An explicit-solvent hybrid QM and MM approach for predicting pKa of small molecules in SAMPL6 challenge

Samarjeet Prasad[1,2] · Jing Huang[3] · Qiao Zeng[1] · Bernard R. Brooks[1]

Abstract

In this work we have developed a hybrid QM and MM approach to predict pKa of small drug-like molecules in explicit solvent. The gas phase free energy of deprotonation is calculated using the M06-2X density functional theory level with Pople basis sets. The solvation free energy difference of the acid and its conjugate base is calculated at MD level using thermodynamic integration. We applied this method to the 24 drug-like molecules in the SAMPL6 blind pKa prediction challenge. We achieved an overall RMSE of 2.4 pKa units in our prediction. Our results show that further optimization of the protocol needs to be done before this method can be used as an alternative approach to the well established approaches of a full quantum level or empirical pKa prediction methods.

Introduction

Computational prediction of pKa values is of considerable interest for a number of fields including pharmaceutical and material sciences [1–3]. Even though several methods have been developed to predict this value, the problem still remains a challenge [4–6]. Most prediction methods can be divided into two broad categories—empirical and ab initio ones.

The first set of methods use a cheminformatics based approach [7–9]. In this approach the compound is represented as a vector of molecular descriptors including constitutional, topological, electrostatic and quantum descriptors

✉ Samarjeet Prasad
samar.samarjeet@nih.gov

1 Laboratory of Computational Biology, National Heart, Lung and Blood Institute, National Institutes of Health, Bethesda, MD 20814, USA

2 Biophysics and Biophysical Chemistry, The Johns Hopkins University, School of Medicine, Baltimore, MD 21205, USA

3 School of Life Sciences, Westlake University, 18 Shilongshan Street, Xihu District, Hangzhou, Zhejiang, China

[10]. Machine learning models for specific functional groups are trained based on these descriptors [10]. Notably, these methods ignore the three dimensional conformation of the compound explicitly [11]. Although training the models might be expensive in terms of curating experimental pKa data for generating appropriate models, subsequent pKa prediction using trained models can be very fast and inexpensive.

Ab initio methods use a thermodynamic cycle combining with quantum mechanics (QM) calculations to compute the solvent-phase pKa [12–20] . It consists of the calculation of dissociation free energy in gas phase [21] along with solvation free energy of the acid and the conjugate base using dielectric continuum solvation models (DCSMs) [12, 22–25]. These methods have been very successful in calculating pKa. However, DCSMs cannot model the hydrogen bonding between solute and water, which can be important in the protonation or deprotonation process [26]. Their accuracy in describing the short-range electrostatics of polar solutes and ions is also limited [12]. Moreover, typically only one conformation is used for the estimation of free energy although an ensemble of conformations is required for a complete statistical mechanics treatment of the free energy [27]. Even if multiple low lying conformations are included in the calculation, the entropic variations associated with the deprotonation process still cannot be completely accounted

for without explicitly considering the solvent dynamics and extensively exploring the potential energy landscape of the solute–solvent systems.

Calculation of solvation free energy during pKa estimation remains one of the bottlenecks in getting accurate values. An alternative way of calculating solvation free energy is to use molecular dynamics simulations with empirical force field [28–30]. Shirts et al. were able to do a very precise measurement of solvation free energy with 0.85 kcal/mol RMSE [31]. Gilson and co-workers used double decoupling method and achieved 1.3 kcal/mol RMSE. König et al. [29] used the annihilation approach and obtained accuracy on par with the quantum calculations. Mobley et al. have created the FreeSolv [30] database to catalog molecules with known experimental solvation free energy and assist in development of new methods from these resources.

Given the large number of diverse methods available for predicting pKa, the Statistical Assessment of the Modeling of Proteins and Ligands (SAMPL) [32] blind prediction challenge was organized to assess the methods on a common set of small drug-like molecules. Previous iterations of the SAMPL competitions have focussed on assessing methods for solvation free energy calculations [33], distribution coefficient and other challenges [34–37]. We note that in the SAMPL5 distribution coefficient competition, Pickard and coworkers have calculated pKa values with QM methods, and used computed pKa to further correct their prediction of distribution coefficients [34].

In this work we have presented a new method to computationally predict the pKa of small drug-like molecules in explicit solvent. This is a hybrid QM and MM approach that allows ab initio prediction of absolute pKa values and supports any chemistry. Since calculation of pKa requires relative solvation free energy between the acid (protonated species) and the conjugate base (deprotonated species), our method calculates this quantity directly rather than computing the absolute solvation free energies of both by employing two thermodynamic cycles.

This paper is organized as follows. In "Theory" section, we describe the theory behind the prediction of the microscopic and macroscopic pKa values. "Method" section covers the details of the description of the QM and MM methods that we used to carry out calculations. Next in "Result and discussion" section, we present our results that submitted to the SAMPL6 competition and analyze the accuracy of the results. Finally in "Conclusion" section, a brief conclusion is provided.

Theory

SAMPL6 pKa challenge involved blind computational prediction of pKa of 24 small drug-like molecules (Fig. 1). These molecules were similar to kinase inhibitors and were chosen for experimental tractability. All the molecules were polyprotic in nature i.e. there were multiple sites on each molecule where the molecule could lose a proton. For further details, please refer Isik et al. [38] where the organizers have described the rationale for choosing the molecules as well as the methods used for experimental pKa prediction.

In order to compare the computational and experimental pKa predictions, it is important to understand the difference between the microscopic and macroscopic pKa of a molecule. The chemical environment around a functional group (in this case, the protonation state of other titrable moieties) affect the propensity of the group to lose its proton. This is referred to as the microscopic pKa, i.e. pKa for deprotonation at a site at a fixed protonation state of all other titrable sites in the molecule. This differs from the macroscopic pKa which is related to the dissociation constant of losing a proton from the molecule as a whole and can be experimentally measured. Converting microscopic pKas to macroscopic pKas or vice versa is complicated due to the large number of equilibrium processes involved [8, 39]. If, for a specific charge transition, the microscopic pKas are fairly well separated (ex. more than one pKa unit), the smallest pKa can be considered as the macroscopic pKa. However, if they are close, the macroscopic pKa is shifted as multiple microscopic transitions contribute to the macroscopic value. Several studies [40, 41] discuss this in greater detail. In our method, we calculate microscopic pKa value for each acid–base pair of microscopic states. We then assign one dominant microscopic pKa as the macroscopic pKa for each titration process, which can be directly compared with the experimental observables.

To calculate the microscopic pKa of a particular acid–base pair, let us consider the dissociation of acid HA
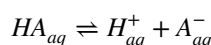
$$HA_{aq} \rightleftharpoons H^+_{aq} + A^-_{aq}$$

Here the subscripts ''aq'' indicate that the species are solvated in water. The dissociation constant and pKa value for this dissociation are given by the following relations:

$$K_a = \frac{[H^+]_{aq}[A^-]_{aq}}{[HA]_{aq}}$$

$$pKa = \frac{\Delta G^*_{aq}}{RTln(10)}$$

where

$$\Delta G^*_{aq} = G^*(H^+_{aq}) + G^*(A^-_{aq}) - G^*(HA_{aq})$$

Here, G refers to the absolute Gibbs free energy of the solvated species. The superscript * implies that the standard state of 1 mol/L and 298.15 K have been used. R and T are
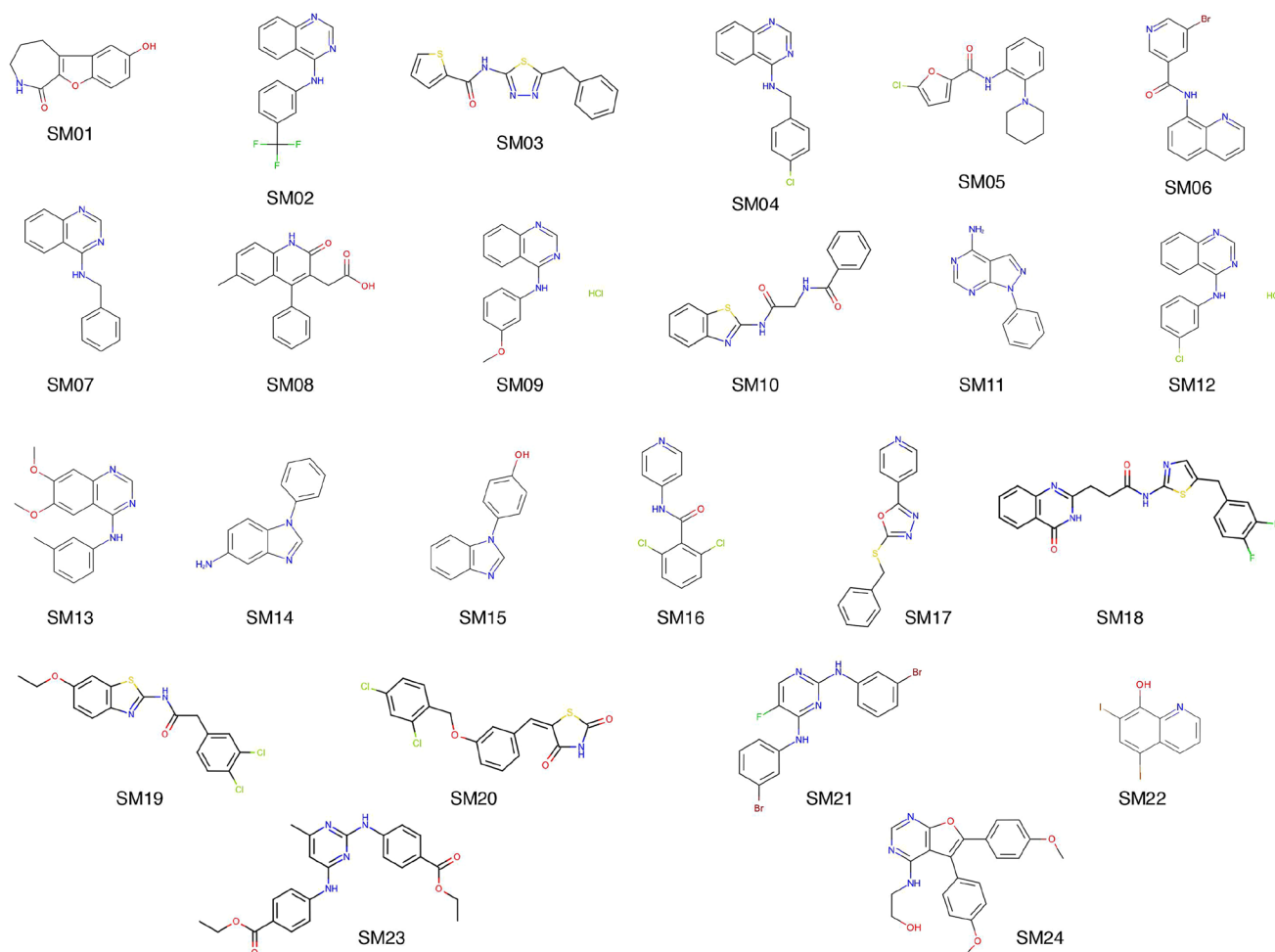
Fig. 1 Molecules in the SAMPL6 prediction challenge

the gas constant and the absolute temperature respectively. Thus, to calculate pKa we need to calculate aqueous phase deprotation free energy ΔG_{aq}.

Rather than calculating the absolute free energies in the aqueous phase directly, the aqueous phase calculations are coupled with gas phase calculation using the following thermodynamic cycle (Fig. 2a). The two vertical lines in the figure refer to the solvation of the species into aqueous phase. Thus, the ΔG_{aq} can be calculated as

$$\Delta G_{aq}^* = \Delta G_g^* + \Delta G_{solv}^*(H^+) + \Delta G_{solv}^*(A^-)$$
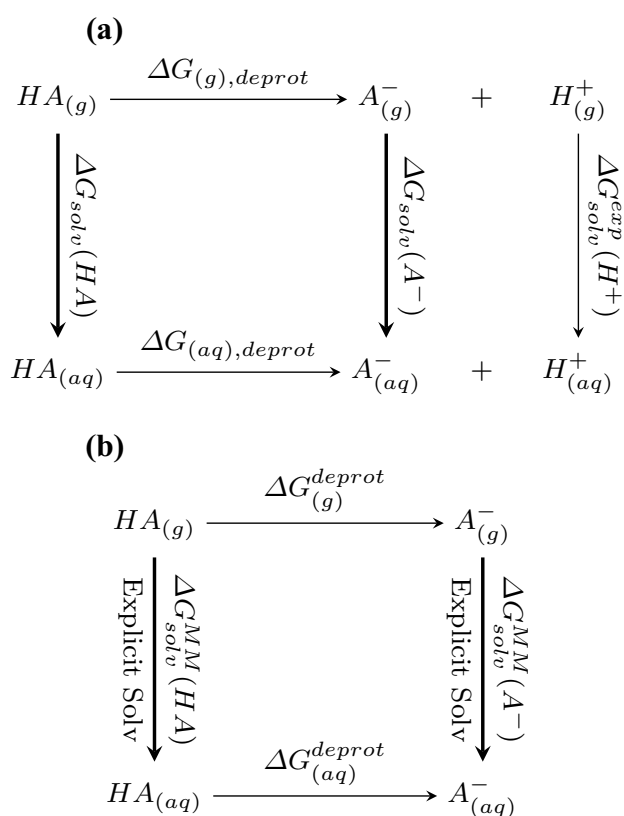$$- \Delta G_{solv}^*(HA)$$

The absolute free energy for proton H^+ in the gas phase at standard temperature and pressure is calculated by Sackur–Tetrode equation and has been previously calculated as − 6.28 kcal/mol [42]. Solavtion free energy of proton (− 264.5 kcal/mol) has been taken from Tissandier et al. [43]. The gas phase calculations are done at standard gas

conditions i.e. one atmosphere of pressure. Converting them to 1 mol/L further involves a standard state correction of − 1.89 kcal/mol.

The above equation involves the calculation of solvation free energies of the deprotonated $\Delta G_{solv}^*(A^-)$ and of the protonated species $\Delta G_{solv}^*(HA)$, respectively. Most ab initio pKa prediction methods compute them in implicit solvent using quantum chemistry and continuum solvent approaches. We note that, however, the only relevant quantity for pKa prediction is the difference of solvation free energies

$$\Delta\Delta G_{solv}^* = \Delta G_{solv}^*(A^-) - \Delta G_{solv}^*(HA)$$

In the present work, we directly compute this solvation free energy *difference* in explicit solvent. The calculation is done at the force field level in order to be computationally tractable. Furthermore we consider a second thermodynamic cycle (Fig. 2b) that alchemically change HA into A^- in the gas and the aqueous phases. As we are interested in only the free energy difference between the two species HA and

(a)

$$HA_{(g)} \xrightarrow{\Delta G_{(g),deprot}} A^-_{(g)} \quad + \quad H^+_{(g)}$$

$$\Big\downarrow \Delta G_{solv}(HA) \qquad \Big\downarrow \Delta G_{solv}(A^-) \qquad \Big\downarrow \Delta G^{exp}_{solv}(H^+)$$

$$HA_{(aq)} \xrightarrow{\Delta G_{(aq),deprot}} A^-_{(aq)} \quad + \quad H^+_{(aq)}$$

(b)

$$HA_{(g)} \xrightarrow{\Delta G^{deprot}_{(g)}} A^-_{(g)}$$

$$\text{Explicit Solv} \Big\downarrow \Delta G^{MM}_{solv}(HA) \qquad \text{Explicit Solv} \Big\downarrow \Delta G^{MM}_{solv}(A^-)$$

$$HA_{(aq)} \xrightarrow{\Delta G^{deprot}_{(aq)}} A^-_{(aq)}$$

Fig. 2 Thermodynamic cycles used in the pKa calculations **a** chemical reaction of acid dissociation. This relates the free energy of dissociation in the aqueous phhase as with the gas phase free energy of dissociation and solvation free energies of the acid, base and proton. **b** Alchemical cycle for deprotonation. This cycle relates the solavtion free energy *difference* of the HA and A⁻ with *difference* in free energy for deprotonation in the aqueous and gas phases

A^- and free energy is a state function so that its sum over a thermodynamic cycle equals zero, we can rewrite $\Delta\Delta G^*_{solv}$ as

$$\Delta\Delta G^*_{solv} = \Delta G^*_{solv}(A^-) - \Delta G^*_{solv}(HA)$$
$$= \Delta G^*_{deprot,aq}(HA) - \Delta G^*_{deprot,g}(HA),$$

where $\Delta G^*_{deprot}(HA)$ can be calculated using free energy perturbation (FEP) methods such as the thermodynamics integration (TI) method. By introducing a number intermediate λ states that alchemically connecting two states 0 and 1, the free energy difference between the two end state is computed by TI as

$$\Delta G = \int_0^1 \left\langle \frac{dU}{d\lambda} \right\rangle_\lambda d\lambda$$

It's worth pointing that for each acid–base pair only one relative free energy in the aqueous phase is computed, rather than two absolute solvation free energies. It has previously

been shown by Jorgensen et al. [44] that this allows the cancellation of errors in MM calculations such as inaccuracy of force field parameters and inadequate conformational samplings. In their work they calculated the relative solvation free energy of methanol and ethane using alchemical transformation of methanol to ethane and vice versa and got results close to experimental relative solvation free energy value. The major advantage of using such a secondary thermodynamic cycle (Fig. 2b) is that the alchemical FEP only involves changing *HA* into *A⁻* in the gas and the aqueous phase, instead of annihilating whole molecules in the aqueous phase. This greatly improves the efficiency, accuracy and the throughput of our calculations.

In summary, we calculate the ΔG^*_{aq} by the following equation

$$\Delta G^*_{aq} = \Delta G^*_g + \Delta G^*(H^+) + \Delta G^*_{deprot,aq}(HA) - \Delta G^*_{deprot,g}(HA),$$

where ΔG^*_g is calculated in the gas phase at the QM level, $\Delta G^*(H^+)$ is obtained from experimental value reported in literature, $\Delta G^*_{deprot,aq}(HA)$ is calculated using FEP in condensed phase at the MM level and $\Delta G^*_{deprot,g}(HA)$ in gas phase at the MM level.

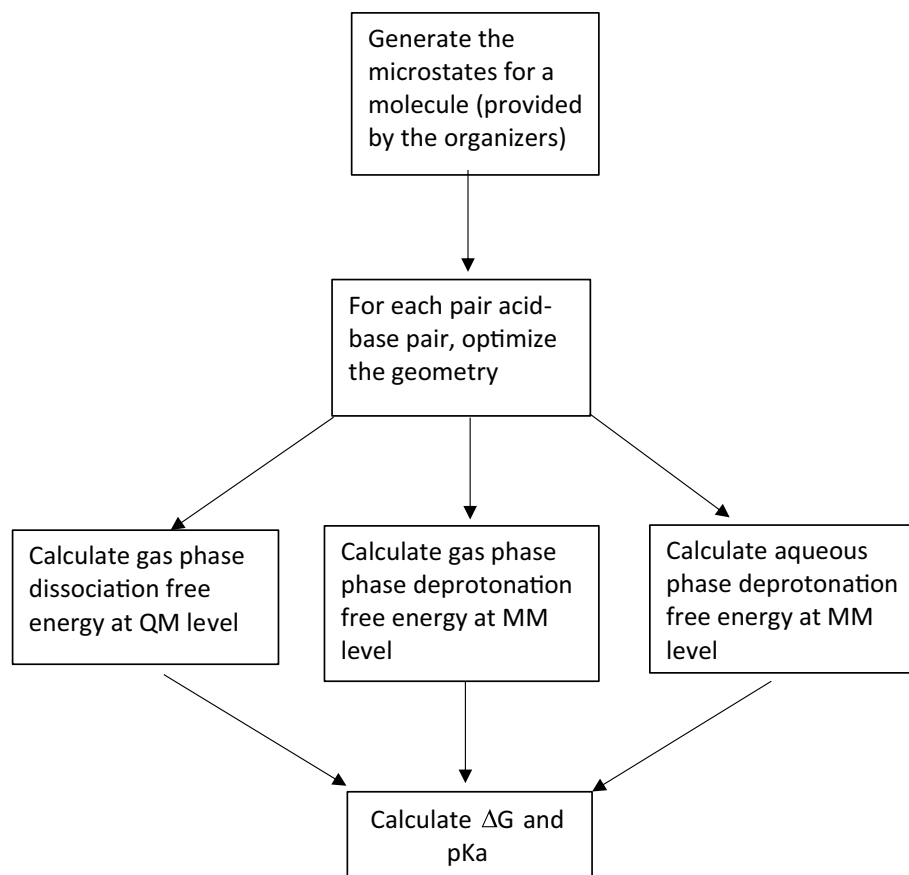# Method

The work flow for the complete method is shown in Fig. 3. First the geometry of each microstate was optimized in gas phase. Then for each acid (protonated)–base (deprotonated) pair, ΔG for deprotonation in gas phase was calculated at the QM level. To carry out the MM simulations, force field parameters were generated for each of the microstates. Next, the gas phase and aqueous phase alchemical free energy difference between each acid–base pair were computed using FEP and MD simulations. All the QM calculations were performed with Gaussian16 [45] , while all the MD simulations were done with CHARMM [46, 47].

Geometry optimization and gas phase QM calculation

SAMPL6 pKa challenge had 24 molecules, each with different number of microstates. SMILES [48] string of the microstates were converted to PDB files using OpenBabel [49]. Geometry optimization and gas phase deprotonation energy ΔG^*_g was calculated with the M06-2X density functional theory [50] and 6-31G* basis set for neutral–cationic microstate pairs and 6-31+G* for neutral–anionic microstate pairs. "Ultrafine grid" and "Tight" convergence criteria were used in all calculations.

Fig. 3 Workflow for the hybrid QM and MM pKa prediction approach

```
┌─────────────────┐
│ Generate the    │
│ microstates for a│
│ molecule (provided│
│ by the organizers)│
└─────────────────┘
         │
         ▼
┌─────────────────┐
│ For each pair acid-│
│ base pair, optimize│
│ the geometry    │
└─────────────────┘
```

Calculate gas phase dissociation free energy at QM level

Calculate gas phase phase deprotonation free energy at MM level

Calculate aqueous phase deprotonation free energy at MM level

Calculate ΔG and pKa

We would like to point out that as the computed pKa are directly related to the calculated electronic energies, higher-level methods such as MP2 and larger basis sets such as cc-pVTZ would improve calculation results. These, however, have not been pursued in this study. We also did not test other functionals, which might potentially lead to better pKa prediction results.

Parameterization of microstates

In order to carry out molecular dynamics simulations, we first generated force field parameters for the microstates based on the fixed-charge molecular mechanics potential energy functions used in CHARMM [51]. The potential energy is given by a sum of bonded and non-bonded components:

$$U = U_{bonded} + U_{non\text{-}bonded}$$

where,

$$U_{bonded} = \Sigma_{bond}K_b(r_{ij} - r_0)^2 + \Sigma_{angle}K_\theta(\theta_{ij} - \theta_0)^2$$
$$+ \Sigma_{dihedrals}K_\chi(1 + cos(n\chi - \delta)) + \Sigma_{improper}K_{imp}(\phi - \phi_0)^2$$

$$U_{non-bonded} = \Sigma\frac{q_iq_j}{4\pi\epsilon_0 r_{ij}} + \epsilon_{ij}\left[\left(\frac{R_{min}}{r_{ij}}\right)^{12} - 2\left(\frac{R_{min}}{r_{ij}}\right)^6\right]$$

Here, K_b and r_0 are bond force constant and equilibrium bond-length for each atom type pair. K_θ and θ_0 are angle force constant and equilibrium angle for each angle type triplet. K_{imp} and ϕ_0 are improper angle force constant and equilibrium improper angle for each improper angle. K_χ, n, and δ are the force constant, periodicity, and phase for each torsional degree of freedom. The non-bonded potential energy terms involve Coulombic interactions between partial charge q_i and q_j, and the van der Waals (VdW) interactions modeled by the ϵ_{ij} and R_{min} parameters.

We used Antechamber to generate GAFF parameters. Single point calculation was done on the optimized geometry mentioned above using Gaussian16 at MP2 level of theory with 6-31G* basis set. RESP charges were calculated using the protocol mentioned in Jakalian et al. [52]. Electrostatic potential was written in a data file using the option IOp(6/33=2) in Gaussian, and the RESP charges were fitted. Other parameters—bonded (bond, angle and torsion) and non-bonded (van der Waals) were assigned as per the General Amber Force Field (GAFF) [53] using the Antechamber [52] program in the AmberTools16 software. CHARMM formatted parameter and topology files were produced. These parameters were modified by in-house scripts to make the formats compatible with CHARMM molecular dynamics package. If the residues did not have an integer charge in

the generated topology file (typically off by $\pm 0.0 - 0.003$), an ad-hoc fix was done by adjusting the charge on a random non-hydrogen atom to round up the total charge of residue.

Free energy simulations

All molecular dynamics simulations were carried out with CHARMM [47] and parameter sets mentioned in the previous subsection. Thermodynamic Integration calculations were carried out using the PERT module of CHARMM. 12 λ windows were used (0.0. 0.075, 0.15, 0.25, 0.35, 0.45, 0.55, 0.65, 0.75, 0.85, 0.95, 1.00) for transforming the partial charges of the acid into those of the conjugate base, with the charge on the dissociating proton transforming to zero. Each λ window was equilibrated for 1 ps followed by 10 ps MD simulations for sampling.

MD simulations in the gas phase were carried out with Langevin dynamics at a temperature of 298 K and using a time step of 2 fs with a friction coefficient of 5 ps^{-1} on all the atoms. No cutoffs were used in calculation of nonbonded interactions for gas phase simulations. For aqueous phase simulations, we used 2022 water molecules to solvate the solute molecule, consituting a 38 Å cubic water box to start with. 50 ps NPT simulations were run at 298 K and 1 atm, after which NVT simulations at 298 K were carried out for TI calculations. A Nosé-Hoover thermostat [54] was used to maintain the microcanonical ensemble. Particle mesh ewald [55] was used to calculate the long range electrostatic interactions with a direct space cutoff of 10 Å. Charge was spread on a grid of $48 \times 48 \times 48$ for reciprocal space calculation using 6th order B-spline interpolation method [56]. A cutoff of 12 Å was applied for van der Waals interactions, and the integration time step is 1 fs.

Result and discussion

The results discussed in this report are the ones that we submitted for the SAMPL6 competition [submission id: 0wfzo]. We submitted only the microscopic pKas for all acid–base pairs of all the 24 molecules. These results were compared to macroscopic pKas using two different approaches—closest and Hungarian. This analysis was done with the assumption that experimentally observed pKas with only one observed pKa or fairly-distant pKas (separated by more than 3 units) are equal to the microscopic pKa of the corresponding microscopic pKa. Only two molecules—SM14 and SM18— did not satisfy this criterion and hence they were excluded from this analysis. Detailed analysis of the results can be found at https://github.com/MobleyLab/SAMPL6/tree/master/physical_properties/pKa/analysis/analysis_of_typeI_predictions.

Table 1 Statistics of the performance of the method using Hungarian and closest schemes

Evaluation scheme	RMSE	MAE	r^2	m
Hungarian	2.89	1.88	0.48	0.99
Closest	2.42	1.61	0.53	1.08

RMSE root mean square error, *MAE* maximum absolute error, r^2 correlation coefficient of determination, *m* slope of the linear regression line

Table 2 Comparison of experimental and calculated values using the closest scheme

Molecule	Exp. value	Calculated value
SM01	9.53 ± 0.01	10.5
SM02	5.03 ± 0.01	8.68
SM03	7.02 ± 0.01	8.81
SM04	6.02 ± 0.01	5.23
SM05	4.59 ± 0.01	1.31
SM06(1)	3.03 ± 0.04	4.84
SM06(2)	11.74 ± 0.01	10.96
SM07	6.08 ± 0.01	4.52
SM08	4.22 ± 0.01	5.85
SM09	5.37 ± 0.01	4.89
SM10	9.02 ± 0.01	8.83
SM11	3.89 ± 0.01	5.52
SM12	5.28 ± 0.01	5.78
SM13	5.77 ± 0.01	5.48
SM15(1)	4.7 ± 0.01	$- 4.16$
SM15(2)	8.94 ± 0.01	12.46
SM16	5.37 ± 0.01	2.86
SM17	3.16 ± 0.01	1.22
SM19	9.56 ± 0.01	9.6
SM20	5.7 ± 0.03	5.03
SM21	4.1 ± 0.01	4.12
SM22(1)	2.4 ± 0.02	3.79
SM22(2)	7.43 ± 0.01	6.29
SM23	5.45 ± 0.01	4.78
SM24	2.6 ± 0.01	2.74

In the closest analysis approach, the experimentally observed pKa is matched with the microscopic pKa which minimizes the absolute error i.e. the one that is closest to the observed pKa (Table 1). We achieved a root mean squared error (RMSE) of 2.42 pKa units with respect to the experimental values. The mean absolute error (MAE) was 1.61 pKa units. The corresponding R^2 for regression fit was 0.53 and the slope of line was 1.08.

In the hungarian approach [57], an optimum global match between experimentally observed pKa and predicted set of pKas is found by minimizing the linear sum of squared

Fig. 4 Plot of the closest analysis scheme and experimental pKa values. Plot courtesy of the organizers https://github.com/MobleyLab/SAMPL6/blob/master/physical_properties/pKa/analysis/analysis_of_typeI_predictions/analysis_outputs_closest/pKaCorrelationPlots/0wfzo.pdf
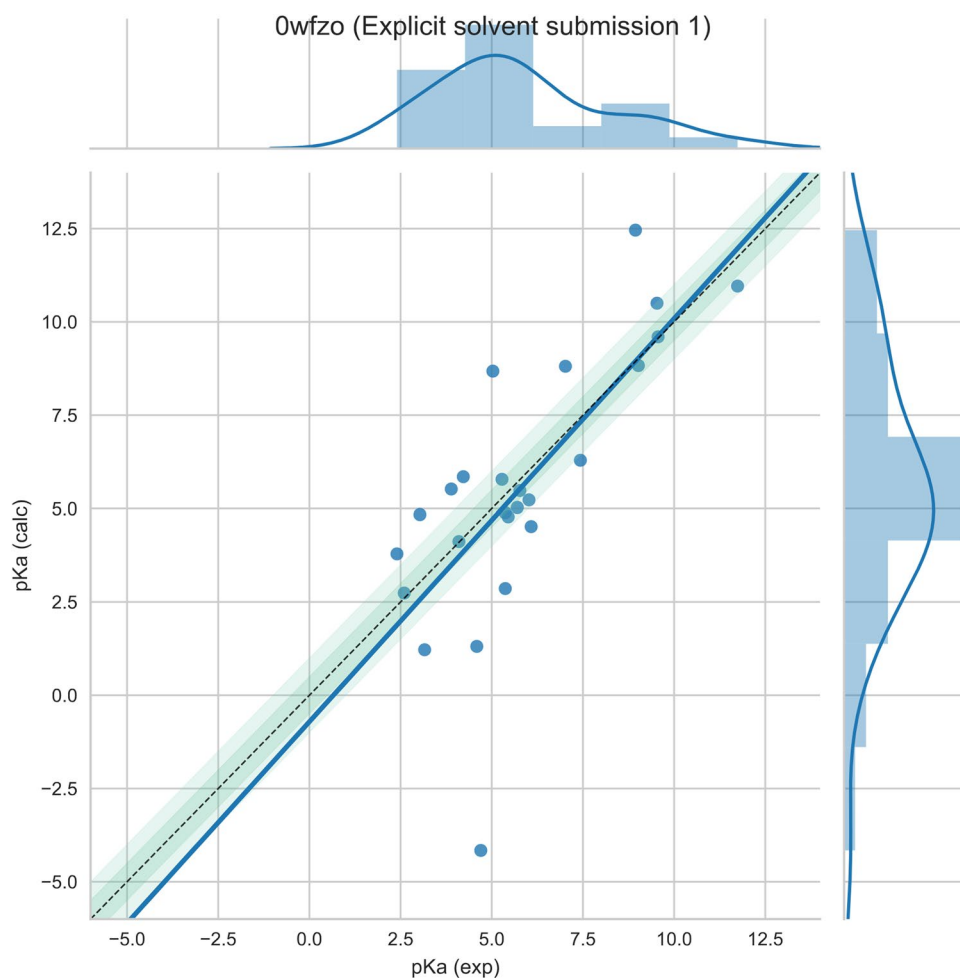
errors of the paired match (Table 1). We achieved a RMSE of 2.89 pKa units with respect to the experimental values. The MAE was 1.88 pKa units. The corresponding R^2 for regression fit was 0.48 and the slope of line was 0.99.

Out of the 22 molecules whose results were compared to experimental results, 3 of the molecules (SM06, SM15 and SM22) had 2 macroscopic pKas in the 2–12 pKa range while the other molecules had just 1 pKa in this range. Among these 25 comparisons, only five predictions were more than 2 pKa units away from the experimental values (Table 2). The most erroneous one concerns SM15, of which the fist predicted pKa underestimated the experimental measurement by 8.86 pKa units, and the second pKa overestimated by 3.52 pKa units (Fig. 4).

In general, our results compare less favorably to some of the more-established methods of pKa prediction, as used by other submissions in the SAMPL6 challenge. By carefully examining our calculations after the submission, a few mistakes were spotted, which are further analyzed and discussed here.

One major error is that the standard state correction was missed in our submission. The QM level gas phase calculation are done at standard state of gas while the aqueous phase species are at 1M concentration. This standard state correction needs to be applied while calculation of the overall free energy difference. This contribution is equal to − 1.89 kcal/mol, i.e. 1.4 pKa units.

Another source of error comes from the inconsistency with GAFF protocol. Standard AMBER and GAFF force fields scale the electrostatic interaction between third-neighbors (1–4 interactions) by 0.833, while CHARMM force fields on the other hand do not scale the electrostatic 1–4 interactions. In the CHARMM program, an option e14fac (electrostatic 1–4 interaction scaling factor) should be set to 0.833 to use GAFF force fields, however its default value of 1.0 was used in our simulations by mistake. Furthermore, the CHARMM modified TIP3P parameter were used for water molecules which place a small ϵ value on the water hydrogen atom. These deviations to the standard GAFF practice render the force field parameters used in this work less optimal.

Other methods to generate more CHARMM-like force field parameters for the microstates have been attempted. The Paramchem server [58], which generates CGENFF force field parameters, was found to report error messages

when parametrizing several charged species. The force field ToolKit (ffTK) [59], which is a plugin in VMD that generates CHARMM parameters, was found to be difficult in automatically generating parameters for all the microstates. Since we needed a method that could parameterize all the microstates in a high throughput fashion, we instead opted for using for Antechamber from AmberTools package.

From the absolute error analysis (Supplementary Fig. 2) we can assume that SM15 parameters are not optimal as the errors for both pKa are very high for this molecule. Force field parameterization for small molecules is indeed difficult due to the very large chemical space of these molecules as compared to the amino acids [60]. The latter have seen several decades of work for a very limited number of species. The general strategy of optimization of parameters of molecules involves the use experimental hydration free energy data [61]. Optimization with this parameter would also be helpful as we indeed need to predict the solvation free energy difference. However, many of microstates of these molecules are charged species and getting high accuracy experimental hydration free energy data would be difficult. Even Self-Consistent Reaction Field based implcit solvent model calculations have one order of magnitude higher error as compared to neutral species [23, 62]. One way to study the SM15 errors would be to generate parameters with a different force field and compare their relative performance. While Antechamber generates GAFF-based parameters, ffTK can be used to used to generate CHARMM-based parameters.

Our simulation runs also suffered from inadequate sampling of the phase space in the aqueous phase simulation. For the calculation of hydration free energy in SAMPL4 competition with similar system sizes, Gilson and co-workers [28] had simulated each λ point for 5 ns. König et al. [29] for the same set of molecules had used a 0.5–1 ns simulation for each λ state in aqueous phase. In principle much less sampling time would be required in our FEP calculations as relative free energies instead of absolute solvation free energies were being computed. However, the MD simulation time used in this study was still too short (10 ps per λ state), not allowing full water reorganization upon solute deprotonation. The number of simulations that we were performing was much larger (~ 650 in SAMPL6 vs. 24 in SAMPL4) and hence we performed only 0.12 ns simulations for each acid–base pair. Achieving proper sampling is an area of active research in the molecular dynamics field. Indeed, one of the competitions in the SAMPL6 challenge focused on benchmarking this quantity especially in a blind setup. The results from that study would be able to set community-wide guidelines for benchmarking. A heuristic that we should have used to reduce the number of microstate pairs should have been to exclude all microstates that had charges more than 1 or less than − 1 i.e. consider only neutral and singly-charged microstates. Some of the other submissions, have used this strategy to limit the number of microstate pairs that needs to be considered without loss in accuracy.

The FEP scheme we used for alchemical transformation included only the transformation of charges on all atoms from the protonated acid to the its deprotonated conjugate base. This was similar in principle to the strategy used by Lee et al. in their enveloping distribution sampling (EDS) based constant-pH simulations [63], where each state differed from the reference state in only the charges on the residue of deprotonation. The changes in the parameters for VdW interactions as well as the internal degrees of freedom during the solute deprotonation process will also contribute to free energy difference, which is not captured in our FEP calculations. We note that it's feasible to include these effects by interpolating all force field parameters, although the bonded interactions might need to be carefully handled [64].

Another possible source of error comes from the value of $\Delta G^*(H^+)$. Solvation free energy of proton is a contentious value and a range of values from − 259 to − 264 kcal/mol are available in the literature. This can lead to large errors in the absolute prediction of pKa as just an difference of 1.36 kcal/mol is equivalent to 1 pKa unit. One way to handle this error is to use isodesmic reactions with another acid with known experimental pKa and couple two thermodynamic cycles together such that the solvation free energy of proton cancels out. The second acid chosen should also be similar to the original acid that we are interested in. Essentially, the pKa shift is calculated with respect to a simpler model compound with known experiemental pKa values, as being done in most constant pH simulation methods [63, 65, 66]. Our approach instead aims at predicting the absolute pKa, and a fixed value of − 264.5 kcal/mol is used for $\Delta G^*(H^+)$ as derived from cluster-ion solvation data by Tissandier et al. [43]. An alternative way to handle this issue, as well as other systematic errors in absolute pKa calculations, is to perform a linear free energy regression against molecules with known experimental pKa, i.e., to consider $\Delta G^*(H^+)$ as a variable whose value is fitted to best reproduce a set of known pKa values. The empirical correction has been shown to improve the results although the slope of the regression still remains a debatable issue [12]. We have also used the assumption that only one microscopic pKa contributes to the macroscopic pKa if the former are fairly well separated. However, this is an approximation as for a given charge transition, multiple protonated–deprotonated pairs of microstates contribute to the macroscopic pKa [41].

In our approach the ΔG_g^* is computed using QM calculations at the M06-2X level using 6-31G* basis set (6-31+G* for microstate pairs involving anionic species). Higher level of ab initio methods, larger basis set, and including counterpoise correction should improve our results. Although our method allows the sampling of the phase space during the calculation of the solvation free energy difference, only one conformation

(the energy minimized one) is considered for the calculation of ΔG_g^* by QM in the gas phase. This is again an approximation as previous work by Bochevarov et al. [11] have shown that multiple low lying conformations do contribute to the deprotonation free energy. There can be a couple of different strategies to handle this phenomenon. Multiple low lying conformations can be sampled and the deprotonation energy of each important conformation can be calculated separately and combined together in a Boltzmann weighted sum. Another solution for this problem is to use reweighting as used by Tao et al. [67]. Free energy of constraining the geometry to the ones used the calculation of gas-phase QM step, can be calculated separately and will have to be added for the protonated microstate and subtracted for the deprotonated microstate.

One of the key physics behind the free energy of deprotonation and hence pKa is the water reorganization when the solute is protonated or deprotonated, which involves water response to the sudden changes of charge distributions. In this case, polarizable force fields should in principle provide higher accuracy in our approach as fixed charge force-fields are limited in their ability to account for the change in charges during the course of the simulation. A theoretically-promising method to handle this effect is to use polarizable force fields such as AMOEBA [68, 69], Drude [70] or a recently formulated multipole and induced dipole (MPID) model [71]. Any of these polarizable models should improve the pKa prediction results of our method, given high quality polarizable force field parameters for general drug-like molecules are available.

Conclusion

This work reports our submission for the SAMPL6 pKa prediction challenge, where we have attempted to calculate pKa of small drug-like molecules in explicit solvent using a hybrid QM and MM approach. While including multiple solvation shells is difficult in pure ab initio (QM) methods, modeling the dissociation of a proton is difficult at the MM level using conventional force fields. The novel contribution of this work is devising a method to allow the calculation of ΔG in explicit solvent while limiting the cost of the calculations. This is important for a high throughput prediction where a large number of microstates need to be considered.

However, traditional limitations in molecular dynamics simulation approaches limits its competitiveness as compared to a machine learning approach or a full-quantum level implicit solvent approach. At the same time we committed a few avoidable mistakes in carrying out the simulations. Due to these results from the present version of our method did not do very well in the SAMPL6 pKa challenge. More work needs to be done to optimize and automate the protocols.

We are currently working on improving the method. We need to improve force field parameters for the small molecules, ensure proper sampling of the intermediate lambda points during free energy calculations and utilize a higher level of theory for the gas phase QM calculations. Our new version of the method is an open source tool where we can use test the method easily for each of these factors. It will allow the method to be used for not just pKa calculation of small molecules but for larger proteins of interest as well. The open source tool, currently in development, is available at https://github.com/samarjeet/hpka.

References

1. Muckerman JT, Skone JH, Ning M, Wasada-Tsutsui Y (2013) Toward the accurate calculation of pKa values in water and acetonitrile. Biochimica et Biophysica Acta (BBA) Bioenergetics 1827(8–9):882–891. https://doi.org/10.1016/j.bbabio.2013.03.011
2. Seybold PG, Shields GC (2015) Computational estimation of pKa values. Wiley Interdisc Rev: Comput Mol Sci 5(3):290–297. https://doi.org/10.1002/wcms.1218
3. Wang Y, Xing J, Yuan X, Zhou N, Peng J, Xiong Z, Liu X, Luo X, Luo C, Chen K et al (2015) In silico adme/t modelling for rational drug design. Q Rev Biophys 48(4):488–515. https://doi.org/10.1017/S0033583515000190
4. Hajjar E, Dejaegere A, Reuter N (2009) Challenges in pKa predictions for proteins: the case of asp213 in human proteinase 3. J Phys Chem A 113(43):11783–11792. https://doi.org/10.1021/jp902930u
5. Lee AC, Crippen GM (2009) Predicting pKa. J Chem Inf Model 49(9):2013–2033. https://doi.org/10.1021/ci900209w
6. Zevatskii YE, Samoilov DV (2011) Modern methods for estimation of ionization constants of organic compounds in solution. Russ J Org Chem 47(10):1445–1467. https://doi.org/10.1134/s1070428011100010
7. Greenwood JR, Calkins D, Sullivan AP, Shelley JC (2010) Towards the comprehensive, rapid, and accurate prediction of the favorable tautomeric states of drug-like molecules in aqueous solution. J Comput-Aided Mol Des 24(6–7):591–604. https://doi.org/10.1007/s10822-010-9349-1
8. Fraczkiewicz R, Lobell M, Gller AH, Krenz U, Schoenneis R, Clark RD, Hillisch A (2014) Best of both worlds: combining pharma data and state of the art modeling technology to improve in silico pKa prediction. J Chem Inf Model 55(2):389–397. https://doi.org/10.1021/ci500585w
9. Shelley JC, Cholleti A, Frye LL, Greenwood JR, Timlin MR, Uchimaya M (2007) Epik: a software program for pK a prediction and protonation state generation for drug-like molecules. J Comput-Aided Mol Des 21(12):681–691. https://doi.org/10.1007/s10822-007-9133-z
10. Li M, Zhang H, Chen B, Wu Y, Guan L (2018) Prediction of pKa values for neutral and basic drugs based on hybrid

artificial intelligence methods. Sci Rep 8(1):3991. https://doi.org/10.1038/s41598-018-22332-7

11. Bochevarov AD, Watson MA, Greenwood JR, Philipp DM (2016) Multiconformation, density functional theory-based pKa prediction in application to large, flexible organic molecules with diverse functional groups. J Chem Theory Comput 12(12):6001–6019. https://doi.org/10.1021/acs.jctc.6b00805

12. Klamt A, Eckert F, Diedenhofen M, Beck ME (2003) First principles calculations of aqueous pKa values for organic and inorganic acids using cosmors reveal an inconsistency in the slope of the pKa scale. J Phys Chem A 107(44):9380–9386. https://doi.org/10.1021/jp034688o

13. Klici JJ, Friesner RA, Liu S-Y, Guida WC (2002) Accurate prediction of acidity constants in aqueous solution via density functional theory and self-consistent reaction field methods. J Phys Chem A 106(7):1327–1335. https://doi.org/10.1021/jp012533f

14. Thapa B, Bernhard Schlegel H (2017) Improved pKa prediction of substituted alcohols, phenols, and hydroperoxides in aqueous medium using density functional theory and a cluster-continuum solvation model. J Phys Chem A 121(24):4698–4706. https://doi.org/10.1021/acs.jpca.7b03907

15. Ho J (2015) Are thermodynamic cycles necessary for continuum solvent calculation of pKas and reduction potentials? Phys Chem Chem Phys 17(4):2859–2868. https://doi.org/10.1039/c4cp04538f

16. Lian P, Johnston RC, Parks JM, Smith JC (2018) Quantum chemical calculation of pKas of environmentally relevant functional groups: carboxylic acids, amines, and thiols in aqueous solution. J Phys Chem A 122(17):4366–4374. https://doi.org/10.1021/acs.jpca.8b01751

17. Riojas AG, Wilson AK (2014) Solv-ccca: implicit solvation and the correlation consistent composite approach for the determination of pKa. J Chem Theory Comput 10(4):1500–1510. https://doi.org/10.1021/ct400908z

18. Liptak MD, Shields GC (2001) Accurate pKa calculations for carboxylic acids using complete basis set and gaussian-n models combined with cpcm continuum solvation methods. J Am Chem Soc 123(30):7314–7319. https://doi.org/10.1021/ja010534f

19. Liptak MD, Shields GC (2001) Experimentation with different thermodynamic cycles used for pKa calculations on carboxylic acids using complete basis set and gaussian-n models combined with cpcm continuum solvation methods. Int J Quantum Chem 85(6):727–741. https://doi.org/10.1002/qua.1703

20. Tehan BG, Lloyd EJ, Wong MG, Pitt WR, Montana JG, Manallack DT, Gancia E (2002) Estimation of pKa using semiempirical molecular orbital methods. part 1: application to phenols and carboxylic acids. Quant Struct-Act Relat 21(5):457–472. https://doi.org/10.1002/1521-3838(200211)21:5<457::aid-qsar457>3.0.co;2-5.

21. Peverati R, Truhlar DG (2014) Quest for a universal density functional: the accuracy of density functionals across a broad spectrum of databases in chemistry and physics. Philos Trans R Soc A 372(2011):20120476–20120476. https://doi.org/10.1098/rsta.2012.0476

22. Klamt A, Schrmann G (1993) Cosmo: a new approach to dielectric screening in solvents with explicit expressions for the screening energy and its gradient. J Chem Soc Perkin Trans 2(5):799–805. https://doi.org/10.1039/p29930000799

23. Marenich AV, Cramer CJ, Truhlar DG (2009) Universal solvation model based on solute electron density and on a continuum model of the solvent defined by the bulk dielectric constant and atomic surface tensions. J Phys Chem B 113(18):6378–6396. https://doi.org/10.1021/jp810292n

24. Ho J, Ertem MZ (2016) Calculating free energy changes in continuum solvation models. J Phys Chem B 120(7):1319–1329. https://doi.org/10.1021/acs.jpcb.6b00164

25. Barone V, Cossi M (1998) Quantum calculation of molecular energies and energy gradients in solution by a conductor solvent model. J Phys Chem A 102(11):1995–2001. https://doi.org/10.1021/jp9716997

26. Ho J (2014) Predicting pKa in implicit solvents: current status and future directions. Aust J Chem 67(10):1441. https://doi.org/10.1071/ch14040

27. Casasnovas R, Ortega-Castro J, Frau J, Donoso J, Muoz F (2014) Theoretical pKa calculations with continuum model solvents, alternative protocols to thermodynamic cycles. Int J Quantum Chem 114(20):1350–1363. https://doi.org/10.1002/qua.24699

28. Muddana HS, Sapra NV, Fenley AT, Gilson MK (2014) The SAMPL4 hydration challenge: evaluation of partial charge sets with explicit-water molecular dynamics simulations. J Comput-Aided Mol Des 28(3):277–287. https://doi.org/10.1007/s10822-014-9714-6

29. König G, Pickard FC, Mei Y, Brooks BR (2014) Predicting hydration free energies with a hybrid qm/mm approach: an evaluation of implicit and explicit solvation models in SAMPL4. J Comput-Aided Mol Des 28(3):245–257. https://doi.org/10.1007/s10822-014-9708-4

30. Mobley DL, Bayly CI, Cooper MD, Shirts MR, Dill KA (2015) Correction to small molecule hydration free energies in explicit solvent: an extensive test of fixed-charge atomistic simulations. J Chem Theory Comput 11(3):1347–1347. https://doi.org/10.1021/acs.jctc.5b00154

31. Shirts MR, Pitera JW, Swope WC, Pande VS (2003) Extremely precise free energy calculations of amino acid side chain analogs: comparison of common molecular mechanics force fields for proteins. J Chem Phys 119(11):5740–5761. https://doi.org/10.1063/1.1587119

32. Peter J (2009) Guthrie. A blind challenge for computational solvation free energies: introduction and overview. J Phys Chem B 113(14):4501–4507. https://doi.org/10.1021/jp806724u

33. Muddana HS, Fenley AT, Mobley DL, Gilson MK (2014) The SAMPL4 hostguest blind prediction challenge: an overview. J Comput-Aided Mol Des 28(4):305–317. https://doi.org/10.1007/s10822-014-9735-1

34. Pickard FC, Knig G, Tofoleanu F, Lee J, Simmonett AC, Shao Y, Ponder JW, Brooks BR (2016) Blind prediction of distribution in the sampl5 challenge with qm based protomer and pKa corrections. J Comput-Aided Mol Des 30(11):1087–1100. https://doi.org/10.1007/s10822-016-9955-7

35. Yin J, Henriksen NM, Slochower DR, Shirts MR, Chiu MW, Mobley DL, Gilson MK (2016) Overview of the sampl5 hostguest challenge: are we doing better? J Comput-Aided Mol Des 31(1):1–19. https://doi.org/10.1007/s10822-016-9974-4

36. Geballe MT, Guthrie JP (2012) The sampl3 blind prediction challenge: transfer energy overview. J Comput-Aided Mol Des 26(5):489–496. https://doi.org/10.1007/s10822-012-9568-8

37. Rustenburg AS, Dancer J, Lin B, Feng JA, Ortwine DF, Mobley DL, Chodera JD (2016) Measuring experimental cyclohexane-water distribution coefficients for the sampl5 challenge. J Comput-Aided Mol Des 30(11):945–958. https://doi.org/10.1007/s10822-016-9971-7

38. Isik M (2018) pKa measurements for the sampl6 prediction challenge for a set of kinase inhibitor-like fragments. J Comput-Aided Mol Des. https://doi.org/10.1101/368787

39. Szakács Z, Noszál B (1999) Protonation microequilibrium treatment of polybasic compounds with any possible symmetry. J Math Chem 26(1):139

40. Philipp DM, Watson MA, Yu HS, Steinbrecher TB, Bochevarov AD (2018) Quantum chemical prediction for complex organic mole. Int J Quantum Chem 118(12):e25561. https://doi.org/10.1002/qua.25561

41. Darvey IG (1995) The assignment of pKa values to functional groups in amino acids. Biochem Educ 23(2):80–82. https://doi.org/10.1016/0307-4412(94)00150-n

42. McQuarrie DA (2000) Statistical mechanics. University Science Books, Sausalito

43. Tissandier MD, Cowen KA, Feng WY, Gundlach E, Cohen MH, Earhart AD, Coe JV, Tuttle TR (1998) The proton's absolute aqueous enthalpy and gibbs free energy of solvation from cluster-ion solvation data. J Phys Chem A 102(40):7787–7794. https://doi.org/10.1021/jp982638r

44. Jorgensen WL, Ravimohan C (1985) Monte carlo simulation of differences in free energies of hydration. J Chem Phys 83(6):3050–3054. https://doi.org/10.1063/1.449208

45. Frisch MJ, Trucks GW, Schlegel HB, Scuseria GE, Robb MA, Cheeseman JR, Scalmani G, Barone V, Petersson GA, Nakatsuji H, Li X, Caricato M, Marenich AV, Bloino J, Janesko BG, Gomperts R, Mennucci B, Hratchian HP, Ortiz JV, Izmaylov AF, Sonnenberg JL, Williams-Young D, Ding F, Lipparini F, Egidi F, Goings J, Peng B, Petrone A, Henderson T, Ranasinghe D, Zakrzewski VG, Gao J, Rega N, Zheng G, Liang W, Hada M, Ehara M, Toyota K, Fukuda R, Hasegawa J, Ishida M, Nakajima T, Honda Y, Kitao O, Nakai H, Vreven T, Throssell K, Montgomery JA Jr, Peralta JE, Ogliaro F, Bearpark MJ, Heyd JJ, Brothers EN, Kudin KN, Staroverov VN, Keith TA, Kobayashi R, Normand J, Raghavachari K, Rendell AP, Burant JC, Iyengar SS, Tomasi J, Cossi M, Millam JM, Klene M, Adamo C, Cammi R, Ochterski JW, Martin RL, Morokuma K, Farkas O, Foresman JB, Fox DJ (2016) Gaussian16 Revision B.01. GaussianInc., Wallingford, CT

46. Brooks BR, Bruccoleri RE, Olafson BD, States DJ, Swaminathan S, Karplus M (1983) Charmm: a program for macromolecular energy, minimization, and dynamics calculations. J Comput Chem 4(2):187–217. https://doi.org/10.1002/jcc.540040211

47. Brooks BR, Brooks CL, Mackerell AD, Nilsson L, Petrella RJ, Roux B, Won Y, Archontis G, Bartels C, Boresch S et al (2009) Charmm: the biomolecular simulation program. J Comput Chem 30(10):1545–1614. https://doi.org/10.1002/jcc.21287

48. Anderson Eric, Veith GD, Weininger D (1987) SMILES, a line notation and computerized interpreter for chemical structures. U.S. Environmental Protection Agency, Environmental Research Laboratory, Duluth

49. Mazzatorta P, Tran L-A, Schilter B, Grigorov M (2007) Integration of structure activity relationship and artificial intelligence systems to improve in silico prediction of ames test mutagenicity. ChemInform. https://doi.org/10.1002/chin.200715211

50. Zhao Y, Truhlar DG (2007) The m06 suite of density functionals for main group thermochemistry, thermochemical kinetics, noncovalent interactions, excited states, and transition elements: two new functionals and systematic testing of four m06-class functionals and 12 other functionals. Theor Chem Acc 120(1–3):215–241. https://doi.org/10.1007/s00214-007-0310-x

51. MacKerell AD, Bashford D, Bellott M, Dunbrack RL, Evanseck JD, Field MJ, Fischer S, Gao J, Guo H, Ha S, Joseph-McCarthy D, Kuchnir L, Kuczera K, Lau FTK, Mattos C, Michnick S, Ngo T, Nguyen DT, Prodhom B, Reiher WE, Roux B, Schlenkrich M, Smith JC, Stote R, Straub J, Watanabe M, Wiskiewicz-Kuczera J, Yin D, Karplus M (1998) All-atom empirical potential for molecular modeling and dynamics studies of proteins. J Phys Chem B 102(18):3586–3616. https://doi.org/10.1021/jp973084f

52. Jakalian A, Jack DB, Bayly CI (2002) Fast, efficient generation of high-quality atomic charges. am1-bcc model: II. Parameterization and validation. J Comput Chem 23(16):1623–1641. https://doi.org/10.1002/jcc.10128

53. Wang J, Wolf RM, Caldwell JW, Kollman PA, Case DA (2004) Development and testing of a general amber force field. J Comput Chem 25(9):1157–1174. https://doi.org/10.1002/jcc.20035

54. Evans DJ, Holian BL (1985) The nosehoover thermostat. J Chem Phy 83(8):4069–4074. https://doi.org/10.1063/1.449071

55. Darden T, York D, Pedersen L (1993) Particle mesh ewald: an nlog(n) method for ewald sums in large systems. J Chem Phys 98(12):10089–10092. https://doi.org/10.1063/1.464397

56. Essmann U, Perera L, Berkowitz ML, Darden T, Lee H, Pedersen LG (1995) A smooth particle mesh ewald method. J Chem Phys 103(19):8577–8593. https://doi.org/10.1063/1.470117

57. Kuhn HW (1955) The hungarian method for the assignment problem. Naval Res Logist Q 2(12):83–97. https://doi.org/10.1002/nav.3800020109

58. Vanommeslaeghe K, Mackerell AD (2012) Automation of the charmm general force field (cgenff) I: bond perception and atom typing. J Chem Inf Model 52(12):3144–3154. https://doi.org/10.1021/ci300363c

59. Mayne CG, Gumbart JC, Tajkhorshid E (2013) The force field toolkit: software for the parameterization of small molecules from first principles. Biophys J 104(2):31a. https://doi.org/10.1016/j.bpj.2012.11.209

60. Huang L, Roux B (2013) Automated force field parameterization for nonpolarizable and polarizable atomic models based on ab initio target data. J Chem Theory Comput 9(8):3543–3556. https://doi.org/10.1021/ct4003477

61. Oostenbrink C, Villa A, Mark AE, Van Gunsteren WF (2004) A biomolecular force field based on the free enthalpy of hydration and solvation: the gromos force-field parameter sets 53a5 and 53a6. J Comput Chem 25(13):1656–1676. https://doi.org/10.1002/jcc.20090

62. Miguel ELM, Santos CIL, Silva CM, Pliego JR Jr (2016) How accurate is the SMD model for predicting free energy barriers for nucleophilic substitution reactions in polar protic and dipolar aprotic solvents? J Braz Chem Soc 27:2055–2061. https://doi.org/10.5935/0103-5053.20160095

63. Lee J, Miller BT, Brooks BR (2015) Computational scheme for pH-dependent binding free energy calculation with explicit solvent. Protein Sci 25(1):231–243. https://doi.org/10.1002/pro.2755

64. Knig G, Brooks BR (2015) Correcting for the free energy costs of bond or angle constraints in molecular dynamics simulations. Biochim Biophys Acta (BBA) 1850(5):932–943. https://doi.org/10.1016/j.bbagen.2014.09.001

65. Khandogin J, Brooks CL (2005) Constant pH molecular dynamics with proton tautomerism. Biophys J 89:141–157

66. Donnini S, Tegeler F, Groenhof G, Grubmuller H (2011) Constant pH molecular dynamics in explicit solvent with λ-dynamics. J Chem Theory Comput 7:1962–1978

67. Tao P, Sodt AJ, Shao Y, Knig G, Brooks BR (2014) Computing the free energy along a reaction coordinate using rigid body dynamics. J Chem Theory Comput 10(10):4198–4207. https://doi.org/10.1021/ct500342h

68. Ponder JW, Wu C, Ren P, Pande VS, Chodera JD, Schnieders MJ, Haque I, Mobley DL, Lambrecht DS, DiStasio RA, Head-Gordon M, Clark GNI, Johnson ME, Head-Gordon T (2010) Current status of the amoeba polarizable force field. J Phys Chem B 114(8):2549–2564

69. Bradshaw RT, Essex JW (2016) Evaluating parametrization protocols for hydration free energy calculations with the amoeba polarizable force field. J Chem Theory Comput 12(8):3871–3883. https://doi.org/10.1021/acs.jctc.6b00276

70. Baker CM, Lopes PEM, Zhu X, Benoit R, Mackerell AD (2010) Accurate calculation of hydration free energies using pair-specific lennard-jones parameters in the charmm drude polarizable force field. J Chem Theory Comput 6(4):1181–1198

71. Huang J, Simmonett AC, Pickard FC, Mackerell AD, Brooks BR (2017) Mapping the drude polarizable force field onto a multipole and induced dipole model. J Chem Phys 147(16):161702. https://doi.org/10.1063/1.4984113