Methods

The pKa value of a titratable residue can be calculated from the one-half point of the probability of ionization states as a function of pH (titration curve), or from the shift of the residue solvent reference pKa (eq.1).

$$pKa_i(\text{protein}) = pKa_{i,ref}(\text{solvent}) + \Delta pKa_i\ (\text{solvent} \Rightarrow \text{protein}) \quad (1)$$

In both approaches, it is essential to calculate the electrostatic free energy of the titratable residue in its protonated and deprotonated states. The DelPhiPka program contains 4 modules to calculate pKa values and they are described in the following.

Protonation

A residue topology based method is applied here to generate the hydrogen positions for the atoms in structural files. For each residue type, the corresponding heavy atom bond connectivity, hydrogen positions are included in the topology file, as well as the reference pKa value for each titratable residue group. We also extend the structural information to include nucleic acids for pKa calculations of RNA and single stranded DNA. Taking into account that the extra hydrogen of carboxyl groups (glutamic acid and aspartic acid) can be bound to either oxygen, an option is provided to choose either hydrogen conformation. The default choice is set to be OE1 (Glu) and OD1 (Asp), which is determined based on the benchmark results.

The program utilizes pre-calculated force-field parameters for atomic charges and radii, and now it supports AMBER, CHARMM, PARSE and GROMOS force fields.

Electrostatic energy calculation

The electrostatic energy contributing to the pKa can be further divided into three terms: polar energy, desolvation energy and charge-charge pairwise interaction energy. And each energy term is calculated via the modified Delphi program with the Gaussian dielectric model. The Gaussian dielectric model that implemented in Delphi has been discussed in our previously published article [1], here we do not describe it any more. There are three adjustable parameters associated with the

Gaussian dielectric model, which are ε_{ref} (reference dielectric constant for the protein), ε_{water} (dielectric constant for the water) and δ (the variance of the Gaussian function). These parameters are inherited and designed to be the key adjustable parameters in the DelPhiPKa program that affect the pKa calculation results.

To calculate the electrostatic energy of the ith ionizable residue, we first charge the side-chain atoms of the ith residue only and leave the rest of the structure uncharged. Then we use the FRC function of Delphi energy module to obtain the electrostatic potentials generated by the charged atoms on the side-chain of ith residue in the protein. Then three focusing calculations are performed to reach a final resolution of 4 grids/Å.

The reaction field energy $G_{i,charged}^{rxn}(protein)$ of the ionizable residue i embedded in the protein is calculated as the total grid energy generated by Delphi energy module as previously described [2]. In order to obtain the desolvation energy, we move the charged side-chain of the ith residue to the water and apply the same computational box with the same grid resolution to perform three focusing calculations again. Thus, the reaction field energy $G_{i,charged}^{rxn}(water)$ of the ionizable residue i in the water is obtained as the total grid energy difference from Delphi calculation. Thus, the desolvation energy of the residue i in its charged state is expressed as:

$$\Delta G_{i,charged}^{desol} = G_{i,charged}^{rxn}(protein) - G_{i,charged}^{rxn}(water) \quad (2)$$

The polar energy term of the electrostatic interactions between the charged residue i and other residues is calculated as:

$$G_{i,charged}^{polar} = \sum_{j \in ionizable} q_{j,backbone} \emptyset_{j,backbone} + \sum_{j \notin ionizable} q_j \emptyset_j \quad (3)$$

where $q_{j,backbone}$ and $\emptyset_{j,backbone}$ are atomic charges and electrostatic potentials for backbone atoms of ionizable residues including the ith residue itself. And q_j and $\emptyset_j$ are atomic charges and electrostatic potentials for the backbone and side-chain atoms of non-ionizable residues, respectively.

The charge-charge pairwise interaction energy between the side-chain of the ith residue and other ionizable residues is calculated as:

$$G_{i,j\ (charged)}^{pairwise} = \sum_{j \in ionizable, j \neq i} q_{j,sidechain}\, \emptyset_{j,sidechain} \quad (4)$$

where $q_{j,sidechain}$ and $\emptyset_{j,sidechain}$ represent the atomic charges and electrostatic potentials for the side-chain atoms of ionizable residues (excluding the ith residue itself).

We then turn the side-chain of the ith residue to its neutral state and follow the same protocol to calculate another three energy components $G_{i,neutral}^{polar}$, $G_{i,j\ (neutral)}^{pairwise}$ and $\Delta G_{i,neutral}^{desol}$. By extracting them from the energies of charged state,

$$\Delta G_i^{polar} = G_{i,charged}^{Polar} - G_{i,\ neutral}^{Polar}, \quad (5)$$

$$\Delta\Delta G_i^{desol} = \Delta G_{i,charged}^{desol} - \Delta G_{i,\ neutral}^{desol}, \quad (6)$$

we obtain the total electrostatic energy shift due to the change of protonation state, which is expressed as:

$$\Delta G_i = \gamma(i)\left[2.3 k_b T\left(pH - pKa_i^{ref,solvent}\right)\right] + \left(\Delta G_i^{polar} + \Delta\Delta G_i^{desol}\right) + \sum_{j=1, j\neq i}^{N} \Delta G_{i,j}^{pairwise} \quad (7)$$

Titration – calculating the probability of ionization states

We use the distribution of microstate electrostatic energy to determine the probability of ionization of the ith residue at the given pH. For the system with total M microstates and the energy $G_m(\mathrm{pH})$ at its mth microstate, the probability of ith residue to be ionized at a particular pH is given by the Boltzmann distribution:

$$P_i(\mathrm{pH}) = \frac{\sum_{m=1}^{M} \chi(i) \cdot e^{-G_m(\mathrm{pH})/kT}}{\sum_{m=1}^{M} e^{-G_m(\mathrm{pH})/kT}} \quad (8)$$

$\chi(i)$ is 1 if the ith residue is ionized and 0 if it is neutral. k is the Boltzmann constant. Then the Boltzmann distribution of ionized states is calculated as a function of pH, resulting the titration

curve where the *i*th residue possesses 50% probability of being protonated is designated as the pK_a value.

Each ionizable residue can be in two microstates: protonated and deprotonated. For the system with N ionizable residues, it has a total microstates $M = 2^N$. The Boltzmann sum needs to be calculated 2^N times per ionizable residues and $2^N \cdot N$ for the entire system. If the system has more than 30 ionizable residues, even for the modern computer and clusters, it is still extremely difficult to compute. Therefore, a cluster or partition approach is required to simplify the modeling, as described below.

Network Partition

Here, we introduce the Network partition algorithm. Networking is a geometrical distance based clustering protocol, which allows duplicate ionizable residues to appear in more than one partition. This eliminates the errors associated with wrong partitioning of strongly interacting groups. To partition the macromolecule with N ionizable residues into groups, we first label the geometric center of the side-chain of each ionizable residue as the representing point (RP) to obtain N RPs. Each RP locates its neighboring RPs within a given radius (a threshold that is set up by the input parameter) and constitutes a network. For efficiency, the ordering within each network is maintained based on the distance and the amount of RPs within a network is limited to be 20. If two networks consist of the same elements, one of them will be eliminated. The duplicate RP is tolerable within different networks. For the residue with a network, the change of its protonation states will be explicitly taken into account. For the residue not in the network, its protonation state is identified by the previous calculation and the microstate is fixed with a particular energy configuration.

Reference

1. Li, L., et al., *On the dielectric "constant" of proteins: smooth dielectric function for macromolecular modeling and its implementation in Delphi.* Journal of chemical theory and computation, 2013. **9**(4): p. 2126-2136.
2. Li, L., C. Li, and E. Alexov, *On the modeling of polar component of solvation energy using smooth Gaussian-based dielectric function.* Journal of Theoretical and Computational Chemistry, 2014. **13**(03): p. 1440002.

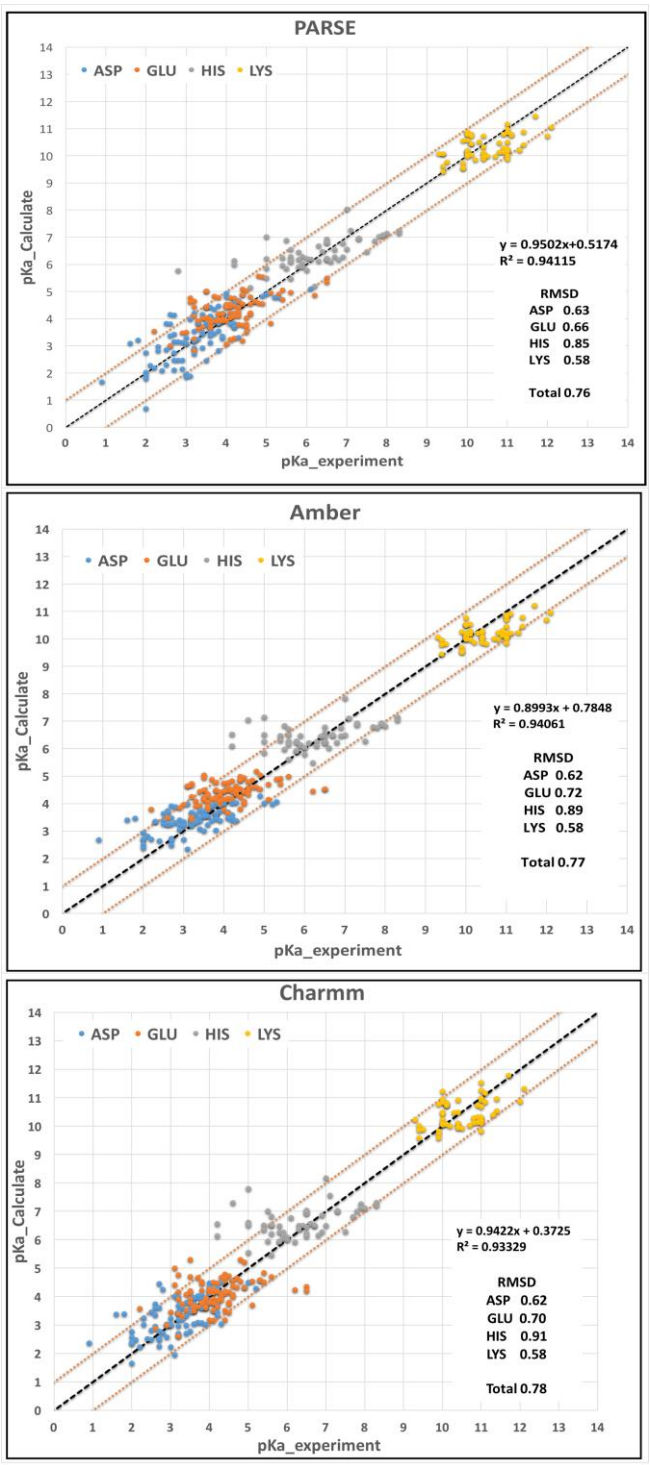

Figure 1S. Comparison of calculated pKa values for 302 ionizable residues from 32 proteins with experimentally measured values of the Protein pKa Dataset (PPD) with PARSE, AMBER and CHARMM force fields. Total RMSD is calculated as 0.76, 0.77 and 0.78, respectively. Correlation coefficients are 0.94, 0.94 and 0.93, respectively. RMSDs for individual residue are marked as well. Yellow lines are +/- 1.0 pK error compared with experimental values.

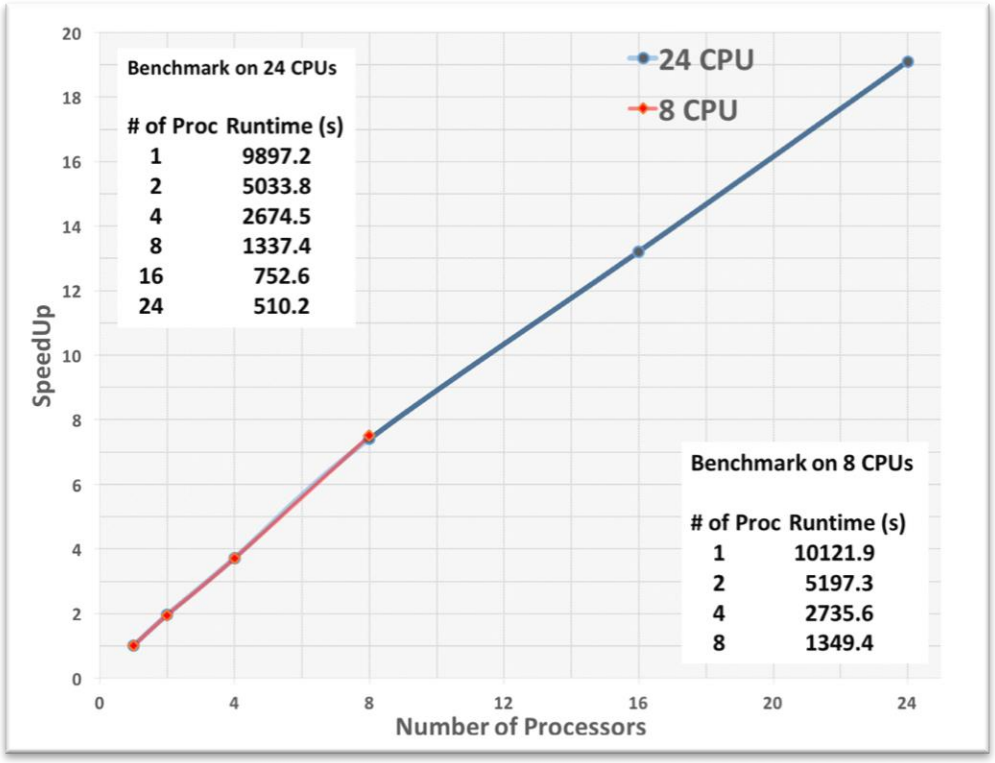

Figure 2S. Results of speed performance tests. The tests are performed on two machines, one of which has 8 CPUs and the other has 24 CPUs. The sample we selected is a large-scale protein 6-Phosphogluconate dehydrogenase (6PGDH, PDB ID: 2zyg), which contains 467 residues and 128 ionizable residues with the dimension of 119x113x113 Å. The benchmark was performed on the nodes with specifications of AMD Opteron 6176 (24 cores and 2.3GHz) and AMD Opteron 2356 (8 cores and 2.3GHz). The same job was performed 5 times and calculated the average runtime for the benchmark. Each job was calculated with 1, 2, 4, 8, 16, 24 CPUs. The speedup vs number of processors was plot.

Table 1S. Statistics of RMSDs and residue positions for results of accuracy test.

(i)	Number / Percentage		
	PARSE	**AMBER**	**CHARMM**
0.0 < RMSD < 1.0	271 / 89.7%	258 / 85.4%	275 / 91.1%
1.0 < RMSD < 2.0	29 / 9.6%	41 / 13.6%	22 / 7.3%
2.0 < RMSD	2 / 0.6%	3 / 1.0%	5 / 1.7%

(ii)	RMSD			Number / Percentage
	PARSE	**AMBER**	**CHARMM**	
Exposed (surface)	0.53	0.58	0.55	218 / 72.2%
< 50% Buried	0.82	0.81	0.88	53 / 17.5%
> 50% Buried	1.11	1.09	1.22	31 / 10.3%

Table 2S. Comparison of calculated nucleotide $\mathbf{pK_a}$ values in RNA with NMR measured results. The calculation was performed on two RNAs, Branch-point helix (BPH) and Lead-dependent ribozyme (LDZ). The mean±standad deviation of the calculated pKa values is given for 12 NMR structures for BPH (PDB ID: 17ra) and 25 NMR structures for LDZ (PDB ID: 1ldz).

Nucleotide	NMR measured $\mathbf{pK_a}$	Calculated $\mathbf{pK_a}$
Branch-point helix (BPH)		
A6	<5.0	4.5±0.6
A7	6.1	5.3±0.7
A10	<5.0	4.1±0.5
A13	5.5	4.9±0.7
A17	<5.0	4.1±0.5
Lead-dependent ribozyme (LDZ)		
A4	≤3.1	3.9±0.8
A8	4.3±0.3	4.7±0.5
A12	≤3.1	4.0±0.3
A16	3.8±0.4	4.3±0.7
A17	3.8±0.4	3.8±0.7
A18	3.5±0.6	4.1±0.3
A25	6.5±0.1	5.7±0.5

Table 3S. Topology information and parameters used to calculate pKa values for nucleic acids

Residue Type	Reference pKa	Force Field	Gaussian Variance	Reference Dielectric	External Dielectric
Adenosine	3.80	AMBER	0.7	8.0	80.0
Cytidine	4.35				

Topology of nucleic acid:

```
#  res atom obtal conf batm batm batm batm
$   A   P      sp3       SD   OP1  OP2  O3'  O5'
$   A   OP1  s           SD   P
$   A   OP2  s           SD   P
$   A   O5'   sp3       SD   P    C5'
$   A   C5'   sp3       SD   O5'  C4'  H5'  H5"
$   A   C4'   sp3       SD   C5'  O4'  C3'  H4'
$   A   O4'   sp3       SD   C4'  C1'
$   A   C3'   sp3       SD   C4'  O3'  C2'  H3'
$   A   O3'   sp3       SD   C3'  P
$   A   C2'   sp3       SD   C3'  O2'  C1'  H2'
$   A   O2'   sp3       SD   C2' HO2'
$   A   C1'   sp3       SD   O4'  C2'  N9   H1'
$   A   N9    sp2       SD   C1'  C8   C4
$   A   C8    sp2       SD   N9   N7   H8
$   A   N7    sp2       SD   C8   C5
$   A   C5    sp2       SD   N7   C6   C4
$   A   C6    sp2       SD   C5   N6   N1
$   A   N6    sp2       SD   C6   H61  H62
$   A   N1    sp2       SD   C6   C2   H1
$   A   C2    sp2       SD   N1   N3   H2
$   A   N3    sp2       SD   C2   C4
$   A   C4    sp2       SD   N9   C5   N3
$   A   H5'   s           SD   C5'
$   A   H5"   s           SD   C5'
$   A   H4'   s           SD   C4'
$   A   H3'   s           SD   C3'
$   A   H2'   s           SD   C2'
$   A  HO2'  s           SD   O2'
$   A   H1'   s           SD   C1'
$   A   H8    s           SD   C8
$   A   H61   s           SD   N6
$   A   H62   s           SD   N6
$   A   H2    s           SD   C2
$   A   H1    s           SD   N1


#  res atom obtal conf batm batm batm batm
$   C   P    sp3   SD   O5'  OP1  OP2  O3'
$   C   OP1  s     SD   P
$   C   OP2  s     SD   P
$   C   C5'  sp3   SD   C4'  O5'  H5'  H5"
```

```
$  C   O5'  sp3   SD   C5'  P
$  C   C4'  sp3   SD   C3'  O4'  C5'  H4'
$  C   O4'  sp3   SD   C1'  C4'
$  C   C3'  sp3   SD   C2'  C4'  O3'  H3'
$  C   O3'  sp3   SD   C3'  P
$  C   C2'  sp3   SD   C1'  O2'  C3'  H2'
$  C   O2'  sp3   SD   C2' HO2'
$  C   C1'  sp3   SD   N1   C2'  O4'  H1'
$  C   N1   sp2   SD   C2   C6   C1'
$  C   C2   sp2   SD   N1   N3   O2
$  C   N3   sp2   SD   C2   C4   H3
$  C   C4   sp2   SD   N3   C5   N4
$  C   C5   sp2   SD   C4   C6   H5
$  C   C6   sp2   SD   N1   C5   H6
$  C   O2   s     SD   C2
$  C   N4   sp2   SD   C4   H41  H42
$  C   H1'  s     SD   C1'
$  C   H2'  s     SD   C2'
$  C   H3'  s     SD   C3'
$  C   H4'  s     SD   C4'
$  C  HO2'  s     SD   O2'
$  C   H5'  s     SD   C5'
$  C  H5"  s     SD   C5'
$  C   H41  s     SD   N4
$  C   H42  s     SD   N4
$  C   H3   s     SD   N3
$  C   H5   s     SD   C5
$  C   H6   s     SD   C6


#  res atom obtal conf batm batm batm batm
$  G   P    sp3   SD   OP1  OP2  O3'  O5'
$  G   OP1  s     SD   P
$  G   OP2  s     SD   P
$  G   O5'  sp3   SD   P    C5'
$  G   C5'  sp3   SD   O5'  C4'  H5'  H5"
$  G   C4'  sp3   SD   C5'  O4'  C3'  H4'
$  G   O4'  sp3   SD   C4'  C1'
$  G   C3'  sp3   SD   C4'  O3'  C2'  H3'
$  G   O3'  sp3   SD   C3'  P
$  G   C2'  sp3   SD   C3'  O2'  C1'  H2'
$  G   O2'  sp3   SD   C2' HO2'
$  G   C1'  sp3   SD   O4'  C2'  N9   H1'
$  G   N9   sp2   SD   C1'  C8   C4
$  G   C8   sp2   SD   N9   N7   H8
$  G   N7   sp2   SD   C8   C5
$  G   C5   sp2   SD   N7   C6   C4
$  G   C6   sp2   SD   C5   O6   N1
$  G   O6   s     SD   C6
$  G   N1   sp2   SD   C6   C2   H1
$  G   C2   sp2   SD   N1   N3   N3
$  G   N2   sp2   SD   C2   H21  H22
```

```
$  G   N3   sp2   SD   C2   C4
$  G   C4   sp2   SD   N9   C5   N3
$  G   H1'  s     SD   C1'
$  G   H2'  s     SD   C2'
$  G   H3'  s     SD   C3'
$  G   H4'  s     SD   C4'
$  G  HO2'  s     SD   O2'
$  G   H5'  s     SD   C5'
$  G  H5"  s     SD   C5'
$  G   H8   s     SD   C8
$  G   H1   s     SD   N1
$  G   H21  s     SD   N2
$  G   H22  s     SD   N2


#  res atom obtal conf batm batm batm batm
$  U   P    sp3   SD   O5'  OP1  OP2  O3'
$  U   OP1  s     SD   P
$  U   OP2  s     SD   P
$  U   O5'  sp3   SD   C5'  P
$  U   C5'  sp3   SD   C4'  O5'  H5'  H5"
$  U   C4'  sp3   SD   C3'  O4'  C5'  H4'
$  U   O4'  sp3   SD   C1'  C4'
$  U   C3'  sp3   SD   C2'  C4'  O3'  H3'
$  U   O3'  sp3   SD   C3'  P
$  U   C2'  sp3   SD   C1'  C3'  O2'  H2'
$  U   O2'  sp3   SD   C2' HO2'
$  U   C1'  sp3   SD   N1   C2'  O4'  H1'
$  U   N1   sp2   SD   C2   C6   C1'
$  U   C2   sp2   SD   N1   N3   O2
$  U   N3   sp2   SD   C2   C4   H3
$  U   C4   sp2   SD   N3   C5   O4
$  U   C5   sp2   SD   C4   C6   H5
$  U   C6   sp2   SD   N1   C5   H6
$  U   O2   s     SD   C2
$  U   O4   s     SD   C4
$  U   H1'  s     SD   C1'
$  U   H2'  s     SD   C2'
$  U  HO2'  s     SD   O2'
$  U   H3'  s     SD   C3'
$  U   H4'  s     SD   C4'
$  U   H5'  s     SD   C5'
$  U  H5"  s     SD   C5'
$  U   H3   s     SD   N3
$  U   H5   s     SD   C5
$  U   H6   s     SD   C6
```