Breaking the Polar-Nonpolar Division in Solvation Free Energy Prediction

Bao Wang,[a]* Chengzhang Wang,[b]* Kedi Wu ,[a]* and Guo-Wei Wei [a,c,d]

Implicit solvent models divide solvation free energies into polar and nonpolar additive contributions, whereas polar and nonpolar interactions are inseparable and nonadditive. We present a feature functional theory (FFT) framework to break this *ad hoc* division. The essential ideas of FFT are as follows: (i) representability assumption: there exists a microscopic feature vector that can uniquely characterize and distinguish one molecule from another; (ii) feature-function relationship assumption: the macroscopic features, including solvation free energy, of a molecule is a functional of microscopic feature vectors; and (iii) similarity assumption: molecules with similar microscopic features have similar macroscopic properties, such as solvation free energies. Based on these assumptions, solvation free energy prediction is carried out in the following protocol. First, we construct a molecular microscopic feature vector that is efficient in characterizing the solvation process using quantum mechanics and Poisson–Boltzmann theory. Microscopic feature vectors are combined with macroscopic features, that is, physical observable, to form extended feature vectors. Additionally, we partition a solvation dataset into queries according to molecular compositions. Moreover, for each target molecule, we adopt a machine learning algorithm for its nearest neighbor search, based on the selected microscopic feature vectors. Finally, from the extended feature vectors of obtained nearest neighbors, we construct a functional of solvation free energy, which is employed to predict the solvation free energy of the target molecule. The proposed FFT model has been extensively validated via a large dataset of 668 molecules. The leave-one-out test gives an optimal root-mean-square error (RMSE) of 1.05 kcal/mol. FFT predictions of SAMPL0, SAMPL1, SAMPL2, SAMPL3, and SAMPL4 challenge sets deliver the RMSEs of 0.61, 1.86, 1.64, 0.86, and 1.14 kcal/mol, respectively. Using a test set of 94 molecules and its associated training set, the present approach was carefully compared with a classic solvation model based on weighted solvent accessible surface area. © 2017 Wiley Periodicals, Inc.

DOI: 10.1002/jcc.25107

Introduction

Life is associated with water—about 65–90% of human mass is water. The understanding of solvation process is of fundamental importance for the study of more sophisticated physical, chemical, and biological processes, such as protein binding, protein DNA and RNA binding, protein–protein interaction, and signal transduction.[1–4] As such, the modeling and analysis of the solvation effects have drawn considerable attention in the past few decades.[1–3,5–7] Since the most important solvation observable is solvation free energies, the accurate prediction of solvation free energies is the major task in solvation modeling and analysis. Solvation and binding are intrinsically connected. Therefore, the accurate solvation analysis has a direct application to the binding free energy prediction, which is crucial for computer aided drug design.[8] The availability of a large amount of solvation data makes it possible to rigorously validate solvation analysis methods. As a result, solvation analysis has become an attractive research topic in computational biophysics. Nevertheless, the accurate prediction of the solvation free energy remains a very challenging issue.[9]

Many theoretical approaches have been developed in the past few decades for solvation free energy predictions. In general, these approaches can be categorized into physical models, knowledge models, and combined physics and knowledge models. In fact, even for physical models, a number of fitting parameters are introduced to match experimental data. In this sense, all predictive solvation models have certain knowledge components. Physical models are attractive as they are, in principle, able to reveal the physical nature of the solvation process. There are two types of physical models based on the treatment of the solvent molecules: explicit and implicit. Typical explicit solvent models in solvation analysis include molecular mechanics (MM)[8] and hybrid quantum mechanics (QM)/molecular mechanics approaches.[10] In contrast, there is a

[a] *B. Wang, K. Wu, Guo-Wei Wei*
Department of Mathematics, Michigan State University, Michigan 48824
E-mail: wei@math.msu.edu
[b] *C. Wang*
School of Statistics and Mathematics, Central University of Finance and Economics, Beijing 100081, China
[c] *Guo-Wei Wei*
Department of Electrical and Computer, Engineering Michigan State University, Michigan 48824
[d] *Guo-Wei Wei*
Department of Biochemistry and Molecular, Biology Michigan State University, Michigan 48824

*These authors contributed equally to this work.
Contract grant sponsor: NSF; Contract grant numbers: DMS-1721024 and IIS-1302285; Contract grant sponsor: MSU Center for Mathematical Molecular Biosciences Initiative

large variety of implicit solvent models, which are some of the most popular solvation models due to their balance between accuracy and efficiency. Commonly used implicit solvent models include the generalized Born (GB) model, which is a generalization of the Born dielectric sphere model, including many variants.[11–16] GBSA[17] and SM.x[18,19] are two typical examples. Polarizable continuum model (PCM) is a more accurate approach, which incorporates the solvent–solute polarization effects.[20–22] The most popular implicit solvent model is based on the Poisson–Boltzmann (PB) theory, which retains an atomistic description of the solute molecule, while treating the solvent and includes possible ions and cofactors as a dielectric continuum.[23–28] More recently, Gaussian-based smooth dielectric functions have also shown success for computing solvation energy of both small molecules and proteins.[29,30] The PB model is generally accepted to be one of the most accurate implicit solvent models. In fact, it can be combined with the density functional theory (DFT) for a more accurate description of solvent polarization and solute response.[31–33] In most implicit solvent approaches, the solvation free energy is split into polar and nonpolar contributions. The polar part can be calculated from the aforementioned models, while the nonpolar component can be modeled by numerous approaches. The scaled-particle theory (SPT) is popular for modeling nonpolar solutes in aqueous solution.[34,35] Within the SPT theory, the solvent-accessible surface area (SASA) is used to model nonpolar solvation free energy. It is shown that a solvent-accessible volume (SAV) term is also relevant to the nonpolar solvation free energy in large length scale regimes.[36,37] Recent studies indicate that SASA-based solvation models may not describe van der Waals (vdW) interactions near solvent–solute interface.[38–41] A combination of surface area, surface enclosed volume, and vdW potential has been shown to provide very accurate nonpolar solvation predictions.[42,43]

In classical implicit solvent models, the polar and nonpolar components are decoupled. Recently, the coupling of polar and nonpolar components has been considered in several models.[44–46] One representative model for this coupling is based on differential geometry theory, variational approach and geometric measure theory. These mathematical apparatuses give rise to an elegant dynamical coupling of polar and nonpolar solvation components.[45–48] By applying constrained optimization to nonpolar parameter selections, this model provides some of the best solvation free energy fitting and cross-validation results for a large amount of solute molecules.[43]

Despite recent effort in coupling the polar and nonpolar models,[44–46] when it comes to the total solvation free energy calculation, the *ad hoc* assumption that polar and nonpolar free energies are independent, linear and additive is still applied.[33,47] However, in realistic solvation processes, polar and nonpolar interactions are coupled and their free energies are dependent, nonlinear, and nonadditive.

The objective of this work is to completely break the polar and nonpolar division used in implicit solvent models. We propose a feature functional theory (FFT) for solvation free energy modeling to capture the physics of the solvation process. In this approach, instead of treating the solvation free energy as two separated parts, namely, polar and nonpolar ones, we consider the solvation free energy as a unity that is modeled as a mathematical functional of microscopic and macroscopic features. We assume that there exists a microscopic feature vector that can uniquely characterize and distinguish one solute molecule from another (i.e., representability assumption). These microscopic features seek an atomic level representation of molecule properties based on quantum mechanical calculations. We also assume that such fine-scale features are able to accurately capture molecular macroscopic features, namely, physical and chemical properties, including solvation free energies and binding affinities. In other words, we assume that solvation free energy is a functional of microscopic and macroscopic features (i.e., feature-function relationship assumption). Finally, we assume that molecules with similar microscopic features have similar macroscopic properties, such as solvation free energies (i.e., similarity assumption). Based on the above assumptions, we introduce an FFT procedure for predicting solvation free energy. First, we construct microscopic feature vectors for molecules in the database. Both quantum mechanics and Poisson–Boltzmann theory are utilized for the extraction of microscopic features. Macroscopic properties, that is, physical observables, are added to form extended feature vectors. Then, we apply a machine learning algorithm to search for the nearest neighbors of a target molecule based on microscopic feature vectors. We further learn a solvation free energy functional based on the extended feature vectors of nearest neighbors. Finally, we predict the solvation free energy of the target molecule by using the learned energy functional. The proposed machine learning model has been extensively validated by a set 668 molecules and SAMPL0-SAMPL4 challenging sets.[49–51] Our results are state-of-the-art in the field. The advantages of the proposed FFT approach are twofold. First, the present FFT approach does not depend on the conventional polar and nonploar division. Second, there is no need to provide an explicit form of solvation free energy functional, which can be constructed through an optimization procedure. Finally, although nonpolar features as not as important as polar ones for the solvation analysis of 668 molecules, their inclusion leads to more accurate solvation predictions.

This article is structured as follows: Methods and algorithms section is devoted to methods and algorithms. We elaborate on three basic assumptions in basic assumptions section, followed by a description of feature selection in microscopic feature selection section. The machine learning algorithm is described in LTR algorithm section, which is divided into three parts: (i) query construction, which incorporates the previous nearest neighbor search results; (ii) feature selection; and (iii) machine learning for molecular neighbor detection. The nearest neighbor information-based algorithm for solvation free energy prediction is presented in feature-function relationship for solvation free energy prediction section. Numerical results and discussions section presents numerical results and discussions. After describing the dataset and force fields in dataset and feature parametrization section, we offer the leave-one-out validation of the proposed model in leave-one-out

prediction section. SAMPLx challenges are presented in solvation energy prediction of SAMPLx challenges section. Some of the best results in solvation free energy prediction are obtained. In particular, we show that the present approach compares well with a classic solvation model based on weighted solvent accessible surface area.[52] This article ends with some concluding remarks.

Methods and Algorithms

Basic assumptions

To overcome the drawback of decoupled polar and nonpolar solvation models, we present a theory based a few assumptions.

Representability assumption. We consider a total of N molecules $\{M_i\}_{i=1}^N$. Each molecule can be distinguished by a combination of its chemical formulas, chemical name, and geometric structure. We assume that there exists an n-dimensional microscopic feature vector, denoted as $\mathbf{x}_i=(\mathbf{x}_{i1},\mathbf{x}_{i1},\cdots,\mathbf{x}_{in})$ to uniquely characterize and distinguish the ith solute molecule. Here, the vector components include various microscopic features, such as atomic types and numbers, atomic charges, atomic reaction field energies, atomic dipoles, atomic quadrupoles, and pairwise atomic vdW interactions.

For ith molecule, apart from its n microscopic features, there are l macroscopic features, or physical observables $\mathbf{o}_i=(\mathbf{o}_{i1},\mathbf{o}_{i1},\cdots,\mathbf{o}_{il})$, such as density, pressure, boiling point, enthalpy of formation, heat of combustion, solvation free energy, pKa, pH, viscosity, permittivity, and electrical conductivity Whenever these macroscopic features are available, they can be listed as part of an extended feature vector $\mathbf{v}_i=(\mathbf{x}_i,\mathbf{o}_i)$ for the ith molecule.

Extended feature vectors $\{\mathbf{v}_i\}_{i=1}^N$ span a vector space $\mathcal{V}$. As a vector space, it satisfies the commonly required eight axioms for addition and multiplication, such as associativity, commutativity, identity element, inverse elements of addition, and compatibility of scalar multiplication with field multiplication. However, no notion of nearness, angles or distances are defined for the extended feature space—these tasks are achieved via machine learning algorithms. The construction of microscopic feature vectors or the selection of microscopic features depends on what physical or chemical prediction is interested in this work. For example, for solvation free energy prediction, we select features that are derived from implicit solvent models. This issue is discussed in more detail in microscopic feature selection section.

Note that based on our assumption, microscopic features play the unique role in characterizing and distinguishing molecules. Therefore, for a given task, say solvation free energy prediction, there is no need to include all the macroscopic features in the feature vector $\mathbf{o}_i$. One needs only to select $\mathbf{o}_i=(\mathbf{o}_{i1})=\Delta G_i,\forall i=1,\cdots,N$, where $\{\Delta G_i\}$ are known solvation free energies. For this reason, the selection of macroscopic features is described in the dataset preparation, that is, dataset and feature parametrization section.

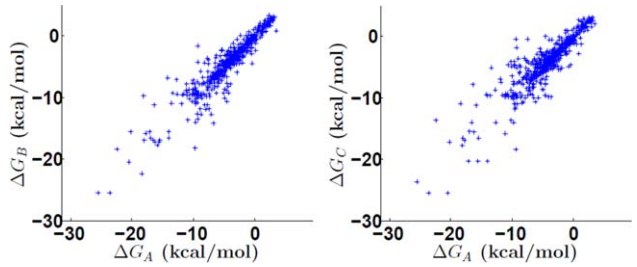

Figure 1. The plot of solvation free energies of the central and its neighbor molecules, the left chart for the first nearest neighbor, the right chart for the second nearest neighbor. In both cases, the horizontal axis represent the solvation free energy for the central molecule, the vertical axis stands for that of the nearest neighbor molecule. [Color figure can be viewed at wileyonlinelibrary.com]

Feature-function relationship assumption. In this work, we are interested in the prediction of solvation free energies based on an existing dataset. The information from the dataset includes molecular identities and corresponding solvation free energies. We construct a feature space for the dataset and the solvation free energy of target molecule A is expressed as a functional of extended feature vectors

$$\Delta G_{\mathrm{A}}=\mathbf{f}_{\mathrm{sol}}(\mathbf{x}_{\mathrm{A}},\mathbf{v}_1,\mathbf{v}_2,\cdots,\mathbf{v}_N) \tag{1}$$

where ΔG_{A} is the solvation free energy of molecule A, $\mathbf{f}_{\mathrm{sol}}$ is an unknown functional for modeling the relationship between solvation free energy and extended features, and $\mathbf{x}_{\mathrm{A}}$ is the microscopic feature vector of the target solute molecule.

In general, we assume a general feature-function relationship to the jth physical observable $\mathbf{o}_j$ of target molecule A

$$\mathbf{o}_{\mathrm{A}j}=\mathbf{f}_j(\mathbf{x}_{\mathrm{A}},\mathbf{v}_1,\mathbf{v}_2,\cdots,\mathbf{v}_N), \tag{2}$$

where $\mathbf{f}_j$ is an unknown functional modeling the jth physical observable of molecule A. Therefore, the present approach can be used for the prediction of other physical and chemical properties as well. Obviously, the determination of $\mathbf{f}_{\mathrm{sol}}$ is a major task of this work and will be discussed in more detail in feature-function relationship for solvation free energy prediction section.

Similarity assumption. We have observed that the solvation free energy of a target molecule is quite close to that of its nearest neighbors. Figure 1 depicts the correlation between experimental solvation free energy from a molecule and that of its nearest neighbors. The root-mean-square errors (RMSEs) of solvation free energies between molecules and their first and second nearest neighbors are 1.44 and 1.77 kcal/mol, respectively.

Motivated by the above observation, we assume that molecules with similar microscopic features have similar solvation free energies. In other words, molecules with similar microscopic features have similar macroscopic features, or physical observable. This assumption underpins the use of learning-to-rank (LTR) algorithm for the nearest neighbor search discussed further in LTR algorithm section.

Table 1. Microscopic features with high correlations to the solvation free energy.

Feature name	Feature name
Sum of atomic reaction field energy	Sum of the absolute value of atomic reaction field energy
Sum of H atomic reaction field energy	Sum of the absolute value of H atomic reaction field energy
Sum of O atomic reaction field energy	Sum of the absolute value of O atomic reaction field energy
Minimum value of atomic reaction field energy	Maximum of the absolute value of reaction field energy
Minimum value of H atomic reaction field energy	Maximum of the absolute value of H atomic reaction field energy
Mean of atomic reaction field energy	Mean of the absolute value of atomic reaction field energy
Variance of atomic reaction field energy	Variance of the absolute value of reaction field energy
Variance of H atomic reaction field energy	Variance of the absolute value of H atomic reaction field energy
Sum of the absolute value of atomic charge	Sum of H atomic charge
Sum of the absolute value of H atomic charge	Sum of O atomic charge
Sum of the absolute value of O atomic charge	Minimum of atomic charge
Maximum of the absolute value of atomic charge	Maximum of H atomic charge
Maximum of the absolute value of H atomic charge	Mean of the absolute value of atomic charge
Variance of the atomic charge	Variance of the absolute value of atomic charge
Variance of the absolute value of H atomic charge	Variance of H atomic charge

Solvation prediction protocol. For a given molecule, we can predict its solvation free energy in two steps, LTR and learning-to-predict:

- construct microscopic feature vectors for all molecules, including the target one;
- find nearest neighbor molecules to the target molecule in the database with the known solvation free energies using a machine learning algorithm;
- learn the functional relation between the solvation free energy and extended features (i.e., a feature functional) according to a group of nearest neighbor molecules, and then predict the solvation free energy for the target molecule by the feature functional.

In the following sections, we provide detailed descriptions of feature selection, nearest neighbor search algorithm, and feature functional construction for solvation free energy prediction.

Microscopic feature selection

A fundamental assumption of our approach is that there exists a microscopic feature vector that can uniquely characterize and distinguish one molecule from another. Obviously, finding such a feature vector is one of our most important tasks. In our previous hybrid physical and knowledge (HPK) model,[33] the goal of the feature selection is to find the closest set of molecules to a given target molecule in the sense of functional group similarity. Therefore, we selected microscopic features that can distinguish molecules with different functional groups, and designated molecules having the same functional groups as similar. Desirable microscopic features should reflect the similarity in solvation free energies. In other words, microscopic features should be most important to the solvation process. To this end, we first construct a set of microscopic features whose Pearson correlation coefficients to solvation free energies are larger than 0.65 or smaller than −0.65. Table 1 lists these features.

As Table 1 shows, all highly correlated features are of polar type, whereas the traditional assumption of implicit solvent models states that nonpolar features also play an important role in solvation process. As a consequence, we also combine atomic surface areas along with the aforementioned polar features. To make the present model scalable to different molecules, atomic surface area features are constructed in an element-wise manner. Specifically, for each element type, atomic surface areas are summed together as a feature. As a comparison, we use all available features to train models. Therefore, we provide two sets of results, one for polar features listed in Table 1, one for all features (both polar and nonpolar) provided in Supporting Information. We examine the performance of these two sets of features on their prediction of solvation free energies.

A major subset of features in Table 1 is derived from implicit solvent models, such as atomic reaction field energy of the *i*th atom

$$\Delta G_{\mathrm{RF},i}=\frac{1}{2}q_i(\phi(\mathbf{r}_i)-\phi_h(\mathbf{r}_i)), \tag{3}$$

where q_i is the partial charge of the *i*th atom at position $\mathbf{r}_i$, and $\phi(\mathbf{r}_i)$ and $\phi_h(\mathbf{r}_i)$ are, respectively, electrostatic potential and homogenous electrostatic potential from the Poisson equation

$$-\nabla\cdot(\varepsilon(\mathbf{r})\nabla\phi(\mathbf{r}))=\sum_{i=1}^{N_m}Q_i\delta(\mathbf{r}-\mathbf{r}_i), \tag{4}$$

with the interface conditions

$$[\phi]|_\Gamma=0, \tag{5}$$

and

$$[\varepsilon\phi_{\mathbf{n}}]|_\Gamma=0, \tag{6}$$

where N_m is the number of atoms, ϕ is the electrostatics potential over the whole solvent solute domain, Q_i is the

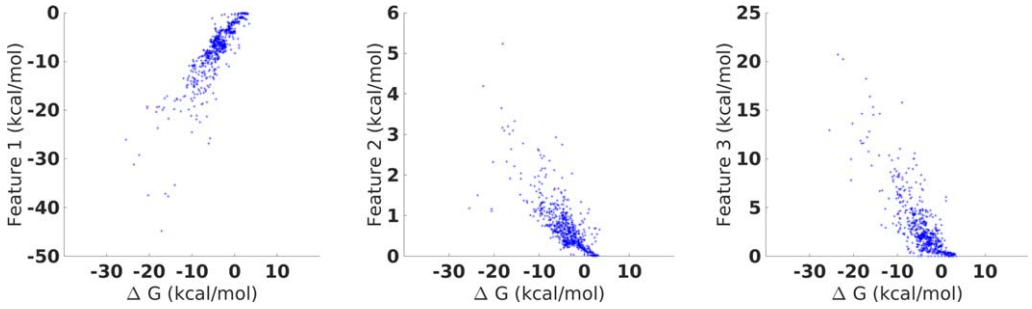

Figure 2. Correlations between features and experimental solvation free energies of 668 molecules. The horizontal axes represent the experimental solvation free energies. From left to right, three charts in the vertical axes represent total reaction field energies, the absolute value of the mean reaction field energies of all atoms, and the absolute value of the total reaction field energy of hydrogen atoms, respectively. [Color figure can be viewed at wileyonlinelibrary.com]

partial charge located at $\mathbf{r}_i$ and $\delta(\mathbf{r}-\mathbf{r}_i)$ is the delta function at point $\mathbf{r}_i$. The permittivity function $\epsilon(\mathbf{r})$ is given by

$$\epsilon(\mathbf{r})=\begin{cases}\epsilon_m=1, & \mathbf{r}\in\Omega^m\\ \epsilon_s=80, & \mathbf{r}\in\Omega^s\end{cases} \tag{7}$$

where Ω^m and Ω^s are solute and solvent domains, respectively, which is separated by the molecular surface Γ.

The following Debye-Huckel type of boundary condition is imposed to make the Poisson model well posed

$$\phi(\mathbf{r})=\sum_{i=1}^{N_m}\frac{Q_i}{4\pi\epsilon_s|\mathbf{r}-\mathbf{r}_i|},\ \text{if}\ \mathbf{r}\in\partial\Omega, \tag{8}$$

where $\Omega=\Omega^m\cup\Omega^s$. The homogenous electrostatic potential $\phi_h(\mathbf{r})$ is obtained by setting $\varepsilon(\mathbf{r})=1$ in the whole computational domain.

For given molecule A, the sum of atomic reaction field energy is defined as

$$\Delta G_{\mathrm{RF}}=\sum_{i=1}^{N_m}\Delta G_{\mathrm{RF},i},$$

which is the same as the electrostatic solvation free energy of the solute molecule A, where N_m is the number of atoms in solute M, $\Delta G_{\mathrm{RF},i}$ is the reaction field energy contributed from the ith atom.

The sum of the absolute value of atomic reaction field energy is

$$\Delta G_{\mathrm{RF}}^{\mathrm{abs}}=\sum_{i=1}^{N_m}|\Delta G_{\mathrm{el},i}|.$$

The other features can be defined mathematically in the same manner.

The above microscopic features are calculated by the following methods.

- Atomic charges and dipoles can be computed by using quantum mechanical theory.
- Atomic reaction field energies can be computed by using PB theory.
- The calculations of maximum, minimum, sum, mean, and variance are based on straightforward statistical theory.

Figure 2 plots some representative features compared to experimental solvation free energies. From left to right, three charts are the correlations of experimental solvation free energies with total reaction field energies, the absolute value of the mean reaction field energies of all atoms, and the absolute value of the total reaction field energy of hydrogen atoms, respectively. Their Pearson correlation coefficients are 0.87, −0.76, and −0.80, respectively.

Remark 1. *The high correlation of reaction field energy calculated by the PB model with the solvation free energy indicates that the PB is an effective approach for modeling the solvation effects. The reaction field energy calculated by the PB is consistent with the experimental solvation free energy.*

LTR algorithm

In this section, we introduce the list-wise LTR algorithm for ranking molecules. In the training procedure of the LTR algorithm, the solvation free energy of the molecule is used as the molecular label, which is consistent with our basic ansatz. A scoring function is learned in the list-wise LTR method on the set of training molecules and is utilized for ranking the molecules in the set of testing molecules. The nearest neighbor search can be regarded as a top-*N* recommendation problem, which is mathematically the same as the item search in the world-wide-web.

Query construction. In our FFT model, we use solvation free energy as a label. Based on our assumption that molecules with similar feature vectors have similar solvation free energies, its reverse statement is not true in general. To deal with this deficiency in our LTR algorithm for nearest neighbor search, we partition the whole dataset (which contains a total 668 molecules as described later) into a number of subsets, where each subset is regarded as a query in the LTR terminology. The basic requirement of the query construction is that molecules in each query should have some chemical similarity. Additionally, we require that each query is invariant to the

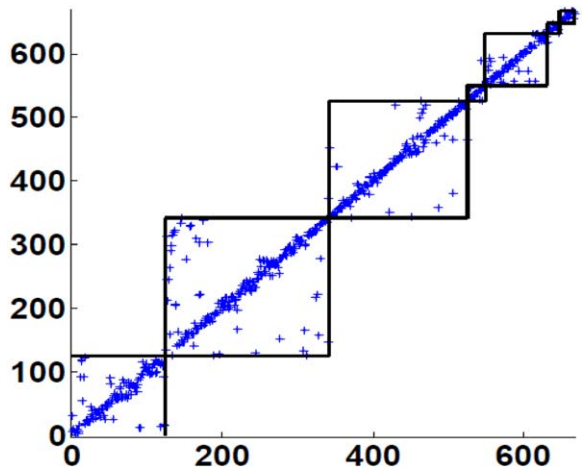

Figure 3. Localization of nearest neighbor molecules. The horizontal axis stands for the index of a target molecule and the vertical axis denotes the index of the nearest neighbor of the target molecule. Each block contains a query of molecules. Note that axis labels reflect molecules in the whole set, a total of 668 molecules. [Color figure can be viewed at wileyonlinelibrary.com]

nearest molecule detection based on the cosine similarity of the microscopic features proposed in our earlier work.[33] To achieve this, the most straight forward approach is to make each query of molecules contain the same functional group. However, the complexity of the molecules in the dataset makes this partition impractical. A direct relaxation is that molecules of each query have the same element types, which will be used for query construction in this work.

We first construct seven groups of molecules, according to element type: (i) H, C; (ii) H, C, O; (iii) H, C, N/H, C, N, O; (iv) H, C, Cl; (v) H, C, O, Cl; (vi) H, S; and (vii) anything else, respectively. The third group contains molecules either with H, C, and N elements or with H, C, N, and O. This classification is due to the fact that based on the cosine similarity, molecules in these two categories may have their nearest neighbors overlap. For the remaining molecules, we iteratively add them into the above six groups based on their nearest neighbor's class label. The molecules that cannot be classified into any of the above categories are regarded as a new query. We label molecules in the dataset from 1 to 668. Figure 3 shows that the queries constructed based on the above procedure are invariant to the nearest neighbor search based on the measure proposed in our earlier work,[33] where each block denotes a query of molecules. It is easy to see that molecules' nearest neighbors are localized into each block. This invariance indicates that our query construction preserves the molecular chemical similarity, that is, each query of molecules is of some similarity in the physical sense. Based on the above query construction, we can approximately regard that close solvation free energies indicate similar molecules in each query, which makes the LTR-based nearest neighbor search physically sound. We list all the queries in Supporting Information.

Since our partition of the dataset is based on our similarity assumption with chemical constraints, we discuss the similarity measure, microscopic feature selection based on chemical and physical properties that facilitate the measure and LTR algorithm for ranking the molecules. For nearest neighbor searches in each query, we emphasize that the nearest neighbor is measured based on the nearness of the solvation free energies, instead of the similarity measure used before.

Gradient boosted decision tree algorithm. In this work, we choose gradient boosted decision tree (GBDT), a multiple additive regression tree (MART), as our ranking strategy. In this part, we provide a brief overview of GBDT algorithm and also discuss how to apply the GBDT algorithm to our solvation modeling.

An overview of GBDT algorithm. GBDT is essentially an ensemble method that has been widely used for biological modeling. It naturally takes care of the correlation between descriptors, usually does not need a feature selection procedure and is generally insensitive to parameters. For more details about this algorithm, the reader is referred to the literature.[53,54]

GBDT for Molecules Ranking. Now let us turn to the application of GBDT to the solvation prediction. In each query of the molecules, the solvation free energies themselves are regarded as the labels of molecules, and the corresponding features are discussed in the next subsection. Our method can be summarized as ranking the nearest neighbors of a target molecule based on their solvation free energies, and then learning a relation between features and solvation free energies for predicting the solvation free energy of the target molecule.

Feature-function relationship for solvation free energy prediction

In this section, we discuss the solvation free energy prediction for a given target molecule. Based on our assumption that solute solvation free energy is a functional of the feature vector, solvation free energy prediction is should actually construct a feature functional around the target molecule. This feature functional will be utilized for solvation free energy prediction for the target molecule.

Consider the solvation free energy for target molecule A characterized by its feature vector $\mathbf{x}_A=(\mathbf{x}_{A1},\mathbf{x}_{A2},\cdots,\mathbf{x}_{An})$, where n is the dimension of the microscopic feature space, that is, the space of all microscopic feature vectors. From the LTR algorithm, we find m nearest neighbors with extended feature vectors $\{\mathbf{v}_i=(\mathbf{x}_i,\Delta G_i)\}_{i=1}^{m}$. Note that in general, the number of nearest neighbors found is far less than the dimension of the feature space, that is, $m \ll n$.

In this work, we assume the functional relation between microscopic features and solvation free energies for target molecule A can be approximated locally by

$$\Delta G_A = b + \sum_{i=1}^{n} w_i \mathbf{x}_{Ai}, \tag{9}$$

where $w_i=w_i(\mathbf{v}_1,\mathbf{v}_2,\cdots\mathbf{v}_m)$ is the weight for feature $\mathbf{x}_{Ai}$ and $b=b(\mathbf{v}_1,\mathbf{v}_2,\cdots\mathbf{v}_m)$ can be intuitively understood as the height

of hyperplane embedded in the Euclidean space. Equation (9) can be regarded as a linear approximation of the solvation free energy functional $\Delta G_A = \mathbf{f}(\mathbf{x}_A, \mathbf{v}_1, \mathbf{v}_2, \cdots \mathbf{v}_m)$.

Since $m \ll n$, the direct regression based on the least square approach may leads to over-fitting. To avoid over-fitting, there are generally two strategies for determining $\{w_i\}$ and b:

- sparse solution via a compressed sensing approach;
- Tikhonov regularization-based least square fitting.

In this work, we use the second strategy for training the local regression model for solvation free energy prediction. The local regression problem is equivalent to solve the linear system in eq. (10) in the L_2 sense

$$\begin{pmatrix} \Delta G_1 \\ \Delta G_2 \\ \vdots \\ \Delta G_m \end{pmatrix} = \begin{pmatrix} x_{11} & x_{12} & \cdots & x_{1n} \\ x_{21} & x_{22} & \cdots & x_{2n} \\ \vdots & \vdots & \vdots & \vdots \\ x_{m1} & x_{m2} & \cdots & x_{mn} \end{pmatrix} \begin{pmatrix} w_1 \\ w_2 \\ \vdots \\ w_n \end{pmatrix} + \begin{pmatrix} b \\ b \\ \vdots \\ b \end{pmatrix}. \quad (10)$$

Equation (10) can be written as

$$\Delta\mathbf{G} = \mathbf{x}\mathbf{w} + b\mathbf{1}, \quad (11)$$

where $\Delta\mathbf{G} = (\Delta G_1, \Delta G_2, \cdots, \Delta G_m)^T$, $\mathbf{w} = (w_1, w_2, \cdots, w_n)^T$, $\mathbf{1}$ is a m-dimensional column vector with all elements equal 1, and matrix $\mathbf{x}$ is given by

$$\mathbf{x} = \begin{pmatrix} x_{11} & x_{12} & \cdots & x_{1n} \\ x_{21} & x_{22} & \cdots & x_{2n} \\ \vdots & \vdots & \vdots & \vdots \\ x_{m1} & x_{m2} & \cdots & x_{mn} \end{pmatrix}.$$

To avoid overfitting, we add the L_2 penalty to the weight vector $\mathbf{w}$, and thus eq. (11) can be solved by the following optimization problem

$$\min_{\mathbf{w},b} ||\Delta\mathbf{G} - \mathbf{x}\mathbf{w} - b\mathbf{1}||_2^2 + ||\lambda\mathbf{w}||_2^2 := \min_{\mathbf{w},b} \mathbf{F}, \quad (12)$$

where λ is the regularization parameter, which is set to 1000 in this work, $||*||_2$ denotes the L_2 norm of the quantity $*$.

By solving $\frac{\partial \mathbf{F}}{\partial \mathbf{w}} = 0$, we have

$$\mathbf{w} = \left(\mathbf{x}^T\mathbf{x} + \mathbf{I}\right)^{-1}\left(\mathbf{x}^T\Delta\mathbf{G} - \mathbf{x}^T(b\mathbf{1})\right), \quad (13)$$

where I is $m \times m$ identity matrix.

To find the value b that solves the optimization problem eq. (12), we relax $b\mathbf{1}$ to arbitrary vector $\mathbf{b} = (b_1, b_2, \cdots, b_m)^T$, by solving $\frac{\partial \mathbf{F}}{\partial \mathbf{b}} = 0$, we have

$$\mathbf{b} = \Delta\mathbf{G} - \mathbf{x}\mathbf{w}. \quad (14)$$

Therefore, we obtain the unbiased estimation of b as

$$b = \frac{\sum_{i=1}^m (\Delta\mathbf{G} - \mathbf{x}\mathbf{w})_i}{m}, \quad (15)$$

where $(\Delta\mathbf{G} - \mathbf{x}\mathbf{w})_i$ is the ith component of the vector $\Delta\mathbf{G} - \mathbf{x}\mathbf{w}$.

We can solve the optimization problem eq. (12) by alternating iterations between eq. (13) and (15), which is essentially an expectation-maximization (EM) algorithm. We summarize the algorithm for solving eq. (12) in Algorithm 1.

Algorithm 1. EM algorithm for solving the optimization problem eq. (12)

1: **Initialize**: $\mathbf{w} = \mathbf{0}$, $b = \frac{\sum_{i=1}^m \Delta\mathbf{G}}{m}$, $\mathbf{F}_1 = ||\Delta\mathbf{G} - \mathbf{x}\mathbf{w} - b\mathbf{1}||_2^2 + ||\mathbf{w}||_2^2$, $\mathbf{F}_2 = \mathbf{F}_1 + 1000$.
2: **do while** $(|\mathbf{F}_1 - \mathbf{F}_2| > \epsilon_0)$
3: Update $\mathbf{F}_2$: $\mathbf{F}_2 \leftarrow \mathbf{F}_1$.
4: Update $\mathbf{w}$: $\mathbf{w} \leftarrow (\mathbf{x}^T\mathbf{x} + \mathbf{I})^{-1}(\mathbf{x}^T\Delta\mathbf{G} - \mathbf{x}^T(b\mathbf{1}))$.
5: Update b: $b \leftarrow \frac{\sum_{i=1}^m (\Delta\mathbf{G} - \mathbf{x}\mathbf{w})_i}{m}$.
6: Update $\mathbf{F}_1$: $\mathbf{F}_1 \leftarrow ||\Delta\mathbf{G} - \mathbf{x}\mathbf{w} - b\mathbf{1}||_2^2 + ||\mathbf{w}||_2^2$.
7: **enddo**

In Algorithm 1, ε_0 is the threshold parameter for control the convergence of the solution to the optimization problem and is set to 0.01 in this work.

After obtaining optimized parameters $\mathbf{w}$ and b, the solvation free energy of target molecule A, is predicted by eq. (9).

Numerical Results and Discussions

Dataset and feature parametrization

Dataset. To assess the performance of the present method, we consider the same dateset that has been constructed in our earlier work.[33] With a total of 668 molecules, this dataset is referred as the 668 set and is the largest for solavtion free energies to the best of our knowledge. It contains both monofunctional group and polyfunctional group molecules. Experimental solvation free energies are collected from the literature.[55–57] The main part of our dataset, that is, 589 molecules, overlaps with Mobley's solvation database (http://mobleylab.org/resources.html). All the structures of this dataset are downloaded from the Pubchem project (https://pubchem.ncbi.nlm.nih.gov/). More detailed description of the dataset can be found in our earlier work.[33]

Our dataset of 668 molecules contains all molecules of SAMPL0, SAMPL1, SAMPL2, SAMPL3, and SAMPL4, except for 5-iodouracil in SAMPL2. Molecule 5-iodouracil involves one iodine atom whose charge density cannot be evaluated using the many force fields considered in the present work. Thus, it is excluded in the present work.

Microscopic feature parametrization. In microscopic feature generation, atomic charges and atomic dipoles are calculated via the distributed multipole analysis method,[58] in which the charge density is originally computed by the DFT with B3LYP and 6–31G basis selection in Gaussian quantum chemistry software.[59–61] Atomic reaction field energies (i.e., atomic electrostatic solvation energies) are calculated by our in-house MIBPB

Table 2. The RMSEs and MEs of the solvation free energy prediction of 668 molecules with different parametrizations and different numbers of nearest neighbors involved using selected polar features.

Parametrization	Error	1	2	3	4	5	6	7	8	9	10
BCC+Amber6	RMSE	1.155	1.188	1.201	1.186	1.184	1.185	1.189	1.197	1.201	1.201
	ME	−0.026	−0.029	−0.024	−0.015	−0.005	0.005	0.011	0.016	0.023	0.031
BCC+Amber Bondi	RMSE	1.278	1.286	1.242	1.237	1.258	1.269	1.255	1.256	1.268	1.271
	ME	−0.034	−0.036	−0.024	−0.020	−0.008	0.008	0.029	0.039	0.053	0.065
BCC+Amber MBondi2	RMSE	1.230	1.229	1.186	1.179	1.208	1.222	1.203	1.201	1.207	1.211
	ME	−0.024	−0.017	−0.006	−0.002	0.006	0.023	0.040	0.052	0.065	0.078
GAS+Amber6	RMSE	1.412	1.419	1.430	1.421	1.417	1.424	1.439	1.449	1.462	1.474
	ME	−0.035	−0.035	−0.035	−0.032	−0.015	−0.000	0.014	0.025	0.031	0.036
GAS+Amber Bondi	RMSE	1.334	1.331	1.349	1.354	1.339	1.364	1.380	1.396	1.422	1.433
	ME	−0.018	−0.018	−0.018	−0.021	−0.003	0.010	0.026	0.035	0.044	0.053
GAS+Amber MBondi2	RMSE	1.292	1.298	1.299	1.326	1.331	1.335	1.357	1.380	1.400	1.411
	ME	−0.002	−0.002	0.003	0.001	0.019	0.035	0.047	0.060	0.070	0.080
MUL+Amber6	RMSE	1.513	1.504	1.528	1.522	1.554	1.579	1.532	1.523	1.536	1.542
	ME	−0.029	−0.029	−0.015	−0.008	−0.002	0.011	0.033	0.047	0.064	0.071
MUL+Amber Bondi	RMSE	1.475	1.465	1.495	1.519	1.540	1.540	1.489	1.478	1.493	1.509
	ME	−0.017	−0.011	−0.008	−0.003	0.012	0.030	0.046	0.056	0.071	0.085
MUL+Amber MBondi2	RMSE	1.490	1.481	1.509	1.537	1.562	1.570	1.517	1.504	1.521	1.543
	ME	−0.004	0.004	0.009	0.014	0.030	0.044	0.059	0.070	0.088	0.096

All errors are in unit kcal/mol.

software (http://weilab.math.msu.edu/MIBPB/)[62–65] with a probe radius of 1.4 Å and dielectric constants being 1 and 80, respectively, for the solute and solvent domains. A uniform grid size of 0.25 Å is used in all atomic reaction field energy calculations. To examine the sensitivity of the present approach to charge force fields, which was a major issue in our earlier HPK model, we utilize three types of atomic radii, namely, Amber 6, Amber bondi, and Amber mbondi2.[66] Additionally, we consider three types of charge assignments, namely, OpenEye-AM1-BCC v1 parameters,[67] Gasteiger,[68] and Mulliken.[66] The combination of radius sets and charge sets gives rise to a total of nine different parametrizations, which have already been utilized in our earlier work[33] to offer some of the best solvation prediction results. For the regularized least square hyperplane fitting, the regularization parameter λ is set to 1000. Atomic surface areas are computed with our in-house ESES software (http://weilab.math.msu.edu/ESES/).[69] Accumulated atomic surface areas of individual element types are used as features.

Leave-one-out prediction

First, we consider the leave-one-out test on the whole dataset of 668 molecules. In this test, we regard the solvation free energy of one molecule as unknown, and use the remaining molecules to predict the solvation free energy for the target molecule. The purpose of the leave-one-out test is two twofold. First, it helps for the parameter selection, that is, the number of nearest neighbors to be used for the prediction of the target molecule's solvation free energy and the parameters used in training the GBDT algorithm. Second, the leave-one-out test can demonstrate the performance of the proposed model for solvation free energy prediction. The performance of the leave-one-out test is measured by both the RMSE and mean error (ME), respectively, defined by

$$\mathrm{RMSE}=\sqrt{\frac{\sum_{i=1}^{N}\left(\Delta G_i^{\mathrm{Pred}}-\Delta G_i^{\mathrm{Expl}}\right)^2}{N}} \tag{16}$$

and

$$\mathrm{ME}=\frac{\sum_{i=1}^{N}\left(\Delta G_i^{\mathrm{Pred}}-\Delta G_i^{\mathrm{Expl}}\right)}{N} \tag{17}$$

where N is the total number of molecules in our dataset, $\Delta G_i^{\mathrm{Expl}}$ and $\Delta G_i^{\mathrm{Pred}}$ stand for the experimental and predicted solvation free energies for the ith molecule, respectively.

The RMSE measures the accuracy of the prediction. A small RMSE indicates the predictions for the whole dataset are uniformly accurate. ME is used to determine whether the prediction is biased or not. If the ME is close to zero, it means that the prediction is unbiased.

Selection of nearest neighbors. In applying our model, one has to determine how many nearest neighbors need to be involved for the solvation free energy prediction. In general, this number depends on the training dataset and parametrization. Numerically, one can use either leave-one-out or fivefold crossvalidation to determine the optimal number of nearest neighbors. Tables 2 and 3 list RMSE and ME of our leave-one-out prediction calculated by a total of nine different combinations of atomic radii and charge force fields with selected features and all polar-nonpolar features. The use of different numbers of nearest neighbors is examined as well. We note that, judging by RMSEs, our method is not sensitive to the number of nearest neighbors. All of the top 10 recommendations have the same level of accuracy. However, when MEs are also taken into consideration, it is found that a large number of nearest neighbors typically contributes to a large ME. We propose to select the number of nearest neighbors based on the following criteria:

Table 3. The RMSEs and MEs of the solvation free energy prediction of 668 molecules with different parametrizations and different numbers of nearest neighbors involved using all polar and nonpolar features.

Parametrization	Error	1	2	3	4	5	6	7	8	9	10
BCC+Amber6	RMSE	1.094	1.214	1.232	1.103	1.134	1.121	1.101	1.084	1.073	1.070
	ME	−0.012	−0.026	−0.027	−0.004	−0.014	−0.012	−0.011	−0.011	−0.004	0.005
BCC+Amber Bondi	RMSE	1.133	1.138	1.263	1.259	1.290	1.304	1.123	1.110	1.101	1.085
	ME	−0.014	−0.010	−0.028	−0.018	−0.021	−0.020	0.003	0.005	0.004	0.012
BCC+Amber MBondi2	RMSE	1.095	1.099	1.229	1.219	1.243	1.271	1.062	1.054	1.056	1.050
	ME	−0.020	−0.014	−0.033	−0.025	−0.024	−0.026	0.007	0.009	0.011	0.026
GAS+Amber6	RMSE	1.259	1.250	1.262	1.260	1.246	1.241	1.266	1.312	1.305	1.323
	ME	−0.026	−0.019	−0.016	−0.017	−0.018	−0.014	−0.013	−0.018	−0.008	−0.005
GAS+Amber Bondi	RMSE	1.235	1.238	1.254	1.262	1.245	1.237	1.257	1.280	1.278	1.288
	ME	−0.022	−0.017	−0.016	−0.010	0.006	0.003	−0.000	0.009	0.021	0.019
GAS+Amber MBondi2	RMSE	1.235	1.238	1.239	1.243	1.244	1.228	1.248	1.256	1.249	1.259
	ME	−0.019	−0.011	−0.008	−0.008	0.002	0.001	−0.001	0.007	0.019	0.032
MUL+Amber6	RMSE	1.308	1.297	1.471	1.468	1.515	1.529	1.310	1.291	1.310	1.277
	ME	−0.024	−0.025	−0.025	−0.024	−0.024	−0.027	−0.010	−0.010	−0.008	0.003
MUL+Amber Bondi	RMSE	1.335	1.330	1.450	1.435	1.444	1.460	1.307	1.310	1.333	1.329
	ME	−0.006	−0.001	−0.018	−0.008	−0.010	−0.015	0.002	0.009	0.008	0.013
MUL+Amber MBondi2	RMSE	1.358	1.353	1.472	1.472	1.470	1.476	1.318	1.297	1.332	1.330
	ME	0.003	0.008	−0.014	−0.005	−0.006	−0.002	0.010	0.010	0.014	0.022

All errors are in unit kcal/mol.

- The RMSE should be as small as possible to give an accurate prediction.
- The ME should be as close to zero as possible to give an unbiased prediction.
- At the same level of RMSE and ME, it is preferred to involve more molecules, which makes it easy to determine the solvation free energy functional.

Usually, there is a tradeoff among the aforementioned criteria in selecting the number of molecules for solvation prediction. As listed in Tables 2 and 3, we emphasize that the proposed method is quite robust with respect to different choices.

Figure 4 illustrates the correlation between experimental solvation free energies and leave-one-out FFT predictions for the set of 668 molecules. The optimal RMSE of selected polar features is 1.18 kcal/mol with four nearest neighbors, while the optimal RMSE using all polar-nonpolar features can be improved to 1.05 kcal/mol with 10 nearest neighbors. The result is by far the lowest for such a large set of molecules, to our best knowledge. For a comparison of FFT and HPK models, we also plot the results for the leave-one-out prediction of the set of 668 molecules. Clearly, the present FFT model outperforms our earlier HPK model with most charge and radius combinations.

Accuracy and sensitivity analysis. A detailed comparison between the present leave-one-out predictions and those of our earlier HPK model[33] can be made under the same radius

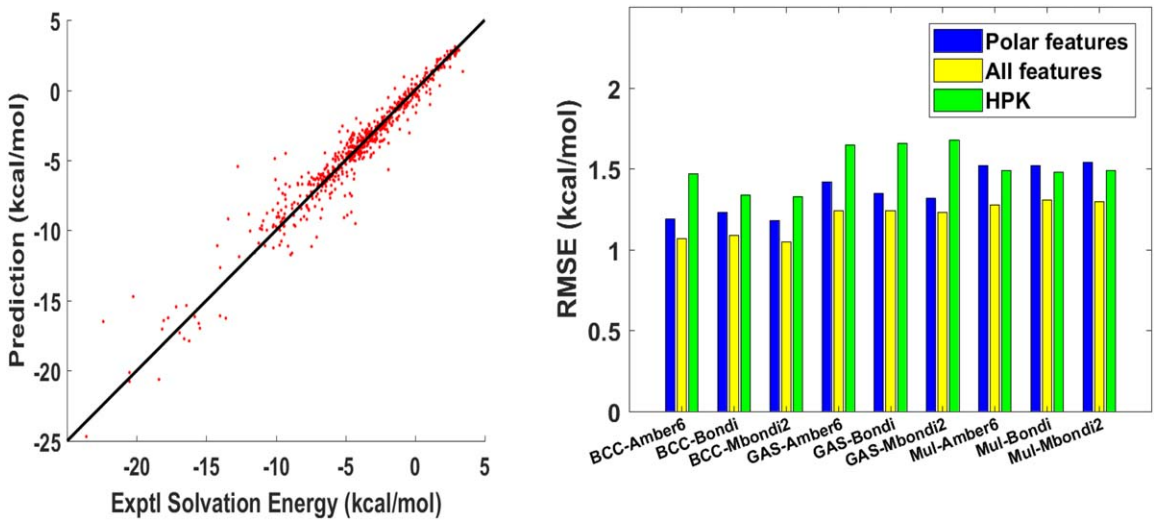

Figure 4. Illustration of leave-one-out predictions for the whole set of 668 molecules. Left chart: Correlation between experimental solvation free energies and FFT predictions obtained by BCC charges and Amber MBondi2 using all polar-nonpolar features. Right chart: Comparison of prediction RMSEs obtained by FFT models with polar features and all features against HPK models. In the plot, GAS and MUL are abbreviations for Gasteiger and Mulliken charges, respectively. [Color figure can be viewed at wileyonlinelibrary.com]

Table 4. The RMSE and ME of the leave-one-out test in the solvation free energy prediction of 668 molecules with polar features FFT (the first position), all features FFT (the second position), and HPK model (the last position).[33]

Radius	Charge	BCC	Mulliken	Gasteiger
Amber 6	RMSE	1.19, 1.07, 1.47	1.52, 1.28, 1.49	1.42, 1.24, 1.65
	ME	−0.01, 0.01, −0.13	−0.01, 0.00, −0.20	−0.03, −0.01, −0.19
Amber Bondi	RMSE	1.23, 1.09, 1.34	1.52, 1.31, 1.48	1.35, 1.24, 1.66
	ME	−0.02, 0.01, −0.14	−0.01, 0.01, −0.21	−0.02, 0.00, −0.13
Amber MBondi2	RMSE	1.18, 1.05, 1.33	1.54, 1.30, 1.49	1.33, 1.23, 1.68
	ME	0.00, 0.03, −0.14	0.0, 0.01, −0.22	0.00, 0.00, −0.22

All errors are in unit kcal/mol.

and charge parametrizations. Table 4 lists the RMSEs and MEs of the current model predictions. For a comparison, corresponding RMSEs obtained by our previous HPK model is also listed in parentheses. From Table 4, we can conclude the following:

- The FFT solvation models are much more accurate than our previous HPK model for this set of 668 molecules. The best prediction with selected polar features and all polar-nonpolar features has an RMSE of 1.18 kcal/mol and 1.05 kcal/mol, respectively, compared to the lowest RMSE of 1.33 kcal/mol achieved by the previous model. Note that the worst earlier result has an RMSE of 1.68 kcal/mol.[33] As a comparison, the worst RMSE of the present prediction has been improved to 1.54 kcal/mol. For a set of 643 molecules, which overlaps the present dataset with 589 molecules, Mobley and Guthrie reported an RMSE of 1.51 kcal/mol.[70] Therefore, the present FFT-SP represents a major advance in solvation free energy prediction.
- The FFT solvation models provide unbiased solvation predictions, as indicated from ME results. The predictions with different molecular parametrizations all achieve near zero MEs. The MEs of the previous model are almost ten times larger than those of the FFT solvation models. Additionally, we note that no matter what type of molecular parametrization is applied, the previous predictions are biased toward one direction, whereas the present models have MEs of both signs.
- The FFT solvation models are less sensitive to the microscopic feature parametrization compared to the HPK solvation model. The ranges of the RMSEs for FFT calculated with all polar and nonpolar features and HPK models associated with nine different parametrizations are 1.05–1.31 kcal/mol and 1.33–1.68 kcal/mol, respectively. Obviously, having a larger range in prediction RMSEs indicates that the HPK model is more sensitive to microscopic feature parametrization.

Solvation energy prediction of SAMPLx challenges

In this part, we consider the prediction of solvation free energy for the SAMPLx challenge sets. Our FFT solvation model is applied to all of five SAMPL test sets, that is, SAMPL0–SAMPL4. We adopt the same protocol used in our previous leave-SAMPLx-out prediction.[33] Specifically, in each SAMPL test prediction, we exclude all the molecules in the given SAMPL in our FFT process, and use the remaining molecules as our training set to find a set of the nearest neighbors to each molecule in the SAMPL test set. Both RMSE and ME measures are evaluated to assess the performance of the proposed FFT model. The same nine sets of charge and radius parametrizations are implemented in leave-SAMPLx-out tests.

SAMPL0 test. First, let us consider the solvation free energy prediction for the SAMPL0 test set, which contains a total of 17 molecules. All structures of this test set are relatively simple. However, the molecule species of this set is quite diverse. Many researchers have reported their solvation free energy predictions for this challenge set.[71,72] Prior to our work, the optimal prediction for this test set has an RMSE of 1.34 kcal/mol for the whole set.[72] Figure 5 depicts the present FFT results for a total of nine charge and radius combinations. When the BCC charge is used, the RMSEs of our predictions with three radius parametrizations are all smaller than 0.75 kcal/mol. Our optimal prediction has an RMSE of 0.61 kcal/mol, obtained from Amber Bondi radius parametrization in conjugation with the BCC charge assignment with polar features only. When all polar and nonpolar features are used, the results become slightly worse whereas performances over all parametrizations turn out to be more stable especially when the Mulliken charge assignment is used.

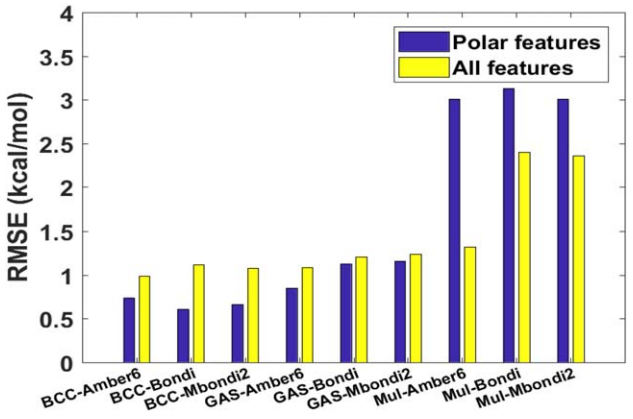

Figure 5. Illustration of prediction RMSEs obtained with different molecular parametrizations by the FFT model for SAMPL0 test set. In the plot, GAS and MUL are abbreviations for Gasteiger and Mulliken charges, respectively. [Color figure can be viewed at wileyonlinelibrary.com]

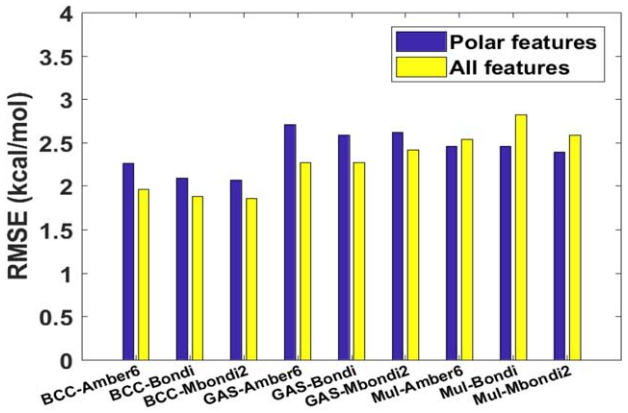

Figure 6. Illustration of prediction RMSEs obtained with different molecular parametrizations by the FFT model for SAMPL1 test set. In the plot, GAS and MUL are abbreviations for Gasteiger and Mulliken charges, respectively. [Color figure can be viewed at wileyonlinelibrary.com]

SAMPL1 test. Having demonstrated the superiority of the proposed FFT model for the prediction of the SAMPL0 challenge set, we further consider the SAMPL1 test set, which is generally believed to be the most difficult one, due to the following two reasons. First, the molecular structures of this test set are extremely complex compared to other molecules with known experimental solvation free energies. Second, the uncertainty of SAMPL1 experimental data is very large. For some molecules the uncertainty is as large as 2.0 kcal/mol.[9,49] Nevertheless, it is extremely desirable to develop an accurate modeling paradigm for this test set because most molecules in this test set are druggable. The best prediction for the whole set has an RMSE of 2.45 kcal/mol.[72] On a subset of the SAMPL1 test set that contains only 56 molecules, the best performance was shown to give an RMSE of 2.4 kcal/mol.[9] Figure 6 illustrates the results of the FFT approach for the whole SAMPL1 test set. It is obvious to see that the FFT model is much more accurate. The optimal prediction with only polar features has an RMSE as small as 2.07 kcal/mol, and adding nonpolar features further improves the RMSE to 1.86 kcal/mol, which is the best to our best knowledge. Additionally, the present FFT model is very robust with respect to the change in force fields. The maximum and minimum prediction RMSEs over nine sets of parametrizations and two feature combinations are 1.86 and 2.82 kcal/mol, respectively. The difference between the maximum and minimum is 0.96 kcal/mol, which is much smaller than experimental uncertainty of 2 kcal/mol for this set.[9,49]

SAMPL2 test. Another difficult test set is SAMPL2, which contains a total of 30 molecules.[73] The experimental uncertainty on these molecules is much less than that of the SAMPL1 test set. Nevertheless, accurate solvation prediction for this set is rare. Using all-atom molecular dynamics simulations and multiple starting conformations for prediction, Klimovich and Mobley reported an RMSE of 2.82 kcal/mol over the whole set and 1.86 kcal/mol over all the molecules except several hydroxyl-rich compounds.[73] Some of the best reported predictions have an RMSE of 1.59 kcal/mol.[72] In our previous test, the molecule containing an I atom (5-iodouracil) is excluded in all calculations due to the lack of appropriate charge force field. In this work, we also ignore this molecule for the same reason. The HPK model gives an optimal prediction with RMSE 1.96 kcal/mol. However, the RMSEs of the prediction vary over a large range, from 1.96 to 4.86 kcal/mol, when different charge and radius force fields are applied. A bar graph of the RMSEs of FFT predictions is given in Figure 7. Parametrizations based on AM1-BCC charge yield the best results among polar features and adding nonpolar features offers a substantial improvement over polar features, as the first three yellow bars are lower than the first three blue bars. The optimal RMSE for SAMPL2 molecules is 1.64 kcal/mol with AM1-BCC charge and AMBER6 radius, when all features are used to train the models. It is also worthy to note that the variation of RMSEs under different parametrizations is 1.42 kcal/mol (1.64 to 3.06 kcal/mol), which indicates the robustness of the present FFT models compared to HPK models.

SAMPL3 test. The SAMPL3 test set, which contains 36 molecules, is relatively easy for prediction. The structures of SAMPL3 molecules are relatively simple, and most molecules in this set are chlorinated hydrocarbon molecules.[51] The best prediction in the literature offers an RMSE of 1.29 kcal/mol.[72] Figure 8 depicts the RMSEs of the predictions by only polar features and all features. Although the optimal result (RMSe of 0.86 kcal/mol) is generated by polar features with Gasteiger charges, all features combination turns out to be more stable over all parametrizations as Figure 8 clearly shows. More specifically, the RMSEs using polar features span over a small range of 0.48 kcal/mol (i.e., from 0.86 to 1.34 kcal/mol) across all nine different parametrizations, while all features yield a variation of 0.24 kcal/mol. This further verifies the robustness of the FFT solvation model.

SAMPL4 test. Finally, we consider the SAMPL4 test set, which is a very popular one. Many explicit, implicit, integral equation, and hybrid QM/MM approaches[10] have been applied to this set.[74] As shown in Figure 9, the overall performance enhances when all features are used as the blue bars are consistently higher than yellow bars, which indicates the predictive power

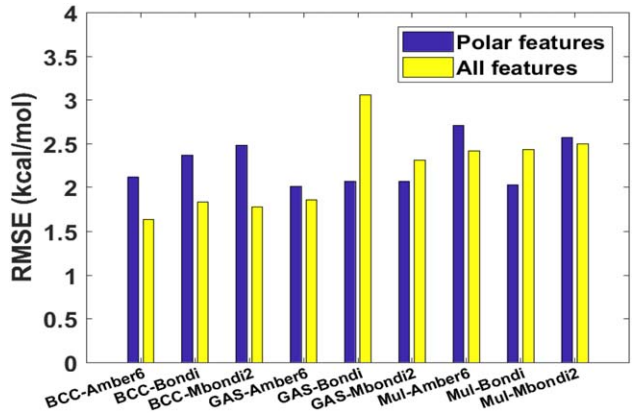

Figure 7. Illustration of prediction RMSEs obtained with different molecular parametrizations by the FFT model for SAMPL2 test set. In the plot, GAS and MUL are abbreviations for Gasteiger and Mulliken charges, respectively. [Color figure can be viewed at wileyonlinelibrary.com]

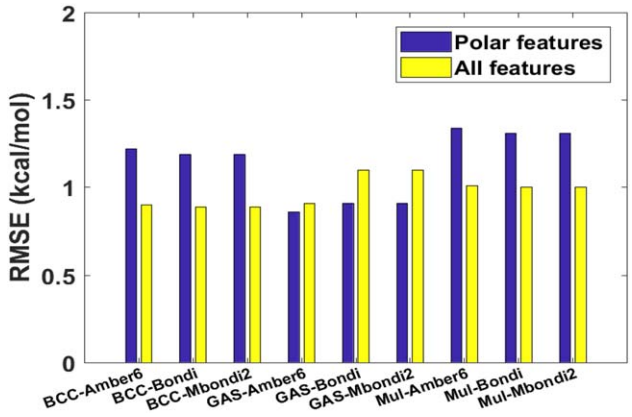

Figure 8. Illustration of prediction RMSEs obtained with different molecular parametrizations by the FFT model for SAMPL3 test set. In the plot, GAS and MUL are abbreviations for Gasteiger and Mulliken charges, respectively. [Color figure can be viewed at wileyonlinelibrary.com]

of nonpolar features. Our FFT model gives an optimal RMSE of 1.14 kcal/mol. It is also easy to see that our FFT approach is quite robust across different force field and charge parametrizations.

Remark. *Table 5 provides a summary of the RMSEs and MEs for all SAMPL0-SAMPL4 test sets. Nevertheless, contrary to the small MEs found in the leave-one-out tests, these errors amplify much in the prediction of SAMPLx test sets, particularly for SAMPL1 and SAMPL2 test sets. Possible explanations for this phenomenon are the complexity of molecules and the lack of physically and chemically similar molecules in our database. We also point out that in the FFT predictions, large RMSEs and MEs occur simultaneously, which indicates that large RMSEs might come from biased predictions. This phenomenon is under our further investigation.*

Discussion

Comparison with a classic solvation model based on weighted solvent accessible surface areas

We would like to further examine whether our FFT approach is competitive with other solvation models on large datasets. To this end, we apply our FFT methods to the prediction of the solvation energy of molecules collected by Wang et al.[52] These authors introduced models based on weighted solvent accessible surface areas and performed both fitting and prediction tasks. In their Model III, the authors further divided 387 molecules (which exclude ions) into a training set (293) and a test set (94) and achieved unsigned average errors of 0.50 and 0.66 kcal/mol for the training set and the test set, respectively,[52] and an unsigned average error of 0.538 kcal/mol for the entire set. It is interesting to compare our results with theirs since compounds used for training and testing are essentially independent, which challenges the predictive power of solvation models. Our FFT models were also trained with scikit-learn package[75] and the average of 50 independent runs yields unsigned average errors of 0.00 and 0.57 kcal/mol, respectively, for the training set and the test set and an unsigned average error of 0.441 kcal/mol for the entire set. In fact, we used a slightly smaller training set of 289 molecules because 4 molecules in the training set have ambiguous chemical names in the PubChem database, while all molecules in the test set are included in our prediction. This comparison indicates that our FFT models have a competitive edge over the classic solvation model based on weighted solvent accessible surface area (SASA).

Remark 2. *It should be noted that there exist discrepancies in experimental solvation free energies for some molecules in Ref. [52] and the 668 set.*[33] *When such discrepancies occur, the experimental values reported by Wang et al.*[52] *are used for training and testing in the above comparison. We provide a list of 177 molecules with inconsistent experimental values in Supporting Information. However, it should be noted that most experimental value differences are within a very small range. Only 23 differences are greater than 0.2 kcal/mol and 5 out of 23 molecules are in the test set. These five compounds and their experimental values corresponding to* table 3 *of Wang et al.*[52] *are listed in Table* 6.

Additionally, four molecules listed in the training set of Ref. [52] *while excluded in our training due to their absence of structures in PubChem have compound ID of 363, 364, 385, and 388 in* table 3 *of Ref.* [52].

Moreover, we have also noticed that there are 11 duplicates in the training set and the test set of Ref. 52. *Their compound IDs and duplicated IDs (Dup-IDs) in the* table 3 *of Ref.* [52] *are listed in Table* 7. *Molecules that are in the test set are marked with a superscript "b" to be consistent with the notation of Ref.* [52]. *When these 11 duplicated molecules in the training set are excluded, the FFT has an RMSE of 0.61 kcal/mol for the test set, which is still smaller than that reported in Ref.* [52] *(i.e., 0.66 kcal/mol).*

Feature importance analysis

Another importance concern for FFT modeling is features importance. To analyze this issue, we rank all features by their

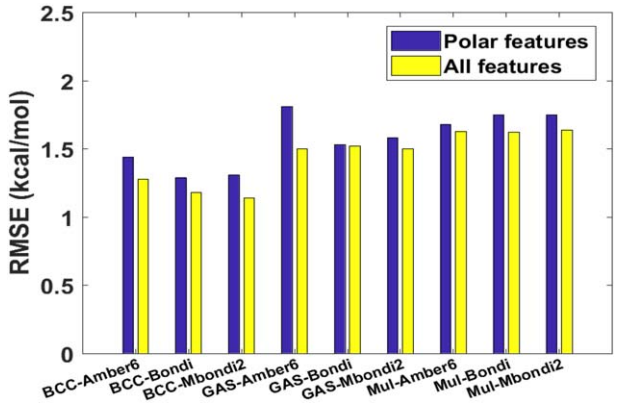

Figure 9. Illustration of prediction RMSEs obtained with different molecular parametrizations by the FFT model for SAMPL4 test set. In the plot, GAS and MUL are abbreviations for Gasteiger and Mulliken charges, respectively. a) SAMPL0, b) SAMPL1, c) SAMPL2, d) SAMPL3, e) SAMPL4. [Color figure can be viewed at wileyonlinelibrary.com]

Table 5. The RMSEs and MEs of the solvation free energy predictions with different parametrizations.

Test set	Radius	Error	AM1-BCC	Mulliken	Gasteiger
SAMPL0	Amber 6	RMSE	0.74, 0.99	3.01, 1.32	0.85, 1.09
		ME	−0.05, −0.17	−0.60, 0.16	−0.14, −0.13
	Amber Bondi	RMSE	0.61, 1.12	3.13, 2.4	1.13, 1.21
		ME	−0.07, −0.10	−0.43, 0.22	0.37, −0.14
	Amber MBondi2	RMSE	0.66, 1.08	3.01, 2.36	1.16, 1.24
		ME	0.04, −0.05	−0.39, 0.31	0.33, −0.22
SAMPL1	Amber 6	RMSE	2.26, 1.96	2.46, 2.54	2.71, 2.27
		ME	−1.10, 0.62	−0.09, 0.14	−0.90, 0.32
	Amber Bondi	RMSE	2.09, 1.88	2.46, 2.82	2.59, 2.27
		ME	−0.90, 0.76	−0.30, 0.35	−0.71, 0.35
	Amber MBondi2	RMSE	2.07, 1.86	2.39, 2.59	2.62, 2.42
		ME	−0.89, 0.49	−0.23, 0.16	−0.92, 0.17
SAMPL2	Amber 6	RMSE	2.12, 1.64	2.71, 2.42	2.01, 1.86
		ME	0.52, −0.28	0.82, −1.59	0.89 −1.23,
	Amber Bondi	RMSE	2.37, 1.83	2.03, 2.43	2.07, 3.06
		ME	0.35, −0.81	0.81, −1.91	0.86, −2.21
	Amber MBondi2	RMSE	2.48, 1.78	2.57, 2.50	2.07, 2.31
		ME	0.54, −0.63	1.71, −2.07	1.14, −1.68
SAMPL3	Amber 6	RMSE	1.22, 0.90	1.34, 1.01	0.86, 0.91
		ME	0.10, 0.05	0.22, −0.19	−0.11, −0.02
	Amber Bondi	RMSE	1.19, 0.89	1.31, 1.00	0.91, 1.10
		ME	0.11, 0.09	0.22, −0.19	−0.06, 0.07
	Amber MBondi2	RMSE	1.19, 0.89	1.31, 1.00	0.91, 1.10
		ME	0.11, 0.09	−0.22, −0.19	−0.06, 0.07
SAMPL4	Amber 6	RMSE	1.44, 1.28	1.68, 1.63	1.81, 1.50
		ME	0.26, −0.00	0.36, −0.06	0.47, 0.02
	Amber Bondi	RMSE	1.29, 1.18	1.75, 1.62	1.53, 1.52
		ME	0.29, 0.07	0.43, −0.06	0.22, 0.03
	Amber MBondi2	RMSE	1.31, 1.14	1.75, 1.64	1.58, 1.50
		ME	0.14, 0.11	0.42, 0.03	0.33, −0.03

The numbers in the first and the second positions are the results obtained from FFT models with polar features and all features, respectively. All errors are in unit kcal/mol.

feature importance and consequently generate 40 different sets of feature combinations. Note that the feature importance here refers to Gini importance,[76] weighted by the number of trees in a forest calculated by our baseline methods. We train models with different numbers of features to examine their predictive performances on test sets. More specifically, the protocol to select features relies on a series of feature importance cutoffs, equally spaced between 0 and 0.01, with features whose importance is greater than the given cutoff value being selected.

Figure 10 represents the RMSEs of predicted solvation energy of SAMPL molecules against different feature importance cutoffs. When the feature importance cutoff value is large, the number of features is small, and RMSE is typically large too. The performance is getting better when the feature importance cutoff value is relatively small. However, further reduction in the cutoff value does not necessarily improve the prediction accuracy and may result in worse performance. Indeed, a suitable cutoff value can benefit overall performance. Cutoff value of 2.5×10^{-3} appears to be a good choice in our case according to our feature importance analysis.

Concluding Remarks

Implicit solvent models intuitively split the total solvation free energies into polar and nonplar contributions. While, in realistic solvation processes, polar and nonpolar interactions are coupled and interdependent. As a result, their energies are nonadditive. We propose a FFT framework to break the polar-nonpolar division used in implicit solvent models and treat polar and nonpolar contributions on an equal footing. Our FFT has three basic assumptions, namely, representability, feature-

Table 6. Molecules in the test sets with large discrepancies in their experimental solvation free energies.

ID	Exp1[52]	Exp2[33]
46	0.29	0.01
67	−3.15	−3.4
97	−0.78	−1.73
103	−0.64	−1.4
352	−4.71	−5.22

Here, "ID" refers to the ID of table 3 of Ref. [52].

Table 7. Duplicated molecules in Ref. [52].

ID	Dup-ID	ID	Dup-ID
104	119	333	335
334	336	384	389
161	202	82	84
140	142	184	194
97	116	58	59
196	203		

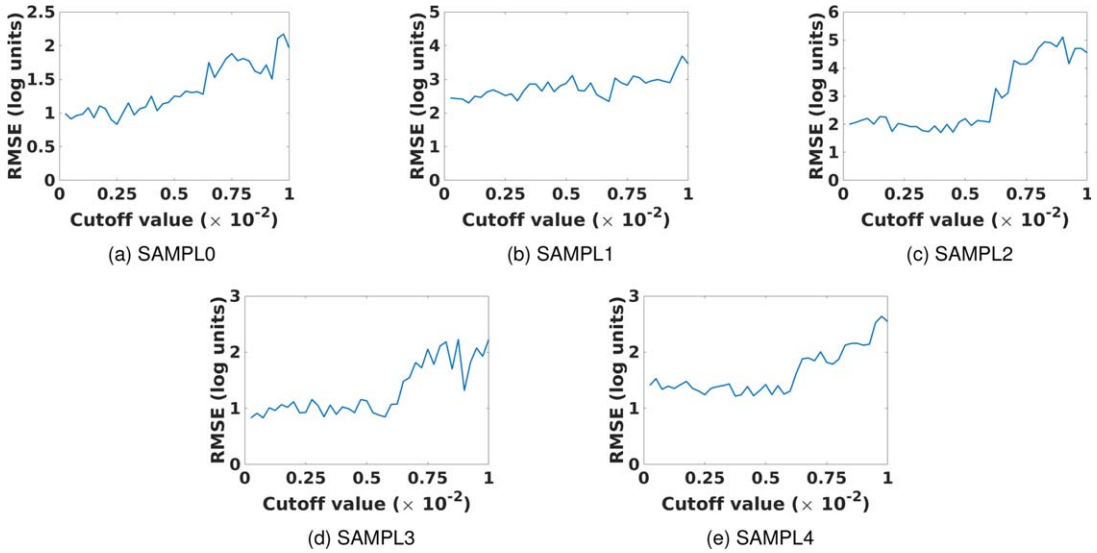

Figure 10. Feature importance cutoff versus RMSE for all test sets with AM1-BCC charge and MBondi2 radius parametrization. The larger cutoff value is, the smaller number of features is selected. [Color figure can be viewed at wileyonlinelibrary.com]

function relationship, and similarity hypotheses. Specifically, we assume that there exists a microscopic feature vector that can uniquely characterize and distinguish one molecule from another. Additionally, the solvation free energy of a molecule is a functional of microscopic feature vectors. Finally, molecules with similar microscopic features have similar macroscopic properties, including solvation free energies.

The performance of the present FFT depends on quality of feature vector construction. Our experience in developing our earlier HPK model[33] enables us to select a set of microscopic features, such as atomic charge, dipole, and reaction field energy, to accurately characterize solvation free energies. These microscopic features are evaluated by quantum mechanics, polarization theory, and Poisson–Boltzmann theory. Additionally, the quality of macroscopic features, namely, the accuracy and reliability of physical measurements, affect the FFT performance as well.

The present FFT method searches the nearest neighbors based on their similarity in solvation free energies. In machine learning terminology, our previous nearest neighbor search method can be regarded as an unsupervised learned method. The present nearest neighbor search is based on the assumption of molecules with similar microscopic features having similar solvation free energies. As such, the nearest neighbor search problem is cast into a supervised learning problem. As a result, the nearest neighbor quality can be improved dramatically, which further improves the accuracy of solvation free energy prediction. To implement our new supervised FFT approach, we first partition molecules into several groups according to their chemical compositions. Each group is regarded as a query in the machine learning terminology. The query construction is designed to constrain the ranking process where similar solvation free energies might not imply similar molecules or similar microscopic features. A state-of-the-art list-wise LTR algorithm, gradient boosted decision trees (GBDT), is adopted for training the FFT method. By using this algorithm, the quality of the nearest neighbor search improves significantly, which is supported from the fact that the difference between the solvation free energies of a target molecule and its neighbors decreased dramatically.

Based on the assumption that molecular solvation free energy is a functional of molecular features, we construct a solvation free energy predictor using a regularized least square-based local hyperplane learning algorithm. To validate the proposed FFT method, we adopt a large dataset of 668 molecules collected in our earlier work.[33]

It is interesting to note that polar features are highly correlated to solvation free energies of this dataset. Nonpolar features, such as molecular area and volume, do not appear on the list of top features. Nevertheless, the inclusion of nonpolar features does improve the overall performance of the present method. Highly accurate solvation free energy prediction is confirmed by both the leave-one-out test over 668 molecules and the prediction of five SAMPL test sets, namely, SAMPL0, SAMPL1, SAMPL2, SAMPL3, and SAMPL4.Finally, we consider a test set of 94 molecules and its associated training set[52] for a comparison of the present method and a classic solvation model based on weighted solvent accessible surface area.[52]

This work is our first attempt in developing an advanced machine learning-based model for solvation free energy prediction. The FFT model can be improved in a number of ways. One improvement is about the current query construction based on molecular element types. We believe that a more sophisticated query construction can further improve the accuracy of the nearest neighbor search. Another potential improvement is a better feature selection. For example, one can select features according to their local correlations with the solvation free energies in a given query. The other improvement can be achieved through better microscopic feature design and more accurate feature evaluation. Many microscopic features were computed via density functional theory (DFT) in the present work. We believe that some other

advanced quantum methodologies for atomic charge, dipole, and quadrupole calculations will significantly improve our prediction. The advantage of the DFT-based polarizable Poisson model has been noticed in our previous work.[33] Therefore, some improvements in the reaction field energy calculation can be valuable as well. Overall, we believe that with a better set of molecular descriptors, molecular parametrization, and molecular partition, the proposed FFT-based solvation free energy prediction can be further improved. The application of the proposed approach to protein–ligand binding affinity prediction is reported elsewhere.[77]

Acknowledgment

We thank Nathan Baker, Michael Gilson, David Mobley, Pengyu Ren, and Weitao Yang for useful discussions.

Keywords: solvation free energy · implicit solvent model · machine learning · microscopic feature functional

How to cite this article: B. Wang, C. Wang, K. Wu, G.-W. Wei. *J. Comput. Chem*. **2017**, DOI: 10.1002/jcc.25107

Additional Supporting Information may be found in the online version of this article.

[1] J. W. Storer, D. J. Giesen, G. D. Hawkins, G. C. Lynch, C. J. Cramer, D. G. Truhlar, D. A. Liotard, In *Structure, Energetics, and Reactivity in Aqueous Solution: Characterization of Chemical and Biological Systems*, Vol. *568*; C. J. Cramer, D. G. Truhlar, Eds.; American Chemical Society: Washington, DC, **1994**; pp. 24–49.
[2] R. Daudel, Quantum Theory of Chemical Reactivity; Reidel: Kufstein, Austria, **1973**.
[3] C. Reichardt, In *Solvents and Solvent Effects in Organic Chemistry*; Wiley-VCH: New York, **1990**.
[4] M. Borisover, M. Reddy, E. R. Graber, *Environ. Sci. Technol.* **2001**, *35*, 2518.
[5] M. Kreevoy, D. Truhlar. In *Investigation of Rates and Mechanisms of Reactions, Part I;* C. Bernasconi, Ed.; Wiley: New York, **1986**; p. 13.
[6] M. E. Davis, J. A. McCammon, *Chem. Rev.* **1990**, *94*, 509.
[7] A. Warshel, A. Papazyan, *Curr. Opin. Struct. Biol.* **1998**, *8*, 211.
[8] S. A. Martins, S. F. Sousa, M. J. Ramos, P. A. Fernandes, *J. Chem. Theory Comput.* **2014**, *10*, 3570.
[9] A. V. Marenich, C. J. Cramer, D. G. Truhlar, *J. Phys. Chem. B* **2009**, *113*, 4538.
[10] G. König, F. C. Pickard, Y. Mei, B. R. Brooks, *J. Comput. Aided Mol. Des.* **2014**, *28*, 245.
[11] B. Jayaram, D. Sprous, D. L. Beveridge, *J. Phys. Chem. B* **1998**, *102*, 9571.
[12] J. A. Grant, B. T. Pickup, M. T. Sykes, C. A. Kitchen, A. Nicholls, *Phys. Chem. Chem. Phys.* **2007**, *9*, 4913.
[13] H. Gohlke, D. A. Case, *J. Comput. Chem.* **2004**, *25*, 238.
[14] M. Feig, W. Im, C. L. Brooks, III, *J. Chem. Phys.* **2004**, *120*, 903.
[15] B. N. Dominy, C. L. Brooks, III, *J. Phys. Chem. B* **1999**, *103*, 3765.
[16] D. Bashford, D. A. Case, *Annu. Rev. Phys. Chem.* **2000**, *51*, 129.
[17] J. J. Tan, W. Z. Chen, C. X. Wang, *J. Mol. Struct. THEOCHEM* **2006**, *766*, 77.
[18] C. J. Cramer, D. G. Truhlar, *Chem. Rev.* **1999**, *99*, 2161.
[19] C. J. Cramer, D. G. Truhlar, *Acc. Chem. Res.* **2008**, *41*, 760.
[20] B. M. J. Tomasi, R. Cammi, *Chem. Rev.* **2005**, *105*, 2999.
[21] J. Tomasi, M. Persico, *Chem. Rev.* **1994**, *94*, 2027.
[22] B. M. E. Cances, J. Tomasi, *J. Chem. Phys.* **1997**, *107*, 3032.
[23] I. Park, Y. H. Jang, S. Hwang, D. S. Chung, *Chem. Lett.* **2003**, *32*, 376.
[24] K. A. Sharp, B. Honig, *J. Phys. Chem.* **1990**, *94*, 7684.
[25] K. A. Sharp, B. Honig, *Annu. Rev. Biophys. Biophys. Chem.* **1990**, *19*, 301.
[26] L. Li, C. Li, S. Sarkar, J. Zhang, S. Witham, Z. Zhang, L. Wang, N. Smith, M. Petukh, E. Alexov, *BMC Biophys.* **2012**, *5*, 9.
[27] B. Honig, A. Nicholls, *Science* **1995**, *268*, 1144.
[28] M. K. Gilson, M. E. Davis, B. A. Luty, J. A. McCammon, *J. Phys. Chem.* **1993**, *97*, 3591.
[29] L. Li, C. Li, E. Alexov, *J. Theor. Comput. Chem.* **2014**, *13*, 1440002.
[30] L. Li, C. Li, Z. Zhang, E. Alexov, *J. Chem. Theory Comput.* **2013**, *9*, 2126.
[31] M. L. Wang, C. F. Wong, *J. Phys. Chem. A* **2006**, *110*, 4873.
[32] M. L. Wang, C. F. Wong, J. H. Liu, P. X. Zhang, *Chem. Phys. Lett.* **2007**, *442*, 464.
[33] B. Wang, Z. Zhao, G. W. Wei, *J. Chem. Phys.* **2016**, *145*, 124110.
[34] F. H. Stillinger, *J. Solution Chem.* **1973**, *2*, 141.
[35] R. A. Pierotti, *Chem. Rev.* **1976**, *76*, 717.
[36] K. Lum, D. Chandler, J. D. Weeks, *J. Phys. Chem. B* **1999**, *103*, 4570.
[37] D. M. Huang, D. Chandler, *Proc. Natil. Acad. Sci. USA* **2000**, *97*, 8324.
[38] E. Gallicchio, L. Y. Zhang, R. M. Levy, *J. Comput. Chem.* **2002**, *23*, 517.
[39] E. Gallicchio, R. M. Levy, *J. Comput. Chem.* **2004**, *25*, 479.
[40] N. Choudhury, B. M. Pettitt, *J. Am. Chem. Soc.* **2005**, *127*, 3556.
[41] J. A. Wagoner, N. A. Baker, *Proc. Natl. Acad. Sci. USA* **2006**, *103*, 8331.
[42] Z. Chen, S. Zhao, J. Chun, D. G. Thomas, N. A. Baker, P. B. Bates, G. W. Wei, *J. Chem. Phys.* **2012**, *137*, 084101.
[43] B. Wang, G. W. Wei, *J. Chem. Phys.* **2015**, *143*, 134119.
[44] J. Dzubiella, J. M. J. Swanson, J. A. McCammon, *Phys. Rev. Lett.* **2006**, *96*, 087802.
[45] G. W. Wei, *Bull. Math. Biol.* **2010**, *72*, 1562.
[46] Z. Chen, N. A. Baker, G. W. Wei, *J. Comput. Phys.* **2010**, *229*, 8231.
[47] Z. Chen, G. W. Wei, *J. Chem. Phys.* **2011**, *135*, 194108.
[48] Z. Chen, N. A. Baker, G. W. Wei, *J. Math. Biol.* **2011**, *63*, 1139.
[49] J. P. Guthrie, *J. Phys. Chem. B* **2009**, *113*, 4501.
[50] M. T. Geballe, A. G. Skillman, A. Nicholls, J. P. Guthrie, J. P. Taylor, *J. Comput. Aided Mol. Des.* **2010**, *24*, 259.
[51] M. T. Geballe, J. P. Guthrie, *J. Comput. Aided Mol. Des.* **2012**, *26*, 489.
[52] J. Wang, W. Wang, S. Huo, M. Lee, P. A. Kollman, *J. Phys. Chem. B* **2001**, *105*, 5055.
[53] J. H. Friedman, *Ann. Statist.* **2001**, *29*, 1189.
[54] C. Burges, T. Shaked, E. Renshaw, A. Lazier, M. Deeds, N. Hamilton, G. Hullender, *Proc. ICML* **2005**, 89.
[55] S. Cabani, P. Gianni, V. Mollica, L. Lepori, *J. Solution Chem.* **1981**, *10*, 563.
[56] J. Wang, W. Wang, S. Huo, M. Les, P. A. Kollman, *J. Phys. Chem. B* **2001**, *105*, 5055.
[57] J. Li, T. Zhu, G. D. Hawkins, P. Winget, D. A. Liotard, C. J. Cramer, D. G. Truhlar, *Theor. Chem. Acc.* **1999**, *103*, 9.
[58] A. J. Stone, *Chem. Phys. Lett.* **1981**, *83*, 233.
[59] M. J. Frisch, G. W. Trucks, H. B. Schlegel, G. E. Scuseria, M. A. Robb, J. R. Cheeseman, G. Scalmani, V. Barone, B. Mennucci, G. A. Petersson, H. Nakatsuji, M. Caricato, X. Li, H. P. Hratchian, A. F. Izmaylov, J. Bloino, G. Zheng, J. L. Sonnenberg, M. Hada, M. Ehara, K. Toyota, R. Fukuda, J. Hasegawa, M. Ishida, T. Nakajima, Y. Honda, O. Kitao, H. Nakai, T. Vreven, J. A. Montgomery, Jr., J. E. Peralta, F. Ogliaro, M. Bearpark, J. J. Heyd, E. Brothers, K. N. Kudin, V. N. Staroverov, R. Kobayashi, J. Normand, K. Raghavachari, A. Rendell, J. C. Burant, S. S. Iyengar, J. Tomasi, M. Cossi, N. Rega, J. M. Millam, M. Klene, J. E. Knox, J. B. Cross, V. Bakken, C. Adamo, J. Jaramillo, R. Gomperts, R. E. Stratmann, O. Yazyev, A. J. Austin, R. Cammi, C. Pomelli, J. W. Ochterski, R. L. Martin, K. Morokuma, V. G. Zakrzewski, G. A. Voth, P. Salvador, J. J. Dannenberg, S. Dapprich, A. D. Daniels,. Farkas, J. B. Foresman, J. V. Ortiz, J. Cioslowski, D. J. Fox, *Gaussian09 Revision E.01*; Gaussian Inc.: Wallingford, CT, **2009**.
[60] A. D. Becke, *J. Chem. Phys.* **1993**, *98*, 5648.
[61] C. Lee, W. Yang, R. G. Parr, *Phys. Rev. B* **1988**, *37*, 785.
[62] D. D. Nguyen, B. Wang, G. W. Wei, *J. Comput. Chem.* **2017**, *38*, 941.
[63] D. Chen, Z. Chen, C. Chen, W. H. Geng, G. W. Wei, *J. Comput. Chem.* **2011**, *32*, 657.
[64] S. N. Yu, W. H. Geng, G. W. Wei, *J. Chem. Phys.* **2007**, *126*, 244108.
[65] W. Geng, S. Yu, G. W. Wei, *J. Chem. Phys.* **2007**, *127*, 114106.
[66] D. A. Case, J. T. Berryman, R. M. Betz, D. S. Cerutti, T. E. Cheatham, III, T. A. Darden, R. E. Duke, T. J. Giese, H. Gohlke, A. W. Goetz, N. Homeyer, S. Izadi, P. Janowski, J. Kaus, A. Kovalenko, T. S. Lee, S. LeGrand, P. Li, T. Luchko, R. Luo, B. Madej, K. M. Merz, G. Monard, P.

Needham, H. Nguyen, H. T. Nguyen, I. Omelyan, A. Onufriev, D. R. Roe, A. Roitberg, R. Salomon-Ferrer, C. L. Simmerling, W. Smith, J. Swails, R. C. Walker, J. Wang, R. Wolf, X. Wu, D. M. York, P. A. Kollman, *Amber 2015*; University of California: San Francisco, **2015**.
[67] A. Jakalian, B. L. Bush, D. B. Jack, C. I. Bayly, *J. Comput. Chem.* **2000**, *21*, 132.
[68] J. Gasteiger, M. Marsili, *Tetrahedron* **1980**, *36*, 3219.
[69] B. Liu, B. Wang, R. Zhao, Y. Tong, G. W. Wei, *J. Comput. Chem.* **2017**, *38*, 446.
[70] D. L. Mobley, J. P. Guthrie, *J. Comput. Aided Mol. Des.* **2014**, *28*, 711.
[71] A. Nicholls, D. L. Mobley, J. P. Guthrie, J. D. Chodera, C. I. Bayly, M. D. Cooper, V. S. Pande, *J. Med. Chem.* **2008**, *51*, 769.
[72] C. W. Kehoe, C. J. Fennell, K. A. Dill, *J. Comput. Aided Mol. Des.* **2012**, *26*, 563.
[73] P. V. Klimovich, D. L. Mobley, *J. Comput. Aided Mol. Des.* **2010**, *24*, 307.
[74] D. L. Mobley, K. L. Wymer, N. M. Lim, J. P. Guthrie, *J. Comput. Aided Mol. Des.* **2014**, *28*, 135.
[75] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, E. Duchesnay, *J. Mach. Learn. Res.* **2011**, *12*, 2825.
[76] L. Breiman, *Mach. Learn.* **2001**, *45*, 5.
[77] B. Wang, Z. Zhao, D. D. Nguyen, G. W. Wei, *Theor. Chem. Acc.* **2017**, *136*, 55.

Received: 31 May 2017
Revised: 2 September 2017
Accepted: 22 October 2017
Published online on 00 Month 2017