Computational and Structural Biotechnology Journal 21 (2023) 1383–1389

Contents lists available at ScienceDirect

Computational and Structural Biotechnology Journal

journal homepage: www.elsevier.com/locate/csbj

Software/web server article

HIT-2: Implementing machine learning algorithms to treat bound ions in biomolecules

Shengjie Sun [a], Honglun Xu [a], Yixin Xie [b], Jason E. Sanchez [a], Wenhan Guo [a], Dongfang Liu [c], Lin Li [a,d,*]

[a] Computational Science Program, The University of Texas at El Paso, 500 W University Ave, TX 79968, USA
[b] Department of Information Technology, College of Computing and Software Engineering, Kennesaw State University, 1000 Chastain Rd NW, Kennesaw, GA 30144, USA
[c] Department of Computer Engineering, Rochester Institute of Technology, Rochester, NY 14623, USA
[d] Department of Physics, the University of Texas at El Paso, 500 W University Ave, TX 79968, USA

ARTICLE INFO

Article history:
Received 28 October 2022
Received in revised form 6 February 2023
Accepted 6 February 2023
Available online 8 February 2023

Keywords:
Bound ions
Explicit solvent model
Implicit solvent model
Electrostatic calculation
Delphi

ABSTRACT

Electrostatic features are fundamental to protein functions and protein-protein interactions. Studying highly charged biomolecules is challenging given the heterogeneous distribution of the ionic cloud around such biomolecules. Here we report a new computational method, Hybridizing Ions Treatment-2 (HIT-2), which is used to model biomolecule-bound ions using the implicit solvation model. By modeling ions, HIT-2 allows the user to calculate important electrostatic features of the biomolecules. HIT-2 applies an efficient algorithm to calculate the position of bound ions from molecular dynamics simulations. Modeling parameters were optimized by machine learning methods from thousands of datasets. The optimized parameters produced results with errors lower than 0.2 Å. The testing results on bound Ca^{2+} and Zn^{2+} in NAMD simulations also proved that HIT-2 can effectively identify bound ion types, numbers, and positions. Also, multiple tests performed on HIT-2 suggest the method can handle biomolecules that undergo remarkable conformational changes. HIT-2 can significantly improve electrostatic calculations for many problems in computational biophysics.

1. Introduction

Electrostatic features are essential for the proper functioning of biomolecules because the electrostatic character of the molecule plays significant roles in biomolecular binding/repelling [1–3], ion transport [4], and structural stability [5]. Nevertheless, electrostatic calculation *in silico* is a challenging topic in computational biophysics due to the presence of ions. Ions contribute to the complexity of the environment around biomolecules. Highly charged biomolecules, such as nucleic acids, globular proteins, and motor proteins tend to attract ions with opposite charges to balance the net charge in a local environment. The bound ions significantly affect electrostatic potential on biomolecular surfaces, which further influences the interactions of this molecule with other molecules.

At present, there are two models that can handle ions and water surrounding biomolecules: the explicit solvent model and the implicit solvent model. The explicit model is widely applied in all-atom molecular dynamics (MD) simulations when the goal is to simulate ions and H_2O (TIP3P [6], TIP4P [7]) explicitly—that is, with coordinates for each water and ion atom. The explicit model neutralizes highly charged biomolecules by adding unbalanced amounts of cations and anions in the modeled system. However, the electrostatic calculations by the explicit model may consider billions of atoms in hundreds of frames (each frame including the coordinates of the protein, water, and ions). This approach is extremely memory-intensive and time-consuming. Therefore, the implicit solvent model seems to be more suitable to handle ions and water surrounding biomolecules. Implicit solvent models include Poisson-Boltzmann (PB) model [8] and Generalized Born (GB) model [9]. In both models, the ionic environment of biomolecules is treated homogeneously by setting dielectric constants for biomolecules and solutions. The implicit method avoids the energy calculations for ions and water in solutions [10,11]. Compared with the explicit method, the implicit

* Corresponding author at: Computational Science Program, The University of Texas at El Paso, 500 W University Ave, TX 79968, USA.
E-mail address: lli5@utep.edu (L. Li).

https://doi.org/10.1016/j.csbj.2023.02.013

model has the obvious advantage of accelerating electrostatic calculations on biomolecules. However, the homogeneous treatment of the solvent ignores local effects of bound ions. The homogenous solvent is not realistic in many situations—especially for highly charged biomolecules. In some cases, even though the net charge of a biomolecule is neutral, the charge distribution on the surface of the biomolecule is not. In detail, the interfaces between biomolecules are often highly charged, for binding or repelling other molecules. Those charged areas in biomolecules may cause nonhomogeneous ionic distribution. Overall, the homogeneous treatment of solutions has the limitation of ignoring this feature of highly charged biomolecules.

To solve the above difficulties when treating highly charged biomolecules, we proposed a Hybridizing Ions Treatment (HIT) method [12,13] for representing ions in implicit solvation calculations. HIT method adds explicitly bound ions into the implicit solvent model for the electrostatic calculation. The core idea behind this method is using the frames of MD simulations to calculate the position of bound ions via clustering. Compared with prediction methods by coordination numbers [14], geometries [15], or electrostatic potentials [11,16], HIT calculations are more reliable. This is because HIT uses the results from all-atom MD simulations to do calculations. Most popular MD programs, such as NAMD [17] and GROMACS [18], take the majority of intermolecular forces acting on a system, including Van der Waals (VDW) forces and electrostatic forces, into consideration. Nevertheless, the HIT method has difficulties handling biomolecules with significant movement. It also has trouble treating multi-component systems simultaneously. Overall, HIT worked well in previous test cases, but sometimes calculations were computationally expensive and program input was unfriendly for users.

Here we report HIT-2, a new version of the Hybridizing Ions Treatment method to combine bound ions and implicit solvent solutions for accurate electrostatic calculations. This method does not require the setting of a cube size or the type and number of bound ions, unlike the first version of HIT [12]. We observed two issues when we developed HIT at first stage: redundancy and incorrect calculation [12]. Redundancy means two or more calculated ion positions are close to one real position of bound ions while an incorrect calculation means the calculated position is far away (over 5 Å) from the real position. Redundancy occurs because HIT cuts a binding site into several pieces equally or approximately equally, and then the two or more segments, which are treated as multiple binding sites. Because HIT only selected bound ions according to ranking of occupancy, redundancy further causes the last correct results to be abandoned. This results in incorrect calculations. In HIT-2, we applied an iteration technique to handle redundancy and incorrect calculations. Iteration ensures a binding site is fully encompassed within a cube by increasing the cube size, so eliminating redundancy and subsequent incorrect calculations. Furthermore, we defined several independent parameters and prepared tools to train and test HIT-2 by machine learning methods.

Several machine learning methods were applied to optimize HIT-2 parameters. We developed a Random Ions Generation Tool (RIGT) to generate 3888 cases for parameter optimization via machine learning. To ensure the accuracy of ion distributions in RIGT, the Maxwell Boltzmann distribution was applied to simulate ionic distributions. After classification and optimization, the best parameters were found and applied to test real datasets (real explicit MD simulations). We tested HIT-2 in proteins with bound Ca^{2+}/Zn^{2+} to further validate the accuracy of the bound ions' positions. Also, we applied HIT-2 on proteins and nucleic acids to validate the broad applicability. Moreover, we improved HIT-2 so that it is also able to handle biomolecules with significant conformational changes. Lastly, we applied HIT-2 to predict the binding position for signal ions (Ca^{2+}) in troponins to exhibit the wide applicability of this method. The results showed it is a very promising tool in computational biophysics and related fields.

2. Methods

2.1. Dataset

2.1.1. Random ions training sets

A water box with ions simulation was prepared by NAMD [17] to observe the velocity distribution of ions. The velocity of ions in the x, y, and z directions was calculated in pm/ns (Å/0.01 ns) and fitted to a Maxwell Boltzmann distribution [19]. The Maxwell Boltzmann distribution was first defined and used for describing particle speeds in idealized gases. Here we applied it to fit ions moving in solvation boxes.

$$v = \sqrt{v_x^2 + v_y^2 + v_z^2} \tag{1}$$

Where the v is the ion's speed and v_x, v_y, and v_z is the speed component in the x, y, and z directions. The v conforms to a Chi-squared distribution while the v_x, v_y, and v_z conform to a normal distribution. Here we took the unit of pm/ns as the ion's speed in simulation. The result was applied for ion generation in the Random Ions Generation Tool (RIGT).

RIGT is the tool designed to generate abundant ionic simulations for HIT-2 testing. It can quickly generate a series of trajectories of ions' simulation for a certain time. The ions trajectories were further combined by RIGT to generate an ionic cloud file (In PDB format). Here in our experiment, the solution was 150 mM NaCl. 10 Na^+ and 10 Cl^- were trapped in a sphere with random diameters of 0–5 Å. Maxwell Boltzmann distribution was applied to the velocity of ions to better simulate each frame of the ionic cloud. Here, the interval is 0.01 ns for saving frames of the ionic cloud. 54 datasets were generated with random simulation times (0.05–40 ns).

2.1.2. Testing sets in MD simulations

A random ions testing set is solvated by a 150 Å × 150 Å × 150 Å solvation box with 150 mM NaCl with 10 K^+ and 10 Cl^- restrained to simulate the bound ions. Cysteine dioxygenase (PDB: 5L0S) [20] and Factor VIIa with bound Ca^{2+} (PDB: 5PB2) [21] were used for Ca^{2+} testing while insulin with bound Zn^{2+} (PDB: 1ZEH) [22] was used for Zn^{2+} testing. Moreover, DNA(PDB: 5J2M) [23] and RNA(PDB: 4TNA) [24] with bound ions are were also tested by HIT-2. The membrane protein BamA (PDB: 4K3B) [25] was simulated to include biomolecules which exhibit marked conformational changes. Additionally, we also conducted actin filament simulations to further explore the potential ability of HIT-2 for the search of signal ions and corresponding binding sites. A piece of actin filament with troponins (PDB: 6KN8) [26] was chosen as the model for simulations. The missing loops in these structures were made up by Swiss-model [27]. The solvation was achieved by VMD [28]/CHARMM-GUI [29] using TIP3P [6]. The membrane of BamA was constructed by the CHAMRMM-GUI membrane builder [30]. MD Simulations were performed on NAMD 2.12 [17]. The electrostatic potential was calculated by Delphi [10] using the CHARMM36m [31] forcefield while charges were assigned by pdb2pqr [32]. More details about the MD simulations and electrostatic calculations are included in the supporting information.

2.2. Algorithm

2.2.1. Preparation, solvation box cutting, and ion counting

After MD simulations, the relative positions of ions are calculated using the mass center of biomolecules in the corresponding frame. Then ions in all frames are aligned together by their relative positions, forming an ionic cloud (Fig. 1: a and b). The cube size (the side

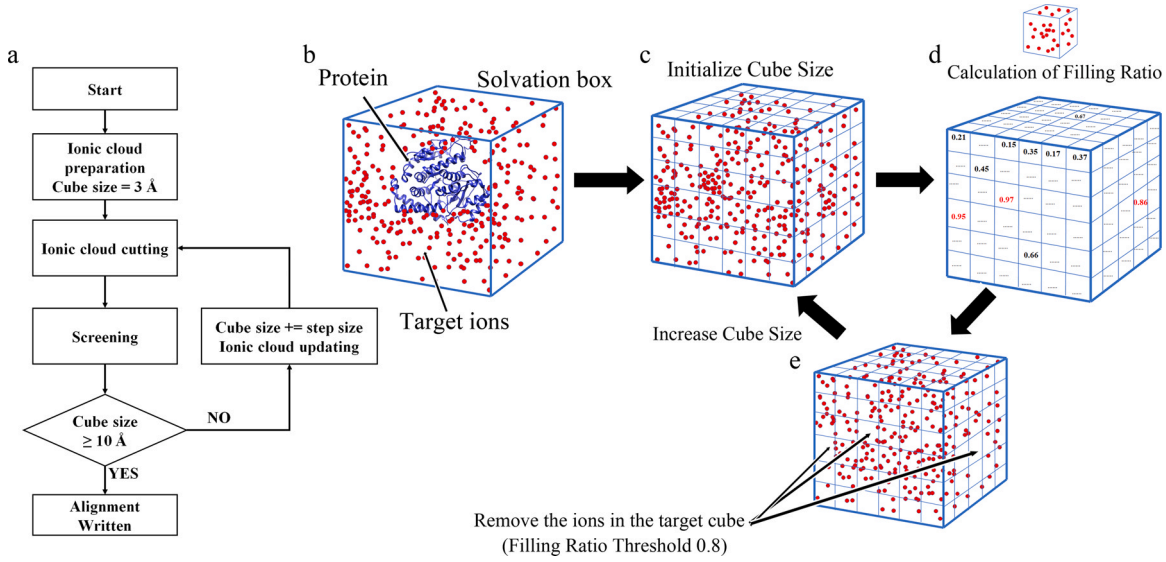

Fig. 1. The flow chart and schematic graph of the algorithm. a. The flow chart of HIT-2. b. The preparation of the ionic cloud. C. The ionic cloud cutting partition step. D. The calculation of the filling ratio for each cube. E. Removing ions from the ionic cloud after binding site calculations and then increasing the cube size for the next iteration. In this diagram, the filling ratio threshold is set as 0.8. In e, the cubes with filling ratio higher 0.8 are regarded as binding sites and all ions in corresponding cubes will be removed from ionic cloud before the next iteration. The iterations will run until the cube size is bigger than 10 Å.

length of each cube) is initialized to 3 Å. Based on the initialized cube size, the ionic cloud is cut into several cubes (Fig. 1: c). The ions in each cube are counted to calculate the filling ratio (Eq. 1 and Fig. 1: d).

$$R_f = \frac{n_{ic}}{n_f} \tag{2}$$

Where the R_f represents the filling ratio, n_{ic} represents the number of ions in the corresponding cube, and the n_f represents the number of frames.

2.2.2. Screening and ionic cloud updating

If the filling ratio is higher than a given threshold (filling ratio threshold), the corresponding cube is selected as a binding site, where the mass center is calculated by Eq. 2, representing the position of the bound ion (Eq. 1). The ions in the selected cubes are then removed from the ionic cloud (Fig. 1e). Then the cube size increases by a given step size (Å). The step size means the increment of the cube size in iterations. The updated ionic cloud and cube size are further used in the iterations until the cube size is bigger than 10 Å. The filling ratio threshold and step size are important parameters to be optimized by machine learning methods.

$$\vec{P} = \frac{\Sigma \vec{v_i}}{n_{ic}} \tag{3}$$

Where the $\vec{P}$ represents the position of bound ions and the $\vec{v_i}$ represents the positions of all ions in the selected cubes.

2.2.3. Structural alignment

After iterations of screening and ionic cloud updating, the positions of bound ions are logged and aligned to the biomolecular structure in the target frame by Eq. 3. The output is the biomolecule with bound ions in PDB format in the target frame (n^{th} frame decided by users).

$$\vec{P_o} = \vec{P_b} + \vec{P} \tag{4}$$

Where the $\vec{P_o}$ and $\vec{P_b}$ represent the output positions of bound ions and the mass center of the biomolecules.

2.3. Algorithm testing

Based on the RIGT, 54 simulation results (0.05–40 ns) are prepared for testing. In each dataset, HIT-2 applied different step sizes (0.05–1 Å) and filling ratio thresholds (0.5–0.99) for testing. In total, 3888 cases were generated including the results generated by HIT-2. We define success and failure in the section below and for all successful cases, the error was calculated for further optimization.

2.4. Classification and optimization

First, we defined strict criteria for success to describe the results from HIT-2. These are two conditions: 1, the number of calculated bound cations/anions should be the same as the real bound cation/anions; 2, the distance between each pair (calculated bound ion and real bound ion) should be smaller than 5 Å. If these conditions are not met, the result is a failure. To find what parameters (simulation time, step size, and filling ratio thresholds) lead to success, we applied Logistic Regression (LR) [33], Classification and Regression Tree (CART) [34], Random Forest (RF) [35], and Artificial Neural Networks (ANN) [36] to address the binary classification problem (success/failure). methods. The number of cases is 3888 with a train/test split of 0.7/0.3. For all successful cases with simulation time over 2 ns, we further tested the distance between our calculations and targets. The average distance of all pairs was measured and regarded as an average error for further analysis. The contributions made by simulation time, step size, and filling ratio thresholds to average error are analyzed. Additionally, the run time of HIT-2 is highly related to step size and simulation time so we also consider these criteria when determining optimal parameters. The run times of HIT-2 for different cases are also measured and compared with the average error for step size and simulation time optimization.

After classifications, quantitative analysis was applied to understand the relationship between error and related parameters. In successful cases, the number and type of calculated bound ions are same as that of real bound ions. Each cases includes 20 pairs of real bound ions and calculated bound ions. We consider the distance between the real and calculated ions as the error and calculate this value for the 20 pairs. The average error was plotted against simulation time, filling ratio and step size for quantitative analysis. Moreover, we also calculated the running time of HIT-2 to further

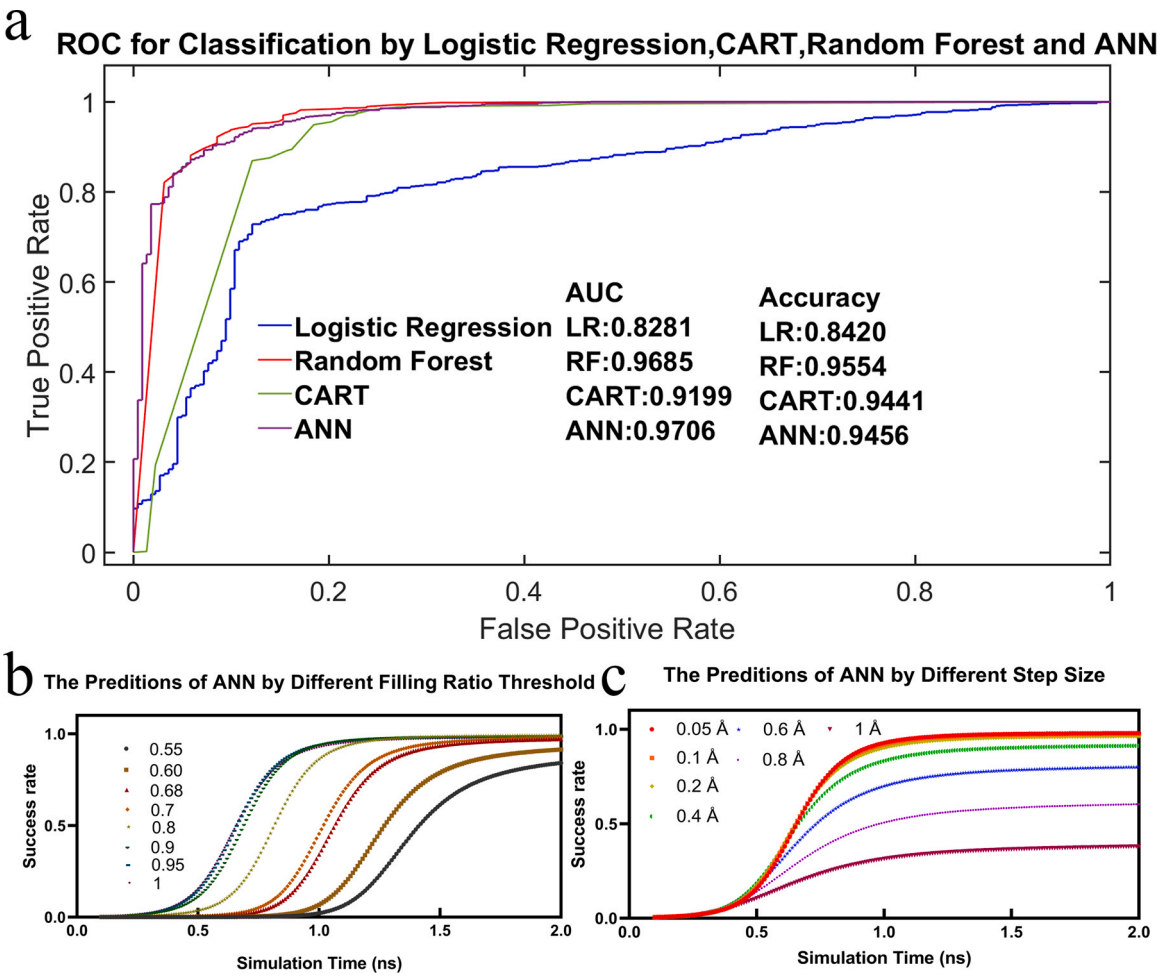

Fig. 2. The comparison among several machine learning methods and the results generated by ANN. a. The ROC results of Logistic Regression, Random Forest, CART, and ANN. b. The success rate predicted by ANN with different filling ratio thresholds (step size = 0.05). c. The success rate predicted by ANN with different step sizes (The filling ratio is 1.00).

optimize the parameters for users. The PC used for this optimization was a Dell-XPS with Intel I7 processor (i7–11700–2.5 GHz) and 16 GB memory.

3. Results

3.1. The workflow of HIT-2

HIT-2 involves 4 steps: preparation, ionic cloud cutting, screening, and position alignment (Fig. 1a). In the preparation step, the relative positions of all ions in all frames from MD simulations are calculated and combined into an ionic cloud (Fig. 1b). In the ionic cloud cutting step, the ionic cloud is cut into several cubes (Fig. 1c). Afterward, the ions are counted and divided by the number of frames to calculate the filling ratio of each cube (Fig. 1d). If the filling ratios are larger than the threshold (filling ratio threshold), those cubes are regarded as binding sites. The mass centers of the binding area are the positions of bound ions (Fig. 1e). The ions in the binding sites are removed from the ionic cloud before the next iteration. In the next iteration, the size of the cube is increased by the step size. In the alignment step, the positions of bound ions are calculated by the relative positions of bound ions and the mass center of biomolecules (Fig. 1a). The simulation time, the filling ratio threshold, and the step size are three crucial parameters affecting the results of HIT-2. We further optimized these parameters by employing several classification machine learning methods with different parameters.

3.2. Classification

HIT-2 requires certain conditions to find bound ions from simulations. The three parameters, including filling ratio, step size, and simulation times are set as inputs, and the success/failure is set as an output. Several machine learning methods were applied to obtain a better model to search the parameter space for successful results. Here, we define two very strict criteria for success: 1, the number of calculated bound cations/anions must be the same as the number of real bound cation/anions; 2, the distance between each pair (a calculated bound ion and the corresponding real bound ion) must be shorter than 5 Å.

Fig. 2a illustrates the ROC of Logistic Regression (LR) [33], Classification and Regression Tree (CART) [34,37], Random Forest (RF) [35], and Artificial Neural Network (ANN) [36]. LR has an accuracy of 84% with 0.83 Area Under the Curve (AUC) while the RF, CART, and ANN have high accuracy of around 95%. Among these methods, the ANN possesses the highest AUC (0.97) and the highest accuracy (95%). We further used the ANN model to predict the result from different input parameters (simulation time, filling ratio threshold, and step size). Millions of cases were generated by the ANN model to show the relationship between different parameters and results. Fig. 2b shows the predicted success rate by different filling ratio thresholds and simulation times. After the 5 ns simulation, the success rate for different filling ratio thresholds (≥0.55) is nearly 100% (Fig. S2). The success rate is highly related to the simulation time. This is because ideal ionic clouds are formed when simulations

reach equilibrium. The longer simulations produce more stable equilibria. Among different filling ratio thresholds, the thresholds of 0.95 and 1.00 first reached a 100% success rate after 1 ns. Fig. 2c and S3 show the success rate predicted by ANN with different step sizes and simulation times. The step size is the increment of cube size in iterations. Intuitively, a smaller step size should lead to more accurate results. The result shows that when the step size is increased to 1 Å, only 50% of calculations were correct (even when the simulation time is greater than 20 ns). With the decrease in step size, a higher success rate is achieved. When the step size is lower than 0.2 Å, there is no significant difference. Indeed, all simulations with a step size lower than 0.2 Å reached a near 100% success rate after 1 ns. In summary, for most simulations a filling ratio of 0.95 and step size of 0.2 Å is enough to get a nearly 100% success rate after running for 1 ns.

3.3. Optimization

Due to the very strict criteria for success, the condition of success is enough for HIT-2 to get reliable results. However, we are interested in improving HIT-2 performance as much as possible and we were not satisfied with "just success". The error (distance between calculated bound ions and real bound ions) should be further quantified to optimize the parameters for users. The average error shown in Fig. 4 demonstrates how HIT-2 may still be improved. The relationship between the filling ratio threshold and average error is shown in Fig. 4a. The average error is decreased when filling ratio threshold increases. Average error reaches a minimum when the filling ratio threshold is 0.95. When the filling ratio is increased to 0.99, error slightly increases. In the generated datasets, there is no escape or rebinding of ions in the binding sites. In this case, the ideal situation (equilibrium system) will cause the lowest error to appear at the point of filling ratio equal to 100%. However, in most cases, an equilibrium simulation cannot be practically achieved because such simulations are time-consuming. In our error analysis, the lowest value appeared when the filling ratio equaled 0.95. The filling ratio is approximately equal to the occurrence frequency of bound ions. In nature, bound ions sometimes are too mobile to be bound tightly, resulting in a frequency of occurrence lower than 100%. In this case, researchers can choose a better filling ratio for their studies. Additionally, testing with a range of filling ratio thresholds is also a good idea to distinguish the strength of different bound ions. With that said, we remark that the filling ratio should always be higher than 0.5; otherwise, a binding site would be treated as two binding sites and this would likely cause incorrect calculations [12].

Step size and simulation time are highly related to the accuracy and the running time of HIT-2. With a decrease in step size, the average error linearly decreased until 0.01 Å at the point of step size = 0.2 Å. When the step size is lower than 0.2 Å, the average error begins to increase (Fig. 3b). This is because the screening step (Algorithm section) may count several free ions or miss several bound ions from the ionic cloud. A small step size (< 0.2 Å) leads to more screening iterations and may cover more unrelated ions into consideration, causing the bias. This error can be reduced by performing long MD simulations. Moreover, a smaller step size also requires calculations be performed for many small cubes which significantly increases the computational cost of HIT-2. From Fig. 3b., we see that when step size is lower than 0.2 Å, the run time increase from 3 min to a maximum of 10 min (Fig. 3b). Therefore, the 0.2 Å step size is the best choice for most users. By contrast, after 5 ns, longer simulation contributes less to error minimization. The average error decreases with an increase of simulation time (Fig. 3c). The simulation time and the HIT-2 run time are linearly related because simulation time is equivalent to the number of frames, which affects the preparation of the ionic cloud (Algorithm section). Compared to the filling ratio and step sizes, the simulation time is not that highly related to error minimization. However, simulation time is highly related to the success rate, which is fundamental to error minimization. So at least 5 ns simulations with a frequency of less than 10,000 fs/frame are suggested to get accurate results.

The heat map showing average error against simulation time and filling ratio is shown in Fig. 3d. Simulations longer than 30 ns have an average error lower than 0.02 Å. The average error decreases dramatically with when the filling ratio increases because the ions in the dataset are 100% bound. Overall, the highest average error is 0.03 Å, which means the average error for any pair is lower than 0.6 Å (0.03 Å x 20 = 0.6 Å, all error happened at a bound ion) if the calculation succeeds. Also, we observe that the presence of red peaks diminishes with an increase in simulation time. Additionally, at higher simulation times, the error decreases more smoothly with an increase in the filling ratio. Thus, longer simulations can reduce the standard deviations for average errors, making the results more reliable.

3.4. Testing and application

In the random ions simulations from NAMD, 10 Na^+ and 10 Cl^- are restrained to simulate bound ions. The calculations by HIT-2 tested the ionic cloud from 0.5, 1, 2, 3, 4, 5, 6, 7, 8, 9, and 10 ns. The comparison between calculations and targets is shown in Fig. S4. Notably, from 0.5 ns onward, the calculations from HIT-2 are successful. We further analyzed the average errors and the associated standard deviations (Fig. 4a). We observe the errors and standard deviations decreased with the increase in simulation time. These results are consistent with the above analysis, further proving the accuracy of HIT-2.

In X-ray crystallography experiments, ions are co-crystalized with a macromolecule and are thus resolved as a part of the structure. For our tests, we chose the X-ray structure from the PDB bank (5L0S [20], 5PB2 [21], and 1ZEH [22]) to do accuracy testing for HIT-2 to determine the positions of Ca^{2+} and Zn^{2+}. The errors associated with our predictions for Ca^{2+} are 3.26 Å and 0.88 Å (Fig. 4bc) and those for Zn^{2+} are 0.18 Å and 1.26 Å (Fig. 8 A). All calculations are successful (error < 5 Å), but the error exceeds our expectations. The main reason is that the binding area may be bigger than the training sets, causing the bound ions to have a wider active region. In addition to this result, we observe HIT-2 also works for other biomolecules, including DNA/RNA (Fig. S5) and biomolecules with significant conformational changes (Fig. S6). In short, without inputting the number and name of bound ions, HIT-2 directly found the correct number and positions of the ions with an error lower than 4 Å for various in vitro systems.

We further colored the surface of 1ZEH by electrostatic potential (Fig. 4e-g), with and without bound ions. The binding sites without bound Zn^{2+} are negatively charged (Fig. 8e). *In vitro*, the bound Zn^{2+} imparted a positively charge to the surface of the protein (Fig. 4 f). Similarly, with HIT-2, the bound calculated Zn^{2+} also applied a positive charge (Fig. 4 g).

Some bound ions such as Ca^{2+} also play vital roles in intracellular signaling, such as muscle contractions. Here we simulate with a piece of actin-filament with a troponin complex (including troponin I, troponin C, and troponin T), globular-actins (G-actin), and tropomyosin. Three bound calcium ions (Fig. 4 h) were found by HIT-2 with two of these being located inside the G-actins and tropomyosin and the other attached to the troponin. These results are consistent with previous studies that show troponin is a receptor for calcium ions [38]. Moreover, the calcium is not only bound to troponin but also attached to the whole actin filament. The conformational changes of the actin filament could be affected by the coverage of calcium ions on the actin filament. It requires further deep analysis and experimental support.

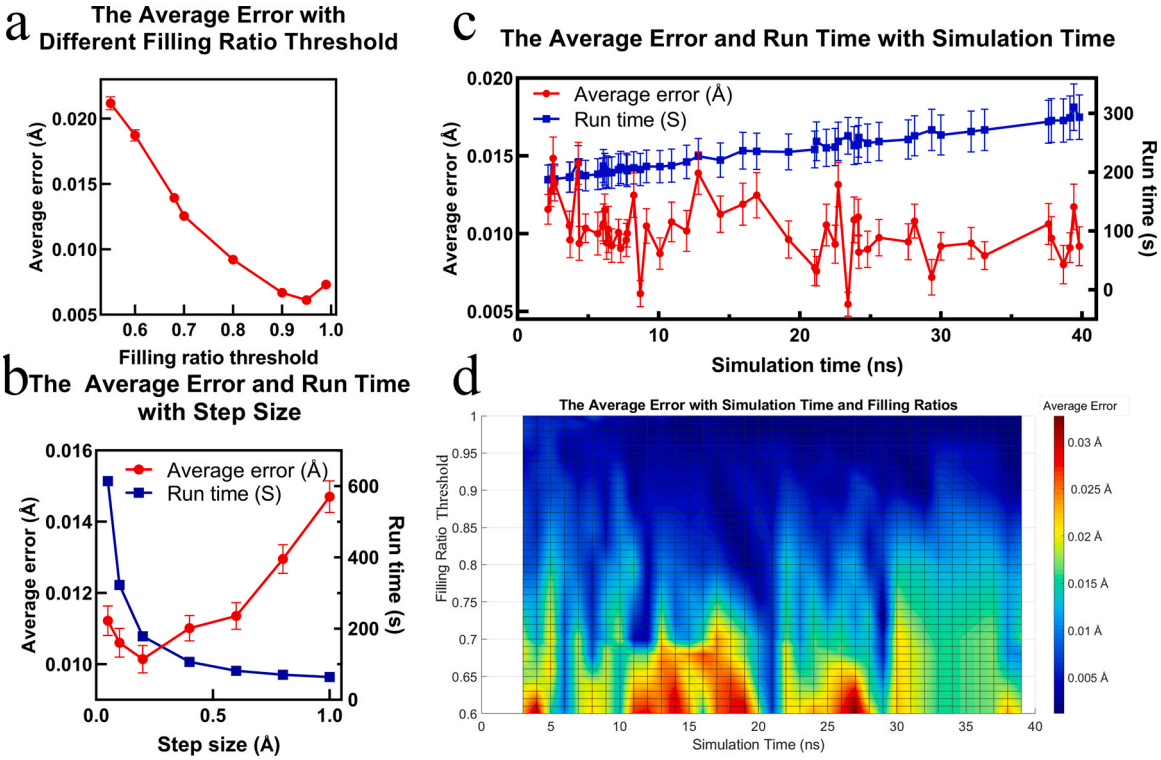

Fig. 3. The average error and running time plotted against filling ratio, step size, and simulation time. a. The average error against the filling ratio. b. The run time and average error against the step size (the filling ratio is 0.99). c. The run time and average error against simulation time (the step size is 0.2 Å). d. The heat map of average error with different filling ratio thresholds and simulation times (the step size is 0.05 Å).

4. Discussion

Bound ions are crucial for the function of highly charged biomolecules. The current implicit solvation model has many limitations when applied to highly charged biomolecules. To solve this problem, our previous work, HIT, was developed, and works by hybridizing bound ions with the implicit solvent model to improve electrostatic calculations.

Because the clustering algorithm in our previous version of HIT is computationally expensive, we sought to make improvements. Here,

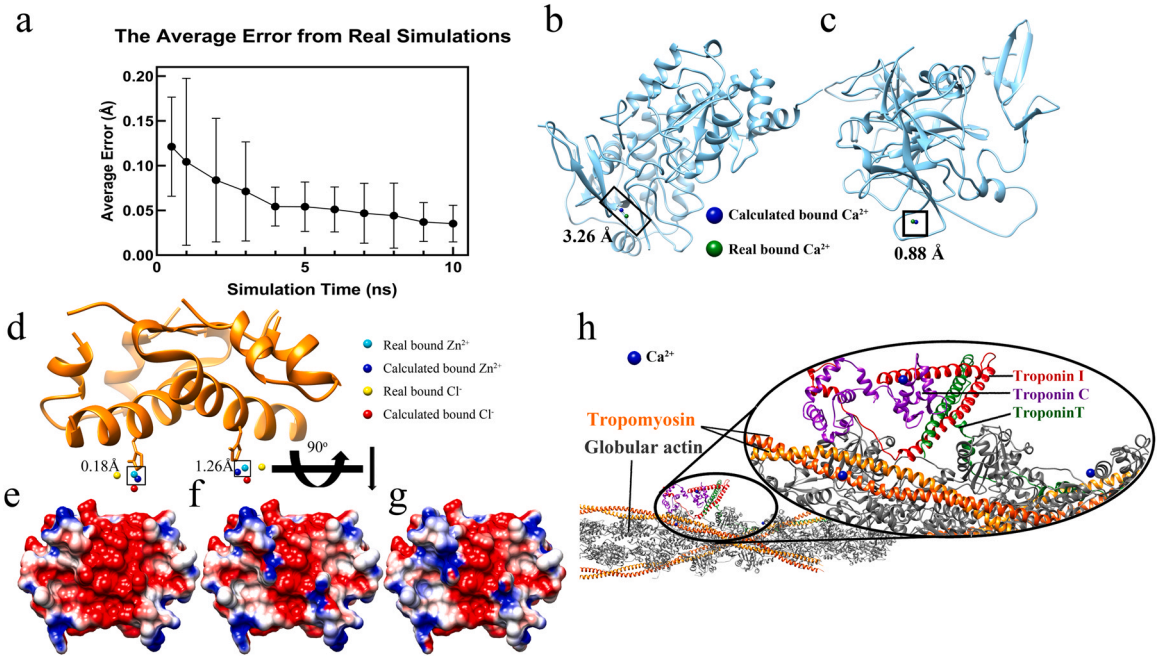

Fig. 4. Tests and applications of HIT-2. a. The results from the NAMD testing set (average error is the average distance among 20 pairs of calculated bound cations/anions). b. The calculated and real bound Ca^{2+} in protein (PDB: 5L0S). c. The calculated and real bound Ca^{2+} in protein (PDB: 5PB2). d. The results from bound Zn^{2+} testing for the protein (PDB: 1ZEH) e. The electrostatic surface without bound ions. f. The electrostatic surface with real bound ions. g. The electrostatic surface with calculated bound ions. h. The structure of the actin filament with bound calcium ions was produced by HIT-2.

we abandoned the clustering method in HIT and employed machine learning approaches in HIT-2. Also, the newest HIT-2 tool is more user-friendly by eliminating the need to specify cube size and the type and number of bound ions.

To validate our work, we tested HIT-2 on DNA, RNA, and membrane-bound proteins. Also, we showed HIT-2 can be applied to biomolecules that undergo significant conformational changes, and biomolecules with multiple components. Expectedly, our work shows HIT-2 can also be used for predicting the location of signal ions and associated receptors. Specifically, we demonstrated HIT-2 can locate the position of Ca^{2+} in a troponin complex.

In conclusion, HIT-2 is useful for dealing with difficult biological problems that cannot be resolved by experimentation. However, there are some limitations of HIT-2. First, it performs very well in binary systems with one type of cation and one type of anion, but when multiple types of cations and anions are involved, the problem would require multiple uses of HIT-2. Second, while HIT-2 is able to handle the movement of biomolecules in simulations, it cannot handle rotations. The rotated biomolecules should be aligned first in VMD or other software before the use of HIT-2. In spite of these problems, HIT-2 is a useful tool for users who are studying highly charged biomolecular systems.

Data availability

The data that support the findings of this study are openly available in http://compbio.utep.edu/software/

Conflicts of Interest Statement

The authors whose names are listed immediately below certify that they have NO affiliations with or involvement in any, organization or entity with any financial interest (such as honoraria; educational grants; participation in speakers' bureaus, membership, employment, consultancies, stock ownership, or other equity interest; and expert testimony or patent-licensing, arrangements), or non-financial interest (such as personal or professional relationships, affiliations, knowledge or beliefs) in, the subject matter or materials discussed in this manuscript.

Acknowledgments

This work was supported by the National Institutes of Health under Grant No. SC1GM132043 and by the National Institute on Minority Health and Health Disparities under Grant No. 5U54MD007592, a component of the NIH. The calculations and analyses were performed at the Texas Advanced Computing Center.

Appendix A. Supporting information

Supplementary data associated with this article can be found in the online version at doi:10.1016/j.csbj.2023.02.013.

References

[1] Xie Y, Karki CB, Du D, Li H, Wang J, Sobitan A, Teng S, Tang Q, Li L. Spike proteins of SARS-CoV and SARS-CoV-2 utilize different mechanisms to bind with human ACE2. Front Mol Biosci 2020:392.
[2] Xie Y, Du D, Karki CB, Guo W, Lopez-Hernandez AE, Sun S, Juarez BY, Li H, Wang J, Li L. Revealing the mechanism of SARS-CoV-2 spike protein binding with ACE2. Comput Sci Eng 2020;22:21–9.
[3] Karki C, Xian Y, Xie Y, Sun S, Lopez-Hernandez AE, Juarez B, Wang J, Sun J, Li L. A computational model of ESAT-6 complex in membrane. J Theor Comput Chem 2020;19:2040002.
[4] Sun S, Karki C, Aguilera J, Lopez Hernandez AE, Sun J, Li L. Computational study on the function of palmitoylation on the envelope protein in SARS-CoV-2. J Chem Theory Comput 2021;17:6483–90.
[5] Matousek WM, Ciani B, Fitch CA, Kammerer RA, Alexandrescu AT. Electrostatic contributions to the stability of the GCN4 leucine zipper structure. J Mol Biol 2007;374:206–19.
[6] Price DJ, Brooks III CL. A modified TIP3P water potential for simulation with Ewald summation. J Chem Phys 2004;121:10096–103.
[7] Abascal JL, Vega C. A general purpose model for the condensed phases of water: TIP4P/2005. J Chem Phys 2005;123:234505.
[8] Nicholls A, Honig B. A rapid finite difference algorithm, utilizing successive over-relaxation to solve the Poisson–Boltzmann equation. J Comput Chem 1991;12:435–45.
[9] Jayaram B, Sprous D, Beveridge D. Solvation free energy of biomacromolecules: parameters for a modified generalized Born model consistent with the AMBER force field. J Phys Chem B 1998;102:9571–6.
[10] Li L, Li C, Sarkar S, Zhang J, Witham S, Zhang Z, Wang L, Smith N, Petukh M, Alexov E. DelPhi: a comprehensive suite for DelPhi software and associated resources. BMC Biophys 5. 2012. p. 1–11.
[11] Petukh M, Zhenirovskyy M, Li C, Li L, Wang L, Alexov E. Predicting nonspecific ion binding using DelPhi. Biophys J 2012;102:2885–93.
[12] Sun S, Karki C, Xie Y, Xian Y, Guo W, Gao BZ, Li L. Hybrid method for representing ions in implicit solvation calculations. Comput Struct Biotechnol J 2021;19:801–11.
[13] Sun S, Lopez JA, Xie Y, Guo W, Liu D, Li L. HIT web server: a hybrid method to improve electrostatic calculations for biomolecules. Comput Struct Biotechnol J 2022;20:1580–3.
[14] Katz AK, Glusker JP, Beebe SA, Bock CW. Calcium ion coordination: a comparison with that of beryllium, magnesium, and zinc. J Am Chem Soc 1996;118:5752–63.
[15] Babor M, Gerzon S, Raveh B, Sobolev V, Edelman M. Prediction of transition metal-binding sites from apo protein structures. Proteins 2008;70:208–17.
[16] Shashikala HM, Chakravorty A, Panday SK, Alexov E. BION-2: predicting positions of non-specifically bound ions on protein surface by a Gaussian-based treatment of electrostatics. Int J Mol Sci 2020;22:272.
[17] Phillips JC, Braun R, Wang W, Gumbart J, Tajkhorshid E, Villa E, Chipot C, Skeel RD, Kale L, Schulten K. Scalable molecular dynamics with NAMD. J Comput Chem 2005;26:1781–802.
[18] Van Der Spoel D, Lindahl E, Hess B, Groenhof G, Mark AE, Berendsen HJ. GROMACS: fast, flexible, and free. J Comput Chem 2005;26:1701–18.
[19] Condon E. A simple derivation of the Maxwell-Boltzmann Law. Phys Rev 1938;54:937.
[20] Li Z, Fischer M, Satkunarajah M, Zhou D, Withers SG, Rini JM. Structural basis of Notch O-glucosylation and O–xylosylation by mammalian protein–O-glucosyltransferase 1 (POGLUT1). Nat Commun 2017;8:1–12.
[21] Snarski-Adamski J, Werwiński M. Effect of transition metal doping on magnetic hardness of CeFe12-based compounds. J Magn Magn Mater 2022;554:169309.
[22] Whittingham JL, Edwards DJ, Antson AA, Clarkson JM, Dodson GG. Interactions of phenol and m-cresol in the insulin hexamer, and their effect on the association properties of B28 Pro→ Asp insulin analogues. Biochemistry 1998;37:11516–23.
[23] Salie ZL, Kirby KA, Michailidis E, Marchand B, Singh K, Rohan LC, Kodama EN, Mitsuya H, Parniak MA, Sarafianos SG. Structural basis of HIV inhibition by translocation-defective RT inhibitor 4′-ethynyl-2-fluoro-2′-deoxyadenosine (EFdA). Proc Natl Acad Sci USA 2016;113:9274–9.
[24] Hingerty B, Brown R, Jack A. Further refinement of the structure of yeast tRNAPhe. J Mol Biol 1978;124:523–34.
[25] Noinaj N, Kuszak AJ, Gumbart JC, Lukacik P, Chang H, Easley NC, Lithgow T, Buchanan SK. Structural insight into the biogenesis of β-barrel membrane proteins. Nature 2013;501:385–90.
[26] Yamada Y, Namba K, Fujii T. Cardiac muscle thin filament structures reveal calcium regulatory mechanism. Nat Commun 2020;11:1–9.
[27] Guex N, Peitsch MC. SWISS-MODEL and the Swiss-Pdb Viewer: an environment for comparative protein modeling. electrophoresis 1997;18:2714–23.
[28] Humphrey W, Dalke A, Schulten K. VMD: visual molecular dynamics. J Mol Graph 1996;14:33–8.
[29] Jo S, Kim T, Iyer VG, Im W. CHARMM-GUI: a web-based graphical user interface for CHARMM. J Comput Chem 2008;29:1859–65.
[30] E.L. Wu, X. Cheng, S. Jo, H. Rui, K.C. Song, E.M. Dávila-Contreras, Y. Qi, J. Lee, V. Monje-Galvan, R.M. Venable, CHARMM-GUI membrane builder toward realistic biological membrane simulations, Wiley Online Library, 2014.
[31] Huang J, Rauscher S, Nawrocki G, Ran T, Feig M, De Groot BL, Grubmüller H, MacKerell AD. CHARMM36m: an improved force field for folded and intrinsically disordered proteins. Nat Methods 2017;14:71–3.
[32] Dolinsky TJ, Nielsen JE, McCammon JA, Baker NA. PDB2PQR: an automated pipeline for the setup of Poisson–Boltzmann electrostatics calculations. Nucleic Acids Res 2004;32:W665–7.
[33] Hilbe JM. Logistic regression models. Chapman and hall/CRC; 2009.
[34] Breiman L, Friedman JH, Olshen RA, Stone CJ. Classification and regression trees 432. Belmont, CA: Wadsworth, International Group; 1984. p. 9.
[35] Breiman L. Random forests. Mach Learn 2001;45:5–32.
[36] Hopfield JJ. Artificial neural networks. IEEE Circuits Syst Mag 1988;4:3–10.
[37] Breiman L, Friedman JH, Olshen RA, Stone CJ. Classification and regression trees. Routledge; 2017.
[38] Vinogradova MV, Stone DB, Malanina GG, Karatzaferi C, Cooke R, Mendelson RA, Fletterick RJ. Ca2+-regulated structural changes in troponin. Proc Natl Acad Sci USA 2005;102:5038–43.