Article

Ab initio characterization of protein molecular dynamics with AI^{2}BMD

https://doi.org/10.1038/s41586-024-08127-z

Received: 31 March 2023

Accepted: 26 September 2024

Published online: 6 November 2024

Open access

Check for updates

Tong Wang[1,2,3]✉, Xinheng He[1,2], Mingyu Li[1,2], Yatao Li[1,2], Ran Bi[1], Yusong Wang[1], Chaoran Cheng[1], Xiangzhen Shen[1], Jiawei Meng[1], He Zhang[1], Haiguang Liu[1], Zun Wang[1], Shaoning Li[1], Bin Shao[1,3]✉ & Tie-Yan Liu[1]

Biomolecular dynamics simulation is a fundamental technology for life sciences research, and its usefulness depends on its accuracy and efficiency[1–3]. Classical molecular dynamics simulation is fast but lacks chemical accuracy[4,5]. Quantum chemistry methods such as density functional theory can reach chemical accuracy but cannot scale to support large biomolecules[6]. Here we introduce an artificial intelligence-based ab initio biomolecular dynamics system (AI^{2}BMD) that can efficiently simulate full-atom large biomolecules with ab initio accuracy. AI^{2}BMD uses a protein fragmentation scheme and a machine learning force field[7] to achieve generalizable ab initio accuracy for energy and force calculations for various proteins comprising more than 10,000 atoms. Compared to density functional theory, it reduces the computational time by several orders of magnitude. With several hundred nanoseconds of dynamics simulations, AI^{2}BMD demonstrated its ability to efficiently explore the conformational space of peptides and proteins, deriving accurate ^3J couplings that match nuclear magnetic resonance experiments, and showing protein folding and unfolding processes. Furthermore, AI^{2}BMD enables precise free-energy calculations for protein folding, and the estimated thermodynamic properties are well aligned with experiments. AI^{2}BMD could potentially complement wet-lab experiments, detect the dynamic processes of bioactivities and enable biomedical research that is impossible to conduct at present.

The research paradigm of life sciences is shifting as the accuracy of computational simulation models is becoming indistinguishable from that of wet-lab experiments[1,2]. Among the computational models, molecular dynamics (MD) simulation, as the 'computational microscope', is of particular interest for understanding how life works[3,5,8]. MD simulations study the dynamic evolution of molecules by moving the atoms in a molecular system. They differ in the way that the forces are calculated[3]. In classical MD, forces are calculated using a prescribed interatomic potential function, whereas in ab initio MD (AIMD), forces are calculated using the potential derived from the electronic structure of molecules[6]. AIMD provides accurate characterization of molecules; the main challenge of applying AIMD to biomolecular simulation is scalability. On the one hand, the widely used quantum chemistry methods for AIMD are computationally expensive; for example, with the system size N, the time complexity of density functional theory (DFT) is about $O(N^3)$, and that of the coupled cluster method with the inclusion of single, double and perturbative triple excitations (CCSD(T)) is $O(N^7)$. On the other hand, observing important conformational changes for biomolecules such as proteins usually requires billions of steps with at least cubic time complexity for thousands of atoms[4]. Until now, scalable and accurate AIMD for biomolecules has not existed.

To alleviate the dilemma, machine learning force fields (MLFFs) trained on data generated at the DFT level provide accurate force calculations at a much lower cost and can be applied to small peptides and proteins[7,9,10]. The ability to generalize is the key challenge for the applicability and robustness for biomolecule simulations[11]. First, as the conformational space of a molecule is enormous, training on limited conformations of one kind of molecule and adapting it for conformational space exploration of other kinds of molecule is difficult[5]. Second, as the time and cost for generating data with DFT increase cubically with the size of the molecules, the lack of training data hinders the application of MLFFs for large biomolecules[11]. Furthermore, it is impossible to train a specific model for each kind of protein, and a unified solution with good generalization ability is needed.

In this study, we propose AI^{2}BMD, a generalizable solution for efficiently simulating a wide range of full-atom proteins with ab initio accuracy, surrounded by an explicit solvent modelled by a polarizable force field (Fig. 1). A generalizable protein fragmentation approach splits proteins into overlapped protein units. Simulations are performed by the AI^{2}BMD simulation system. At each simulation step, the AI^{2}BMD potential, based on ViSNet[7], calculates the energy and atomic forces for the protein with ab initio accuracy. Through comprehensive analysis from both kinetics and thermodynamics perspectives, AI^{2}BMD exhibits good alignment with wet-lab experimental data, such as the melting temperature of fast-folding proteins, and detects different phenomena than molecular mechanics (MM).

[1]Microsoft Research, Beijing, China. [2]These authors contributed equally: Tong Wang, Xinheng He, Mingyu Li, Yatao Li. [3]These authors jointly supervised this work: Tong Wang, Bin Shao. ✉e-mail: tongwang.bio@outlook.com; binshao@live.com

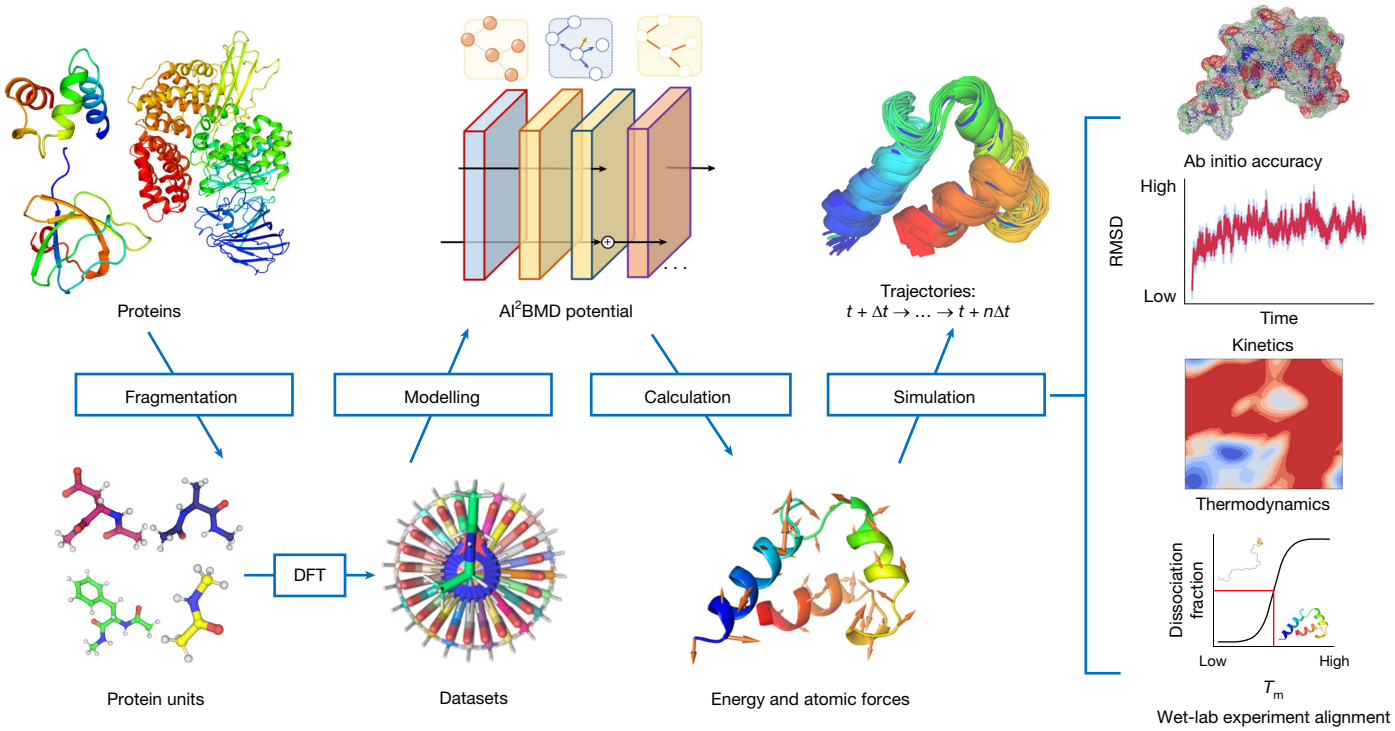


Fig. 1 | The overall pipeline of AI²BMD. Proteins are divided into protein units by a fragmentation process. The AI²BMD potential is designed on the basis of ViSNet, and the datasets are generated at the DFT level. It calculates the energy and atomic forces for the whole protein. The AI²BMD simulation system is built on these components and provides a generalizable solution for simulating the MD of proteins. It achieves ab initio accuracy in energy and force calculations. Through comprehensive analysis from both kinetics and thermodynamics perspectives, AI²BMD exhibits good alignment with wet-lab experimental data and detects different phenomena than MM.

Energy and force calculations

To provide a generalizable solution for accurately simulating proteins, AI²BMD adopts a universal protein fragmentation approach. Although generating samples for a specific kind of protein and training MLFF on them is straightforward, simulating other kinds of protein with the MLFF usually leads to simulation collapse[12] (Supplementary Fig. 1). Furthermore, it is computationally prohibitive to generate training data at the DFT level for large proteins. Thus, we fragment proteins into smaller units, specifically dipeptides, calculate intra- and inter-unit interactions, and then assemble them to determine the protein energy and forces acting on the atoms (see Methods for more details). Our fragmentation approach contains only 21 kinds of protein unit, and all protein units have similar and moderate numbers of atoms (range from 12 to 36), which is convenient for DFT data generation and MLFF training. Moreover, all kinds of protein can be broken down into the 21 kinds of protein unit, indicating that this is a generalizable fragmentation approach.

We built a comprehensively sampled protein unit dataset. During dataset construction, we scanned the main-chain dihedrals of all protein units to cover a wide range of conformations and ran AIMD simulations with the 6-31g* basis set and the M06-2X functional[13], as this functional models dispersion and weak interactions well and has been widely used for biomolecules[14,15]. We obtained 20.88 million samples (see Methods for more details). The whole dataset was split into training, validation and test sets to train ViSNet[7] models as the AI²BMD potential. The model encodes physics-informed molecular representations and calculates four-body interactions with linear time complexity. The model subsequently generates precise force and energy estimations based on the atom types and the coordinates as inputs (Methods and Extended Data Fig. 1). The performance of the AI²BMD potential was compared with that of the conventional MM force field on the test set, with the results presented in Supplementary Table 1. In terms of energy mean absolute error (MAE), the AI²BMD potential outperformed the MM force field by approximately two orders of magnitude (AI²BMD: 0.045 kcal mol^{-1}, MM: 3.198 kcal mol^{-1}). The AI²BMD potential also demonstrated superior performance for the force MAE (0.078 kcal mol^{-1} Å^{-1}) compared to MM (8.125 kcal mol^{-1} Å^{-1}). Overall, the AI²BMD potential offers accurate predictions for both potential energy and atomic forces for protein units.

On the basis of the AI²BMD potential, we developed an MD simulation system with a polarizable solvent described by the AMOEBA force field[16] (see the Methods for further details). Then we conducted simulations for 9 proteins with the number of atoms ranging from 175 to 13,728 (Fig. 2a; see the Methods for more details). Each protein was assessed with 5 folded, 5 unfolded and 10 intermediate structures derived from replica-exchange MD simulations as the initial conformations, and 10 AI²BMD simulation steps were run resulting in 200 structures per protein. The AI²BMD simulation system's ability to reach ab initio accuracy was evaluated by comparing its results to those calculated by DFT. Calculations by MM act as a control (Fig. 2b–e). For evaluation on potential energy (Fig. 2b,c), MM exhibited a broader error distribution and a much higher upper bound of error (that is, the maximum error) than AI²BMD. The average MAE of the MM potential energy consistently hovered around 0.2 kcal mol^{-1} per atom, whereas AI²BMD achieved a much lower value (0.038 kcal mol^{-1} per atom, averaged over the five proteins) (Fig. 2b). As the protein size increased from chignolin (175 atoms) to PACSIN3 (1,040 atoms), the increase of energy errors could be attributed to insufficient modelling for the escalating many-body interactions among protein units. For proteins from SSO0941 with 2,450 atoms to aminopeptidase N with 13,728 atoms, the reference value could be determined only through fragmented DFT (Fig. 2c). For these four proteins, AI²BMD's performance (MAE of 7.18×10^{-3} kcal mol^{-1} per atom) was substantially superior to that of MM (0.214 kcal mol^{-1} per atom). In terms of force (Fig. 2d,e), compared with the MM force field, AI²BMD aligned much more closely

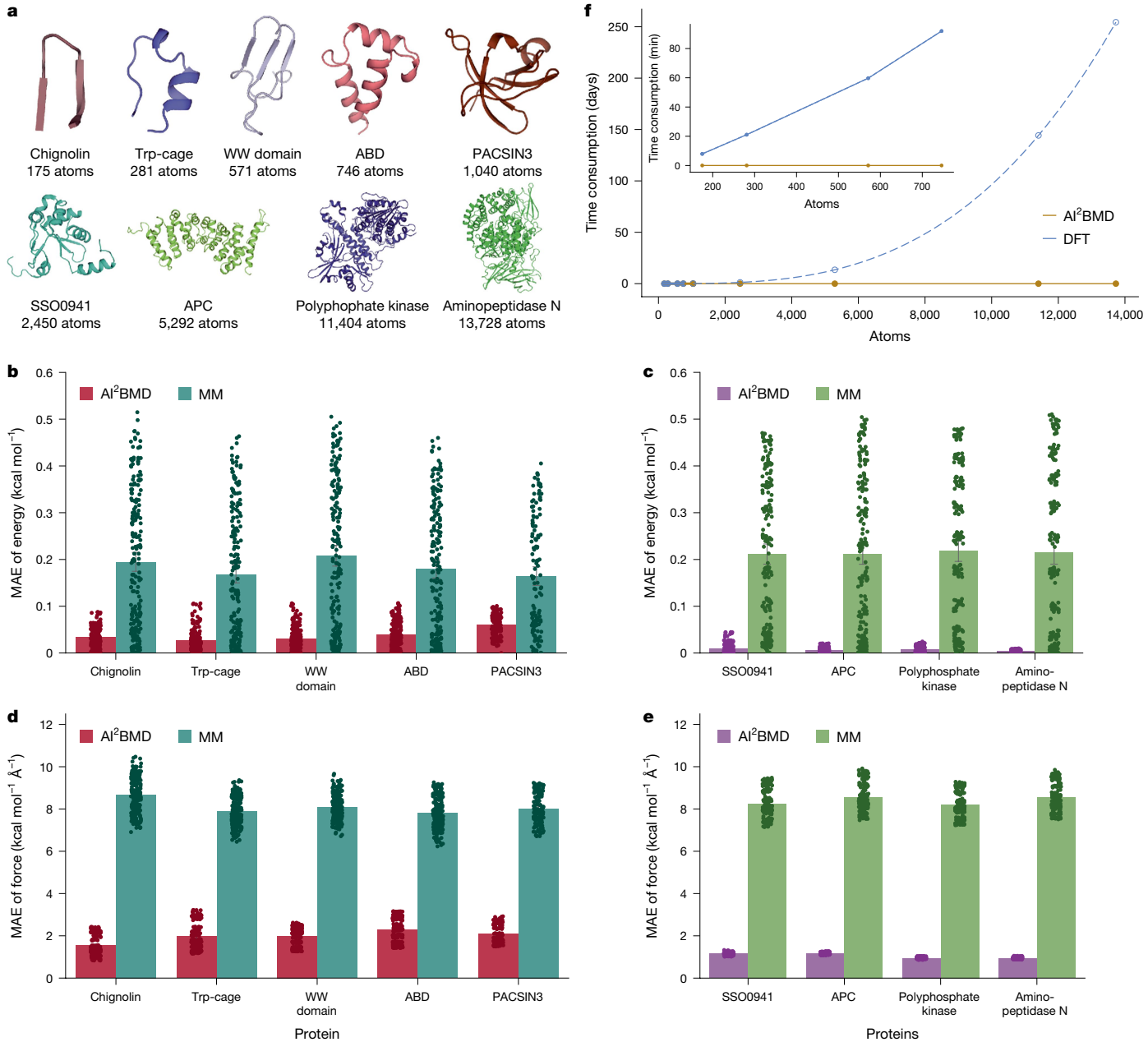


Fig. 2 | Evaluation of energy and force calculations by AI²BMD and MM. **a**, Folded structures of nine evaluated proteins. For these proteins, the number of atoms ranges from 175 to 13,728. **b**–**e**, The MAE of potential energy (**b**,**c**) and atomic force (**d**,**e**). For each protein, we conducted replica-exchange MD and structure clustering to select representative structures, including folded, unfolded or intermediate states. AI²BMD simulations were conducted for the representative structures, and 200 samples in total were selected for evaluation. For the first 5 proteins within 1,040 atoms shown in **b**,**d**, DFT calculation for the whole protein performed by ORCA with the same settings in dataset generation is set as the reference value, whereas for the last 4 proteins shown in **c**,**e**, the reference value is set as the fragment DFT calculation owing to prohibitive computational cost. In **b**,**c**, the potential energy of each structure has that of the initial folded structure subtracted, and then is normalized by the number of atoms. The error bars in **b**–**e** indicate the standard deviations of the potential energy and atomic force of 200 different samples of the protein ($n = 200$), with each sample shown as a filled circle. **f**, Comparison of time consumption of energy calculation for nine proteins. DFT calculations were carried out on a GPU. For the last five proteins, the time consumption by DFT was estimated by the fitting curve from those of the first four proteins and is shown with a dashed line and circles. The inset shows a comparison for the first four proteins.

with DFT results. For the first five proteins directly calculated by DFT, AI²BMD had an average MAE of 1.974 kcal mol^{-1} Å^{-1} compared to MM's 8.094 kcal mol^{-1} Å^{-1} (Fig. 2d). For the last four large proteins, AI²BMD achieved an average MAE of 1.056 kcal mol^{-1} Å^{-1}, whereas MM's value was 8.392 kcal mol^{-1} Å^{-1} across four systems (Fig. 2e). We further compared the performance of AI²BMD for different conformations. As shown in Supplementary Figs. 2–4, the MAE values of the potential energy for unfolded, intermediate and folded conformations of each kind of protein were analysed. The MAE values of the potential energies of different conformations fluctuated among different proteins, whereas those of the atomic forces were slightly increased from unfolded conformations to folded conformations. The minimal MAE across different proteins and conformations underscores the ab initio accuracy of the AI²BMD system.

Furthermore, to examine the efficiency of AI²BMD, we compared the time consumption of the energy calculation for all nine proteins by AI²BMD and DFT calculation software with graphics processing unit (GPU) support. In Fig. 2f, we present the computation time for

AI^{2}BMD and DFT on a desktop with an A6000 GPU card (48-GB GPU memory) and 32 central processing unit cores. It is obvious that AI^{2}BMD achieved ab initio accuracy much faster than DFT. The computational time for AI^{2}BMD exhibited a near-linear increase. AI^{2}BMD took 0.072 s to perform a simulation step for Trp-cage with 281 atoms, compared to 21 min by DFT. For the albumin-binding domain with 746 atoms, the time slightly increased to 0.125 s for AI^{2}BMD compared to 92 min for DFT. For a larger protein, aminopeptidase N with 13,728 atoms, it was 2.610 s, and DFT calculations were not feasible with the estimated time exceeding 254 days, which would be more than 6 orders slower than AI^{2}BMD. We further compared AI^{2}BMD's simulation speed with that of other AI-driven simulation systems, including DPMD[17] and Allegro[18], as well as the AMOEBA force field implemented in Tinker 8 (ref. 19) and ff19SB implemented in Amber. As shown in Extended Data Table 1, AI^{2}BMD exhibits a faster simulation speed, except for the smallest protein chignolin, than DPMD, even though DPMD uses a simpler model architecture. AI^{2}BMD's simulation speed substantially surpassed Allegro and AMOEBA for all cases. Furthermore, both DPMD and Allegro encountered an 'out-of-memory' error on an A6000 GPU card for some large proteins, whereas AI^{2}BMD worked well. In addition, a non-polarizable force field exhibits the fastest simulation, being about one order faster than AI^{2}BMD. In summary, AI^{2}BMD is versatile, is generalizable to various proteins and offers both ab initio accuracy and highly efficient calculation for MD simulation.

Conformational space exploration

To demonstrate the capabilities of AI^{2}BMD for conformational space exploration and protein kinetics, we carried out AI^{2}BMD simulations for both protein dipeptides and proteins. We initially constructed an asparagine dipeptide (Ace-N-Nme) in which the amino acid is capped with acetyl and *N*-methylamino groups at its amino and carboxy termini, respectively, in a 5-Å water box and sampled the hydrogen bonds between the solute and the solvent by carrying out a 500-ps simulation using quantum mechanics (QM)–MM, AI^{2}BMD with polarizable embedding and MM with Amber ff19SB. Then we scanned the distance between the oxygen in the water molecule and the acceptor on the dipeptide and calculated the energy fluctuations for the entire system by pure QM, AI^{2}BMD and MM. As depicted in Extended Data Fig. 2a,b, the distance distributions between the oxygen in the water molecule and the hydrogen-bond acceptor on the main chain, as sampled by QM–MM and AI^{2}BMD, exhibited high similarity. AI^{2}BMD also demonstrated an energy distribution much more consistent with QM–MM than MM in the hydrogen-bond scanning (Extended Data Fig. 2c). Furthermore, AI^{2}BMD showed consistent O–O distance distributions in comparison to QM–MM for the side-chain hydrogen bond with water (Extended Data Fig. 2d–f), with the peaks of both AI^{2}BMD and QM–MM located at identical positions. In conclusion, the hydrogen-bond sampling and scanning experiments suggest that AI^{2}BMD can accurately model the solvent effect and the interactions between the solute and the solvent.

Then we comprehensively sampled the conformation space of different protein units. We first evaluated the accuracy of potential energy and atomic force calculations during the simulations produced by the AI^{2}BMD system. AI^{2}BMD simulations of 10 ns were carried out for each kind of dipeptide with a 10-Å water box, and 200 snapshots with solvent were evenly picked from the simulation trajectory. The energy and force were calculated by QM for the whole protein and the AMOEBA force field for the solvent part as the reference value. The MM calculations are for comparison. Throughout the various simulation trajectories, regardless of the type of protein unit involved, the relative energy and force for the entire system, as calculated by AI^{2}BMD, showed neglectable errors when compared to the reference values. By contrast, the pure MM deviated substantially from the reference values (Fig. 3a–d and Supplementary Figs. 5 and 6). Specifically, for the negatively charged protein unit Ace-E-Nme (Fig. 3a), AI^{2}BMD exhibited slight differences compared with the reference values during the simulation (MAE: 0.183 kcal mol^{-1}), whereas MM presented a distinct difference, with an MAE of 4.111 kcal mol^{-1}. Furthermore, for Ace-R-Nme, the energy calculated by MM also exhibited fluctuations and was noticeably different from the reference value (MAE: 4.286 kcal mol^{-1} versus AI^{2}BMD MAE: 0.477 kcal mol^{-1}; Fig. 3b). In addition, AI^{2}BMD consistently outperformed MM by a large margin for Ace-F-Nme with a benzene ring in the side chain (MM MAE: 2.997 kcal mol^{-1} versus AI^{2}BMD MAE: 0.091 kcal mol^{-1}) (Fig. 3c). With smaller side chains, such as Ace-S-Nme (Fig. 3d), the discrepancy between AI^{2}BMD and the reference value further diminished (MAE: 0.056 kcal mol^{-1}), whereas that of MM remained distinct (MAE: 2.788 kcal mol^{-1}). For evaluations on atomic forces, AI^{2}BMD also demonstrated much greater fidelity to the reference values (MAE: 0.002 kcal^{-1} mol^{-1} Å^{-1}) than MM (MAE: 0.132 kcal mol^{-1} Å^{-1}) for all cases in the 10-ns simulations (Supplementary Fig. 6). Consequently, AI^{2}BMD maintained its accuracy across a diverse range of protein units during simulations.

We further comprehensively sampled the conformation space of different protein units. For each protein unit, we conducted 100 independent AI^{2}BMD simulations. To promote simulation efficiency, 50 initial structures were first derived from comprehensively sampled MM trajectories. Then, each initial structure underwent two independent AI^{2}BMD simulation runs with an explicit solvent for 1,000 ps, resulting in microsecond-level AI^{2}BMD simulations for protein units. We then analysed the conformational space explored by AI^{2}BMD by reproducing the $^3J(H_N, H_\alpha)$ coupling measured by nuclear magnetic resonance (NMR) experiments[13,20]. The $^3J(H_N, H_\alpha)$ coupling can accurately reflect the ϕ angle distributions for peptides[13] and thus was adopted from an experimental view to measure the conformational space exploration by simulations for the protein units. With the proline and histidine dipeptides excluded, we calculated the $^3J(H_N, H_\alpha)$ coupling values for the other 18 kinds of protein unit on the basis of the main-chain ϕ angles derived from AI^{2}BMD simulation trajectories and then averaged the values of two parallel repeats to obtain the final estimates. As a comparison, those made by MM were reported in the ff19SB force field[21]. Figure 3e illustrates that those simulations driven by AI^{2}BMD exhibited a significantly higher Pearson correlation (ρ = 0.924) with NMR experiment measurements than those calculated by MM (ρ = 0.543). Furthermore, AI^{2}BMD considerably outperformed MM for all protein units as shown in Fig. 3f. Such results further highlighted the effectiveness of AI^{2}BMD for conformation exploration and sampling from a wet-lab experimental perspective.

We then carried out AI^{2}BMD simulations for the decapeptide chignolin[22] in an explicit solvent to study the differences of its dynamics sampled by AI^{2}BMD. A total of 60 simulations starting from folded or unfolded structures were conducted, and each simulation was up to 10 ns. AI^{2}BMD captured both the folding and unfolding processes of chignolin during the simulations. In Fig. 4a and Supplementary Fig. 7a, starting from an unfolded structure, the protein formed into a folded hairpin structure with packed β-strands. During the process, the relative energy error of AI^{2}BMD was 3.44 kcal mol^{-1}, whereas that for MM was 15.20 kcal mol^{-1}. By contrast, Fig. 4b and Supplementary Fig. 7b depict the unfolding process from the folded hairpin to an extended unfolded structure. During this process, the value calculated by AI^{2}BMD differed from the reference value by 4.40 kcal mol^{-1}, whereas MM exhibited a deviation of 15.11 kcal mol^{-1}. In both simulation processes, as shown in Supplementary Fig. 8, compared with the force error of 0.614 kcal mol^{-1} Å^{-1} for the folding process and 0.620 kcal mol^{-1} Å^{-1} for the unfolding process made by MM, AI^{2}BMD also exhibited much closer force calculations to the reference values (a force error of 0.063 kcal mol^{-1} Å^{-1} for the folding process and a force error of 0.073 kcal mol^{-1} Å^{-1} for the unfolding process). In addition, projecting the conformations from the simulation shown in Fig. 4a onto the free-energy landscape enables observation of the folding process from an unfolded metastable state to the folded

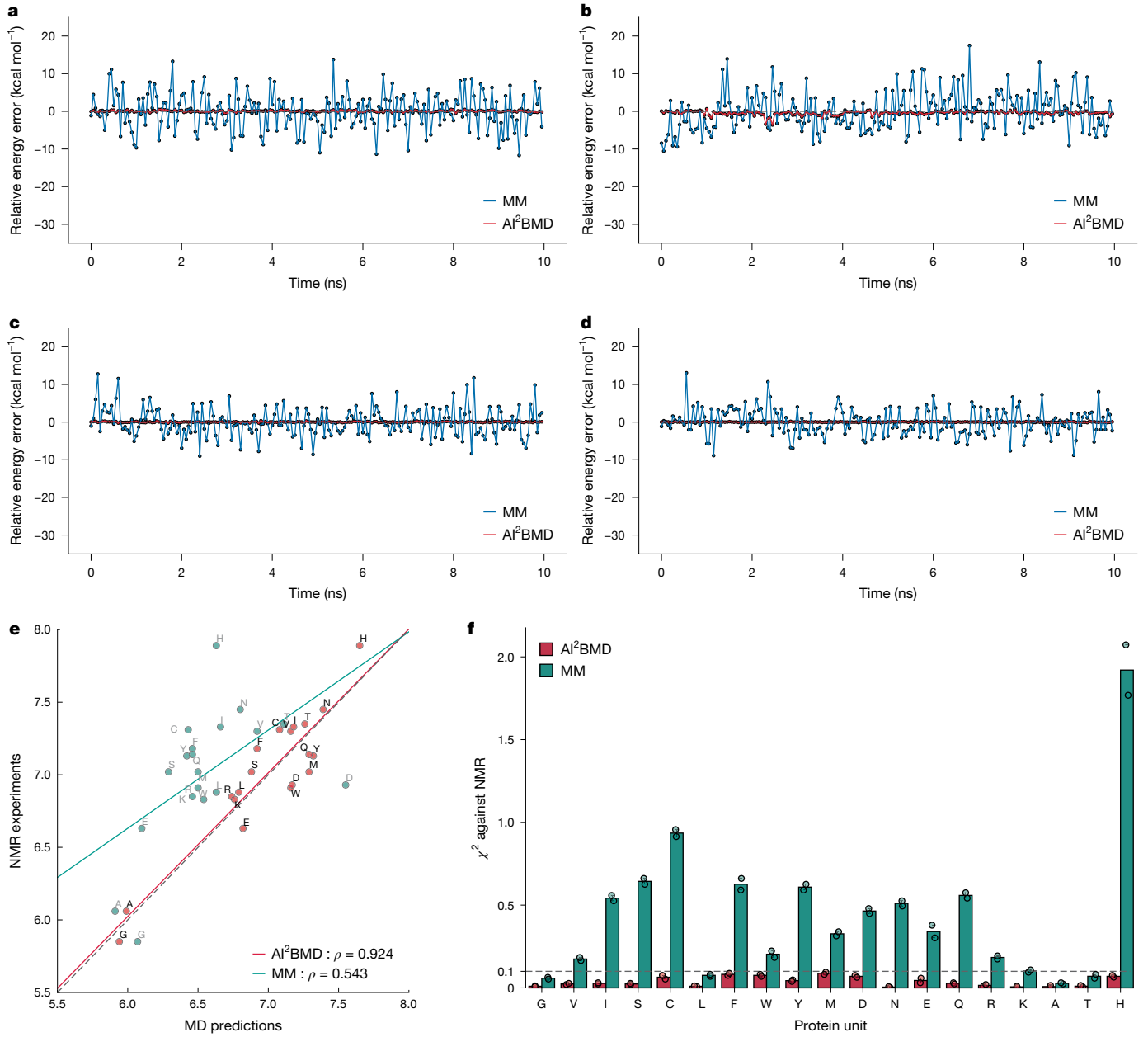


Fig. 3 | AI²BMD simulations for protein units and comparisons with NMR experiments. a–**d**, The errors of potential energies of the entire system during the 10-ns simulations for negatively charged Ace-E-Nme (**a**), positively charged Ace-R-Nme (**b**), aromatic Ace-F-Nme (**c**) and polar Ace-S-Nme (**d**). Calculations for the entire simulation system by QM for the whole protein and MM for the solvent part act as the reference values and those by MM are for comparison. The potential energy of each structure has that of a reference structure subtracted, and then the errors are calculated by comparison with the reference values. The range of the potential energy fluctuation of the entire system during the simulation is about 300 kcal mol^{-1} (see Supplementary Fig. 7 for further details). **e**, Comparison between NMR $^3J(H_N, H_\alpha)$ couplings and those derived from simulations driven by AI²BMD and MM. The $^3J(H_N, H_\alpha)$ couplings derived from AI²BMD and MM are shown as red and green points, respectively. For each approach, a linear regression curve is drawn, and the corresponding Pearson correlation is shown in the legend. **f**, χ^2 errors in reproducing $^3J(H_N, H_\alpha)$ couplings for each protein unit measured by NMR. The results produced by AI²BMD and MM are shown in red and green, respectively. In **e**,**f**, the Ace-*X*-Nme dipeptide is abbreviated as *X* for simplicity. Proline dipeptide is neglected owing to its unique bonding pattern. The error bars in **f** indicate the standard deviations of χ^2 errors from two repeated experiments with different initial simulation configurations ($n = 2$) for both AI²BMD and MM.

metastable state possessing the lowest energy values (Extended Data Fig. 3). These results indicate that AI²BMD can fold proteins by itself and detect meaningful conformational changes to study protein dynamics.

To further explore the differences in chignolin dynamics driven by AI²BMD and MD, we also carried out simulations by MM with the same initial structures and simulation configurations and found that AI²BMD simulations exhibited several distinct features. First, simulations performed by AI²BMD exhibit similar structure fluctuations to those performed by MM. As shown in Fig. 4c, the root-mean-square deviation (RMSD) compared to the initial structure in AI²BMD simulations increased slightly faster than that in classical MD simulations for the first few nanoseconds, and the gap gradually vanished as the simulations converged. After 10 ns, AI²BMD had an average RMSD of 3.378 Å, whereas MM simulations reached 3.454 Å. We also analysed the distances of adjacent Cα atoms during the simulation. As shown in Supplementary Fig. 9, simulations by AI²BMD showed a similar distribution to that by MM. The averaged distances between adjacent Cα

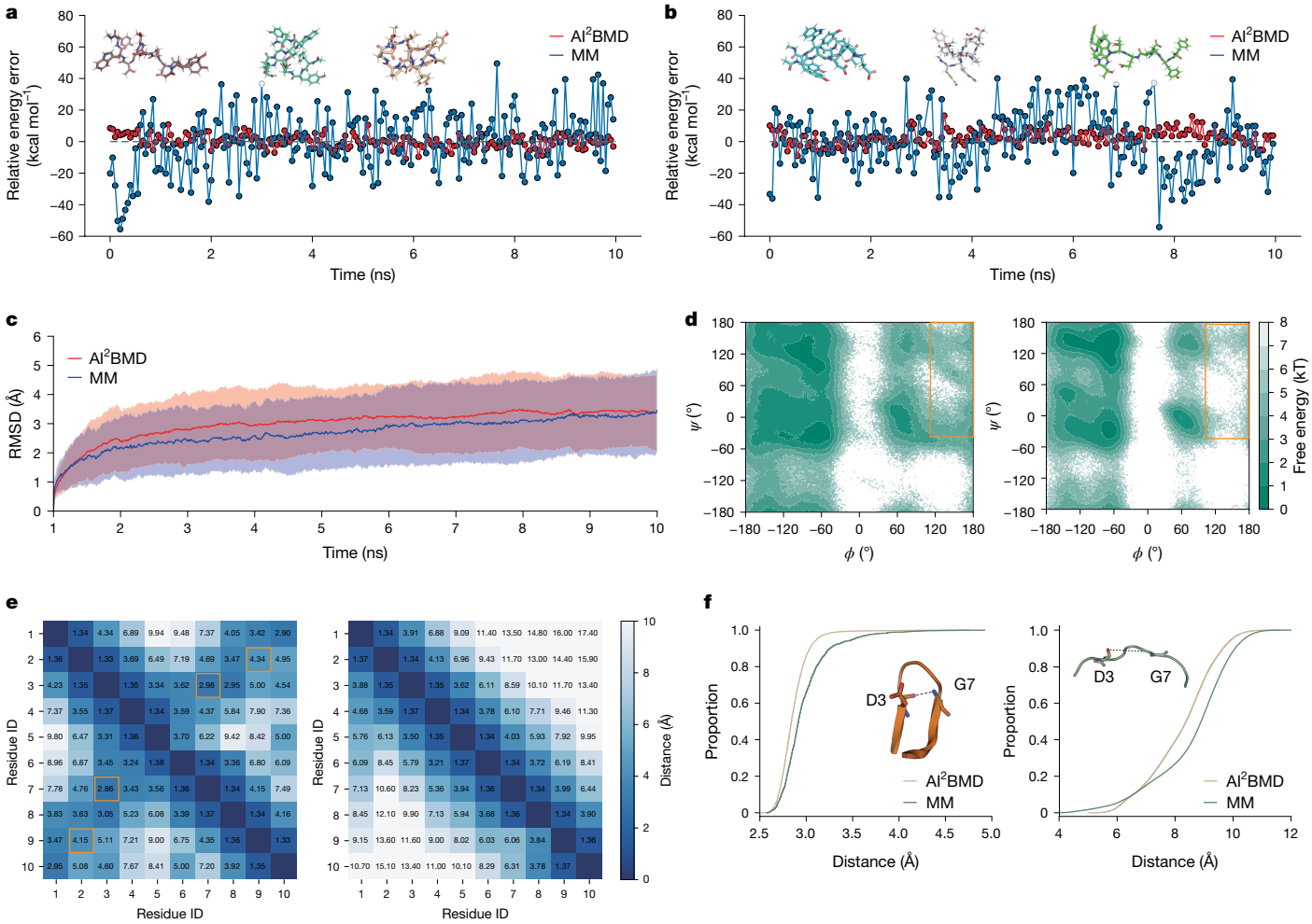


Fig. 4 | Analysis of chignolin dynamics by AI^{2}BMD simulations. a,**b**, The error of relative potential energies of the entire system during the 10-ns simulations for chignolin starting from an unfolded structure (**a**) and a folded structure (**b**). QM calculation for the whole protein and AMOEBA force field calculation for the remaining solvent part act as the reference values. The relative energy is defined as the potential energy of each structure with that of a reference structure subtracted. The initial, intermediate and final structures during the simulations are shown at the top. The range of the potential energy fluctuation of the entire system during the simulation is about 500 kcal mol^{-1} (see Supplementary Fig. 9 for further details). **c**, RMSD during an ensemble of 60 trajectories of 10-ns simulations (the first 1 ns was omitted). The average RMSD is shown as a line, with the ranges of all 60 simulation trajectories shown in shading. **d**, Ramachandran plot of conformations sampled by AI^{2}BMD (left) and MM (right) during an ensemble of 60 trajectories of 10-ns simulations. Different regions are highlighted with yellow rectangles for visualization. **e**, Residue minimum distance map for folded (left) and unfolded (right) structures. AI^{2}BMD and MM are shown in the lower triangle matrix and upper triangle matrix, respectively. A π–π interaction and a hydrogen bond are highlighted with yellow squares for analysis. **f**, The cumulative plots for the distance between the main-chain O of D3 and the main-chain N of G7 for folded (left) and unfolded (right) structures.

atoms are 3.816 Å and 3.863 Å in AI^{2}BMD and MM, respectively, which are close to the empirical value of 3.8 Å, implying that the simulations performed stably, without simulation collapse. Second, regarding the specific main-chain conformations depicted in the Ramachandran plot (Fig. 4d), AI^{2}BMD exhibited a slightly broader distribution than MM, particularly in the region where ϕ ranges from 120° to 180° and ψ ranges from −60° to 180°. We further investigated the differences between AI^{2}BMD and MM by inspecting the Ramachandran plot of each residue and found that such differences mainly came from the ϕ–ψ angle distribution of G7 (Extended Data Fig. 4a,b). As shown in Extended Data Fig. 4c,d, we then clustered all simulation trajectories into 10 clusters according to the ϕ and ψ angle of G7 and present the representative structures for each cluster. AI^{2}BMD showed more diverse ϕ and ψ angles compared to MM for G7. Considering that glycine can explore most of the regions on the Ramachandran plot owing to the lack of a side chain[23], the observations during simulations imply that AI^{2}BMD could explore more possible conformational space without the harmonic constraints on bond lengths applied in MM.

Furthermore, using the Q score, which evaluates protein folding on the basis of contacts during simulations[24], we separated the AI^{2}BMD and MM snapshots into folded and unfolded structures. As shown in the residue minimum distance map (Fig. 4e), the results of AI^{2}BMD and MM were similar in both folded and unfolded structures. They both described the stable interactions between chignolin's N-terminal and C-terminal aromatic residues (Y2–W9) and the hydrogen bonds between D3 and G7. For the strongest hydrogen bond between D3 and G7 (refs. 22,25), the folded conformations sampled by AI^{2}BMD depicted it as more stable (minimum distance of 2.86 Å in AI^{2}BMD versus 2.98 Å in MM). The cumulative plots for the distance between the main-chain O of D3 and the main-chain N of G7 (Fig. 4f) also illustrate that AI^{2}BMD stabilized the hydrogen bond more than MM. For unfolded structures, the trends of AI^{2}BMD and MM were similar in that both described a much longer distance than that in the folded structures. These results indicate that by performing simulations with ab initio accuracy, AI^{2}BMD can detect both meaningful conformational changes and detailed interatomic interactions to study protein dynamics. The observed

difference may be attributed to the fact that AI^{2}BMD does not impose harmonic constraints on bonds and angles, thereby offering larger flexibility and closer approximation to real-world conditions.

Protein property estimation

To examine the usefulness of AI^{2}BMD's application in protein property estimation, we first investigated the thermodynamic properties of various fast-folding proteins simulated by AI^{2}BMD. From the comprehensively sampled trajectories of proteins by ref. 4, we evenly procured 100,000 snapshots for each protein and used the Q score to categorize them into folded and unfolded states. The representative folded and unfolded structures of the seven proteins (that is, BBA, WW domain, NTL9, homeodomain, protein G, α3D and λ-repressor) are shown in Extended Data Fig. 5a. These proteins consist of 504 to 1,258 atoms and showcase a variety of secondary structures. Clustering analysis shows that for each protein, the structure from the centre of the folded cluster exhibits close alignment with the corresponding folded crystal structures (Supplementary Fig. 10). We first calculated the potential energy values by AI^{2}BMD and displayed the potential energy surface using time-lagged independent components. As shown in Extended Data Fig. 5b for the case of NTL9, for the results produced by both MM and AI^{2}BMD, folded and unfolded structures were clearly separated with distinct potential energies, although the difference in potential energy between them was smaller in AI^{2}BMD's result. By reweighting on potential energy values, we estimated the free-energy difference during protein folding (ΔG) and the melting temperature (T_m) derived from simulation trajectories (Extended Data Fig. 5c,d and Supplementary Table 2). As such protein simulations were conducted near the corresponding melting temperatures[4], ΔG is expected to be closer to 0 and the calculated T_m is expected to be closer to the simulation temperature. Both MM and AI^{2}BMD achieved small free-energy differences and comparable calculated melting temperatures to simulation temperatures, although AI^{2}BMD slightly outperformed MM for these cases.

Notably, the WW domain is the only protein dominated by β-sheets in the evaluation[26]. Compared with the experimentally determined melting temperature of 371 ± 2 K for the WW domain[26], the estimation of T_m provided by AI^{2}BMD was closer (359.06 ± 0.07 K) than that of MM (353.69 ± 0.38 K). NTL9 has a folded structure with an α-helix and β-sheets[27]. With the simulation temperature of 355 K nearly the same as the experimental T_m of 354.75 ± 1.7 K for NTL9 (ref. 28), AI^{2}BMD achieved a better estimation of ΔG than MM (−0.34 kcal mol^{-1} versus −1.54 kcal mol^{-1}). Furthermore, AI^{2}BMD estimated the T_m to be 351.84 ± 0.11 K, which is more accurate than the value of 349.47 ± 0.35 K made by MM.

In addition, for all α-proteins, the homeodomain features a folded structure with three α-helices and loops[29]. Although AI^{2}BMD's ΔG estimation (−0.18 kcal mol^{-1}) is smaller than MM's (−0.73 kcal mol^{-1}), their T_m predictions were nearly the same (AI^{2}BMD: 359.61 ± 0.14 K; MM: 359.60 ± 0.13 K) and were comparable to the experimental T_m of >372 K (ref. 30). α3D is an artificial protein composed mainly of α-helices[31]. AI^{2}BMD's ΔG estimation is −0.098 kcal mol^{-1}, whereas MM's is 1.33 kcal mol^{-1}. Both of the melting temperatures of 369.67 ± 0.06 K and 366.94 ± 0.26 K generated by AI^{2}BMD and MM, respectively, were comparable to the experimental value (>363 K)[32] for α3D. λ-repressor is the largest protein in the evaluation and is dominated by α-helices[33]. AI^{2}BMD and MM both deviated from 0 in the ΔG estimation (AI^{2}BMD: 0.79 kcal mol^{-1}; MM: 1.09 kcal mol^{-1}), but AI^{2}BMD is closer. Their T_m estimations were nearly the same (AI^{2}BMD: 349.55 ± 0.21 K; MM: 349.48 ± 0.21 K) and were close to the experimental value (347 K)[33].

For α- and β-proteins, BBA features a folded structure comprising an α-helix and a β-hairpin[34]. Lacking a known melting temperature, simulations were conducted at 325 K. AI^{2}BMD's ΔG calculation (0.057 kcal mol^{-1}) and T_m estimation (323.94 ± 0.22 K) were similar to MM's results (1.22 kcal mol^{-1} and 322.34 ± 0.31 K). Protein G consists of an α-helix and a four-fold β-sheet group[35]. AI^{2}BMD's ΔG prediction (0.14 kcal mol^{-1}) is closer to 0 than MM's (0.74 kcal mol^{-1}). In the absence of experimental T_m, AI^{2}BMD's estimated T_m (349.49 ± 0.12 K) demonstrates a 3-K improvement over MM's (346.49 ± 0.66 K) concerning the simulation temperature of 350 K for protein G.

We further estimated the changes of enthalpy and heat capacity during protein folding. The two-state proteins 110-residue barnase and 84-residue CI2 were chosen for evaluation. Twenty simulations starting from the folded structures and unfolded structures were carried out. For each protein, simulations were conducted at three different temperatures under an NPT ensemble. Potential energy values of conformations sampled during simulations were calculated by AI^{2}BMD. As shown in Fig. 5 and Supplementary Table 3, AI^{2}BMD outperformed MM in achieving closer calculations to experimental values in all evaluation metrics. For barnase, the changes in enthalpy and heat capacity calculated by AI^{2}BMD are 116.5 ± 0.43 kcal mol^{-1} and 1.4 ± 0.04 kcal mol^{-1} K^{-1}, and those calculated by MM are 110.4 ± 3.1 kcal mol^{-1} and 1.0 ± 0.1 kcal mol^{-1} K^{-1} (ref. 36), respectively. Compared with the experimentally determined values of 118.7 ± 4.9 kcal mol^{-1} and 1.4 ± 0.1 kcal mol^{-1} K^{-1} (refs. 36,37),

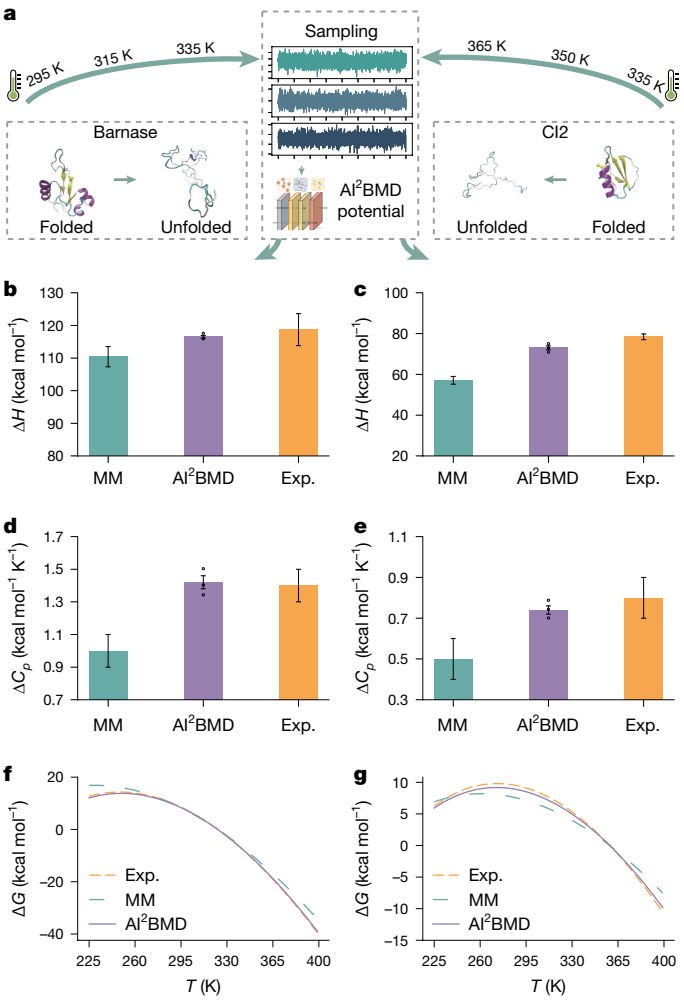


Fig. 5 | Comparison of the change of enthalpy, heat capacity and free energy of two-state proteins, barnase and CI2. The values were calculated by MM, AI^{2}BMD and experiments (Exp.). **a**, The overall calculation scheme of AI^{2}BMD. **b**–**e**, The changes in enthalpy (ΔH) (**b**,**c**) and heat capacity (ΔC_p) (**d**,**e**) during protein unfolding for barnase (**b**,**d**) and CI2 (**c**,**e**). In **b**–**e**, the error bars indicate the standard errors of ΔH and ΔC_p values from three repeated experiments with different initial simulation configurations ($n = 3$) for AI^{2}BMD. The mean and standard errors of MM and experimental values are from ref. 36. **f**,**g**, The free-energy fluctuation against temperature for barnase (**f**) and CI2 (**g**).

the results achieved by AI^{2}BMD were much closer than those of MM (Fig. 5a,c). Similar results were also observed on CI2 as shown in Fig. 5b,d. The changes in enthalpy and heat capacity calculated by AI^{2}BMD are 73.0 ± 0.97 kcal mol^{-1} and 0.7 ± 0.02 kcal mol^{-1} K^{-1}, which are close to the experimental measurements 78.4 ± 1.4 kcal mol^{-1} and 0.8 ± 0.1 kcal mol^{-1} K^{-1} (refs. 36,38). As a comparison, MM's enthalpy change is only 57.1 ± 1.9 kcal mol^{-1} and the heat capacity change is 0.5 ± 0.1 kcal mol^{-1} K^{-1} (ref. 36). For the free-energy fluctuation against temperature, AI^{2}BMD's results are more aligned with the experimental results than MM's results (Fig. 5f,g). These observations further consolidate that the ab initio energy calculation made by AI^{2}BMD can provide a better description of properties in protein folding.

In addition, we also demonstrated AI^{2}BMD's utility in alchemical free-energy calculations through the case study of the negative base 10 logarithm of the acid dissociation constant (pK_a) estimation for thioredoxin Asp26 using thermodynamic integration[39]. Initially, AI^{2}BMD was applied to reweight the potential energy values across simulation trajectories for both thioredoxin with AspH26 and Asp26 and the dipeptide AspH–Asp. Subsequently, thermodynamic integration was utilized to compute the free-energy difference ($\Delta\Delta G$) for the protonation state change from AspH26 to Asp26 in thioredoxin, as depicted in Extended Data Fig. 6. The estimated pK_a value of 7.61 by AI^{2}BMD closely corresponded with the experimental benchmark of 7.5 and surpassed the accuracy of established methodologies, including force-field-, empirical- and QM–MM-based approaches. Given the protein property evaluations on diverse proteins, AI^{2}BMD exhibits accurate calculations on conformational ensembles, leads to reasonable estimations of protein folding thermodynamics and advances its utility in biochemical research.

Discussion

Simulating biomolecular dynamics with ab initio accuracy is a long-standing challenge as it is difficult to achieve accuracy, efficiency and generalization ability at the same time[5,40]. We have developed a generalizable, efficient and close-to-ab initio simulation program for a wide range of proteins, AI^{2}BMD, that offers considerable improvements over MM in energy and force calculations and protein property estimations. The generalizability of AI^{2}BMD across different protein systems and its robustness showcase its potential for broader applications in protein research.

Compared to QM–MM, which defines a focused region calculated by QM and the remaining part calculated by MM, AI^{2}BMD expands ab initio calculation from a small preset QM region to the whole full-atom protein without any prior knowledge. Furthermore, AI^{2}BMD eliminates the potential incompatibility of QM and MM mechanics on the boundary for proteins and accelerates QM region calculation by several orders. In addition, for some complex biomolecular dynamics that QM–MM cannot deal with, such as the focused regions that require QM treatment and dynamically evolve during simulations, or large QM regions as in processes of some allosteric regulations or intrinsically disordered protein simulations, AI^{2}BMD could offer opportunities with new perspectives for future studies.

The generalization ability of AI^{2}BMD originates from the fundamental principle that most proteins are composed of common kinds of amino acid. This understanding allows AI^{2}BMD to serve as a versatile and adaptable algorithm that can be applied to a diverse range of proteins. By incorporating this knowledge into AI^{2}BMD, it can serve for various conformations of a protein and accommodate proteins of different sizes and compositions, enabling researchers to explore and investigate the complex world of proteins with greater confidence and precision.

Furthermore, in the realm of protein dynamics, without harmonic constraints on bond lengths and angles, AI^{2}BMD's reasonable incorporation of flexibility into protein movements, grounded in calculations that approach ab initio accuracy for full-atom proteins, will create more opportunities to study protein dynamics.

In addition, although AI^{2}BMD has a much faster computational speed than DFT methods, it still lags classical MD simulations in terms of efficiency. To bridge this gap, several strategies can be used in future. For example, implementing additional engineering optimization could lead to substantial improvements in the efficiency of AI^{2}BMD. Furthermore, applying AI^{2}BMD for a broader range of systems, including lipids, nucleotides, nanomaterials and solute–solvent interfaces, will broaden the scope of AI^{2}BMD's applicability to more complex biomolecular systems to unlock new insights into the intricate world of biomolecular systems[41,42], paving the way for more accurate and efficient simulations in a variety of contexts, such as drug discovery, protein design and enzyme engineering.

Online content

Any methods, additional references, Nature Portfolio reporting summaries, source data, extended data, supplementary information, acknowledgements, peer review information; details of author contributions and competing interests; and statements of data and code availability are available at https://doi.org/10.1038/s41586-024-08127-z.

1. Brini, E., Simmerling, C. & Dill, K. Protein storytelling through physics. *Science* https://doi.org/10.1126/science.aaz3041 (2020).
2. Groenhof, G. Solving chemical problems with a mixture of quantum-mechanical and molecular mechanics calculations: Nobel Prize in Chemistry 2013. *Angew. Chem. Int. Ed. Engl.* **52**, 12489–12491 (2013).
3. Karplus, M. & McCammon, J. A. Molecular dynamics simulations of biomolecules. *Nat. Struct. Biol.* **9**, 646–652 (2002).
4. Lindorff-Larsen, K., Piana, S., Dror, R. O. & Shaw, D. E. How fast-folding proteins fold. *Science* **334**, 517–520 (2011).
5. Schlick, T. & Portillo-Ledesma, S. Biomolecular modeling thrives in the age of technology. *Nat. Comput. Sci.* **1**, 321–331 (2021).
6. Iftimie, R., Minary, P. & Tuckerman, M. E. Ab initio molecular dynamics: concepts, recent developments, and future trends. *Proc. Natl Acad. Sci. USA* **102**, 6654–6659 (2005).
7. Wang, Y. et al. Enhancing geometric representations for molecules with equivariant vector-scalar interactive message passing. *Nat. Commun.* **15**, 313 (2024).
8. Schlick, T., Collepardo-Guevara, R., Halvorsen, L. A., Jung, S. & Xiao, X. Biomolecularmodeling and simulation: a field coming of age. *Q. Rev. Biophys.* **44**, 191–228 (2011).
9. Smith, J. S., Isayev, O. & Roitberg, A. E. ANI-1: an extensible neural network potential with DFT accuracy at force field computational cost. *Chem. Sci.* **8**, 3192–3203 (2017).
10. Unke, O. T. et al. Biomolecular dynamics with machine-learned quantum-mechanical force fields trained on diverse chemical fragments. *Sci. Adv.* **10**, eadn4397 (2024).
11. Anstine, D. M. & Isayev, O. Machine learning interatomic potentials and long-range physics. *J. Phys. Chem. A* **127**, 2417–2431 (2023).
12. Wang, Z. et al. Improving machine learning force fields for molecular dynamics simulations with fine-grained force metrics. *J. Chem. Phys.* **159**, 035101 (2023).
13. Hohenstein, E. G., Chill, S. T. & Sherrill, C. D. Assessment of the performance of the M05-2X and M06-2X exchange-correlation functionals for noncovalent interactions in biomolecules. *J. Chem. Theory Comput.* **4**, 1996–2000 (2008).
14. Jakobsen, S., Kristensen, K. & Jensen, F. Electrostatic potential of insulin: exploring the limitations of density functional theory and force field methods. *J. Chem. Theory Comput.* **9**, 3978–3985 (2013).
15. Robertson, M. J., Tirado-Rives, J. & Jorgensen, W. L. Improved peptide and protein torsional energetics with the OPLSAA force field. *J. Chem. Theory Comput.* **11**, 3499–3509 (2015).
16. Shi, Y. et al. The polarizable atomic multipole-based AMOEBA force field for proteins. *J. Chem. Theory Comput.* **9**, 4046–4063 (2013).
17. Zhang, L., Han, J., Wang, H., Car, R. & Weinan, E. Deep potential molecular dynamics: a scalable model with the accuracy of quantum mechanics. *Phys. Rev. Lett.* **120**, 143001 (2018).
18. Musaelian, A. et al. Learning local equivariant representations for large-scale atomistic dynamics. *Nat. Commun.* **14**, 579 (2023).
19. Rackers, J. A. et al. Tinker 8: software tools for molecular design. *J. Chem. Theory Comput.* **14**, 5273–5289 (2018).
20. Avbelj, F., Grdadolnik, S. G., Grdadolnik, J. & Baldwin, R. L. Intrinsic backbone preferences are fully present in blocked amino acids. *Proc. Natl Acad. Sci. USA* **103**, 1272–1277 (2006).
21. Tian, C. et al. ff19SB: amino-acid-specific protein backbone parameters trained against quantum mechanics energy surfaces in solution. *J. Chem. Theory Comput.* **16**, 528–552 (2020).
22. Honda, S., Yamasaki, K., Sawada, Y. & Morii, H. 10 residue folded peptide designed by segment statistics. *Structure* **12**, 1507–1518 (2004).
23. Ho, B. K. & Brasseur, R. The Ramachandran plots of glycine and pre-proline. *BMC Struct. Biol.* **5**, 14 (2005).
24. Best, R. B., Hummer, G. & Eaton, W. A. Native contacts determine protein folding mechanisms in atomistic simulations. *Proc. Natl Acad. Sci. USA* **110**, 17874–17879 (2013).
25. Satoh, D., Shimizu, K., Nakamura, S. & Terada, T. Folding free-energy landscape of a 10-residue mini-protein, chignolin. *FEBS Lett.* **580**, 3422–3426 (2006).

26. Piana, S. et al. Computational design and experimental testing of the fastest-folding β-sheet protein. *J. Mol. Biol.* **405**, 43–48 (2011).
27. Cho, J. H. et al. Energetically significant networks of coupled interactions within an unfolded protein. *Proc. Natl Acad. Sci. USA* **111**, 12079–12084 (2014).
28. Horng, J.-C., Moroz, V. & Raleigh, D. P. Rapid cooperative two-state folding of a miniature α–β protein and design of a thermostable variant. *J. Mol. Biol.* **326**, 1261–1270 (2003).
29. Shah, P. S. et al. Full-sequence computational design and solution structure of a thermostable protein variant. *J. Mol. Biol.* **372**, 1–6 (2007).
30. Gillespie, B. et al. NMR and temperature-jump measurements of de novo designed proteins demonstrate rapid folding in the absence of explicit selection for kinetics. *J. Mol. Biol.* **330**, 813–819 (2003).
31. Walsh, S. T., Cheng, H., Bryson, J. W., Roder, H. & DeGrado, W. F. Solution structure and dynamics of a de novo designed three-helix bundle protein. *Proc. Natl Acad. Sci. USA* **96**, 5486–5491 (1999).
32. Zhu, Y. et al. Ultrafast folding of α3D: a *de novo* designed three-helix bundle protein. *Proc. Natl Acad. Sci. USA* **100**, 15486–15491 (2003).
33. Yang, W. Y. & Gruebele, M. Folding at the speed limit. *Nature* **423**, 193–197 (2003).
34. Sarisky, C. A. & Mayo, S. L. The ββα fold: explorations in sequence space. *J. Mol. Biol.* **307**, 1411–1418 (2001).
35. Nauli, S., Kuhlman, B. & Baker, D. Computer-based redesign of a protein folding pathway. *Nat. Struct. Biol.* **8**, 602–605 (2001).
36. Galano-Frutos, J. J., Nerín-Fonz, F. & Sancho, J. Calculation of protein folding thermodynamics using molecular dynamics simulations. *J. Chem. Inf. Model.* **63**, 7791–7806 (2023).
37. Vuilleumier, S. & Fersht, A. R. Insertion in barnase of a loop sequence from ribonuclease T1: Investigating sequence and structure alignments by protein engineering. *Eur. J. Biochem.* **221**, 1003–1012 (1994).
38. Jackson, S. E. & Fersht, A. R. J. B. Folding of chymotrypsin inhibitor 2. 1. Evidence for a two-state transition. *Biochemistry* **30**, 10428–10435 (1991).
39. Simonson, T., Carlsson, J. & Case, D. A. Proton binding to proteins: pKa calculations with explicit and implicit solvent models. *J. Am. Chem. Soc.* **126**, 4167–4180 (2004).
40. Shen, L., Wu, J. & Yang, W. Multiscale quantum mechanics/molecular mechanics simulations with neural networks. *J. Chem. Theory Comput.* **12**, 4934–4946 (2016).
41. Lier, B., Poliak, P., Marquetand, P., Westermayr, J. & Oostenbrink, C. BuRNN: buffer region neural network approach for polarizable-embedding neural network/molecular mechanics simulations. *J. Phys. Chem. Lett.* **13**, 3812–3818 (2022).
42. Manzhos, S. & Carrington, T. Jr. Neural network potential energy surfaces for small molecules and reactions. *Chem. Rev.* **121**, 10187–10217 (2021).

Methods

Protein fragmentation approach

Generally speaking, proteins are composed of 20 kinds of amino acid, each of which has a common main chain consisting of Cα, C, O, N and H, and a different side chain (termed the R group). A dipeptide is an amino acid capped with Ace and Nme groups at its N and C termini, respectively. As amino acids are the fundamental units of proteins, we designed the generalizable protein fragmentation approach on the basis of dipeptides and trained AI^{2}BMD potential accordingly, which ensures the generalization ability to all proteins.

The concept of peptide fragmentation has been around for years, and previous studies have demonstrated its accuracy and efficiency to proteins[10,43]. As shown in Extended Data Fig. 7a, each dipeptide consists of: all atoms of the main chain and the side chain of the amino acid; the Cα, H connected to Cα, C, O of the main chain of the previous amino acid; and the N, H connected to N, Cα and H connected to Cα of the main chain of the next amino acid. We cut the polypeptide chains with a sliding window, and thus the Ace-Nme fragments act as the overlapping regions between two successive dipeptides (Extended Data Fig. 7b). The extra hydrogens for the terminal Cα were added to dipeptides and Ace-Nme fragments, according to the C–H bond length and the direction of the bond connected to the Cα in the whole peptide chain. If the first or last amino acid was glycine, we added only one hydrogen connected to Cα according to the C–H bond length. If the latter amino acid is proline, we also added a hydrogen connected to N according to the N–H bond length where the N is connected to Cδ. Then, the limited-memory Broyden–Fletcher–Goldfarb–Shanno quasi-Newton algorithm[44] was applied to optimize the positions of the added hydrogens while the other parts were constrained.

We first calculated the total energy and force for all protein units by summing up the energies of dipeptides, subtracting the energies of all overlapping Ace-Nme fragments (equation (1)).

$$E^{\text{prot\_units}} = \sum_{i=1}^{n} E_i^{\text{dipeptide}} - \sum_{i=1}^{n-1} E_i^{\text{Ace-Nme}} \tag{1}$$

in which n is the number of amino acids or dipeptides.

The force for atoms in the same dipeptide and Ace-Nme is calculated following equation (2).

$$F_i^{\text{prot\_units}} = \sum_{j=1}^{m} F_{ij}^{\text{dipeptide}} - \sum_{j=1}^{n} F_{ij}^{\text{Ace-Nme}} \tag{2}$$

in which i denotes the atom for force calculation, m represents all of the dipeptides the atom i belongs to, n represents all of the Ace-Nme fragments the atom i belongs to, and j represents any other atom that coexists with atom i in the same dipeptide or Ace-Nme.

We further complemented the extra interactions among non-overlapped protein units. Supplementary Fig. 18 shows the extra interactions among different protein units of tetrapeptide. Two parts of interactions were not calculated. Extended Data Fig. 8a illustrates the extra interactions between the group of CH^1, C^1, O^1 and NH^1 (outlined in purple) and the last part of the tetrapeptide (also outlined in purple). Furthermore, Extended Data Fig. 8b exhibits extra interactions between the beginning part of the tetrapeptide that includes $CH3^0$, C^0, O^0 and NH^1 (outlined in brown) and another part beginning from the second side chain to the C terminus (also outlined in brown).

Considering that the interactions in such non-overlapped regions are dominated by electrostatic force and van der Waals interactions, we used the Coulomb equation and the Lennard-Jones potential to describe them. Then, we used the corresponding parameters derived from the Amber ff19SB force field[21] and the distance between atoms to calculate the potential energy and atomic forces for the extra interactions (equations (3) and (4)). The selection of ff19SB was informed by its superior performance in evaluating the relative energy when compared to another widely used force field, CHARMM36 (ref. 45), as illustrated in Supplementary Fig. 11.

$$E^{\text{prot}} = E^{\text{prot\_units}} + \sum_{\substack{i=1 \\ i\in A}}^{n-1} \sum_{\substack{j=i+1 \\ j\notin A}}^{n} E_{ij}^{\text{Coulomb}} + \sum_{\substack{i=1 \\ i\in A}}^{n-1} \sum_{\substack{j=i+1 \\ j\notin A}}^{n} E_{ij}^{\text{VDW}} \tag{3}$$

$$F_i^{\text{prot}} = F_i^{\text{prot\_units}} + \sum_{\substack{j=i+1 \\ j\notin A}}^{n} F_{ij}^{\text{Coulomb}} + \sum_{\substack{j=i+1 \\ j\notin A}}^{n} F_{ij}^{\text{VDW}} \tag{4}$$

in which the energy and force with the superscript 'units' represent the values obtained from equation (1) to equation (2), and A denotes the atom set in the corresponding unit. To avoid double counting, the sum traverses all of the atoms with the index after the current atom.

Given the protein fragmentation approach, all proteins can be converted into 21 kinds of protein unit (that is, 20 kinds of dipeptide and another Ace-Nme), which substantially reduced the number of specific types of protein unit, facilitated dataset construction and model training, contributed to exploring the whole conformational space, avoided holes in the potential energy surface, and thus improved the generalization, efficiency and robustness of the MD simulation.

Protein unit dataset

The training dataset for the AI^{2}BMD potential was generated through the following protocols. First, the 'Sequence' command in the tleap module of AmberTools20 (ref. 46) was used to generate the topology and coordinate files for the initial 20 kinds of dipeptide and Ace-Nme. Then, ϕ (that is, the dihedral of C–N–Cα–C) and ψ (that is, the dihedral of N–Cα–C–N) were two-dimensionally scanned over ranges of −180° to 175° with an interval of 5°. For the proline dipeptide, the ϕ dihedral was refined from −180° to 120° owing to its ring conformation. The rotation of the dihedral was accomplished by the 'rotatedihedral' command in CPPTRAJ[47]. For each non-proline dipeptide, 5,184 anchors were generated. For Ace-Nme, scanning over ranges of −180° to 175° with an interval of 5° was applied on the axis of C of Ace and N of Nme resulting in 72 anchors.

Each anchor first encountered a geometry optimization ('GO') process to obtain a reasonable structure. The solvation model density (SMD) solvent model was used during GO. The ϕ and ψ dihedrals were also constrained the same as for anchor generation. For each anchor, the last structure of the GO process was used as the input structure for AIMD simulations. SMD was used to sample conformations by taking the solvent effect in QM into consideration. For dipeptides, 225-fs simulations were applied for each anchor, and the last 200-fs structures were extracted. Simulations of 2,025 fs were carried out for Ace-Nme, and the last 2,000 fs was extracted for each trajectory. As an explicit solvent was used during MD simulation driven by our AI^{2}BMD potential, after AIMD simulations, we recalculated the single-point energy and forces for all extracted conformations without SMD, which were used for MLFF training.

During GO, AIMD simulations and single-point energy calculation, DFT with the M06-2X density functional with the 6-31g* basis set were used[48]. This basis set and functional are generally suitable for biomolecular sampling[14,15,49,50]. We set tight convergence conditions in the simulation processes, and convergence was mandatory for the next calculation step. Systems encountered a canonical sampling through velocity rescaling thermostat at 290 K (ref. 51). Such simulations were performed by ORCA 5.0.1 (ref. 52). The charge of each system was set according to the charge for the sum of all amino acids at pH 7. The GO, AIMD simulations and single-point energy calculation took about 12,928,993 central processing unit (CPU) core hours (1,476 CPU core years) for calculation. As a result, 1,036,800 conformations were

sampled and calculated at the DFT level for each kind of dipeptide and 144,000 conformations were sampled for Ace-Nme. The distributions of energy and the norm of force for each kind of protein unit are shown in Supplementary Tables 4 and 5 and Supplementary Figs. 12 and 13. The whole protein unit dataset consists of 20 million conformations that comprehensively captured the conformational space of the protein units and provided a solid guarantee for machine learning potential training and AI^{2}BMD simulation.

ViSNet as AI^{2}BMD potential

ViSNet is a versatile geometric deep learning model[7,53] that can predict potential energy and atomic forces, as well as various quantum chemical properties, by taking atomic coordinates and atomic numbers as inputs. As shown in Supplementary Fig. 2a, the ViSNet model is composed of an embedding block and multiple stacked ViSNet blocks, followed by an output block. The atomic number and coordinates are fed into the embedding block followed by ViSNet blocks to extract and encode geometric representations. The geometric representations are then used to predict molecular energy and force through the output block. Supplementary Fig. 2b demonstrates the ViSNet block, which consists of a message block and an update block. These blocks work together as parts of a vector scalar interactive message-passing mechanism, referred to as ViS-MP. The rich geometric information passed via ViS-MP is extracted by the runtime geometric calculation module with linear complexity. The operations in Supplementary Fig. 2b can be summarized as follows:

$$m_i^l = \sum_{j\in\mathcal{N}(i)} \phi_m^s\left(h_i^l, h_j^l, f_{ij}^l\right) \tag{5}$$

$$\mathbf{m}_i^l = \sum_{j\in\mathcal{N}(i)} \phi_m^v\left(m_{ij}^l, \mathbf{r}_{ij}, \mathbf{v}_j^l\right) \tag{6}$$

$$h_i^{l+1} = \phi_{un}^s\left(h_i^l, m_i^l, \left\langle \mathbf{v}_i^l, \mathbf{v}_i^l \right\rangle\right) \tag{7}$$

$$f_{ij}^{l+1} = \phi_{ue}^s\left(f_{ij}^l, \left\langle \mathrm{Rej}_{\mathbf{r}_{ij}}(\mathbf{v}_i^l), \mathrm{Rej}_{\mathbf{r}_{ji}}(\mathbf{v}_j^l) \right\rangle\right) \tag{8}$$

$$\mathbf{v}_i^{l+1} = \phi_{un}^v\left(\mathbf{v}_i^l, m_i^l, \mathbf{m}_i^l\right) \tag{9}$$

in which h_i^l represents the scalar feature of node i in the lth layer, $\mathbf{v}_i^l$ represents the vectorized node feature and f_{ij} represents the scalar edge feature between node i and node j. ϕ_m^s, ϕ_m^v are nonlinear message functions to transform messages from neighbours and $\phi_{un}^s, \phi_{ue}^s, \phi_{un}^v$ are nonlinear update functions to update the corresponding feature according to the message and geometric features. More details about runtime geometric calculation and ViS-MP can be found in ref. 7.

For each kind of protein unit, ViSNet was trained as an energy-conserving potential model; that is, the predicted atomic forces were derived from the negative gradients of the potential energy with respect to the atomic coordinates. We randomly split each protein unit dataset into a training set, a validation set and a test set with the ratio of 8:1:1. Hyperparameters were tuned on the validation set of the alanine dipeptide and directly applied to other protein units. Concretely, all ViSNet models trained for protein units were relatively light with only 6 hidden layers and 128 embedding dimensions for node and edge representations. To better capture geometric information, we expanded the raw three-dimensional coordinates of molecules by adapting higher-order spherical harmonics[54]. The cutoff of the edge connection was set to 5 Å for all protein units, and the maximum number of neighbours for each atom was 32. We leveraged a combined mean squared error loss for energy and force training with the weight of 0.05 and 0.95, respectively. We adopted a learning rate of 2×10^{-4} with 1,000 warm-up steps[55] using the AdamW optimizer[56]. The learning rate decays if the validation loss stopped decreasing. The patience was set to 15 epochs, and the decay factor was set to 0.8. We also adopted an early-stopping strategy to prevent over-fitting[57]. The maximum number of epochs was set to 6,000, and the early-stopping patience was 150 epochs. All models were trained on a GPU cluster with 16 NVIDIA 32G-V100 GPUs per cluster node, and the batch size was 64 or 128 per GPU according to the size of the protein units. To make the model converge better, for the training set, we subtracted the sum of atomic reference energies from the total energy and then normalized them with Z-score normalization. More details on the hyperparameters of ViSNet can be found in Supplementary Table 6.

AI^{2}BMD simulation program

To carry out simulations with the AI^{2}BMD potential, we designed an AI-driven MD simulation program based on the atomic simulation environment[58]. Extended Data Fig. 9 illustrates the overview block diagram of the program. On program start, the initial protein structure is fed into the preprocessing module, where the solvent and ions are added, and the structure is relaxed. The entire simulation system is then sent into the MD loop, the main logic component. For each iteration in the MD loop, the protein is first decomposed into fragments by the protein fragmentation module and then partitioned by the work scheduler. The partitioning scheme is dictated by a tunable device strategy, and a user can choose to, depending on the size of the simulation system, instruct the work scheduler to maximize the utilization of all the GPUs by oversubscribing, and to reduce the memory pressure on a particular device by balancing the computation on different fragments across the GPU cards. The partitioned fragments and the solvent atoms are then asynchronously sent to different computation servers running in separate processes. This asynchronous client–server paradigm helps to alleviate a substantial limitation in the Python runtime: that only a single thread can execute Python code at a time in the same process. After the workload is distributed from the main component to the computation servers, it will be processed in parallel, and the main Python process can immediately resume processing other tasks such as persisting trajectory data, without being blocked by the servers.

Considering that cloud computing is a popular and cost-efficient way to support scientific computing workloads, we designed the simulation process to be cloud-oriented. The software configuration is fully defined with a Docker image and remains invariant across different machines, which allows us to not only effortlessly deploy the software system to the cloud, but also fine-tune the program against a fixed set of supporting libraries. As cloud-based machines may be pre-empted, and the machine-local storage is volatile during a long-time simulation, we implemented a job scheduling component that periodically persists the computation results to cloud-based storage and resumes the simulation.

System configuration in simulation

We prepared the biomolecular systems using the Amber20 package with the AMOEBA 13 force field[16]. The protein was first solvated in a cubic TIP3P[59] water box and then was relaxed in energy minimization cycles. Then, NaCl atoms as counterions and another 0.15 mol l^{-1} buffer were added. We used classical Amber Coulombic potential-based methods to add ions. Initially, a grid of 1 Å bin size was generated, and all grids point Coulombic potentials were calculated. Then, the ions were placed on the grid where the contrast types of Coulombic potential were the highest. If an ion had a steric conflict with a solvent molecule, the ion was moved to the centre of that solvent molecule, and the latter was removed.

We adopted a hybrid calculation strategy for the simulation system; that is, the proteins were calculated by the AI^{2}BMD potential with ab initio accuracy, whereas the AMOEBA 13 force field was used to deal

with the solvent. The total energy of the system (E^{total}) is computed as the sum of the deep learning (DL) energy calculated by ViSNet (E_{DL}^{prot}) for the protein, and the energy from the MM calculation for the entire system ($E_{\text{MM}}^{\text{total}}$). Then, to avoid double counting the energy contribution from the protein, the energy of protein atom interactions ($E_{\text{MM}}^{\text{prot}}$) is subtracted from the total energy as shown in the following equation. Such calculations were based on the classical integrated molecular orbital + MM model implanted in the atomic simulation environment package[60,61].

$$E^{\text{total}} = E_{\text{DL}}^{\text{prot}} + E_{\text{MM}}^{\text{total}} - E_{\text{MM}}^{\text{prot}} \tag{10}$$

Similarly, the force F_i^{total} for atom i is initially set as the forces from the interactions between atom i and all other atoms in the protein (F_i^{prot}), which is depicted in equation (4). To account for the solvent effect, an additional force was calculated between atom i and the other atoms in the system by the AMOEBA force field (the second item in equation (11)), and then the solute items were subtracted (the third item in equation (11)).

$$F_i^{\text{total}} = F_i^{\text{prot}} + \sum_{\substack{j\neq i \\ j\in B}}^{n} F_{ij} - \sum_{\substack{j\neq i \\ j\in C}}^{n} F_{ij} \tag{11}$$

in which B represents all atoms in the entire system, and C represents the atoms in the solute. Furthermore, to calculate $E_{\text{DL}}^{\text{prot}}$ and $F_{\text{DL}}^{\text{prot}}$, we first split the protein into protein units, calculated the potential energies and atomic forces by ViSNet models and then combined all protein units by equation (3). More details about the protein fragmentation and ViSNet potential calculation can be found in the above sections. A simulation carried out under an NVE ensemble demonstrates the conserved total energy and the counterbalanced forces, thereby further substantiating the validity of subsequent sampling procedures (Supplementary Figs. 14 and 15). Furthermore, we also carried out the simulation for the same arginine dipeptide under an NVT ensemble and calculated the heat capacity by the following equation:

$$C_{\text{V}} = \left(\frac{\partial U}{\partial T}\right)_{\text{V}} = \frac{\langle E^2\rangle - \langle E\rangle^2}{k_{\text{B}}\langle T\rangle^2} \tag{12}$$

in which $\langle E^2\rangle$ denotes the ensemble average of the square value of the system energy and $\langle E\rangle^2$ denotes the square value of the ensemble average of the system energy. The heat capacity values made by MM and AI^{2}BMD are 0.052 kcal mol^{-1} K^{-1} and 0.053 kcal mol^{-1} K^{-1}, which are similar and comparable to those of previous experimental studies. The subsequent simulations in this study were run in the Berendsen NVT ensembles with initial velocities randomly drawn from a Maxwell–Boltzmann distribution. The time step in this study was set to 1 fs. During the simulation, the trajectory would be written to a high-precision XYZ file.

Simulation details

In the evaluation of protein energy and force calculation, protein structures (Protein Data Bank (PDB) IDs: chignolin, 5AWL; Trp-cage, 2JOF; WW domain, 2F21; albumin-binding domain, 1PRB; PACSIN3, 6F55; SSO0941, 5VFK; APC, 5IZA; polyphosphate kinase, 1XDO; aminopeptidase N, 4XN9) were solvated in a generalized Born implicit solvent model. The alteration on the WW domain follows the GTT mutation in the previous study[4], and the first five flexible residues in the albumin-binding domain were removed. The Amber program makeCHIR_RST was used to create chiral restraint files during replica-exchange molecular dynamics (REMD) simulation to preserve chiral properties at high temperatures. After 1,000 steps of minimization, equilibration runs of 200 ps were conducted at temperatures ranging from 300 K to 1,000 K with a stride of 100 K. The final equilibrated structures were used for REMD simulations at the corresponding temperatures. During the simulation, each replica ran for 2 ps before exchanging with neighbouring temperatures and 5,000 exchanges occurred in each production run. REMD trajectories were divided into three states according to the Cα RMSD against the crystal structure. Specifically, for chignolin, the folding structures have an RMSD of 0–2.5 Å, the intermediate structures have an RMSD of 2.5–7.5 Å, and the unfolding structures have an RMSD of more than 7.5 Å. For other proteins, the ranges of the three states are 0–5 Å, 5–15 Å and >15 Å. Then, folding and unfolding states were further divided into 5 clusters, and the intermediate structures were divided into 10 clusters via the CPPTRAJ 'cluster' program. We picked the structures of each cluster centre, accumulating 20 initial structures in total. Finally, each initial structure was solvated in a 5-Å TIP3P water box and encountered 10 steps of 1-fs AI^{2}BMD simulation. Simulations were carried out under an NVT ensemble. The simulation temperature (300 K) was controlled by a Berendsen thermostat and τ was 10 fs. The reference energy and force of the corresponding structures were calculated at the M06-2X–6-31g* level. MM energy and force were calculated by the ff19SB force field.

For sampling on Ace-N-Nme, we constructed the system using the 'sequence' command in tleap, and then applied a 10-ns REMD simulation, identical to the one used for protein sampling. From this, we extracted 50 representative structures using the CPPTRAJ 'cluster' program. We then conducted 10-ps simulations for each initial structure using AI^{2}BMD with AMOEBA polarizable embedding, resulting in a cumulative sampling time of 500 ps. We also implemented 10-ps simulations using QM–MM with AMOEBA polarizable embedding and MM with Amber ff19SB on these conformations. Each simulation incorporated a water box of 5 Å. We then examined each snapshot during the simulations to locate any water molecules that formed a hydrogen bond with the main chain or side chain (criteria: frequency >90%, donor atom distance <3.5 Å, O–H–O angle >150°). Subsequently, we delineated the distribution of donor atom distances. Following the formation of one hydrogen bond, we isolated the water and Ace-N-Nme molecules and incrementally pulled the water from 2.5 Å to 4.0 Å to form 150 structures. Finally, we carried out single-point energy evaluation on the system of the water molecule and the dipeptide by QM at the M06-2X–6-31g* level, AI^{2}BMD with AMOEBA solvent and MM with ff19SB.

For AI^{2}BMD simulation on dipeptides, we first generated the conformations of the dipeptides through the 'sequence' command in tleap. Then, the dipeptides were solvated in a 5-Å TIP3P water box, and we ran two repetitive 500-ns classical MD simulations under the ff19SB force field for sufficient sampling. k-means clustering was then applied, and 50 representative structures were picked up. Starting from the representative structures, we carried out 10-ns AI^{2}BMD simulation for the negatively charged protein unit Ace-E-Nme, the positively charged Ace-R-Nme, Ace-F-Nme with a benzene ring in the side chain and Ace-S-Nme with a smaller side chain solvated by a 10-Å water box under an NVT ensemble. Furthermore, for J coupling analysis, 2 independent runs of 1-ns AI^{2}BMD simulations were used, and 10,000 snapshots were saved. Then, ϕ values were estimated from each snapshot. The $^3J(H_N, H_\alpha)$ coupling value was calculated through equation (13).

$$J = 7.09\cos^2(\phi - 60°) - 1.42\cos(\phi - 60°) + 1.55 \tag{13}$$

For AI^{2}BMD simulation on chignolin, we first aligned the structures in a 106-μs comprehensively sampled trajectory to the initial structure. Then, time-lagged independent component analysis was used on raw atom coordinates for projecting the free-energy landscape to a six-dimensional super surface[62]. On the basis of minibatch k-means algorithm, we clustered all conformations and then picked up 60 folded and unfolded structures as the representative structures. Then for each structure, we ran 10-ns AI^{2}BMD and 10-ns MM simulations.

In the Ramachandran plot, ϕ is the dihedral angle determined by C_{n-1}, N_n, $C\alpha_n$ and C_n, and ψ is the dihedral angle determined by N_n, $C\alpha_n$,

C_n and N_{n+1}. The subscript represents the index of a residue in a protein. The energy was estimated according to a Boltzmann distribution based on the density of points in each bin. This estimation was carried out using the potential of mean force. ϕ and ψ were set as two reaction coordinates (x, y). The potential of mean force values were calculated using equation (14).

$$\Delta G(x, y) = k_B T \ln g(x, y) \tag{14}$$

in which k_B represents the Boltzmann constant, T is the temperature of systems (300 K) and $g(x, y)$ represents the normalized joint probability distribution. The free-energy value presented in the plot represents a relative energy value, computed by deducting the minimum free-energy value from the observed value. The Q score was calculated through equation (15) (ref. 24).

$$Q = \frac{1}{N} \sum_{(i,j)} \frac{1}{1 + \exp[5(r_{ij}(X) - 1.8\, r_{ij}^0)]} \tag{15}$$

Native contacts were defined as any pairs of heavy atoms of two residues separated by at least three residues and the distance of which is smaller than 4.5 Å in the native conformation. Equation (14) sums N pairs of native contacts in the crystal structure; r_{ij}^0 is the distance between heavy atom i and atom j of native contacts in the crystal structure, $r_{ij}(X)$ is the distance between atom i and atom j in the conformation X. The thresholds of Q values for folded and unfolded structures were set to >0.82 and <0.03, respectively[24].

For free-energy estimation for fast-folding proteins, we first evenly sampled 100,000 points in the simulation trajectories of ref. 4. Folded and unfolded states were classified by the same thresholds of Q values in the previous study[24]. Structures in the folded state were clustered into 10 clusters. The RMSD values were calculated on the basis of Cα coordinates according to the 'rmsd' method in MDTraj. ΔG, the free energy for the folding process, was calculated according to the ratio between the folded and unfolded structures. Using the re-evaluated energy for each conformation, we determined the folding enthalpy and the heat capacity change for protein folding. The melting temperature was extrapolated from the calculated ΔG, folding enthalpy and the heat capacity change.

For the calculation of changes in enthalpy and heat capacity during protein folding and unfolding, 110-residue barnase (PDB: 1A2P) and 84-residue CI2 (PDB: 2CI2) were selected for evaluation with enthalpy and heat capacity values measured by differential scanning calorimetry and spectroscopy[37,38]. For each protein, besides the folded structure derived from PDB, 20 unfolded structures were also generated for simulation. Following the previous study[36], each conformation was explicitly solvated by a 10-Å water box. For barnase, 20 parallel simulations starting from the folded structure and 20 simulations starting from the unfolded structures were performed by GROMACS 2018 with the CHARMM36 force field at pH 4.1 and at temperatures of 295 K, 315 K and 335 K. The same settings were applied for CI2, except the simulations were carried out at pH 6.3 and temperatures of 335 K, 350 K and 365 K. Each system configuration above was conducted 2-ns simulation under an NPT ensemble. Potential energy values of conformations sampled from simulations were calculated by AI^{2}BMD. The enthalpy change following thermal unfolding (ΔH) was calculated as the difference between the averaged enthalpy of the unfolded ensemble and that of the folded ensemble. We then conducted linear regression to determine the change in heat capacity (ΔC_p) from the slope, as well as the enthalpy change at the melting temperature. Additionally, we also estimated the folding free energy using the Gibbs–Helmholtz equation.

In the pK_a determination using thermodynamics integration, we initially reweighted for all data points in the simulation trajectories provided by The Amber Project (https://ambermd.org/tutorials/advanced/tutorial6/index.php). Subsequently, we focused on the trajectories' converged sections, selecting 2,500 data points per window for the dipeptide and 500 data points per window for thioredoxin to calculate the mean energy values. ΔG was computed using the integral:

$$\Delta G = \int \frac{\partial U}{\partial \lambda} d\lambda = \sum_{\lambda} w_{\lambda} \frac{\partial U}{\partial \lambda} \tag{16}$$

in which w_λ represents the window width, U denotes the internal energy, and λ specifies the sampling window. This approach encapsulates the free-energy variation across different protonation states, facilitating the accurate computation of the pK_a value.

Reporting summary

Further information on research design is available in the Nature Portfolio Reporting Summary linked to this article.

Data availability

The protein unit dataset built for this study is available via GitHub at https://github.com/microsoft/AI2BMD. The proteins for energy and force evaluations in this work are available from PDB (https://www.rcsb.org) with the following accession numbers: chignolin: 5AWL; Trp-cage: 2JOF; WW domain: 2F21; albumin-binding domain: 1PRB; PACSIN3: 6F55; SSO0941: 5VFK; APC: 5IZA; polyphosphate kinase: 1XDO; aminopeptidase N: 4XN9; barnase: 1A2P; and CI2, 2CI2. The proteins for melting temperature estimation are available at PDB with the following accession numbers: α3D: 2A3D; BBA: 1FME; NTL9: 2HBA; protein G: 1MIO; WW domain: 2F21; λ-repressor: 1LMB; and homeodomain: 2P6J. The simulation trajectories for pK_a analysis are available via The Amber Project at https://ambermd.org/tutorials/advanced/tutorial6/index.php. Source data are provided with this paper.

Code availability

The source code and the model checkpoints for the AI^{2}BMD simulation program are available via GitHub at https://github.com/microsoft/AI2BMD.

43. Xu, M., He, X., Zhu, T. & Zhang, J. Z. H. A fragment quantum mechanical method for metalloproteins. *J. Chem. Theory Comput.* **15**, 1430–1439 (2019).
44. Liu, D. C. & Nocedal, J. On the limited memory BFGS method for large scale optimization. *Math. Program.* **45**, 503–528 (1989).
45. Huang, J. & MacKerell, A. D. Jr CHARMM36 all-atom additive protein force field: validation based on comparison to NMR data. *J. Comput. Chem.* **34**, 2135–2145 (2013).
46. Case, D. A. et al. The Amber biomolecular simulation programs. *J. Computat. Chem.* **26**, 1668–1688 (2005).
47. Roe, D. R. & Cheatham, T. E. 3rd PTRAJ and CPPTRAJ: software for processing and analysis of molecular dynamics trajectory data. *J. Chem. Theory Comput.* **9**, 3084–3095 (2013).
48. Zhao, Y. & Truhlar, D. G. 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**, 215–241 (2008).
49. Xu, Z., Zhang, Q., Shi, J. & Zhu, W. Underestimated noncovalent interactions in Protein Data Bank. *J. Chem. Inf. Model.* **59**, 3389–3399 (2019).
50. Wang, T., He, X., Li, M., Shao, B. & Liu, T.-Y. AIMD-Chig: exploring the conformational space of a 166-atom protein Chignolin with ab initio molecular dynamics. *Sci. Data* **10**, 549 (2023).
51. Bussi, G., Donadio, D. & Parrinello, M. Canonical sampling through velocity rescaling. *J. Chem. Phys.* **126**, 014101 (2007).
52. Neese, F., Wennmohs, F., Becker, U. & Riplinger, C. The ORCA quantum chemistry program package. *J. Chem. Phys.* **152**, 224108 (2020).
53. Wang, Y. et al. An ensemble of VisNet, Transformer-M, and pretraining models for molecular property prediction in OGB Large-Scale Challenge @ NeurIPS 2022. Preprint at https://arxiv.org/abs/2211.12791 (2022).
54. Müller, C. *Spherical Harmonics* Vol. 17 (Springer, 2006).
55. Goyal, P. et al. Accurate, large minibatch sgd: training imagenet in 1h. Preprint at https://arxiv.org/abs/1706.02677 (2017).
56. Loshchilov, I. & Hutter, F. Decoupled weight decay regularization. Preprint at https://arxiv.org/abs/1711.05101 (2017).
57. Yao, Y., Rosasco, L. & Caponnetto, A. J. C. A. On early stopping in gradient descent learning. *Constr. Approx.* **26**, 289–315 (2007).

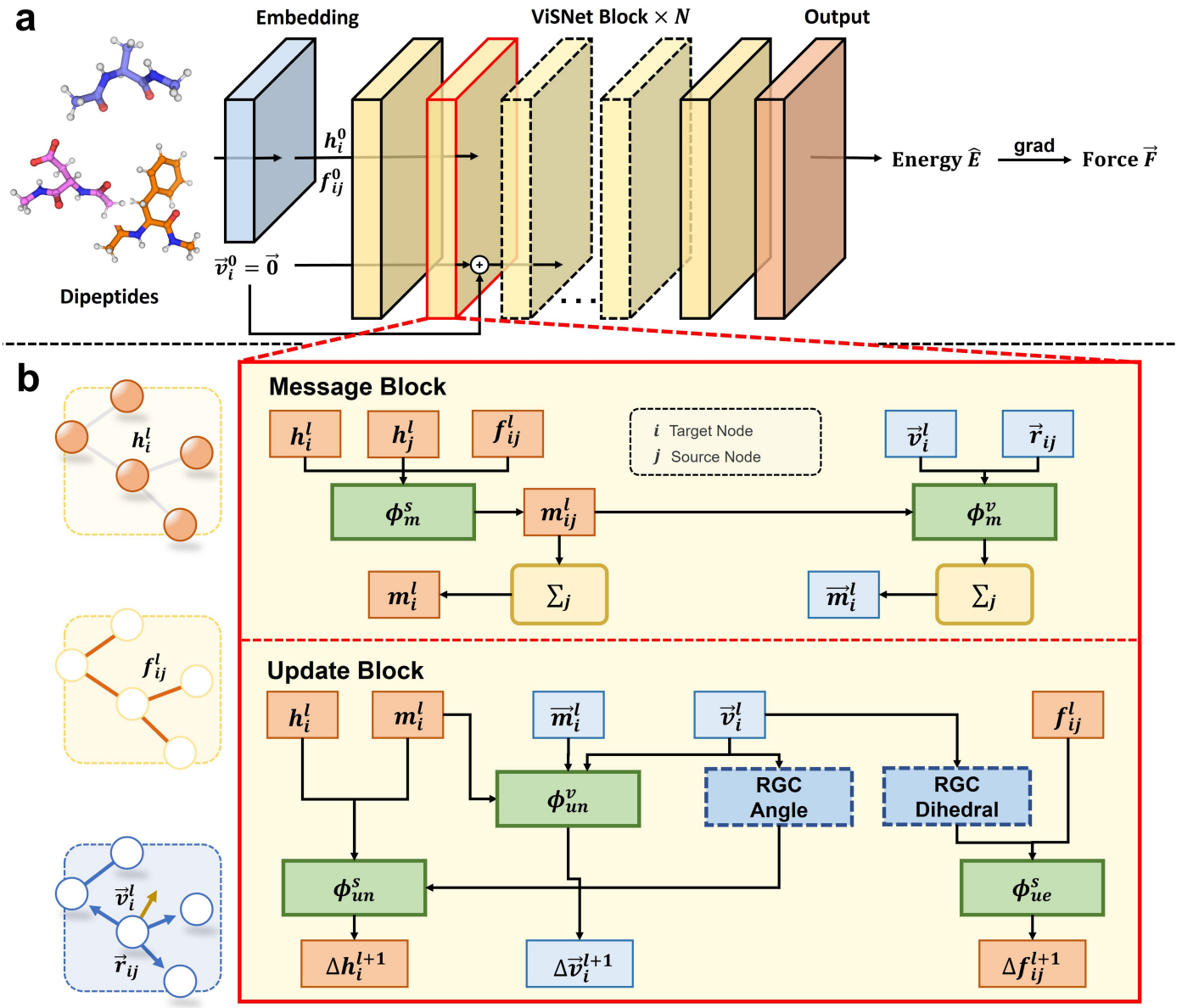
58. Hjorth Larsen, A. et al. The atomic simulation environment—a Python library for working with atoms. *J. Phys. Condens. Matter* **29**, 273002 (2017).
59. Jorgensen, W. L., Chandrasekhar, J., Madura, J. D., Impey, R. W. & Klein, M. L. Comparison of simple potential functions for simulating liquid water. *J. Chem. Phys.* **79**, 926–935 (1983).
60. Svensson, M. et al. ONIOM: a multilayered integrated MO + MM method for geometry optimizations and single point energy predictions. A test for Diels–Alder reactions and $Pt(P(t\text{-}Bu)_3)_2 + H_2$ oxidative addition. *J. Phys. Chem.* **100**, 19357–19363 (1996).
61. Chung, L. W. et al. The ONIOM method and its applications. *Chem. Rev.* **115**, 5678–5796 (2015).
62. Gong, S. et al. Stochastic lag time parameterization for Markov state models of protein dynamics. *J. Phys. Chem. B* **126**, 9465–9475 (2022).

Acknowledgements X.H., M.L., Y.W., C.C., X.S., J.M., H.Z. and S.L. performed the work as part of their internship at Microsoft Research Beijing, China. We acknowledge N. Baker, F. Noe, H. Gong, J. Ponder and W. Im for suggestions and discussions.

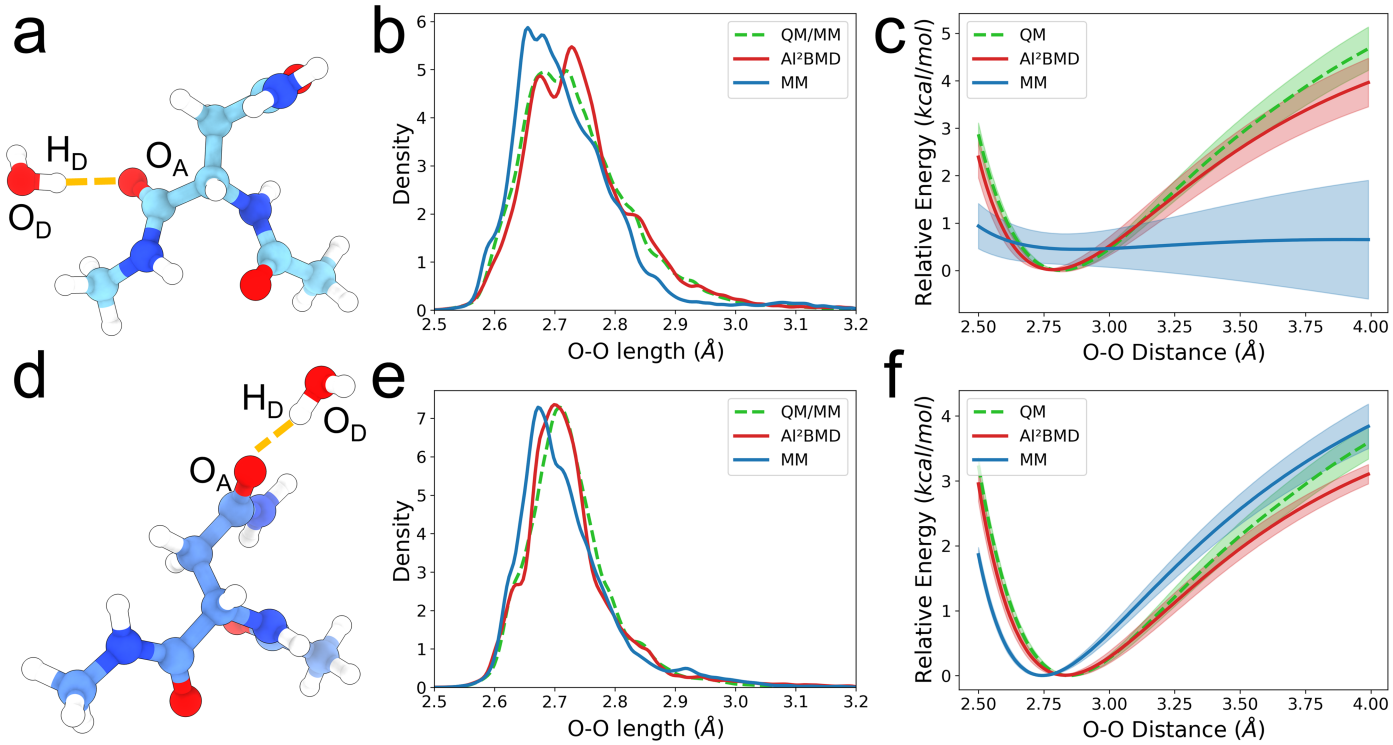
Author contributions T.W. is the primary corresponding author. T.W. (primary) and B.S. led this study. T.W. (primary), B.S. and T.-Y.L. conceived the Al^{2}BMD project. T.W. conceived and designed the end-to-end approach. T.Y.L. advised on this study. T.W., Y.L. and R.B. designed, accelerated, optimized and deployed the Al^{2}BMD simulation program. T.W. designed the generalizable protein fragmentation approach. R.B., X.H. and M.L. implemented and optimized the generalizable protein fragmentation approach. M.L. and X.H. built the protein unit dataset. T.W., Y.W. and S.L. (in the early stage) designed and trained the Al^{2}BMD potential. C.C., X.H. and M.L. implemented the polarizable solvent. X.H. and M.L. evaluated the accuracy of energy and force calculations. T.W., Y.L. and R.B. evaluated the time consumptions of energy and force calculations. Y.L. and X.S. conducted and analysed Al^{2}BMD simulations for dipeptides. M.L. analysed the *J* coupling values. Y.L. and X.S. conducted Al^{2}BMD simulations for chignolin. X.S. analysed energy and force of chignolin simulations. J.M. analysed the RMSD and Ramachandran plot of chignolin simulations. X.H. and M.L. analysed the distance distributions from chignolin simulations. X.S., H.Z. and T.W. conducted simulations and thermodynamic property analysis for two-state proteins. M.L. analysed thermodynamic properties for fast-folding proteins. X.H. conducted pK_a analysis. Z.W. participated in discussions in the early stage. T.W. wrote the manuscript. B.S., H.L. and other authors revised the manuscript. All authors approved the final version of the manuscript.

Competing interests The authors declare no competing interests.

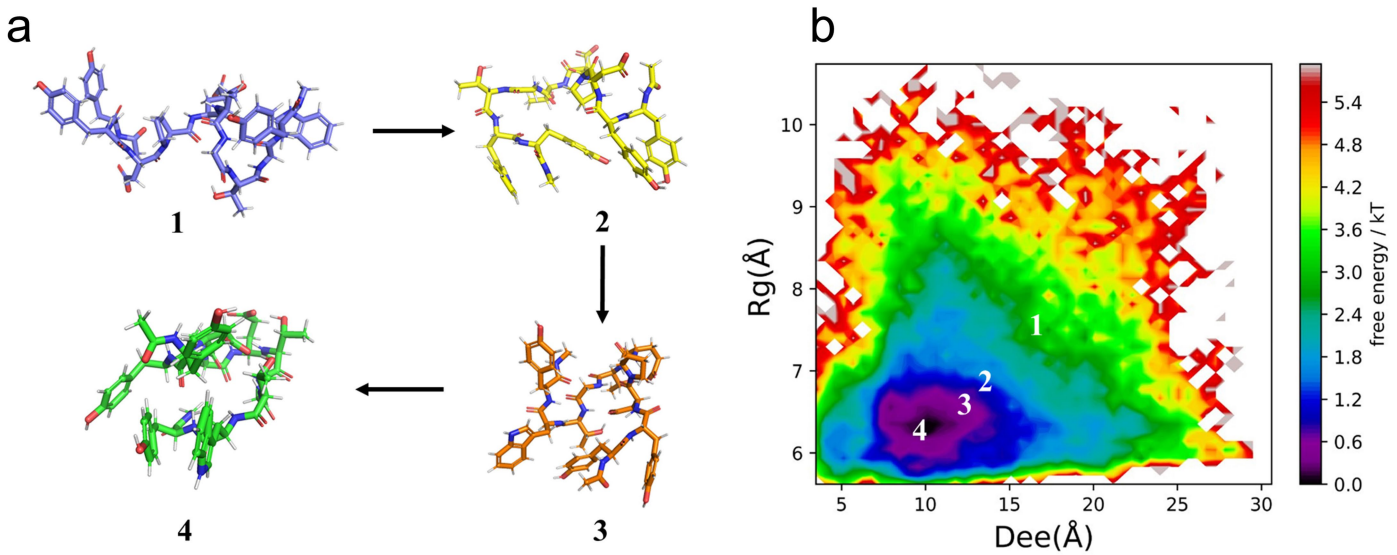
Additional information
Supplementary information The online version contains supplementary material available at https://doi.org/10.1038/s41586-024-08127-z.
Correspondence and requests for materials should be addressed to Tong Wang or Bin Shao.
Peer review information *Nature* thanks Yaoqi Zhou and the other, anonymous, reviewer(s) for their contribution to the peer review of this work.



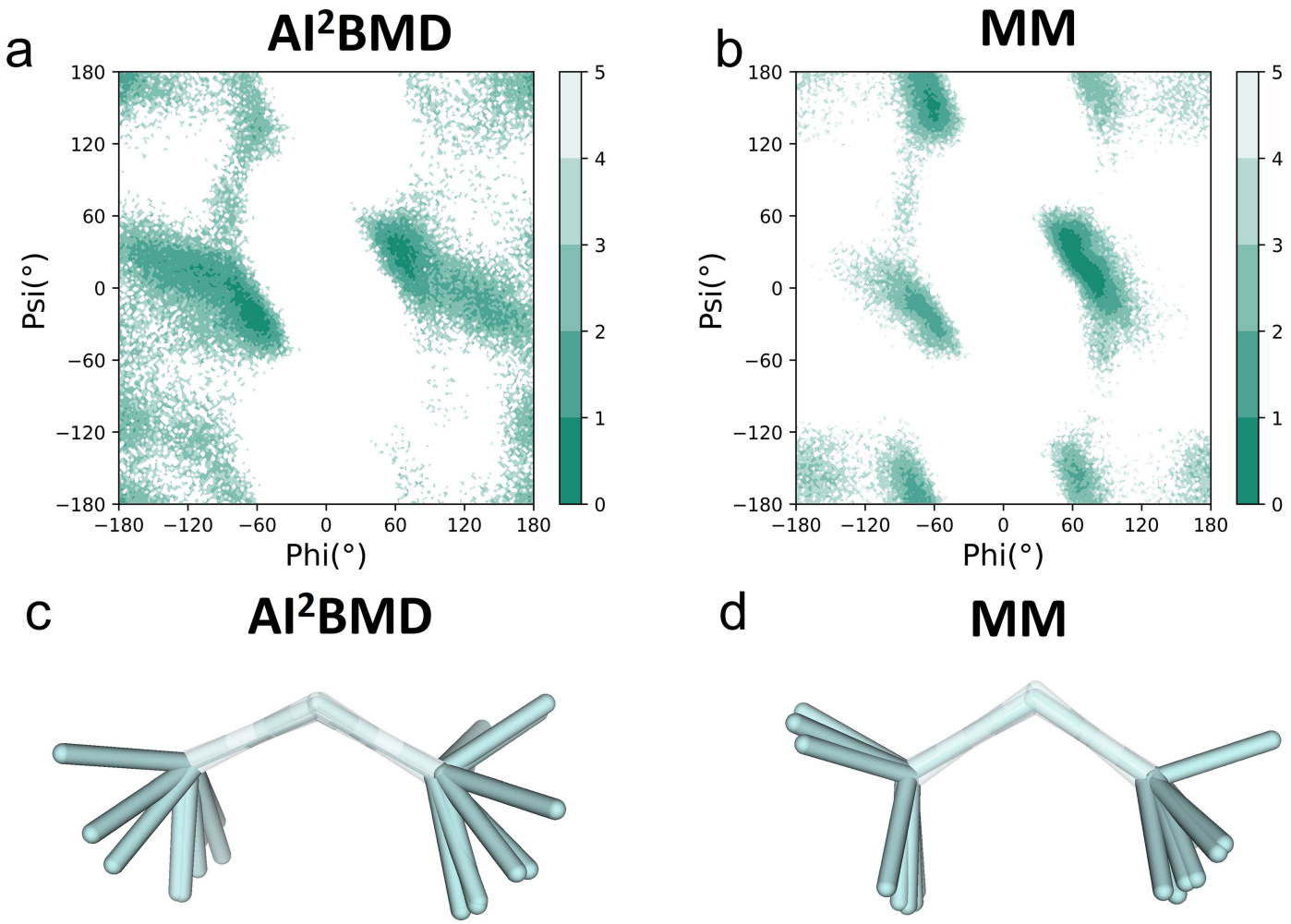
Extended Data Fig. 1 | The architecture of ViSNet. (a) The sketch of ViSNet. The 3D structures of dipeptides are embedded and extracted by an embedding block and multiple stacked ViSNet blocks. The energies are predicted through an output block, and the forces are derived from the negative gradients of the potential energy with respect to the atomic coordinates. (b) The key operations in ViSNet block. A ViSNet block consists of a message block and an update block. Concretely, in the message block, the scalar messages m_{ij} and vector messages $\vec{m}_{ij}$ are first obtained through message functions ϕ_m^s and ϕ_m^v from node scalar feature h_i, edge scalar feature f_{ij}, relative position $\vec{r}_{ij}$, and node vector feature $\vec{v}_i$. Then, they are aggregated to the target node i. In the update block, h_i is updated by the aggregated scalar message m_i and the output of runtime geometry calculation (RGC)-Angle from through an update function ϕ_{un}^s. Then, f_{ij} is updated by the output of RGC-Dihedral and through an update function ϕ_{ue}^s. Finally, $\vec{v}_i$ is updated by both scalar and vector messages through an update function ϕ_{un}^v.



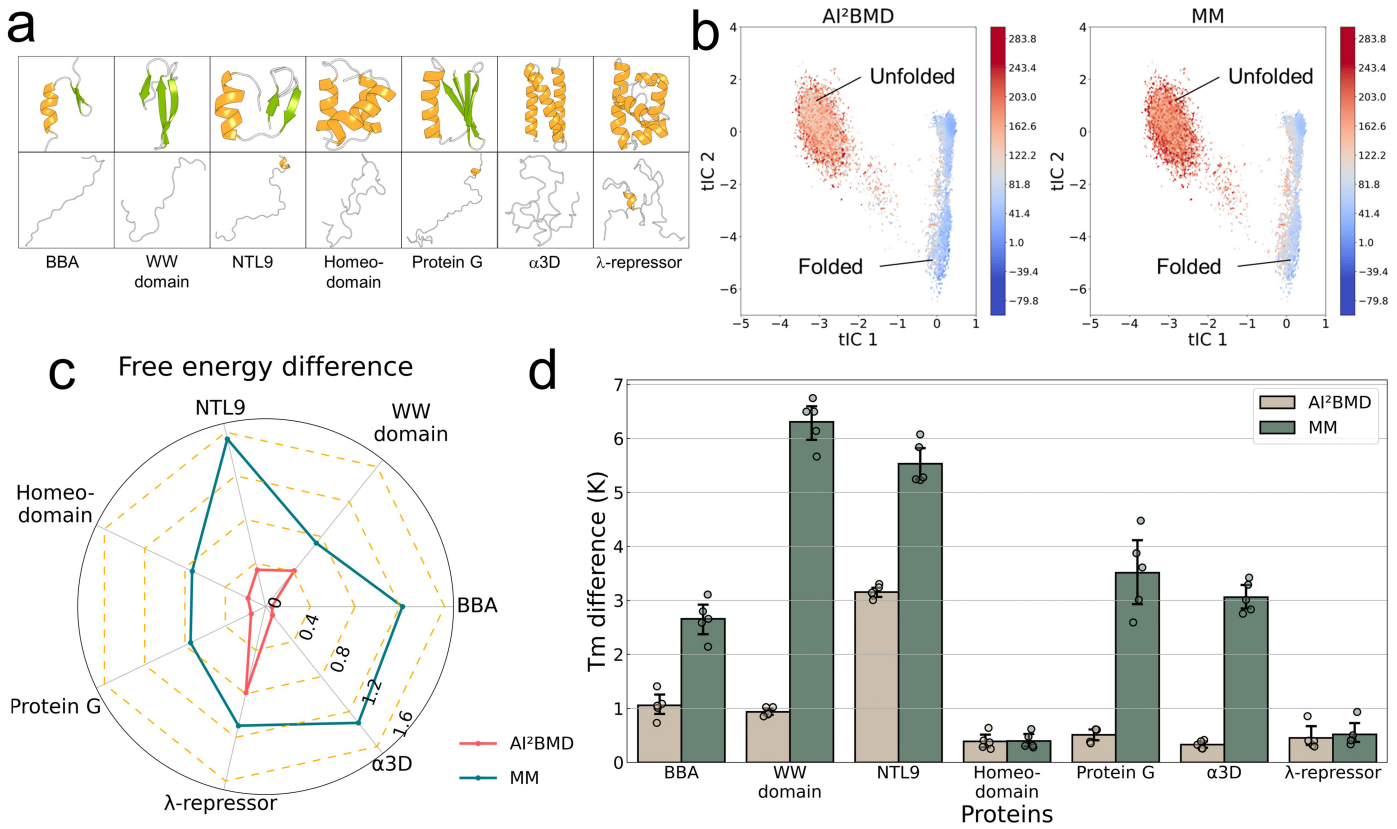
Extended Data Fig. 2 | Examination of hydrogen bonding between water and the asparagine dipeptide (Ace-N-Nme) dipeptide. a, Illustration of a hydrogen bond formed between water and the main chain oxygen. b, Distance distribution of oxygen in the water molecule and the hydrogen bond acceptor on the dipeptide main chain, as determined from a 500 ps sampling using QM/MM, AI^{2}BMD, and MM methods. Such simulations started from 50 initial structures of Ace-N-Nme from a well-sampled conformation ensemble by Replica Exchange Molecular Dynamics (REMD). We then run 10 ps simulations for each initial structure using AI^{2}BMD with AMOEBA polarizable embedding. This resulted in a total sampling time of 500 ps. Meanwhile, QM/MM with Amoeba polarizable embedding and molecular mechanics with Amber FF19SB were also performed on such conformations, respectively. Each simulation includes a water box of 5 Å. (c) Energy fluctuations derived from a scanning of the main chain hydrogen bond distance, calculated using QM, AI^{2}BMD, and MM. (d) Illustration of a hydrogen bond formed between water and ASN side chain oxygen. (e) Distribution of oxygen in the water molecule and the hydrogen bond acceptor on the side chain, as determined from a 500 ps sampling using QM/MM, AI^{2}BMD and MM, respectively. (f) Energy fluctuations derived from a scanning of the side chain hydrogen bond distance, calculated by QM, AI^{2}BMD, and MM, respectively.



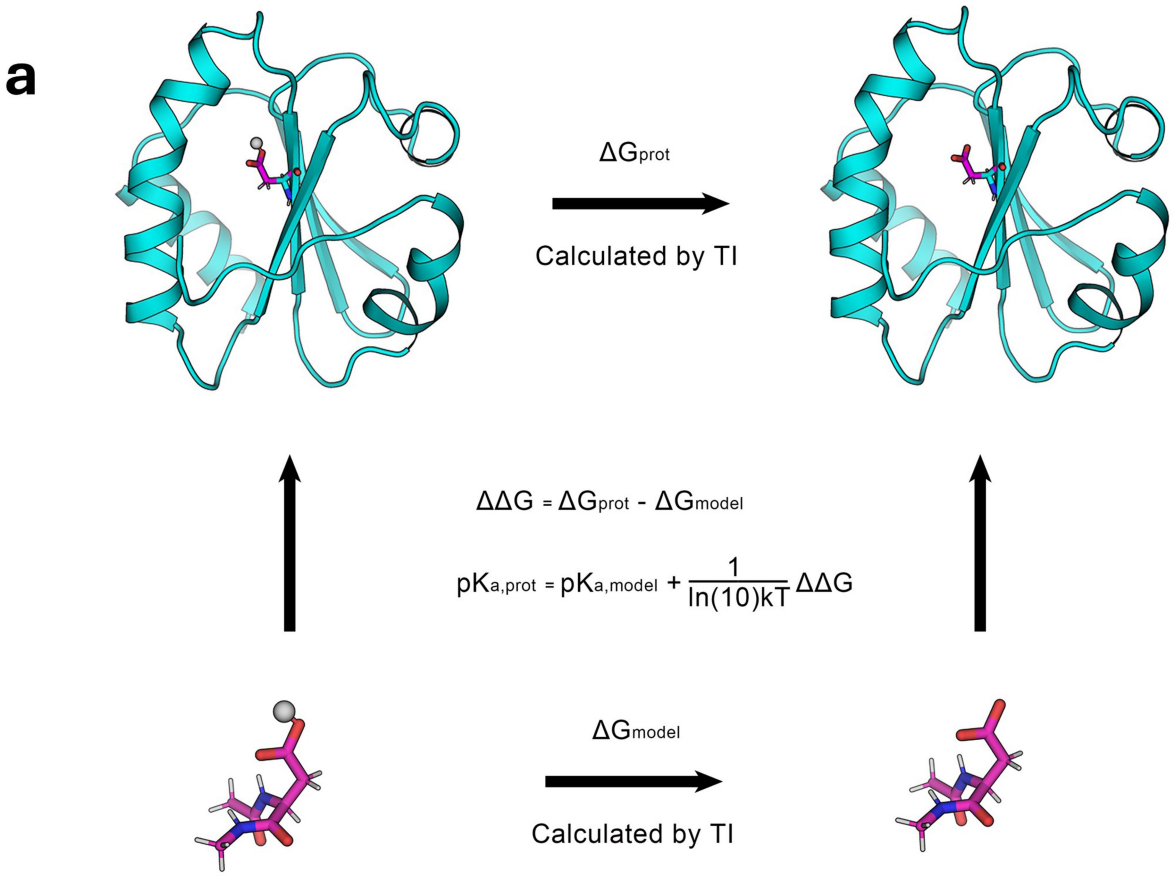
Extended Data Fig. 3 | Analysis of Chignolin simulation process. a, the representative structures during a simulated Chignolin process performed by AI^{2}BMD. b. the free energy landscape of Chignolin. The end-to-end distance of the protein, i.e., the distance between the N terminal and the C terminal and the radius of gyration ("Rg") were chosen as collect variables. The indices of the representative structures shown in a are labeled in the free energy landscape. The Chignolin structure starts from a high energy metastable state and transits to the low energy states. Finally, it sampled to the lowest energy metastable state of the folded structures.

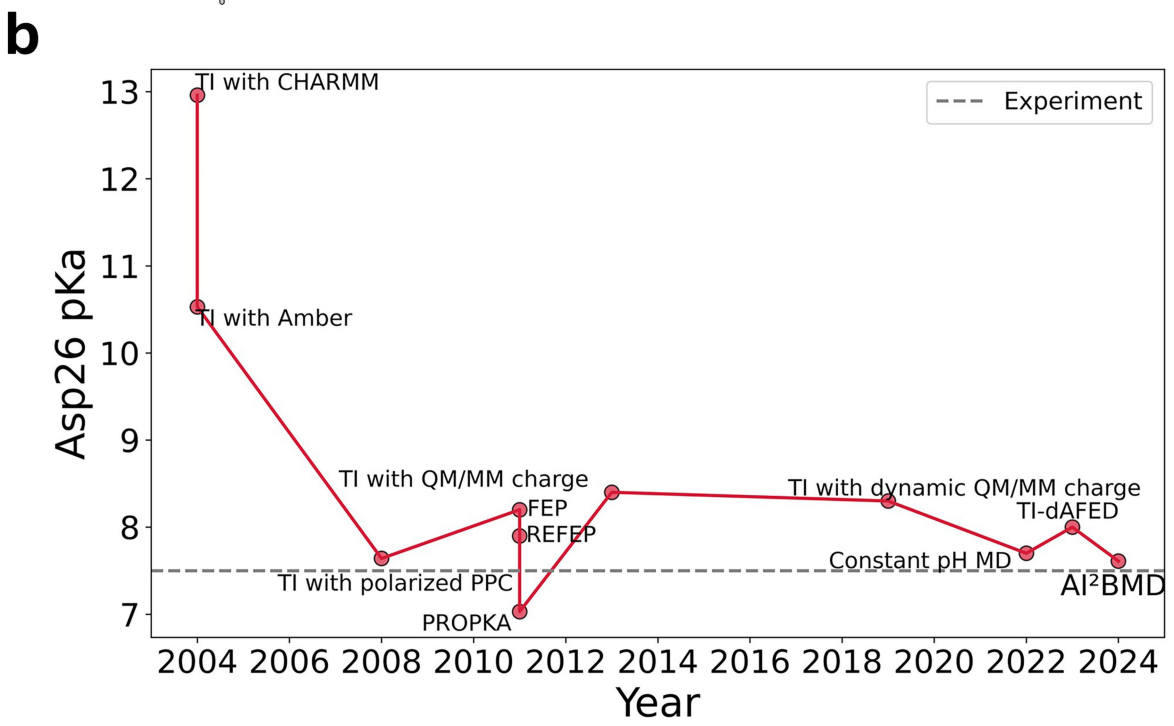


Extended Data Fig. 4 | Structural analysis of the phi and psi angles of G7 in Chignolin. a-b, the Ramachandran plot of conformations sampled by AI^{2}BMD (a) and MM (b) in an ensemble of 60 trajectories of 10 ns simulations. c-d, representative structures of the phi and psi angles of G7 in Chignolin. The backbone atoms of G7 (N, Cα and C) are shown in transparency. The C of T6 and the N of T8 that form torsion angles are colored cyan for visualization.

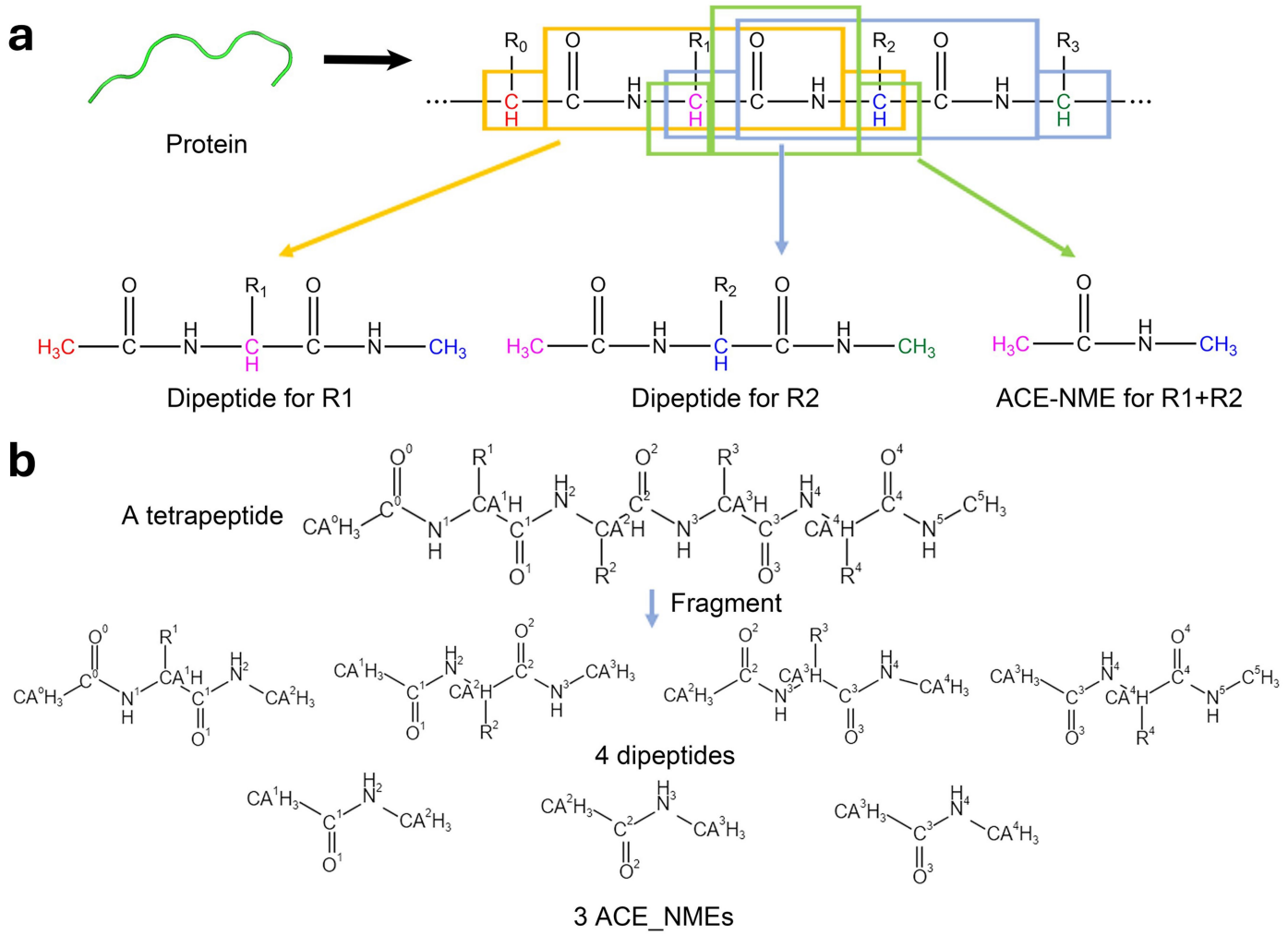


Extended Data Fig. 5 | Analysis of free energy calculation and melting temperature estimation for fast folding proteins. a, the representative folded (top line) and unfolded structures (bottom line) for proteins. b, potential energy surface of NTL9 made by AI²BMD (left) and MM (right), respectively. c, the absolute free energy differences between the folded and unfolded structures for proteins. d, the difference between the simulation temperature and the melting temperature (Tm) calculated from simulations. The error bars in (d) with the averaged value as a measure of center indicate the standard deviations of Tm from 5 independently repeated experiments (n = 5) for both AI²BMD and MM.



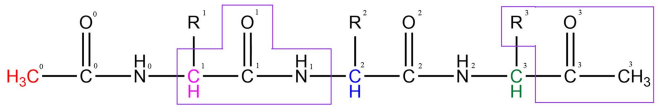


Extended Data Fig. 6 | Comparison of pKa calculation for Asp26 of thioredoxin among AI²BMD and other approaches. a, The pKa calculation process of thioredoxin Asp26 using thermodynamics integration. The free energy difference (ΔΔG) was calculated as the difference between the free energy change (ΔG_{prot}) during the protonation state transition from thioredoxin AspH26 to Asp26, and the ΔG_{model} from the same transition in a dipeptide model. The pKa value is then estimated using the Linear Free Energy Relationship (LFER) approach. B, Comparison of pKa values among different computational methods. The pKa values obtained from various computational methods were plotted against their publication years on the x-axis, with pKa values on the y-axis. The experimental pKa value of 7.5 is marked as a dotted line for reference.

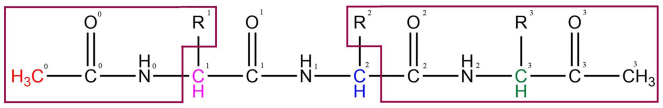


Extended Data Fig. 7 | The sketch of protein fragmentation approach. a. The overall pipeline of the protein fragmentation approach. b. The generated dipeptides and ACE-NMEs by fragmentation on a tetrapeptide. Four successive dipeptides and three ACE-NMEs (i.e., the overlapped regions between two successive dipeptides) are generated. All heavy atoms are labeled with the indices of the corresponding residues for visualization and analysis. Cα atoms are shown as CA for clarity.

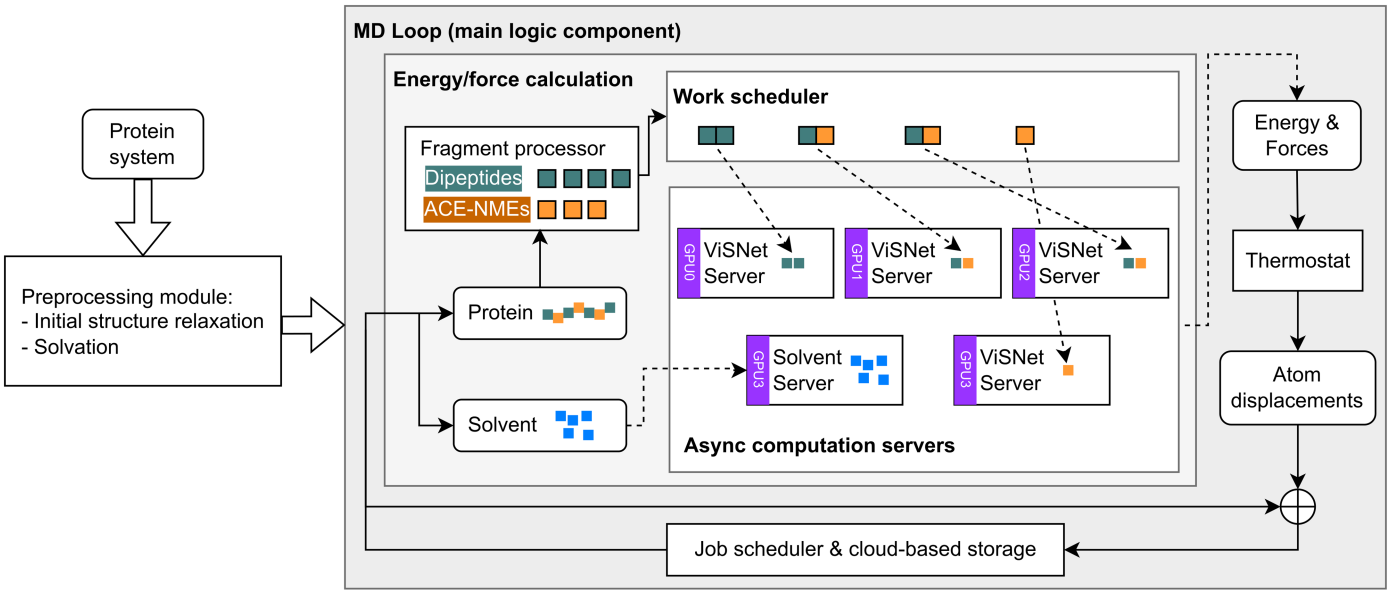
a



b



Extended Data Fig. 8 | Illustration of extra interactions among non-overlapped protein units. a. The extra interactions between the group of CH^1, C^1, O^1, NH^1 (shown in a purple box) and the last part of the tetrapeptide (shown in another purple box). b. The extra interactions between the beginning part of the tetrapeptide including $CH3^0$, C^0, O^0, NH^1 (shown in a brown box) and another part beginning from the second side chain to C-terminus (shown in another brown box).



Extended Data Fig. 9 | The overview block diagram of AI^{2}BMD simulation program. Rounded rectangles denote various types of data, e.g., the initial protein structure in the upper left, the energy and forces of the entire system in the upper right. Rectangles represent computation modules such as the preprocessing module and the molecular dynamics main loop. Solid arrows indicate the data flow within the main Python process, while the dashed ones indicate the communication between the main Python process and various computation servers running in separate processes.

Extended Data Table 1 | Comparison on running time per simulation step for proteins solvated with a 10 Å water box

Protein	Atom number of protein	Atom number of system	AI^{2}BMD (s)	DPMD (s)	Allegro (s)	Tinker Amoeba (s)	Amber FF19SB (s)
Chignolin	175	4,715	0.047	0.040	0.238	0.117	0.004
Trp-cage	281	6,067	0.052	0.055	0.322	0.136	0.005
WW domain	571	10,678	0.070	0.095	0.626	0.196	0.008
ABD	746	11,793	0.085	0.106	0.712	0.208	0.008
PACSIN 3	1,040	17,923	0.106	0.162	-	0.292	0.011
SSO0941	2,450	44,401	0.213	0.414	-	0.699	0.027
APC	5,292	54,999	0.449	0.580	-	0.938	0.033
Polyphosphate Kinase	11,404	97,657	0.966	-	-	1.487	0.058

Evaluations were performed on a desktop with an A6000 GPU and 32 CPU cores. Allegro was trained with a 4 Å radius graph threshold. The dashed lines represent failures due to "out-of-memory" errors.

nature portfolio

Corresponding author(s): Tong Wang, Bin Shao

Last updated by author(s): Aug 14, 2024

Reporting Summary

Nature Portfolio wishes to improve the reproducibility of the work that we publish. This form provides structure for consistency and transparency in reporting. For further information on Nature Portfolio policies, see our Editorial Policies and the Editorial Policy Checklist.

Statistics

For all statistical analyses, confirm that the following items are present in the figure legend, table legend, main text, or Methods section.

n/a	Confirmed
☐	☒ The exact sample size (*n*) for each experimental group/condition, given as a discrete number and unit of measurement
☐	☒ A statement on whether measurements were taken from distinct samples or whether the same sample was measured repeatedly
☒	☐ The statistical test(s) used AND whether they are one- or two-sided *Only common tests should be described solely by name; describe more complex techniques in the Methods section.*
☒	☐ A description of all covariates tested
☒	☐ A description of any assumptions or corrections, such as tests of normality and adjustment for multiple comparisons
☐	☒ A full description of the statistical parameters including central tendency (e.g. means) or other basic estimates (e.g. regression coefficient) AND variation (e.g. standard deviation) or associated estimates of uncertainty (e.g. confidence intervals)
☒	☐ For null hypothesis testing, the test statistic (e.g. *F*, *t*, *r*) with confidence intervals, effect sizes, degrees of freedom and *P* value noted *Give P values as exact values whenever suitable.*
☒	☐ For Bayesian analysis, information on the choice of priors and Markov chain Monte Carlo settings
☒	☐ For hierarchical and complex designs, identification of the appropriate level for tests and full reporting of outcomes
☐	☒ Estimates of effect sizes (e.g. Cohen's *d*, Pearson's *r*), indicating how they were calculated

Our web collection on statistics for biologists contains articles on many of the points above.

Software and code

Policy information about availability of computer code

Data collection	We used Amber20 (https://ambermd.org) and Gromacs 2018 (https://www.gromacs.org) to sample conformations of proteins, ORCA 5.0.1 (https://orcaforum.kofo.mpg.de/app.php/portal) and Gaussian 16 (https://gaussian.com/gaussian16) to calculate single point energy and run quantum molecular dynamics simulations at Density Functional Theory level, Tinker 8/9 (https://dasher.wustl.edu/tinker) to conduct simulations with polarizable effects, and PyTorch (https://pytorch.org) with AdamW optimizer for model training.
Data analysis	We used Python 3.9.0 (https://www.python.org), Numpy 1.21.5 (https://numpy.org), and Matplotlib 3.5.1 (https://matplotlib.org) to analyze data and plot figures.

For manuscripts utilizing custom algorithms or software that are central to the research but not yet described in published literature, software must be made available to editors and reviewers. We strongly encourage code deposition in a community repository (e.g. GitHub). See the Nature Portfolio guidelines for submitting code & software for further information.

Data

Policy information about availability of data

All manuscripts must include a data availability statement. This statement should provide the following information, where applicable:

- Accession codes, unique identifiers, or web links for publicly available datasets
- A description of any restrictions on data availability
- For clinical datasets or third party data, please ensure that the statement adheres to our policy

The protein unit dataset built by this study is available at https://github.com/microsoft/AI2BMD. The proteins for energy and force evaluations in this work are available from PDB (https://www.rcsb.org) with the following accession numbers: Chignolin: 5AWL, Trp-cage: 2JOF, WW domain: 2F21, albumin binding domain (ABD): 1PRB, PACSIN 3: 6F55, SSO0941: 5VFK, APC: 5IZA, polyphosphate kinase: 1XDO, aminopeptidase N: 4XN9, barnase: 1A2P and CI2, 2CI2.The proteins for melting temperature estimation are available at PDB with the indices as follows: α3D: 2A3D, BBA: 1FME, NTL9: 2HBA, Protein G: 1MIO, WW domain: 2F21, λ-repressor: 1LMB, Homeodomain: 2P6J. The simulation trajectories for pKa analysis are provided by Ross Walker & Mike Crowley (https://ambermd.org/tutorials/advanced/tutorial6/index.php).

Human research participants

Policy information about studies involving human research participants and Sex and Gender in Research.

Reporting on sex and gender	N/A
Population characteristics	N/A
Recruitment	N/A
Ethics oversight	N/A

Note that full information on the approval of the study protocol must also be provided in the manuscript.

Field-specific reporting

Please select the one below that is the best fit for your research. If you are not sure, read the appropriate sections before making your selection.

☒ Life sciences ☐ Behavioural & social sciences ☐ Ecological, evolutionary & environmental sciences

For a reference copy of the document with all sections, see nature.com/documents/nr-reporting-summary-flat.pdf

Life sciences study design

All studies must disclose on these points even when the disclosure is negative.

Sample size	We have 20.88 million samples, which were generated by density functional theory.
Data exclusions	No samples were excluded.
Replication	The code, datasets, and computational results were thoroughly reviewed to ensure they could be replicated.
Randomization	The training, validation, and test sets were split randomly.
Blinding	The test sets have never been seen by the model during AI2BMD potential training.

Reporting for specific materials, systems and methods

We require information from authors about some types of materials, experimental systems and methods used in many studies. Here, indicate whether each material, system or method listed is relevant to your study. If you are not sure if a list item applies to your research, read the appropriate section before selecting a response.

Materials & experimental systems

n/a	Involved in the study
☒	☐ Antibodies
☒	☐ Eukaryotic cell lines
☒	☐ Palaeontology and archaeology
☒	☐ Animals and other organisms
☒	☐ Clinical data
☒	☐ Dual use research of concern

Methods

n/a	Involved in the study
☒	☐ ChIP-seq
☒	☐ Flow cytometry
☒	☐ MRI-based neuroimaging