\sim$20 million molecules). Two major drawbacks of this method are the requirement of a pre-trained Mol2Vec embedding and use of SMILES for molecular featurization, which may not capture the pharmacophores appropriately (Jin, Barzilay, and Jaakkola 2018).

In this work, we propose a model that can be applied to any organic solvent and any drug-like molecule for accurate prediction of solvation free energies. We show that the model yields accurate results on test set chosen from a completely different dataset compared to models trained on that external dataset. In addition to predicting pharmacokinetic properties of drug-like molecules, this method can further be extended to calculate drug-receptor binding affinities that is directly related to the efficacy of drug-like molecules, which has far reaching implications in drug discovery tasks.

## 3 Dataset

We use the Minnesota Solvation Database that has 3037 experimental free energies of solvation or transfer free energies of 790 unique solutes in 92 solvents (Marenich et al. 2012). In this work we only consider neutral solutes, which removes 249 charged solute. We also omit entries corresponding to transfer free energy. This makes our dataset contain 2525 unique combinations of solute and solvent.

Table 1: The atom (node) features used for molecular representation.

| Atom Features | Description |
|---|---|
| Atom Type | H, C, N, O, F... (one-hot) |
| Implicit Valence | Has Implicit Valence (Binary) |
| Radical Electrons | Has Radical Electrons (Binary) |
| Chirality | R,S or None (one-hot) |
| Number of Hydrogens | Number of neighbouring Hydrogen atoms (one-hot) |
| Hybridization | sp, $sp^2$, $sp^3$, $sp^3d$ (one-hot) |
| Acidic | Acidic in Nature (Binary) |
| Basic | Basic in Nature (Binary) |
| Aromatic | Part of aromatic group (Binary) |
| Donor | Donates electron (Binary) |
| Acceptor | Accepts electron (Binary) |

The dataset also provides Cartesian coordinates for the solutes, which were used to build molecular graph for solutes. We use PubChemPy to search for SMILES of solvents using their common name provided in the dataset and use the found SMILES to construct molecular graph for solvents.

# 4 Model

In this work, we encode molecules as graphs, where atoms constitute the nodes and bonds the edges. The nodes and the edges are characterized by a set of features. The model proposed in this work named as the CIGIN model, which stands for Chemically Interpretable Graph Interaction Network, can be broken down into three phases - message passing phase, interaction phase and prediction phase. These three phases are explained in detail in the subsequent sections. Figure 1 illustrates the CIGIN model. The source code and supporting material are available here: https://github.com/devalab/cigin

## Feature Representation

For the features to precisely model the free energy of solvation, it should consist of information about the general structures of solute and solvent molecules and should characterize their intermolecular pharmacophores (Hutchinson and Kobayashi 2019). To incorporate these, we use the features given in Table 1 to represent the nodes (atoms) and the ones given in Table 2 to represent the edges (bonds). RDKit library was used for extraction of these features (Landrum ). Note that hydrogen atoms were not considered as explicit nodes and their information is preserved in the node feature of the neighbouring atom (number of neighboring Hydrogen atoms). This is done keeping in view that the explicit consideration of hydrogen atoms does not enhance any information about the intermolecular pharmacophores. Secondly, their explicit consideration would lead to larger graphs that will slow down the training by roughly a factor of 10 (Gilmer et al. 2017).

Table 2: The bond (edge) features used for molecular representation.

| Bond Features | Description |
|---|---|
| Bond Type | Single, double, triple, or aromatic (one-hot) |
| Bond is in Conjugation | Part of conjugation (Binary) |
| Bond is in Ring | Part of ring (Binary) |
| Bond Chirality | E or Z (one-hot) |

## Message Passing Phase

This phase uses the Message Passing Neural Network (MPNN) (Gilmer et al. 2017), which provides a generalized formulation for supervised learning on graph structured data. Consider a molecule represented as an undirected graph $G(V, E)$ with node features $x_v$ and edge features $e_{vw}$. The state of each node at time step $t$ is represented as $h_v^t$, which is initialized to $x_v$ at $t = 0$. The state of a node is updated for $T$ time steps using messages $m_v^{t+1}$ according to following equations:

$$m_v^{t+1} = \sum_{w \epsilon N(v)} M_t(h_v^t, h_w^t, e_{vw}) \quad (1)$$

$$h_v^{t+1} = U_t(h_v^t, m_v^{t+1}) \quad (2)$$

Here, $N(v)$ is the set of neighboring nodes of $v$. $M_t$ and $U_t$ are the message function and vertex update functions respectively. In this work both $M_t$ and $U_t$ are fully connected layers. After T time steps, the final feature vector for each node is obtained using a gather layer-

$$F_v = g(x_v, h_v^t), \forall v \in V \quad (3)$$

$F_v$ is the final atomic feature for each atom $v$ whose context describes the atomic property and as well as the local environment. We experiment with two different choices for $g$, a simple feed-forward layer and a set2set layer (Vinyals, Bengio, and Kudlur 2015). In this work, both solute and solvent are fed through identical MPNN and gather layers with different set of weights. The outputs $F$ of this phase are tensors $A$ (solute) and $B$ (solvent) of sizes $J * L$ and $K * L$ respectively for a solute of $J$ atoms, solvent of $K$ atoms and $L$ atomic features.

## Interaction Phase

Intermolecular interactions between the solute and the solvent atoms is key to determining the solvation free energy. These interactions are influenced by several steric and electronic factors governed by participating atoms and their respective chemical environments. In this phase, pairwise interactions between solute-solvent atoms are modelled in an interaction map. Consider solute features $A$ and solvent features $B$ computed from the message passing phase. The solute-solvent interaction map is computed according to the following equation:

$$I_{nm} = f(A_n, B_m), \forall n = 1, 2, 3..J, \forall m = 1, 2, 3, ..K \quad (4)$$

Here, $I$ is the interaction map which bears the dimensionality of $J * K$. $f$ is a function that computes an interaction

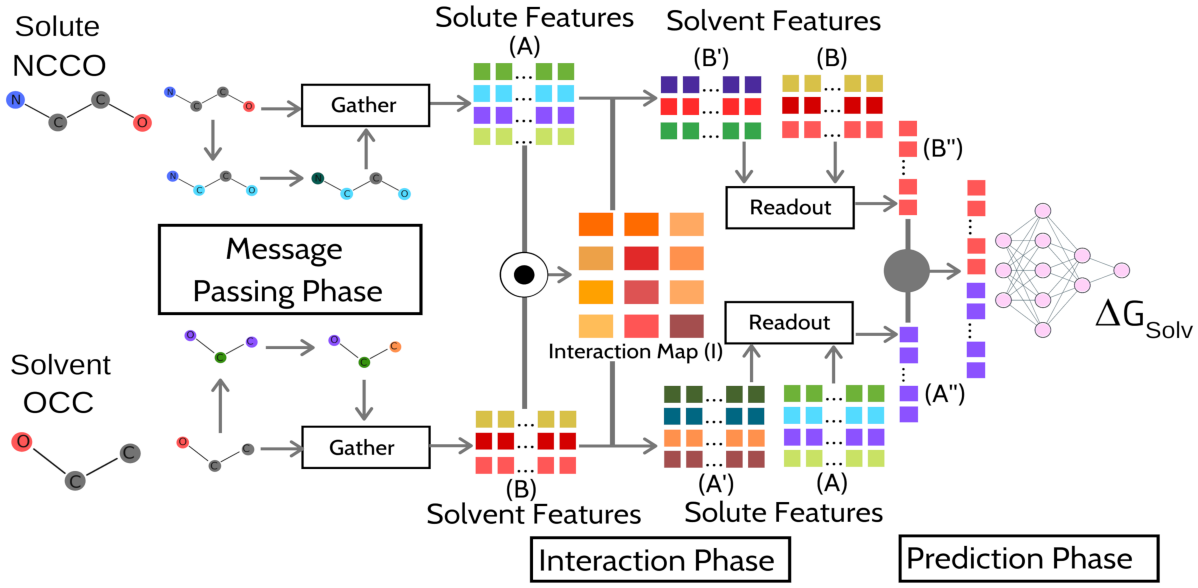


Figure 1: Architecture of CIGIN Model. Two examples one each for solvent (NCCO) and solute (OCC) are used as a test case.

value for each solute-solvent atom pair from their feature vectors. Solvent's influence on solute and vice-versa is computed according to the following equations:

$$A' = IB \tag{5}$$

$$B' = I^T A \tag{6}$$

$A'$ and $B'$ are the solute and solvent features weighted by their contribution to the free energy of solvation. The function $f$ should precisely measure the negative/positive contribution of a specific solute-solvent pair independently. For example, hydrophilic-hydrophobic interactions would decrease the solubility where as a hydrophilic-hydrophilic/hydrophobic-hydrophobic interactions would increase the solubility. To appropriately model this behaviour, the function $f$ is chosen to be :

$$f(A_n, B_m) = tanh(A_n \cdot B_m) \tag{7}$$

### Prediction Phase

In this phase the final solvation free energy is predicted. For both the solute and the solvent, the outputs of the message passing and the interaction phases are combined atom wise. Then, a readout layer $R$ is used to combine the feature vectors across all the atoms in a manner invariant to graph isomorphism in order to obtain a one-dimensional vector.

$$A'' = R_{solute}(A, A') \tag{8}$$

$$B'' = R_{solvent}(B, B') \tag{9}$$

Here, we experiment with two choices of $R$, the first one is sum pooling along the atom dimension and the second is a set2set layer (Vinyals, Bengio, and Kudlur 2015).

The outputs $A''$ and $B''$ are concatenated and are passed through three fully connected layers to predict the free energy of solvation. The intermediate layers have ReLU as the activation function.

## 5 Training

PyTorch framework was used for all our training and validation purposes. 10-fold cross validation scheme was used to assess the model due to the small size of the dataset. The dataset was randomly split in 10 subsets and one of the subsets was used as the test set the remaining 9 were used to train the model. This ensured that training set and test set were in 9:1 ratio. We made 5 such 10 cross validation splits and trained our model independently on each of them to better estimate the accuracy of the proposed model. The $T$ for messaging passing phase was chosen to be 3 as suggested by Gilmer et al.. ADAM optimizer with its default parameters as suggested by Kingma and Ba was used to train the model and mean square error was used as the objective function. The learning rate was decreased on plateau by a factor of $10^{-1}$ from $10^{-2}$ to $10^{-5}$.

## 6 Results

As the baseline for this work, we use two MPNNs (one each for solute and solvent) with their outputs concatenated for prediction of the final free energy of solvation. This choice of baseline allows us to analyze the importance of interaction phase introduced in this work. Further, we train two variants of CIGIN, the first one uses a sum pooling layer whereas the second model uses a set2set layer(Vinyals, Bengio, and Kudlur 2015) as the choice for both the functions $g$ and $R$ in the message passing and prediction phase respectively. Table 3 shows the comparison of RMSE between the baseline model and two variants of CIGIN on the MNSol dataset averaged across 5 independent 10-fold cross validation runs. We find that the proposed model CIGIN (set2set) performs the best with a RMSE of 0.57 kcal/mol on the test set. Further, to show that the embedding methods are effective, we performed two experiments; one by removing the

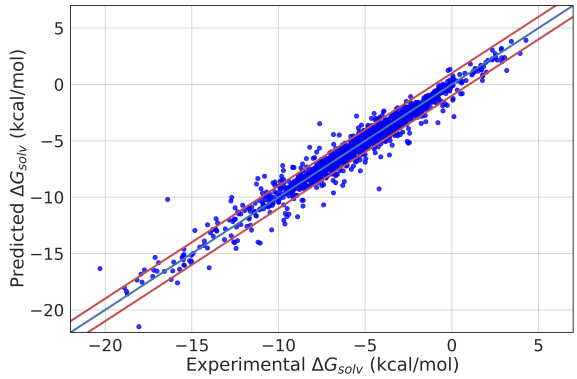


Figure 2: Plot of predicted (averaged over the five independent 10-fold cross validation runs) versus experimental solvation free energies. Predictions with error less than 1 kcal/mol lied between the two red lines.

| Model | RMSE (kcal/mol) |
|---|---|
| Baseline model | $0.65 \pm 0.13$ |
| CIGIN (sum pooling) | $0.61 \pm 0.12$ |
| CIGIN (set2set) | $\mathbf{0.57 \pm 0.10}$ |

Table 3: Average RMSE of five independent 10-fold cross validation runs of models on MNSol Dataset.

messaging passing phase for solvent and another by removing for both solute and solute. The RMSEs were found to be 0.68 and 1.73 kcal/mol. However, these model would not be able to provide chemical insights like the one proposed.

## Validation on External Dataset

FreeSolv is an experimental benchmark dataset that consists of free energies of solvation for 642 different small organic molecules in water solvent(Mobley and Guthrie 2014). The model presented here was tested on the FreeSolv dataset, which was not used in the training. We use the model trained on MNSol dataset to predict the hydration free energy of molecules that are in FreeSolv ($\sim$50%) but not in our training set. We report an RMSE of 1.06 kcal/mol averaged across all models trained during the 5 independent 10 fold cross validation runs. Wu et al. in their work, MoleculeNet, provide a benchmark for prediction of hydration free energy on the FreeSolv dataset using ML methods. Table 4 gives the performance of a single MPNN (the best performing model reported in MoleculeNet) and Delfos method.

| Model | Validation set | Test set |
|---|---|---|
| MPNN (MoleculeNet) | $1.20 \pm 0.02$ | $1.15 \pm 0.12$ |
| Delfos | $1.16 \pm 0.03$ | $1.19 \pm 0.08$ |
| CIGIN (set2set) | $0.73 \pm 0.01$ | $\mathbf{0.91 \pm 0.06}$ |

Table 4: Average RMSE of predicted solvation free energies on FreeSolv dataset employing the current architecture, MoleculeNet and Delfos.

| Solvent | Dielectric constant ($\epsilon$) | RMSE (kcal/mol) |
|---|---|---|
| Methylformamide | 181.56 | $0.29 \pm 0.10$ |
| Nitrobenzene | 34.80 | $0.28 \pm 0.04$ |
| Butanol | 17.33 | $0.30 \pm 0.03$ |
| Heptanol | 11.32 | $0.18 \pm 0.02$ |
| Octanol | 9.86 | $0.85 \pm 0.05$ |
| Chloroform | 4.71 | $0.83 \pm 0.04$ |
| Hexadecane | 2.04 | $0.58 \pm 0.02$ |

Table 5: Average RMSE of the proposed model in solvent holdout test for different solvents with wide ranging polarities. The results are averaged over five independent runs excluding each solvent.

These models were trained explicitly on FreeSolv using a 80-10-10 (train,validation,test) data split as implemented in MoleculeNet. By following data split for train, validation and test similar to the other two benchmarks, we show that this architecture yields an RMSE of 0.91 kcal/mol (see Table 4). This is very similar to the RMSE obtained without retraining the model, which attests to the robustness of the method. Though Delfos method achieves similar accuracy, CIGIN exhibits better transferability to unknown solute-solvent combinations. The above experiments demonstrate that, 1) the proposed model significantly outperforms the baselines reported in MoleculeNet and Delfos (Wu et al. 2018; Lim and Jung 2019), 2) the model is transferable and, 3) explicit consideration of solvent aids the model to better realize the interactions that increase/decreases the solubility of a molecule. As mentioned above, the main advantage of the method proposed here is that it applies to any organic solvent and not just water, and chemical interpretability, which is discussed later.

## Transferability of the Model to Other Solvents

For this model to be applicable to all organic solvents, the model should satisfactorily predict solvation free energies on solvents that are not part of the training set. To demonstrate this, we holdout a particular solvent from the train set and test our model on it. For this particular experiment, we considered all solvents and select ones that cover wide ranging polarity are given in Table 5 (full data is available in the supporting material). The mean RMSE of all holdout tests of 0.41 kcal/mol show that the model is robust for predicting solvation free energies involving any solvent with high accuracy. This indicates that the model is truly transferable to different types of solvents that may mimic biological fluid (polar), cell membrane (non polar) or anything in between that is relevant in other fields of science and technology. The following sections present the chemical interpretability of the model by taking few examples.

## Chemical Insights from the Model

A major drawback of the current state of deep learning methods is their general lack of interpretability, though several research groups are working in this direction. For a model to

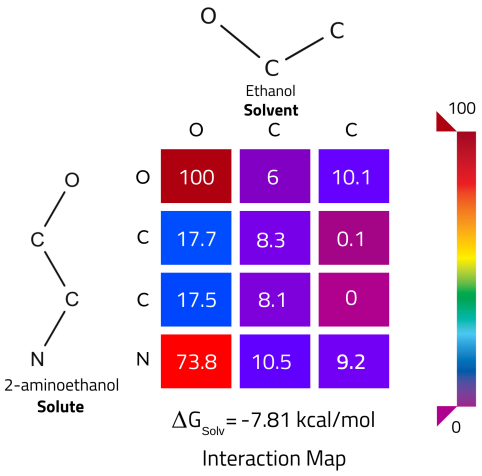


Figure 3: Heat map of the normalized (min-max) interaction map for 2-aminoethanol (solute) and ethanol(solvent) along with the predicted solvation free energy.

be of practical importance and to be widely accepted especially in the field of natural sciences, it is important that the model is not a mere black box but can also provide scientific insights. In our case, we have shown above that the current model is able to predict solubility properties of a wide ranging drug-like molecules involving entire spectrum of organic solvents. We examine if the model is able to to explain the contributing factors that can provide molecular and atomistic level interpretations of the predicted free energies of solvation using the interaction map by taking few examples.

**Factors Affecting Solubility** It is well known that 2-aminoethanol (NCCO) is highly soluble in ethanol (CCO), which is also reflected in the large negative solvation free energy (-7.81 kcal/mol). A chemist would suggest that the major contributing factor to this effect is the possibility of multiple hydrogen bonding between solute (O and N) and solvent (O) atoms via a hydrogen atom connected to one of these atoms due to Coulombic/electrostatic forces. Normalized interaction map showing the pairwise contributions to the overall solvation free energy was analyzed to examine if it captures this chemical insight (see Figure 3). As is evident from the data, highest contribution to the solvation free energy comes from O...O and N...O atom pairs in accordance with the chemical wisdom. As mentioned earlier, hydrogen atoms are not considered explicitly as nodes in the molecular graph representation. However, high contributions corresponding to strong interactions via hydrogen bonding and low contributions corresponding to weak dipole-induced dipole and van der Waals interactions are efficiently captured by the interaction map.

**Impact of Hydrogen Bonding and Steric Factors** The spatial environment of an atom within a molecule determines its extent of accessibility to the solvent environment and in turn the degree of intermolecular interactions. These are referred to as steric factors, which along with the other

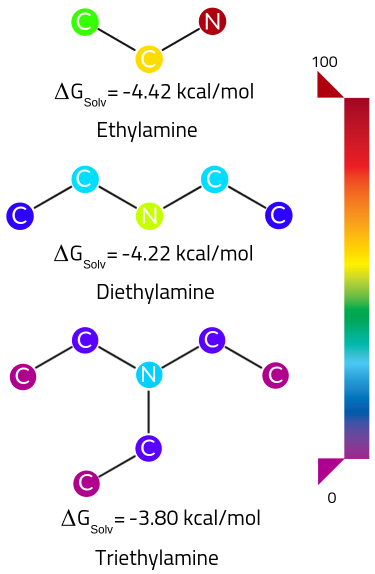


Figure 4: Interaction between the oxygen of water (solvent) and atoms of primary (ethylamine), secondary (diethylamine) and tertiary amines (triethylamine). The fractional interaction of each atom were obtained and were normalized (min-max) across the three solutes. The predicted hydration free energy values are also given.

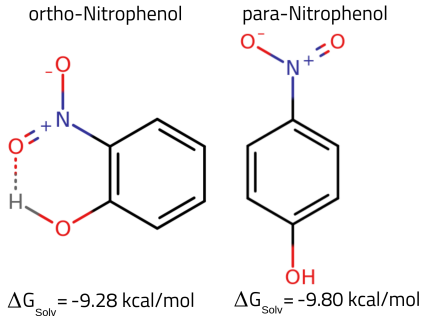


Figure 5: The structures of ortho- and para-nitrophenols showing the possibility of intramolecular hydrogen bond only in the ortho isomer along with the predicted hydration free energies.

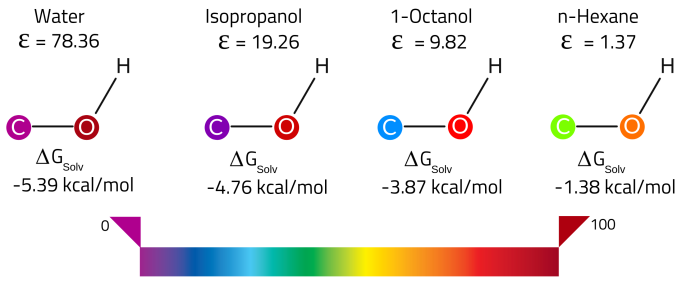


Figure 6: Predicted values of the solvation free energies of methanol in four different solvents with varying polarities ($\epsilon$). The contribution of the interaction of C and O atoms of methanol with the each of these solvents were calculated and the min-max normalized values of the fractional interactions are given.

intermolecular interactions such as hydrogen bonding are vital in dictating the free energy of solvation. The interaction maps of primary, secondary and tertiary amine solutes with water as the solvent are analyzed to examine if it captures these differential effects. Presence of hydrogen atoms in the primary and secondary amines and presence of nitrogen atom in all the three enables hydrogen bonding with water. Additionally, the crowding around the nitrogen atom increase going from primary to tertiary amine which leads to reduction in the hydrogen bonding ability in the same order. This effect is apparent from the predicted hydration free energies (see Figure 4). The pairwise interactions between the solute atoms and O of water from the interaction map is also given in the Figure which predicts that the contribution from the nitrogen atoms gradually reduces from primary to tertiary amines in line with the chemical intuition.

**Intramolecular Effects** Molecular interactions within a molecule may impede the ability of certain chemical groups to form favorable interactions with the solvent. For example, possibility of intramolecular hydrogen bonding (hydrogen bond involving atoms within a given molecule) reduces the solubility compared to when it is absent. A classic example is the hydration free energy difference between para-nitrophenol and ortho-nitrophenol (see Figure 5). Both these molecules are same in terms of the types of atoms/functional groups, but are different in the position of the nitro ($NO_2$) and hydroxyl (OH) groups with respect to each other. The para isomer is predicted to have more favorable hydration free energy compared to the ortho isomer due to the possibility of intramolecular hydrogen bonding in the latter. Further on analyzing the interaction map, we find that interaction of the phenol group (solute) with the oxygen of water is more for para-nitrophenol than ortho-nitrophenol. This shows that the model correctly interprets the formation of intramolecular hydrogen bond in the case of ortho-nitrophenol.

**Solvent Polarity** A bond in a molecule exhibits a dipole moment if the participating atoms have different electronegativity values. Solvent molecules on the basis of the net dipole moments can be broadly classified as polar and nonpolar. Solvents that have large dipole moments are regarded as polar solvents or nonpolar solvents otherwise. 'Like dissolves like' is a common term used by chemists to explain the solubility of polar solutes in polar solvents and nonpolar solutes in nonpolar solvents, and to explain the fact that solubility of a polar solute decreases with respect to decrease in the polarity of the solvent. To validate if the model is in line with this, we consider the solubility of methanol (a polar solute) in various solvents of varying polarity (quantified by $\epsilon$). Figure 6 indicates that decrease in the dielectric constant leads to unfavorable solvation free energies (going from -5.40 to -1.39 kcal/mol). From the interaction map, it is clear that the contribution from the the hydrophobic part (carbon) increases and that of hydrophilic part (oxygen) decreases with respect to decrease in the solvent polarity consistent with the chemical understanding of the effect.

## 7 Conclusion

Solvation free energy as a property has widespread applications in diverse fields of science and technology. The current manuscript reports a novel method based on graph interaction network for predicting solvation free energies involving drug-like/small organic molecules and generic organic solvents. We have demonstrated the robustness of this method by validating it by using different datasets and solvent hold-out tests. Interaction maps calculated for each solvent-solute pair is demonstrated to reveal essential molecular/atomic level details of the inter/intramolecular interactions making the model chemically interpretable. Given that this model accurately captures interaction between solute and solvent molecules, it is possible to extend this approach to chemical and biological problems that involve interactions between two molecular systems. For example, the interaction map used here may be extended to quantify drug-receptor interactions, which can be used for computer enabled identification of new chemical entities that bind to disease relevant protein targets, a key exercise in pharmaceutical industry. Such a model can also be used in lead optimization in drug design, where the molecule is modified to maximize certain contributions in the interaction map to enhance the binding of a drug to a biological receptor.

## 8 Acknowledgment

We thank Dr. Girish Varma for his feedback on the manuscript. We thank DST-SERB (grant no. EMR/2016/007697) for financial assistance.

## References


Behler, J., and Parrinello, M. 2007. Generalized neural-network representation of high-dimensional potential-energy surfaces. *Phys. Rev. Lett.* 98(14):146401.

Butler, K. T.; Davies, D. W.; Cartwright, H.; Isayev, O.; and Walsh, A. 2018. Machine learning for molecular and materials science. *Nature* 559(7715):547.

Cho, H.; Choi, I.; et al. 2018. Three-dimensionally embedded graph convolutional network (3dgcn) for molecule interpretation. *arXiv preprint arXiv:1811.09794*.

Duarte Ramos Matos, G.; Kyu, D. Y.; Loeffler, H. H.; Chodera, J. D.; Shirts, M. R.; and Mobley, D. L. 2017. Approaches for calculating solvation free energies and enthalpies demonstrated with an update of the freesolv database. *J. Chem. Eng. Data* 62(5):1559–1569.

Duvenaud, D. K.; Maclaurin, D.; Iparraguirre, J.; Bombarell, R.; Hirzel, T.; Aspuru-Guzik, A.; and Adams, R. P. 2015. Convolutional networks on graphs for learning molecular fingerprints. In *Advances in Neural Information Processing Systems*, 2224–2232.

Gilmer, J.; Schoenholz, S. S.; Riley, P. F.; Vinyals, O.; and Dahl, G. E. 2017. Neural message passing for quantum chemistry. In *Proceedings of the 34th International Conference on Machine Learning-Volume 70*, 1263–1272. JMLR.org.

Goh, G. B.; Siegel, C.; Vishnu, A.; and Hodas, N. 2018. Using rule-based labels for weak supervised learning: a chemnet for transferable chemical property prediction. In *Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, 302–310. ACM.

Goh, G. B.; Hodas, N. O.; and Vishnu, A. 2017. Deep learning for computational chemistry. *J. Comput. Chem.* 38(16):1291–1307.

Hansen, K.; Biegler, F.; Ramakrishnan, R.; Pronobis, W.; Von Lilienfeld, O. A.; Muller, K.-R.; and Tkatchenko, A. 2015. Machine learning predictions of molecular properties: Accurate many-body potentials and nonlocality in chemical space. *J. Phys. Chem. Lett.* 6(12):2326–2331.

Hutchinson, S. T., and Kobayashi, R. 2019. Solvent-specific featurization for predicting free energies of solvation through machine learning. *J. Chem. Inf. Model.* 59(4):1338–1346.

Jaeger, S.; Fulle, S.; and Turk, S. 2018. Mol2vec: unsupervised machine learning approach with chemical intuition. *J. Chem. Inf. Model.* 58(1):27–35.

Jin, W.; Barzilay, R.; and Jaakkola, T. 2018. Junction tree variational autoencoder for molecular graph generation. In Dy, J., and Krause, A., eds., *Proceedings of the 35th International Conference on Machine Learning*, volume 80 of *Proceedings of Machine Learning Research*, 2323–2332. Stockholmsmssan, Stockholm Sweden: PMLR.

Kearnes, S.; McCloskey, K.; Berndl, M.; Pande, V.; and Riley, P. 2016. Molecular graph convolutions: moving beyond fingerprints. *J. Comput. Aided Mol.* 30(8):595–608.

Kingma, D. P., and Ba, J. 2014. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*.

Klamt, A., and Diedenhofen, M. 2010. Blind prediction test of free energies of hydration with cosmo-rs. *J. Comput. Aided Mol.* 24(4):357–360.

Laghuvarapu, S.; Pathak, Y.; and Priyakumar, U. D. 2019. Band nn: A deep learning framework for energy prediction and geometry optimization of organic small molecules. *ChemRxiv*.

Landrum, G. Rdkit: Open-source cheminformatics.

Li, Y.; Tarlow, D.; Brockschmidt, M.; and Zemel, R. 2015. Gated graph sequence neural networks. *arXiv preprint arXiv:1511.05493*.

Lim, H., and Jung, Y. 2019. Delfos: deep learning model for prediction of solvation free energies in generic organic solvents. *Chem. Sci.* 10:8306–8315.

Liu, Q.; Allamanis, M.; Brockschmidt, M.; and Gaunt, A. 2018. Constrained graph variational autoencoders for molecule design. In *Advances in Neural Information Processing Systems*, 7795–7804.

Marenich, A. V.; Kelly, C. P.; Thompson, J. D.; Hawkins, G. D.; Chambers, C. C.; Giesen, D. J.; Winget, P.; Cramer, C. J.; and Truhlar, D. G. 2012. Minnesota solvation database. *Minnesota Solvation Database version* 20.

Marenich, A. V.; Cramer, C. J.; and Truhlar, D. G. 2009. Performance of sm6, sm8, and smd on the sampl1 test set for the prediction of small-molecule solvation free energies. *J. Phys. Chem. B* 113(14):4538–4543. PMID: 19253989.

Mobley, D. L., and Guthrie, J. P. 2014. Freesolv: a database of experimental and calculated hydration free energies, with input files. *J. Comput. Aided Mol.* 28(7):711–720.

Rupp, M.; Tkatchenko, A.; Müller, K.-R.; and Von Lilienfeld, O. A. 2012. Fast and accurate modeling of molecular atomization energies with machine learning. *Phys. Rev. Lett.* 108(5):058301.

Schütt, K.; Kindermans, P.-J.; Felix, H. E. S.; Chmiela, S.; Tkatchenko, A.; and Müller, K.-R. 2017. Schnet: A continuous-filter convolutional neural network for modeling quantum interactions. In *Advances in Neural Information Processing Systems*, 991–1001.

Vinyals, O.; Bengio, S.; and Kudlur, M. 2015. Order matters: Sequence to sequence for sets. *arXiv preprint arXiv:1511.06391*.

Wu, Z.; Ramsundar, B.; Feinberg, E. N.; Gomes, J.; Geniesse, C.; Pappu, A. S.; Leswing, K.; and Pande, V. 2018. Moleculenet: a benchmark for molecular machine learning. *Chem. Sci.* 9(2):513–530.

You, J.; Liu, B.; Ying, Z.; Pande, V.; and Leskovec, J. 2018. Graph convolutional policy network for goal-directed molecular graph generation. In *Advances in Neural Information Processing Systems*, 6410–6421.

Zhou, Z.; Kearnes, S.; Li, L.; Zare, R. N.; and Riley, P. 2019. Optimization of molecules via deep reinforcement learning. *Sci. Rep* 9(1):10752.