Learning Atomic Interactions through Solvation Free Energy Prediction Using Graph Neural Networks

by

Yashaswi Pathak, Sarvesh Mehta, U Deva Priyakumar

in

JCIM

Report No: IIIT/TR/2021/-1

Centre for Computational Natural Sciences and Bioinformatics
International Institute of Information Technology
Hyderabad - 500 032, INDIA
May 2021

JCIM JOURNAL OF CHEMICAL INFORMATION AND MODELING

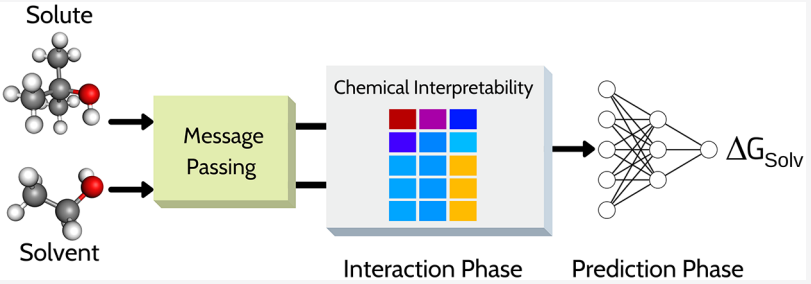
pubs.acs.org/jcim Article

Learning Atomic Interactions through Solvation Free Energy Prediction Using Graph Neural Networks

Yashaswi Pathak, Sarvesh Mehta, and U. Deva Priyakumar*

Cite This: J. Chem. Inf. Model. 2021, 61, 689−698 Read Online

ACCESS | Metrics & More | Article Recommendations | Supporting Information



ABSTRACT: Solvation free energy is a fundamental property that influences various chemical and biological processes, such as reaction rates, protein folding, drug binding, and bioavailability of drugs. In this work, we present a deep learning method based on graph networks to accurately predict solvation free energies of small organic molecules. The proposed model, comprising three phases, namely, message passing, interaction, and prediction, is able to predict solvation free energies in any generic organic solvent with a mean absolute error of 0.16 kcal/mol. In terms of accuracy, the current model outperforms all of the proposed machine learning-based models so far. The atomic interactions predicted in an unsupervised manner are able to explain the trends of free energies consistent with chemical wisdom. Further, the robustness of the machine learning-based model has been tested thoroughly, and its capability to interpret the predictions has been verified with several examples.

INTRODUCTION

Application of modern artificial intelligence/machine learning (AI/ML) methods to problems in fundamental science has surged in the last few years. This seems to alter the nature of computations that are traditionally performed in chemistry, biology, and related areas.[1−3] Typically, computations are performed using computer programs developed based on methods in physics such as quantum mechanics and classical mechanics. However, machine learning algorithms use existing data to build predictive models. Machine learning methods are effective in processing high-dimensional data, which are in general difficult for human perception or existing methods. In the recent past, a number of studies on development of machine learning models to predict quantum mechanical/density functional theory energies, drug properties, retrosynthetic pathways for organic molecules, material properties, etc. have been reported.[4−10] Several successes of machine learning methods within a short span of time seem to give rise to a perception that these methods will become valuable tools in scientific research. However, one of the most important criticisms of machine learning methods is the explainability. There have been efforts in this direction in the area of artificial intelligence to develop algorithms that not only make accurate predictions but also are capable of interpretations that allow for the possible understanding of the phenomenon.[11−13]

Solvation, a process that is primarily driven by the nature of interactions between solute and solvent molecules, is of immense importance in a number of physical, chemical, and biological processes (protein folding, chemical reactivity, protein−substrate binding, colloids, etc.).[14,15] In fact, a large majority of chemical and biological processes occur in solution, and hence, solvation free energy plays a central role. For example, all cellular processes including protein−protein interactions, protein−nucleic acid complex formation, protein−substrate formation, protein folding, etc.[16−18] that are essential for the existence of life occur in aqueous conditions, and these processes are often assisted by solvent molecules. Solvation free energy is the free energy change when a molecule in its gaseous state is transferred to a given solvent. Solvation free energy is related to a number of target properties crucial in molecular design, among other important applications. The pharmacokinetic properties, namely, distribution and absorption, can be assessed using the solvation free energies.[19−21]

Received: December 8, 2020
Published: February 5, 2021

https://dx.doi.org/10.1021/acs.jcim.0c01413
J. Chem. Inf. Model. 2021, 61, 689−698

The calculation of solvation free energies is typically carried out using molecular dynamics (MD) simulations and primarily using the alchemical free energy methods.[22−24] MD simulations are performed on a series of nonphysical intermediate states between the solute in the gas phase and in the solution phase. This serves as a path to calculate the free energy difference between the two states. The two most commonly used alchemical methods are free energy perturbation (FEP) and thermodynamic integration (TI).[25] Bennett's acceptance ratio (BAR) and multistate Bennett's acceptance ratio (MBAR) methods are also useful for the calculation of solvation free energies.[23,26] These methods have been shown to yield free energy values that are comparable to experiments.[27−30] Accuracies of free energy calculations mentioned above are mostly limited by the quality of the force field used for both solute and solvent. These methods are also computationally expensive, which makes quantitative quick and accurate estimation of solvation free energies impractical. Additionally, deriving an adequate quality force field involves a large number of quantum chemical calculations and molecular dynamics simulations, which adds to the computational complexity of the problem. Density functional theory and quantum mechanical calculations are also useful in calculating solvation free energies typically using continuum solvation models.[31,32]

In the past few years, several machine learning approaches[2,33,34] have been proposed to predict aqueous solubility of druglike molecules using the FreeSolv data set.[35] These methods are inherently restricted to the prediction of free energy of solvation for a single solvent (water) and cannot be generalized to all organic solvents. Some factors that contribute to the extent of solubility of a solute in any solvent, for example, H-bonding and hydrophobic interactions, can be extended to any other solute−solvent pair. Hence, a generic method can be used to determine the solubility of solutes in a new solvent. Recently, Lim and Jung proposed a model "Delfos" that used recurrent neural networks to predict the solubility of molecules in any generic solvent.[36] In the Delfos model, molecular embedding was obtained from SMILES sequence using Mol2Vec[37] featurization technique. Although SMILES representations have been widely used as feature representations for druglike molecules, there has been a widespread shift to chemical graph-based representations lately.[38,39] This can be explained due to the limitations of the SMILES-based methods. First, the SMILES representations do not explicitly account for molecular similarity. There are challenges associated with the learning of the SMILES syntax using deep learning-based models. On the other hand, the chemical graph-based representations appropriately capture the molecular structures and can more readily model the pharmacophores associated with chemical properties.[39,40]

As discussed above, machine learning methods are capable of performing prediction tasks accurately and have started to contribute to the area of molecular science significantly.[13,38] However, there is a strong need to develop methods that also offer explainability in addition to accurate predictions. In the current study, a novel method, namely, chemically interpretable graph interaction network (CIGIN), for the prediction of solvation free energy of small organic molecules with respect to any commonly used organic solvents is proposed. A preliminary version of this method was presented at a conference recently.[41] The method proposed works on molecular graphs, using a message passing neural network (MPNN)[42] and an interaction layer to accurately model the free energy of solvation. This method predicts solvation free energies accurately, and the interaction map calculated as part of the prediction task captures chemical insights that explain the magnitude of the solvation free energies. Extensive analysis has been performed to assess the predictive capability and robustness of the model, and several examples have been used to demonstrate the chemical interpretability of the interaction map. Further, the potential use case of this model is illustrated by taking prodrug development as an example.

METHODS

Data Set. In this work, we use a combination of Solv@TUM database[43] and FreeSolv data set.[35] Solv@TUM database has 5952 experimental values for the free energy of solvation. The molecules in this database contain chemically diverse organic and inorganic molecules in nonaqueous solvents.[43] For our purpose, molecules consisting of elements C, H, N, O, F, P, Cl, S, Br, and I were used and the rest were filtered out, which results in a total of 5597 unique solute−solvent pairs. As this data set only contains solvation free energies in nonaqueous solvents, the FreeSolv data set, which contains 642 hydration free energies of organic molecules, was combined with this set. This results in a total of 6239 unique combinations of solute−solvent pairs with 935 unique solutes and 146 unique solvents. The final data set is processed using RDKit[44] to construct molecular graphs.

Molecular Graph. Molecules, due to their inherent structure can naturally be represented as graphs, where nodes correspond to atoms and edges correspond to bonds. More formally, for a molecule M, we construct an undirected graph $G(V, E)$, where the set of atoms correspond to V and the set of edges correspond to E. x_v and e_{vw} are the node and edge features corresponding to node v and the edge e between nodes v and w, respectively. The choice of these feature vectors depends upon the problem at hand. The feature vectors are chosen to appropriately model the solute−solvent features and their intermolecular pharmacophore, and hence the atomic features are selected to capture the electronic and structural features of the atom. Similarly, the bond features are selected to determine the type of bond, the kind of topology (ring or aliphatic) it is present in, and whether it shows stereochemistry. The node features and bond features are given in Tables 1 and 2, respectively. RDKit[44] was used to extract the given node and edge features, and the deep graph library (DGL)[45] was used to prepare the molecular graph.

Model. As mentioned above, the proposed model consists of three different phases: message passing phase, interaction phase,

Table 1. 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 neighboring 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)

https://dx.doi.org/10.1021/acs.jcim.0c01413
J. Chem. Inf. Model. 2021, 61, 689−698

Table 2. 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 stereochemistry	*E* or *Z* (one-hot)

and prediction phase. The three phases are explained in subsequent sections. Figure 1 illustrates the proposed model.

Message Passing Phase. In this phase, the message passing neural network (MPNN), a general framework for supervised learning of graph-structured data proposed by Gilmer et al., is used.[42] Consider a molecular graph $G(V, E)$ having node features x_v and edge features e_{vw}. The state of each node is represented as h_v^t, which is initialized to x_v at $t = 0$ and are updated for T time steps using messages m_v^{t+1} and vertex update function U_t according to the following equations

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

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

$N(v)$ represents the set of neighboring nodes of v. In this work, both M_t and U_t are fully connected layers. After T time steps, a residual connection[46] from x_v to h_v^t is added nodewise according to the following equation

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

F_v represents the final atomic feature for each atom $v \in G(V, E)$. Each row describes the atomic property and the local environment of the corresponding atom. T was taken as 6 for all of the experiments performed in this work.

Molecular graphs corresponding to both solute and solvents are fed through separate MPNNs. The outputs F of this phase are tensors A (solute) and B (solvent) of sizes $J * L$ and $K * L$,

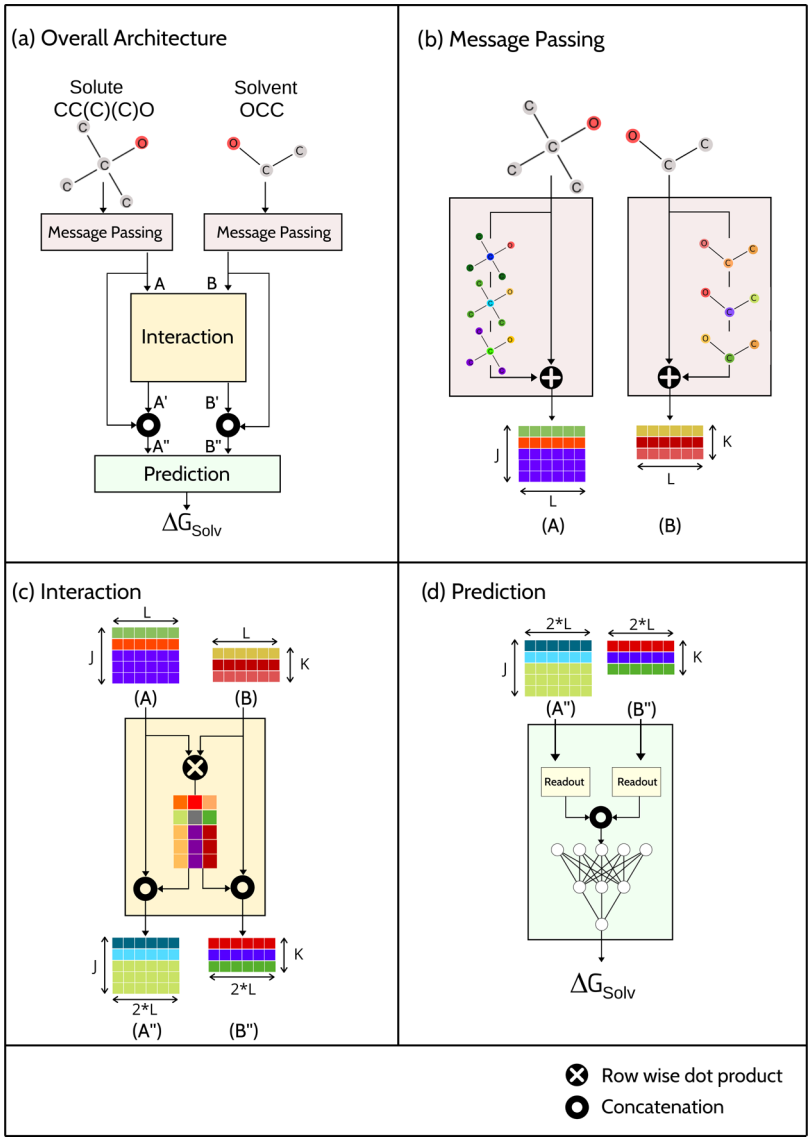


Figure 1. (a) Architecture based on the graph neural network for the prediction of solvation free energies of any generic organic solvent. The three main constituents of the architecture, namely, (b) message passing, (c) interaction, and (d) prediction phases, are also given. For more details, see the Methods section.

https://dx.doi.org/10.1021/acs.jcim.0c01413
J. Chem. Inf. Model. 2021, 61, 689−698

respectively, for a solute of J nonhydrogen atoms, a solvent of K nonhydrogen atoms, and L atomic features.

Interaction Phase. As discussed earlier, the interactions between the atoms of solute and solvent molecules play a key role in determining the solubility of a solute in a given solvent. These interactions are due to the electronic and steric factors of the atoms of the solute and solvent molecules. We capture the pairwise interactions between solute and solvent atoms in an interaction map.

For the 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

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

In the above equation, I denotes the computed interaction map using solute features (A) and solvent features (B). The function f models the interatomic interaction between every solute−solvent atom pair from their feature vectors. The function f needs to precisely measure the negative/positive contribution of a specific solute−solvent pair independently. For example, hydrophilic−hydrophobic interactions would decrease the solubility whereas hydrophilic−hydrophilic/hydrophobic−hydrophobic interactions would increase the solubility. To model these, different choices of function f are used (see Table 3).

Table 3. Accuracy of the CIGIN Model as Assessed by the Mean Absolute Error (MAE, kcal/mol) with Respect to Its Different Variants

model	interaction function	message passing	MAE (kcal/mol)
CIGIN	$\tan h(A.B)$	√	0.16 ± 0.002
CIGIN	$\tan h(W(A, B))$	√	0.16 ± 0.003
CIGIN	$\tan h(A.B)$	×	0.20 ± 0.002
CIGIN	×	√	0.23 ± 0.006

Further, the influence of solvent on solute and that of solute on solvent is calculated as

$$A' = I \cdot B \tag{5}$$

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

A' and B' are the solute and solvent features weighted by their contribution to the free energy of solvation.

Prediction Phase. The outputs A and B of the message passing phase represent the intramolecular context of both solute and solvent molecules, whereas the outputs A' and B' of the interaction layer represent the effect of solvent on solute and vice-versa, respectively. The free energy of solvation is governed by both these factors, i.e., the intramolecular and the intermolecular effects.

Hence, the outputs of both the message passing and the interaction phases are concatenated atomwise for both solute (A'') and solvent (B'') molecules. These are then transformed into a one-dimensional vector by combining the feature vectors across all atoms using a readout layer R according to equations given below

$$A'' = R_{\text{solute}}(A, A') \tag{7}$$

$$B'' = R_{\text{solvent}}(B, B') \tag{8}$$

A key thing to note is that the transformation function R should be invariant to graph isomorphism, and hence, a set2set layer,[47] which is invariant to permutation, is used. A''′ and B''′ represent the final feature vectors of both solute and solvent molecules, respectively. These are then concatenated and passed through three fully connected layers with dimensions 360:256:128 to predict the free energy of solvation.

Training. The molecular graphs were constructed using RDkit[44] and Deep Graph Library (DGL).[45] All of the training, validation, and analysis were performed using the PyTorch framework. A 10-fold cross-validation strategy was used for training the models. For this, the data set was split into 10 subsets; one of them was used as the test set and the remaining 9 became the training set. Hence, the train−test split was in the ratio of 9:1, and 10 independent train−test runs were performed. Further, to ensure minimum variance on the test set, five independent 10-fold cross-validation runs were performed. Mean squared error was used as the objective function, and along with this, an L2 regularization was also added on the interaction map. The ADAM optimizer[48] with its default parameter as suggested by Kingma and Ba were used,[48] and the learning rate was decreased on plateau by a factor of 10^{-1} from 10^{-3} to 10^{-5}. DGL was used to batch molecular graphs and train them together in batches of size 16. The rectified linear unit was used as the activation function in all of the layers except the last layer, where no activation function was used. The maximum epoch was set to 100, and early stopping was employed, so as to terminate the training when the model started to overfit. The architecture of the model is given in the Model Architecture section in the Supporting Information. All of the codes associated with the proposed method are made available at https://github.com/devalab/CIGIN/tree/master/CIGIN_V2

RESULTS AND DISCUSSION

In this section, the statistical fitness and the robustness of the CIGIN model are demonstrated by carrying out different experiments. Following this, the capability of the model to adapt to solvents unseen by the model during training is shown. In the subsequent subsections, several examples are discussed to illustrate the ability of the model to learn the underlying chemistry that affects the free energy of solvation. The potential use case of the model in prodrug development is demonstrated in the last subsection.

The CIGIN model presented in this work consists of three different phases, namely, the message passing phase, the interaction phase, and the prediction phase (see Figure 1). In the message passing, the node embedding of each atom in the solute and solvent molecules is computed; these are then used to measure the interatomic interactions between solute and solvent atoms in the interaction phase. Finally, the features from message passing and interaction phases are combined in the prediction phase to estimate the free energy of solvation for a given solute−solvent pair. The model is trained end-to-end so that the interatomic interactions are jointly learned while predicting the free energies of solvation.

CIGIN toward Chemical Accuracy. Figure 2 depicts the correlation between the experimental and predicted solvation free energies with respect to different solvent−solute pairs. A near-perfect linear relationship was observed between the two with an R^2 value of 0.98. More than 98% of the predictions are well within 1 kcal/mol of the experimental solvation free energy values. Table 3 lists the MAEs corresponding to the different variants of the CIGIN model. The table shows that the CIGIN model achieves a high accuracy of predicting solvation free energy with a mean absolute error of 0.16 ± 0.01 averaged upon

https://dx.doi.org/10.1021/acs.jcim.0c01413
J. Chem. Inf. Model. 2021, 61, 689−698

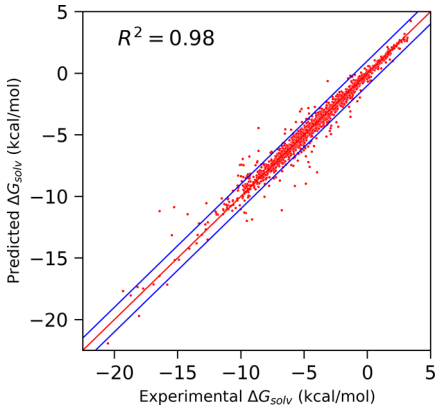


Figure 2. Plot of predicted (averaged over five independent 10-fold cross-validation runs) versus experimental solvation free energies.

five independent 10-fold cross-validation runs, indicating a very high accuracy of the model.

A comprehensive comparison of solvation free energies calculated with different methods and measured using experiments for a set of solutes and for a set of solvent molecules would be ideal. In the absence of such extensive data, we compare with previously published work. Several studies have attempted to reliably calculate/predict solvation free energies of small organic molecules.[21,23,31,32,34,36,49−51] These studies used quantum mechanical methods, molecular dynamics free energy calculations, and machine learning methods. So far, compared to all of the methods including high-level theoretical calculations, the proposed model significantly outperforms others on a diverse set of solute and solvent combinations.[21,23,36,50−52] The performance on another commonly used data set MNSOL Database[53] along with available results from QM methods is given in Table S1. Additionally, this model offers chemical interpretability that is not the case with most of the machine learning-based methods that have been proposed so far.[34,52]

Robustness of the Model. It is important to evaluate the model to test its robustness and to understand the different components (phases) of the architecture toward high accuracy. Hence, different experiments were performed to examine these factors. First, the message passing phase for both solute and solvent molecules was removed to examine the importance of learning graph embedding in an end-to-end fashion. Second, the message passing phase was retained and the interaction phase was removed so that the need to learn the interatomic interactions can be examined. From Table 3, it can be concluded that the learning of molecular embedding through the message passing layer helps the model to better capture the features that affect solubility. Further, we see that the role of the interaction phase helps the model in better capturing the interatomic contributions responsible for solubility or in other words to identify the favorable/unfavorable interacting atom pairs of both solute and solvent molecules.

FreeSolv is one of the widely used data sets, which comprises the hydration free energies of small organic molecules. To further study the importance of jointly learning the interatomic interactions via the interaction map along with the prediction of the solvation free energies, the CIGIN model was trained only on the FreeSolv data set.[35] The model was trained using the 80-10-10 train−test−validation split as given in MoleculeNet.[34] Table 4 compares the performance of CIGIN, Delfos,[36] and MPNN benchmarks provided in MoleculeNet.[34] Table 4 conclusively shows that CIGIN outperforms the other two baselines and obtains a high accuracy of 0.76 ± 0.11 on the test set. This also demonstrates that the joint learning of interaction between atoms aids the model in better performance.

Table 4. Comparison of the Performances of the CIGIN (Current Study), Delfos, and MPNN Models on FreeSolv Data Set Using the Mean Absolute Errors (kcal/mol)

model	validation set	test set
CIGIN	**0.67 ± 0.04**	**0.76 ± 0.11**
Delfos	1.16 ± 0.03	1.19 ± 0.08
MPNN	1.20 ± 0.02	1.15 ± 0.12

Accuracy upon Solvent Holdout. The current model is trained for a diverse range of organic solvents with varying polarities. To assess the ability of this model to predict the solvation free energy involving solvents unseen by the model, the solvent holdout test was performed. Briefly, 146 different experiments were conducted, where each experiment involves holding out solvation free energies involving a given solvent from the training set and using the data of that particular solvent as the test set. This would enable us to test if the model is able to accurately predict the free energies of solvation of the solvent that is not used in the training. The mean absolute error obtained over three independent complete holdout tests over all of the solvents was 0.18 kcal/mol; the distribution of MAE is shown in Figure 3, and the corresponding MAEs are given in Table S2 in the Supporting Information. The low MAE value of 0.18 kcal/mol indicates that the accuracy of the CIGIN model is not significantly affected upon solvent holdout. It is also evident from the distribution that solvation free energies corresponding to all solvents (except water) unseen by the model are predicted with very high accuracy. The model does not seem to adequately learn to predict the free energy of solvation for water when it is held out during the training. This can be explained due to the fact that water has only one heavy atom and hence message passing is not possible. Therefore, the model does not seem to learn from the other solvents, which have at least two heavy atoms.

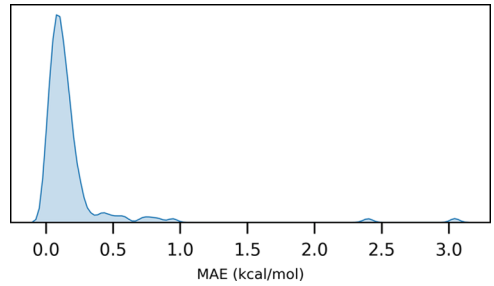


Figure 3. Probability distribution of mean absolute errors of the free energies of solvation corresponding to each solvent in the data set by excluding the given solvent during training (solvent holdout test).

Chemistry Learned by the Model. One of the major criticisms of machine learning applications, especially in natural sciences, is the lack of explainability. For the CIGIN model to be useful, it should not be a mere black box model but should also provide meaningful insights. These meaningful insights can help a chemist determine the cause of solubility (insolubility) of a solute in a solvent. In the following subsections, through a series of experiments, we show the meaningful insights that can be obtained from the proposed model.

https://dx.doi.org/10.1021/acs.jcim.0c01413
J. Chem. Inf. Model. 2021, 61, 689−698

Estimation of Intermolecular Interactions. The interaction phase in the CIGIN model is designed to quantify the interatomic interactions among all of the solute−solvent atom pairs. The interaction map calculated in the interaction phase (see Figure 1) is a $J \times K$ matrix (J and K being the number of nonhydrogen atoms of solute and solvent, respectively). These interactions play a key role in determining the solubility of a molecule and are governed by electronic and steric factors. The interaction map calculated for the t-butanol and ethanol solute−solvent pair is given in Figure 4. The 5×3 matrix corresponds to

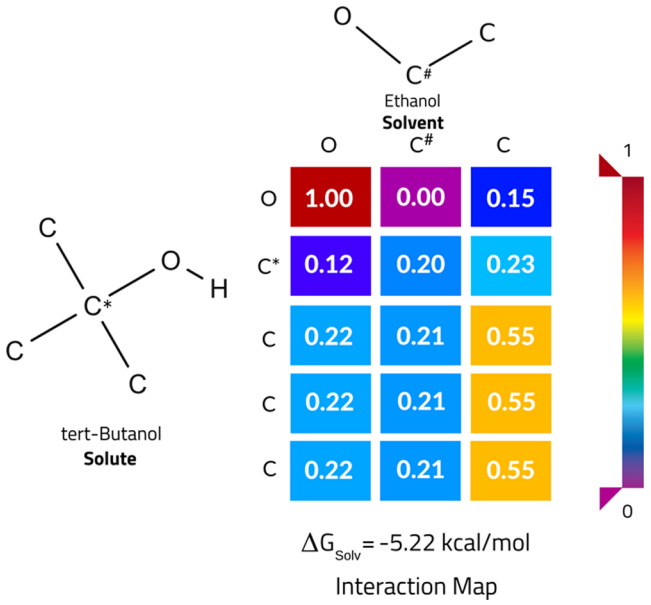


Figure 4. Interaction map between the atoms of *tert*-butanol (solute) and ethanol (solvent) along with the predicted free energy of solvation.

the interaction between all nonhydrogen atoms of the solute and the solvent molecules. Min-max normalized values of the calculated interaction map are depicted as a heat map. The most favorable interaction is observed between the two oxygen atoms, which would mean a hydrogen-bonded interaction in the chemical sense. The least interaction is observed between the oxygen of ethanol and the central carbon of the *t*-butanol. According to conventional wisdom, such an interaction is not favorable due to the inaccessibility of the central carbon atom and its hydrophobic nature. Intermediate values are observed between the terminal carbons, which can be thought of as hydrophobic contributions. Further examples of how the interaction map is able to capture the electronic factors and how the model is able to explain common chemical wisdom is demonstrated below.

Steric and Hydrophobic Factors. The ability of the model to capture the steric factors was tested by predicting the solvation free energies of a series of secondary amines. Starting from diethyl amine, the steric factor around the amine group was changed by introducing additional methyl groups on the α carbon atoms. The solvation free energies calculated for a polar and nonpolar solvent (water and n-hexane) are depicted in Figure 5a. In the polar solvent (water), the solubility of secondary amines is primarily governed by the interaction of the nitrogen atom of the solute and the oxygen atom of the solvent via formation of H-bonds, leading to favorable hydration free energy (−4.15 kcal/mol) for diethyl amine. Such an interaction would be more dominant when the participating atoms, i.e., nitrogen atom, in solute can readily be accessible by the oxygen atom in the solvent. This would mean that as the steric hindrance is increased around the nitrogen atom, it becomes less accessible to the solvent atoms and also that the hydrophobicity of the molecule increases with respect to addition of more methyl groups. This is expected to lead to the lower solubility of the secondary amine in water with respect to the increase in crowding around the nitrogen atom. Figure 5a confirms that the model is able to learn such a steric factor that as the branching is increased on the adjacent or α carbon atoms connected to the nitrogen atom, the solubility decreases. On the other hand, in n-hexane solvent, an increase in solubility with increasing crowding around the nitrogen is expected. This is due to the fact that the hydrophobicity increases and at the same time, the polar NH group becomes less accessible. Figure 5a shows that the solubility of secondary amines increases with an increase in the number of carbon atoms and hence confirms that the proposed model follows the intuition of a chemist.

Intramolecular Interactions. The hydration free energies of 1,2, 1,3, and 1,4 isomers of dimethyl, dihydroxy, and diaminobenzenes along with the heat maps corresponding to the interaction maps are given in Figure 5b. The lack of hydrophilic substituents and major differences in the intramolecular interactions among dimethyl benzene isomers results in hydration free energies close to each other and with values around −1 kcal/mol. The isomers of dihydroxy benzene and diamino benzene are more readily soluble in water than the dimethyl benzenes, and this can be explained due to the possibilities of H-bonding between solute and solvent atoms, which is absent in dimethyl benzene. In addition to the nature of the electronic factors between the solute and solvent atoms, intramolecular interactions affect the solvation free energies. For example, among the three isomers of dihydroxy benzene (1,2, 1,3, 1,4), 1,2-dihydroxy benzene was found to have the least boiling and melting points.[54] This is possibly due to the intramolecular H-bonding in the 1,2-isomer and more stabilizing intermolecular interactions in the other two isomers. This interpretation is demonstrated by the CIGIN model as well. In Figure 5b, the oxygen and nitrogen atoms present in the 1,2-isomer of dihydroxy and diamino benzene, respectively, show less interaction w.r.t the oxygen atom of water when compared to the other two isomers. This ultimately results in higher hydration free energy for 1,2-isomer compared to those for the other two. However, one would expect this intramolecular effect to be less pronounced in diamino benzene than dihydroxy benzene possibly due to the fact that only one of the four hydrogen atoms is involved in intramolecular H-bonding as against one of two in the case of dihydroxy benzene. The same is predicted by the CIGIN model via the interaction map, where the δ contribution of the oxygen atoms between the ortho and para isomers of dihydroxy benzene is higher than that of diamino benzene. Such subtle differences in solubilities influenced by intramolecular molecular interactions are appropriately captured by the CIGIN model, and the interaction maps explain the electronic factors behind the change.

Transfer Free Energies. The examples discussed above demonstrated the ability of the model to learn the electronic and steric factors of the solute molecules that influence the solvation free energies and the ability of the model to capture these effects via the interaction maps. The proposed model due to its design can be used to predict the solvation free energy in any generic solvent. Here, we examine the effect of solvation free

https://dx.doi.org/10.1021/acs.jcim.0c01413
J. Chem. Inf. Model. 2021, 61, 689−698

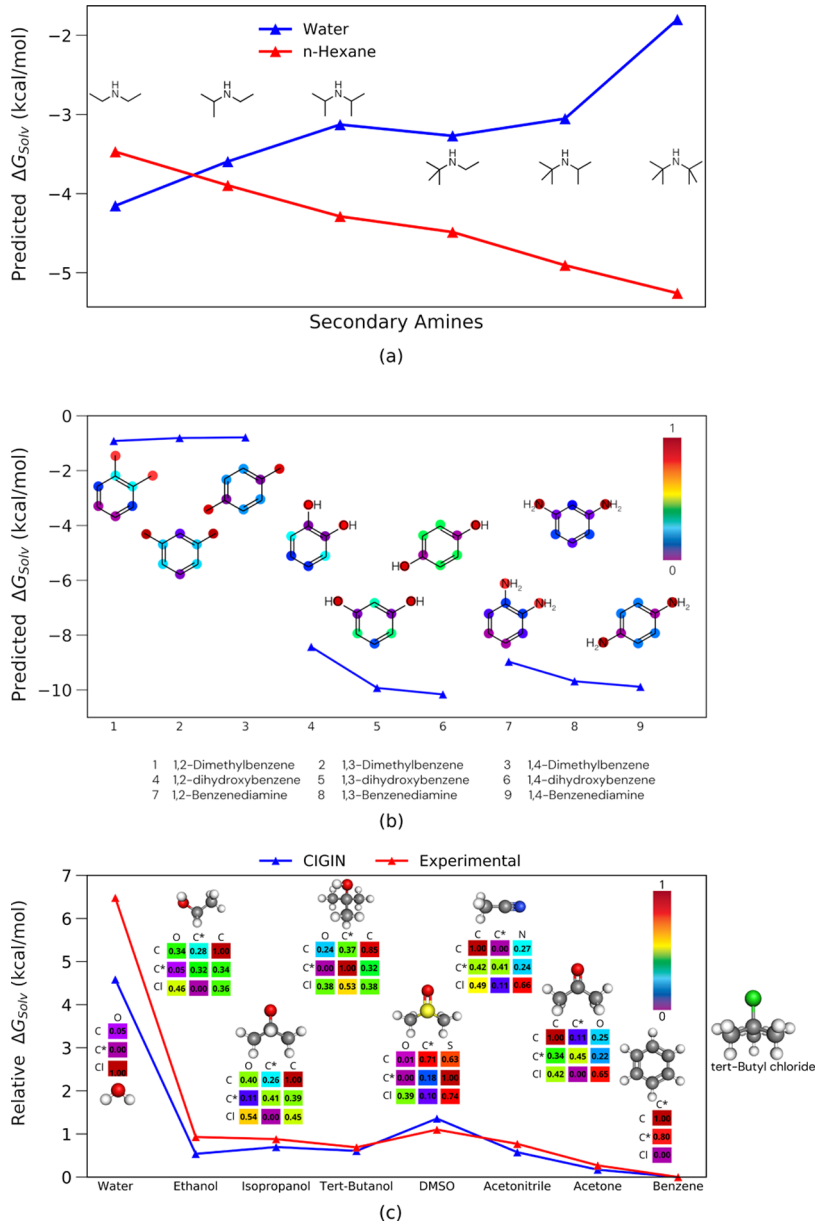


Figure 5. (a) Solvation free energies of secondary amines in water and *n*-hexane. (b) Variation in aqueous solubility w.r.t ortho, meta, and para substitution in benzene. (c) Plot of transfer solvation free energy of *t*BuCl between methanol and different solvents with respect to that in benzene (solvent in which *t*BuCl is most readily soluble) along with the individual interaction map between atoms of solute and solvent molecules. * and # denote the C connected to hetereoatoms in solute and solvent, respectively.

predicting free energy of solvation and estimating the interatomic interactions. Figure 5c compares the experimental observations[55] and the relative solvation free energies of t-butyl chloride in various solvents obtained using the CIGIN model. The figure also depicts the interaction maps for all of the solvents. The figure shows that the CIGIN model is able to follow the trend of the experimental values. As expected, t-butyl chloride is more soluble in nonpolar solvents compared to polar solvents. In addition to capturing the relative solvation free energies reasonably well, the interaction maps capture the possible hydrogen bond interactions with respect to polar groups (e.g., high value for interaction between the O of water and the Cl of t-butyl chloride) and hydrophobic interactions (e.g., high values for terminal methyl carbon of the solute and the methyl/aromatic carbons of the solvent). For nonpolar solvents, the interactions between chlorine and oxygen atoms are less dominating than those between carbon and carbon or chlorine and sulfur atoms; this can be explained due to the presence of stronger van der Waals forces between the atoms of the same period. This analysis further demonstrates that the CIGIN model is able to distinguish different types of chemical interactions that determine solubilities of small druglike organic molecules in diverse solvents.

Applications in Molecular Design. Understanding the structure−activity/property relationships is an important exercise for designing new material/molecules with desired properties. The concept of the interaction map introduced in the CIGIN model can be helpful in understanding the atomic-level details, especially when molecular designs involve two entities (e.g., designing molecules by maximizing the interactions with a certain protein). Here, we demonstrate this by taking prodrugs as an example. Designing prodrugs involves chemical

https://dx.doi.org/10.1021/acs.jcim.0c01413
J. Chem. Inf. Model. 2021, 61, 689−698

modification of high-value drug candidates to improve their pharmacokinetic properties and to decrease their toxicities.[56−58] For example, chemical modifications to a drug can effectively be used to alter its solubility profiles such that its bioavailability can be increased, or if the molecule is too hydrophilic, certain substitutions can be used to increase the lipophilicity so that permeation across cell membranes is improved.[56,58] Two such examples are provided in Figure 6. Penciclovir is a nucleoside

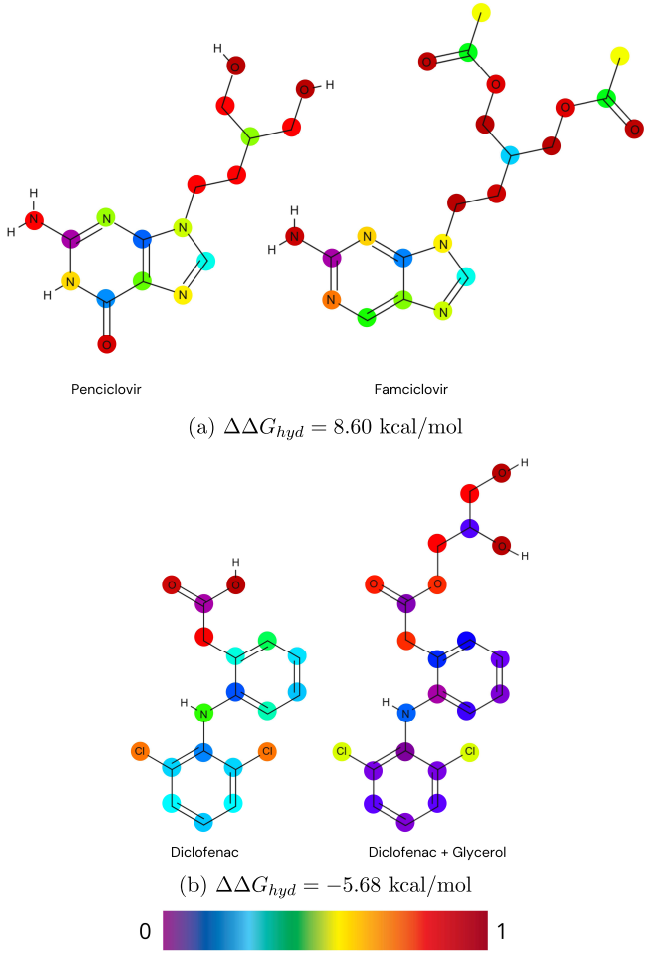


Figure 6. (a) Penciclovir and its prodrug famciclovir. (b) Diclofenac and its prodrug diclofenac and glycerol adduct.

analog that is an effective antiviral drug. However, the oral bioavailability of this drug is 4%, which improves to more than 75% upon chemical modification.[59−61] The prodrug, famciclovir, undergoes conversion to the original drug by esterases and upon oxidation. Another example is diclofenac and the glycerol ester of diclofenac. The glycerol adduct has been shown to improve the transdermal delivery of diclofenac when applied topically.[62,63] The $\Delta\Delta G_{hyd}$ between penciclovir and its prodrug, where the goal is to make the former more lipophilic, is 8.60 kcal/mol and that for diclofenac and its prodrug, where the goal is to increase aqueous solubility, is -5.68 kcal/mol. These predictions made by the CIGIN model confirm the experimental observation very well, and the change in the interaction map while going from the drug to the prodrug molecule explains the atomistic effects. Hence, when a molecule needs to be chemically modified to optimize its solubility profiles, a prediction model along with an interaction map as proposed here may aid in efficient optimization not only based on accurate predictions but also by directed modifications guided by the interaction maps.

CONCLUSIONS

In summary, estimation of solvation free energy is an important task and has diverse use cases. In this work, a novel method based on graph neural networks to predict the solubility of a molecule in any generic organic solvent has been proposed. The proposed framework consists of three phases, namely, message passing, interaction, and prediction phases. The interatomic interactions between the solute and solvent atoms are jointly learned in the end-to-end process via the interaction map. Several examples are used to demonstrate different chemical interactions that are captured in the interaction map. Further, different phases of models are ablated to understand their contribution to the high predictive capability of the CIGIN model. As a practical use case of this model, its potential application in prodrug development is demonstrated. The model proposed here can be used to study the interaction between any two molecular systems, such as drug−target interaction, and the interaction map introduced can be used to assign credit at the atom level along with that at a molecular level when optimizing for a property. Future work in this direction is in progress.

ASSOCIATED CONTENT

Supporting Information

The Supporting Information is available free of charge at https://pubs.acs.org/doi/10.1021/acs.jcim.0c01413.

Comparison of the CIGIN model on MNSOL data set with other machine learning models and quantum mechanical methods; detailed model architecture; transferability of the CIGIN model; and results of the solvent holdout test (PDF)

AUTHOR INFORMATION

Corresponding Author

U. Deva Priyakumar − *Center for Computational Natural Sciences and Bioinformatics, International Institute of Information Technology, Hyderabad 500032, India;* orcid.org/0000-0001-7114-3955; Phone: 0091 40 6653 1161; Email: deva@iiit.ac.in; Fax: 0091 40 6653 1413

Authors

Yashaswi Pathak − *Center for Computational Natural Sciences and Bioinformatics, International Institute of Information Technology, Hyderabad 500032, India*

Sarvesh Mehta − *Center for Computational Natural Sciences and Bioinformatics, International Institute of Information Technology, Hyderabad 500032, India*

Complete contact information is available at: https://pubs.acs.org/10.1021/acs.jcim.0c01413

Notes

The authors declare no competing financial interest.

ACKNOWLEDGMENTS

Siddhartha Laghuvarapu is thanked for his initial involvement in the work. DST-SERB grant (no. EMR/2016/007697) is acknowledged for financial assistance.

REFERENCES

(1) Butler, K. T.; Davies, D. W.; Cartwright, H.; Isayev, O.; Walsh, A. Machine Learning for Molecular and Materials Science. *Nature* **2018**, *559*, 547−555.

(2) Goh, G. B.; Hodas, N. O.; Vishnu, A. Deep Learning for Computational Chemistry. *J. Comput. Chem.* **2017**, *38*, 1291−1307.

(3) Mater, A. C.; Coote, M. L. Deep Learning in Chemistry. *J. Chem. Inf. Model.* **2019**, *59*, 2545−2559.

(4) Senior, A. W.; Evans, R.; Jumper, J.; Kirkpatrick, J.; Sifre, L.; Green, T.; Qin, C.; Žídek, A.; Nelson, A. W. R.; Bridgland, A.; Penedones, H.; Petersen, S.; Simonyan, K.; Crossan, S.; Kohli, P.; Jones, D. T.; Silver, D.; Kavukcuoglu, K.; Hassabis, D. Improved Protein Structure Prediction using Potentials from Deep Learning. *Nature* **2020**, 706−710.

(5) Segler, M. H.; Preuss, M.; Waller, M. P. Planning Chemical Syntheses with Deep Neural Networks and Symbolic AI. *Nature* **2018**, *555*, 604.

(6) Coley, C. W.; Barzilay, R.; Jaakkola, T. S.; Green, W. H.; Jensen, K. F. Prediction of Organic Reaction Outcomes using Machine Learning. *ACS Cent. Sci.* **2017**, *3*, 434−443.

(7) Schreck, J. S.; Coley, C. W.; Bishop, K. J. Learning Retrosynthetic Planning through Simulated Experience. *ACS Cent. Sci.* **2019**, *5*, 970−981.

(8) Pathak, Y.; Singh, K.; Varma, G.; Ehara, M.; Priyakumar, U. D. Deep Learning Enabled Inorganic Material Generator. *Phys. Chem. Chem. Phys.* **2020**, *22*, 26935−26943.

(9) Folmsbee, D.; Hutchison, G. Assessing Conformer Energies using Electronic Structure and Machine Learning Methods. *Int. J. Quantum Chem.* **2021**, *121*, No. e26381.

(10) Laghuvarapu, S.; Pathak, Y.; Priyakumar, U. D. BAND NN: A Deep Learning Framework for Energy Prediction and Geometry Optimization of Organic Small Molecules. *J. Comput. Chem.* **2020**, *41*, 790−799.

(11) Ribeiro, M. T.; Singh, S.; Guestrin, C. In *"Why Should I Trust You?": Explaining the Predictions of Any Classifier*, Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, San Francisco, CA; ACM Digital Library, 2016; pp 1135−1144.

(12) Selvaraju, R. R.; Cogswell, M.; Das, A.; Vedantam, R.; Parikh, D.; Batra, D. In *Grad-CAM: Visual explanations from Deep Networks via Gradient-Based Localization*, Proceedings of the IEEE international conference on computer vision; IEEE. 2017; 618−626.

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

(14) Jalan, A.; Ashcraft, R. W.; West, R. H.; Green, W. H. Predicting Solvation Energies for Kinetic Modeling. *Annu. Rep. Prog. Chem. C* **2010**, *106*, 211−258.

(15) Borhani, T. N.; García-Muñoz, S.; Luciani, C. V.; Galindo, A.; Adjiman, C. S. Hybrid QSPR models for the Prediction of the Free Energy of Solvation of Organic Solute/Solvent Pairs. *Phys. Chem. Chem. Phys.* **2019**, *21*, 13706−13720.

(16) Eisenberg, D.; McLachlan, A. D. Solvation Energy in Protein Folding and Binding. *Nature* **1986**, *319*, 199−203.

(17) Makarov, V.; Pettitt, B. M.; Feig, M. Solvation and Hydration of Proteins and Nucleic Acids: A Theoretical View of Simulation and Experiment. *Acc. Chem. Res.* **2002**, *35*, 376−384.

(18) Hospital, A.; Candotti, M.; Gelpí, J. L.; Orozco, M. The Multiple Roles of Waters in Protein Solvation. *J. Phys. Chem. B* **2017**, *121*, 3636−3643.

(19) Jensen, J. H. Predicting Accurate Absolute Binding Energies in Aqueous Solution: Thermodynamic Considerations for Electronic Structure Methods. *Phys. Chem. Chem. Phys.* **2015**, *17*, 12441−12451.

(20) Roy, D.; Hinge, V. K.; Kovalenko, A. Predicting Blood-Brain Partitioning of Small Molecules Using a Novel Minimalistic Descriptor-Based Approach via the 3D-RISM-KH Molecular Solvation Theory. *ACS Omega* **2019**, *4*, 3055−3060.

(21) Subramanian, V.; Ratkova, E.; Palmer, D.; Engkvist, O.; Fedorov, M.; Llinas, A. Multisolvent Models for Solvation Free Energy Predictions Using 3D-RISM Hydration Thermodynamic Descriptors. *J. Chem. Inf. Model.* **2020**, 2977−2988.

(22) Skyner, R.; McDonagh, J.; Groom, C.; Van Mourik, T.; Mitchell, J. A Review of Methods for the Calculation of Solution Free Energies and the Modelling of Systems in Solution. *Phys. Chem. Chem. Phys.* **2015**, *17*, 6174−6191.

(23) Duarte Ramos Matos, G.; Kyu, D. Y.; Loeffler, H. H.; Chodera, J. D.; Shirts, M. R.; Mobley, D. L. Approaches for Calculating Solvation Free Energies and Enthalpies Demonstrated with an Update of the FreeSolv Database. *J. Chem. Eng. Data* **2017**, *62*, 1559−1569.

(24) Shivakumar, D.; Williams, J.; Wu, Y.; Damm, W.; Shelley, J.; Sherman, W. Prediction of Absolute Solvation Free Energies using Molecular Dynamics Free Energy Perturbation and the OPLS Force Field. *J. Chem. Theory Comput.* **2010**, *6*, 1509−1519.

(25) McDonald, N. A.; Carlson, H. A.; Jorgensen, W. L. Free Energies of Solvation in Chloroform and Water from a Linear Response Approach. *J. Phys. Org.* **1997**, *10*, 563−576.

(26) Shirts, M. R.; Chodera, J. D. Statistically Optimal Analysis of Samples from Multiple Equilibrium States. *J. Chem. Phys.* **2008**, *129*, No. 124105.

(27) Ytreberg, F. M.; Swendsen, R. H.; Zuckerman, D. M. Comparison of Free Energy Methods for Molecular Systems. *J. Chem. Phys.* **2006**, *125*, No. 184114.

(28) Jorge, M.; Garrido, N. M.; Queimada, A. J.; Economou, I. G.; Macedo, E. A. Effect of the Integration Method on the Accuracy and Computational Efficiency of Free Energy Calculations using Thermodynamic Integration. *J. Chem. Theory Comput.* **2010**, *6*, 1018−1027.

(29) Goyal, S.; Chattopadhyay, A.; Kasavajhala, K.; Priyakumar, U. D. Role of Urea-Aromatic Stacking Interactions in Stabilizing the Aromatic Residues of the Protein in Urea-Induced Denatured State. *J. Am. Chem. Soc.* **2017**, *139*, 14931−14946.

(30) Jaganade, T.; Chattopadhyay, A.; Pazhayam, N. M.; Priyakumar, U. D. Energetic, Structural and Dynamic Properties of Nucleobase-Urea Interactions that Aid in Urea Assisted RNA Unfolding. *Sci. Rep* **2019**, *9*, No. 8805.

(31) Tomasi, J.; Mennucci, B.; Cances, E. The IEF version of the PCM Solvation Method: An Overview of a New Method Addressed to Study Molecular Solutes at the QM ab initio level. *J. Mol. Struct. THEOCHEM* **1999**, *464*, 211−226.

(32) Lin, S.-T.; Hsieh, C.-M. Efficient and Accurate Solvation Energy Calculation from Polarizable Continuum Models. *J. Chem. Phys.* **2006**, *125*, No. 124103.

(33) Cho, H.; Choi, I. S. Three-Dimensionally Embedded Graph Convolutional Network (3DGCN) for Molecule Interpretation 2018, arXiv:1811.09794. arXiv.org e-Print archive. http://arxiv.org/abs/1811.09794.

(34) Wu, Z.; Ramsundar, B.; Feinberg, E. N.; Gomes, J.; Geniesse, C.; Pappu, A. S.; Leswing, K.; Pande, V. MoleculeNet: A Benchmark for Molecular Machine Learning. *Chem. Sci.* **2018**, *9*, 513−530.

(35) Mobley, D. L.; Guthrie, J. P. FreeSolv: ADatabase of Experimental and Calculated Hydration Free Energies, with Input Files. *J. Comput. Aided Mol. Des.* **2014**, *28*, 711−720.

(36) Lim, H.; Jung, Y. Delfos: Deep Learning Model for Prediction of Solvation Free Energies in Generic Organic Solvents. *Chem. Sci.* **2019**, *10*, 8306−8315.

(37) Jaeger, S.; Fulle, S.; Turk, S. Mol2vec: Unsupervised Machine Learning Approach with Chemical Intuition. *J. Chem. Inf. Model.* **2018**, *58*, 27−35.

(38) Feinberg, E. N.; Joshi, E.; Pande, V. S.; Cheng, A. C. Improvement in ADMET Prediction with Multitask Deep Featurization. *J. Med. Chem.* **2020**, 8835−8848.

(39) Elton, D. C.; Boukouvalas, Z.; Fuge, M. D.; Chung, P. W. Deep learning for Molecular Design-A Review of the State of the Art. *Mol. Syst. Des. Eng.* **2019**, *4*, 828−849.

https://dx.doi.org/10.1021/acs.jcim.0c01413
J. Chem. Inf. Model. 2021, 61, 689−698

(40) Jin, W.; Barzilay, R.; Jaakkola, T. Junction Tree Variational Autoencoder for Molecular Graph Generation 2018, arXiv:1802.04364. arXiv.org e-Print archive. http://arxiv.org/abs/1802.04364.
(41) 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; AAAI, 2020; 873−880.
(42) Gilmer, J.; Schoenholz, S. S.; Riley, P. F.; Vinyals, O.; Dahl, G. E. *Neural Message Passing for Quantum Chemistry*, Proceedings of the 34th International Conference on Machine Learning; ACM Digital Library, 2017; pp 1263−1272.
(43) Hille, C.; Ringe, S.; Deimel, M.; Kunkel, C.; Acree, W. E.; Reuter, K.; Oberhofer, H. Generalized molecular solvation in non-aqueous solutions by a single parameter implicit solvation scheme. *J. Chem. Phys.* **2019**, *150*, No. 041710.
(44) Landrum, G. RDKit: Open−source Cheminformatics. http://www.rdkit.org, (accessed on May 10,) 2019.
(45) Wang, M.; Yu, L.; Zheng, D.; Gan, Q.; Gai, Y.; Ye, Z.; Li, M.; Zhou, J.; Huang, Q.; Ma, C.; Huang, Z.; Guo, Q.; Zhang, H.; Lin, H.; Zhao, J.; Li, J.; Smola, A. J.; Zhang, Z. Deep Graph Library: Towards Efficient and Scalable Deep Learning on Graphs 2019, arXiv:1909.01315. arXiv.org e-Print archive. http://arxiv.org/abs/1909.01315.
(46) He, K.; Zhang, X.; Ren, S.; Sun, J. *Deep Residual Learning for Image Recognition*, Proceedings of the IEEE conference on computer vision and pattern recognition; IEEE, 2016; pp 770−778.
(47) Vinyals, O.; Bengio, S.; Kudlur, M. Order matters: Sequence to Sequence for Sets. 2015, arXiv:1511.06391. arXiv.org e-Print archive. http://arxiv.org/abs/1511.06391.
(48) Kingma, D. P.; Ba, J. Adam: A Method for Stochastic Optimization. 2014, arXiv:1412.6980. arXiv.org e-Print archive. http://arxiv.org/abs/1412.6980.
(49) Marenich, A. V.; Cramer, C. J.; Truhlar, D. G. Generalized Born Solvation Model SM12. *J. Chem. Theory Comput.* **2013**, *9*, 609−620.
(50) Klamt, A.; Diedenhofen, M. Calculation of Solvation Free Energies with DCOSMO-RS. *J. Phys. Chem. A* **2015**, *119*, 5439−5445.
(51) Kromann, J. C.; Steinmann, C.; Jensen, J. H. Improving Solvation Energy Predictions using the SMD Solvation Method and Semi-empirical Electronic Structure Methods. *J. Chem. Phys.* **2018**, *149*, No. 104102.
(52) Hutchinson, S. T.; Kobayashi, R. Solvent-Specific Featurization for Predicting Free Energies of Solvation through Machine Learning. *J. Chem. Inf. Model.* **2019**, *59*, 1338−1346.
(53) 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*. Retrieved from the Data Repository; University of Minnesota (accessed on June 7, 2019).
(54) HOLLER, A. C. An observation on the relation between the melting points of the disubstituted isomers of benzene and their chemical constitution. *J. Org. Chem.* **1948**, *13*, 70−74.
(55) Janz, G. J. *Nonaqueous Electrolytes Handbook*; Elsevier: New York, NY, 2012; Vol. *1*.
(56) Rautio, J.; Kumpulainen, H.; Heimbach, T.; Oliyai, R.; Oh, D.; Järvinen, T.; Savolainen, J. Prodrugs: Design and Clinical Applications. *Nat. Rev. Drug Discov.* **2008**, *7*, 255−270.
(57) Jornada, D. H.; dos Santos Fernandes, G. F.; Chiba, D. E.; De Melo, T. R. F.; Dos Santos, J. L.; Chung, M. C. The Prodrug Approach: A Successful Tool for Improving Drug Solubility. *Molecules* **2016**, *21*, 42.
(58) Rautio, J.; Meanwell, N. A.; Di, L.; Hageman, M. J. The Expanding Role of Prodrugs in Contemporary Drug Design and Development. *Nat. Rev. Drug Discov.* **2018**, *17*, 559−587.
(59) Vere Hodge, R.; Sutton, D.; Boyd, M.; Harnden, M.; Jarvest, R. Selection of an Oral Prodrug (BRL 42810; famciclovir) for the Antiherpesvirus Agent BRL 39123 [9-(4-hydroxy-3-hydroxymethyl-but-1-yl) guanine; penciclovir]. *Antimicrob. Agents Chemother.* **1989**, *33*, 1765−1773.
(60) Simpson, D.; Lyseng-Williamson, K. A. Famciclovir. *Drugs* **2006**, *66*, 2397−2416.
(61) Stella, V.; Borchardt, R.; Hageman, M.; Oliyai, R.; Maag, H.; Tilley, J. *Prodrugs: Challenges and Rewards*; Springer Science & Business Media, 2007.
(62) Lobo, S.; Li, H.; Farhan, N.; Yan, G. Evaluation of Diclofenac Prodrugs for Enhancing Transdermal Delivery. *Drug Dev. Ind. Pharm.* **2014**, *40*, 425−432.
(63) Peesa, J. P.; Yalavarthi, P. R.; Rasheed, A.; Mandava, V. B. R. A Perspective Review on Role of Novel NSAID Prodrugs in the Management of Acute Inflammation. *J. Acute Dis.* **2016**, *5*, 364−381.

https://dx.doi.org/10.1021/acs.jcim.0c01413
J. Chem. Inf. Model. 2021, 61, 689−698