DiffPIE: Guiding Deep Generative Models to Explore Protein Conformations under External Interactions

Yanbin Wang[1,] **and Ming Chen**[1,*]

[1]Purdue University, Department of Chemistry, West Lafayette, 47906, USA
[*]chen4116@purdue.edu

ABSTRACT

Proteins play crucial roles in life processes such as catalysis, signal transduction, immune response, and molecular transport. Many of these functions are mediated by external interactions between a protein and other chemical species, including proteins, nucleic acids, and organic molecules. Furthermore, in therapeutic applications, protein functions can be influenced by interactions with unnatural chemical species, such as molecular organic linkers, inorganic materials, and polymers. Predicting protein conformational ensembles under the influence of external interactions remains a significant challenge both experimentally and computationally. While deep generative models have been developed to predict protein structural ensembles, their application to systems involving external interactions remains largely unexplored. In this work, we present a method that integrates a deep generative model with physics-based interaction modeling to predict protein conformations under external constraints. Without requiring any retraining or fine-tuning of the generative model, our approach can efficiently and accurately predict protein conformations covalently constrained by organic linkers, as well as protein conformations adsorbed on gold surfaces.

1 Introduction

Proteins are fundamental building blocks of all life forms, performing essential functions such as catalysis, signal transduction, immune response, and transport[1–4]. Although proteins typically operate in environments composed primarily of water and low-concentration ions, many of their biological functions require interactions with additional chemical species, such as small ligand molecules[5], other proteins[6], and nucleic acids[7]. We refer to such interactions collectively as protein-environment interactions (PEI). Beyond these biological contexts, PEIs also encompass interactions with unnatural chemical species, including proteins adsorbed on inorganic material surfaces[8,9], proteins covalently constrained by organic linkers[10,11], and protein-polymer hybrids[12]. These types of interaction are particularly common in biomaterial applications and protein drug manufacturing. Understanding protein conformational changes under PEI is critical for the rational design of biocompatible materials, therapeutic molecules, and drug delivery systems[13–15].

However, accessing protein conformations with PEI is challenging both for experimental approaches and for theoretical modeling. Although experimental methods such as X-ray crystallography[16], NMR[17], and cryo-EM[18] have achieved great success in determining the atomistic structures of proteins, the direct application of these structural determination methods to protein with PEI is challenging. These limited examples of structural studies focus on protein-protein complexes[19,20], protein-nucleic acid complexes[21,22], protein-drug complexes[23], and cyclic peptides[24]. However, resolving protein structures with many other PEI types, such as protein-material interactions, is a challenge. More often, protein conformations with these PEI are measured using ensemble-based methods such as IR, circular dichroism, and SAXS where obtaining atomic structures is still an open question[25–27]. In addition to experimental approaches, theoretical modeling to resolve protein conformations is also challenging. Generic approaches to exploring proteins conformations under external interactions include all-atom molecular dynamic (MD) simulations and coarse-grained MD simulations[28–35]. All-atom MD simulations are often limited by simulation time scale and system length scale. Although CG models are accessible to larger system sizes and longer simulation timescales, improving the accuracy of CG models is still an open problem[36,37].

In recent years, deep learning models, including large language models and deep generative models, have been developed to predict protein conformations[38–41]. These models have achieved remarkable success in predicting the structures of individual proteins and are now expanding their capabilities to the prediction of protein–protein complexes, protein–nucleic acid complexes, and protein–ligand interactions[42,43]. However, many types of external interactions, such as protein–material interactions, remain unexplored by these models, largely due to the scarcity of experimentally determined structures involving such protein-environment interactions (PEI). Given the diversity of PEIs and the experimental and theoretical challenges associated with structure determination, developing deep learning models tailored for different PEI types would require extensive and coordinated efforts across multiple scientific communities. Thus, guiding existing deep learning models to explore protein

conformations under external interactions presents an attractive and promising alternative.

Recent studies have begun to explore similar directions[44], particularly through guided sampling using diffusion models[45–47]. Diffusion models learn by progressively adding noise to data points and can generate new data through a reverse diffusion process[48,49]. In particular, the reverse diffusion process, either its drift or its diffusion term, can be modified to manipulate the data points generated[50,51]. Examples of such modifications include incorporating restraint potentials to ensure that generated all-atom structures are consistent with experimental measurements[51], enforcing the generation of structures containing specific motifs[52], and guaranteeing correct stereochemistry in protein–ligand complexes[53]. Some studies have also used all-atom force fields to refine protein structure generation[45,47]. Building upon these ideas, we present a systematic framework, termed the "Diffusion Model for Proteins Interacting with the Environment (DiffPIE)", to integrate PEIs into an existing diffusion model. The core idea is that a diffusion model implicitly learns a "statistical potential" for protein structures without PEIs. By adding explicit physical interactions between the protein and the surrounding chemical species to this statistical potential, we construct a new effective potential energy landscape. Incorporating this modified potential into the diffusion process allows the model to generate protein structures that account for external interactions.

We will demonstrate the success of the proposed framework with two representative examples: a peptide denatured on a gold surface and a peptide cyclized with an organic linker. In both cases, we show that PEIs can be effectively modeled using MD simulations of small, localized systems, rather than simulating the entire protein with its environment. Furthermore, we illustrate the effectiveness of using physics-based potentials to guide the structure generation process.

2 Results

2.1 DiffPIE: Guiding Protein Structure Generation with Protein-environment Interaction

Assume that a protein structure ensemble can be generated using a pre-trained foundation score-based diffusion model, which learns through a forward noise-adding process described by:

$$d\mathbf{x} = f(\mathbf{x},t)dt + g(t)d\mathbf{w},\ t \in [0,1], \tag{1}$$

and generates new structures through a reverse denoising process:

$$d\mathbf{x} = [f(\mathbf{x},t) - g^2(t)\mathbf{s}(\mathbf{x},t)]dt + g(t)d\bar{\mathbf{w}}. \tag{2}$$

To simplify the discussion, we consider $\mathbf{x}$ in Eqs.(1) and (2) as representing atomic coordinates. As we will demonstrate later, DiffPIE can also operate with diffusion models that hybridize Cartesian coordinates and generalized coordinates. Here, $f(\mathbf{x},t)$ and $g(t)$ define the noise schedule, while $\mathbf{s}(\mathbf{x},t)$ is trained to approximate $\nabla_{\mathbf{x}} \log P_t(\mathbf{x})$, where $P_t(\mathbf{x})$ is the time-dependent marginal distribution in Eq.(2). The resulting generated structures follow a distribution $p(\mathbf{x})$ learned from the training dataset, which aims to approximate the Boltzmann distribution of a protein without PEI, denoted as $p^{\mathrm{NPIE}}\mathrm{B}(\mathbf{x})$. However, as illustrated in Fig. 1(c), $p(\mathbf{x})$ can differ significantly from the true Boltzmann distribution of a protein under PEI, $p^{\mathrm{PIE}}\mathrm{B}(\mathbf{x})$. Consequently, it is challenging to study the thermodynamics of $p_{\mathrm{B}}^{\mathrm{PIE}}(\mathbf{x})$ using structures generated solely from $p(\mathbf{x})$.

To address this, DiffPIE generates a biased distribution $p_{\mathrm{bias}}(\mathbf{x})$ that approximates $p_{\mathrm{B}}^{\mathrm{PIE}}(\mathbf{x})$, enabling rapid relaxation toward an ensemble representative of the true protein-environment system. DiffPIE achieves this by modifying the reverse diffusion process:

$$d\mathbf{x} = [f(\mathbf{x},t) - g^2(t)(\mathbf{s}(\mathbf{x},t) + \lambda\mathbf{F}(\mathbf{x},t)]dt + g(t)d\bar{\mathbf{w}}. \tag{3}$$

Here, $\mathbf{F}(\mathbf{x},t)$ is a biasing drift term that encodes protein-environment coupling, and $\lambda > 0$ controls the strength of the applied bias. A larger λ leads to stronger bias away from $p(\mathbf{x})$. In DiffPIE, $\mathbf{F}(\mathbf{x},t)$ is constructed by propagating a biasing force $\mathbf{F}_0(\mathbf{x})$, obtained from MD simulations, back into the foundation diffusion model, following an approach similar to manifold constraint algorithms[54,55]. Users have flexibility to design $\mathbf{F}(\mathbf{x},t)$ suitable for a wide range of PEI, as illustrated in Fig. 1(a). Different choices of $\mathbf{F}_0(\mathbf{x})$ allow the generation of protein structures consistent with different PEI scenarios. It is important to emphasize that DiffPIE does not require any retraining or fine-tuning of the foundation diffusion model. Instead, it only modifies the generation process by introducing the biasing drift $\mathbf{F}(\mathbf{x},t)$ during sampling. The DiffPIE framework is illustrated in Fig.1(b). We begin by designing a customized biasing force, with specific examples provided in Sec.2.2 and Sec. 2.3. In this study, we use Str2Str[41] as the foundation diffusion model; however, in principle, any diffusion model can be incorporated. A key advantage of Str2Str is its flexibility in controlling the generated ensemble by selecting different seed structures and tuning hyperparameters. From the generated structures, we select representative conformations which can serve as new seeds or inform further hyperparameter adjustments. This iterative process is repeated until no additional new conformations are generated. Finally, the selected structures are subjected to relaxation through either structure minimization or short molecular dynamics (MD) simulations, depending on the application.

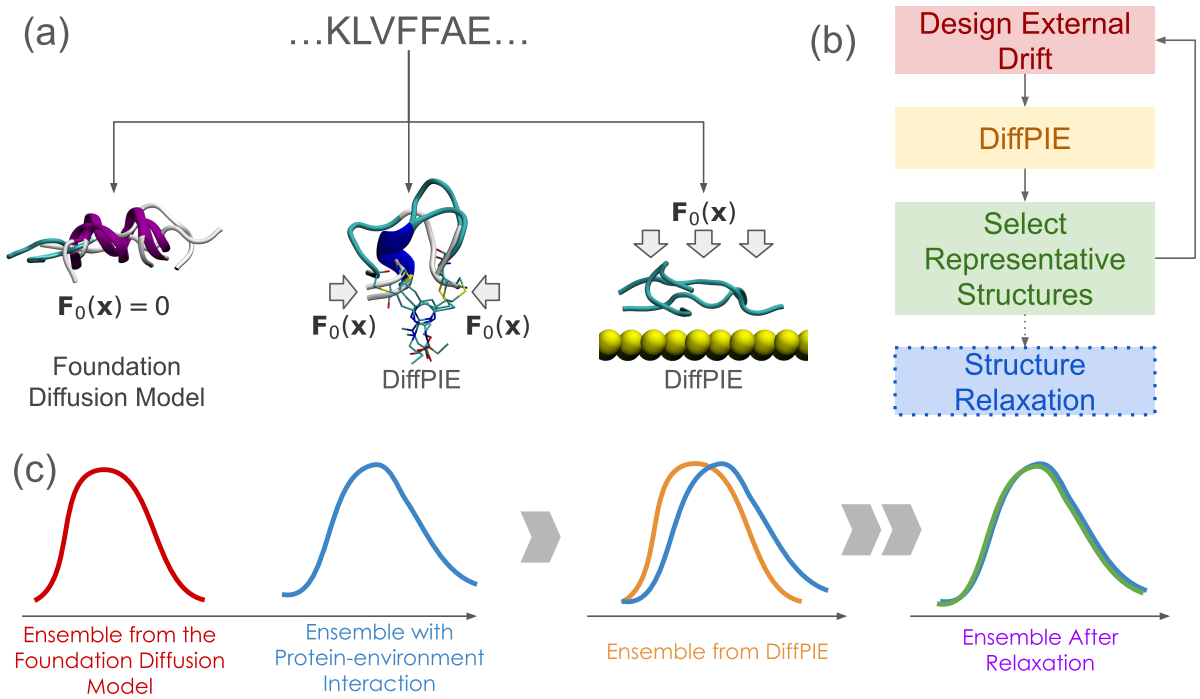

Figure 1. Panel (a) illustrates the concept of biasing protein structure generation using $\mathbf{F}_0(\mathbf{x})$. The left bundle of structures was generated by the foundation diffusion model, Str2Str, with $\mathbf{F}_0(\mathbf{x}) = 0$, corresponding to unbiased generation. The middle bundle shows structures generated with a $\mathbf{F}_0$ designed to mimic the restraint forces imposed by an organic linker. The right bundle presents structures generated with a $\mathbf{F}_0$ simulating adsorption forces, such as those from a surface interaction. Panel (b) summarizes the overall workflow of DiffPIE. Once a biasing force is designed, DiffPIE can be applied without any retraining. The distribution of generated structures can, in turn, inform adjustments to the hyperparameters in DiffPIE or refinements in the design of the biasing force. Although structure relaxation is not the primary focus of this paper, we demonstrate examples using short MD simulations for further refinement. Panel (c) illustrates how DiffPIE alters the ensemble distribution. The ensemble generated by the foundation diffusion model without bias (red) can significantly differ from the true protein ensemble under PEI (blue). By introducing the biasing force through DiffPIE, the generated ensemble distribution (orange) is shifted closer to the true ensemble. Additional MD relaxation can further improve the ensemble (green) to better approximate the true distribution.

2.2 Sampling Cyclic Peptide Structures Containing Organic Linkers

P3-F is a cyclic peptide (Fig. 2(a)) that can interrupt interactions between the Nuclear Factor Erythroid-2-related factor (Nrf2) and the E3 ligase Kelch-like EH-associated protein 1 (Keap1) to counteract oxidative stress[56]. Here, we use P3-F as an example to demonstrate how to use DiffPIE to generate protein structures constrained by organic linkers. The inset of Fig. 2(b) demonstrates how to construct $\mathbf{F}_0$. We first used MD to simulate a system of two cystine dipeptides cross-linked by the linker molecule. The thermodynamic mean force of six distances (labeled with colored dashed lines in Fig. 2(b)) has been used as $\mathbf{F}_0$. Fig. 2(b) highlights a one-dimension projected free energy surface (FES) of the distance between two α-carbons (Cα) in two cystines. It is clear that the FES is highly anharmonic with multiple minima, suggesting that the conventional manifold constraint method, which uses a harmonic potential to apply restrictions to generated structures, is not applicable in this case. Fig.(2)(b) also highlights that the Cα–Cα distance in the PDB structure is not a minimum in the FES, suggesting that stable structures in this cyclic peptide should balance the internal energy of the peptide modeled by the foundation diffusion model and the bias energy of the linker modeled by $\mathbf{F}_0$. Fig.(2)(c)-(e) shows the impact of applying a bias drift. Structures generated by Str2Str without $\mathbf{F}_0$, DiffPIE, and a metadynamic[57] simulation of the entire peptide are mapped to 2D CVs including end-to-end distance and AlphaRMSD[58]. As shown in Fig. 2(c), Str2Str generated a large number of helix-like structures, while these structures are nearly not seen in the metadynamics simulation. DiffPIE significantly reduces the population of helix structures and highlights the conformation with a low AlphaRMSD and a broad range of end-to-end distances, which is consistent with the conformations from metadynamics. In addition, we examine the quality of the generated structures. First, stable conformations are included in the structural library generated by DiffPIE, as shown in Fig.2(f). All structures shown in the panel are comparable to those found in the metadynamics simulations and we want to highlight that DiffPIE only requires hours of computational time to generate all structures. Fig. 2(c) highlights a non-optimal structure generated by DiffPIE. Even

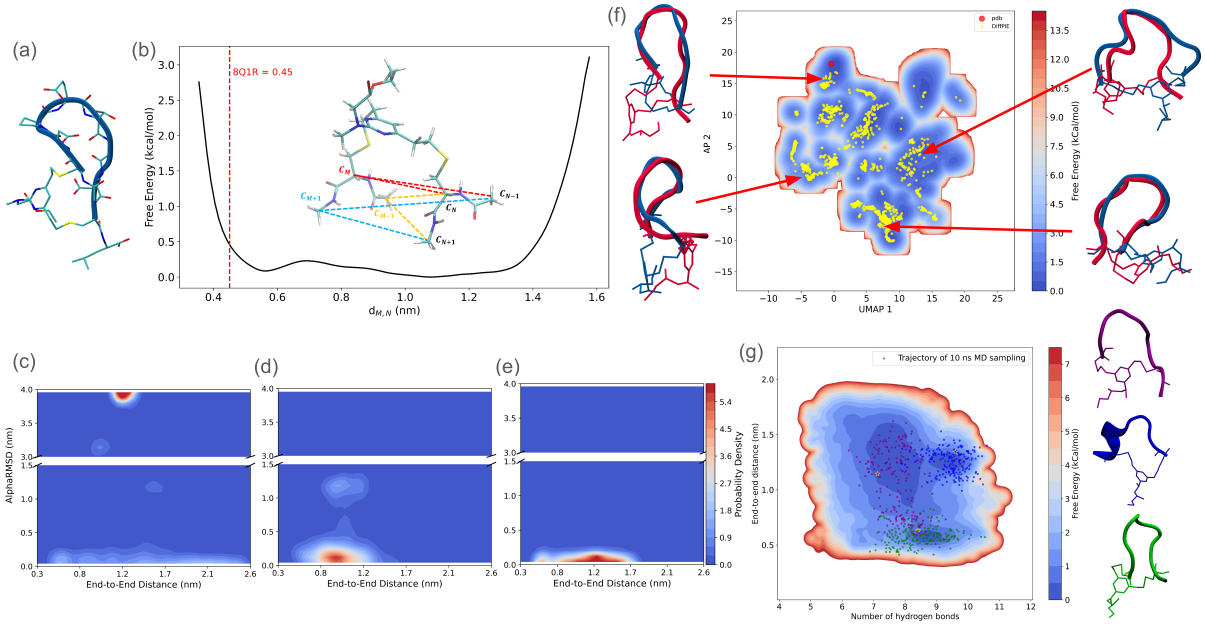

Figure 2. Panel (a) shows the structure of P3-F (PDB ID: 8Q1R). The inset of panel (b) presents the model molecule used for generating $\mathbf{F}_0$, where six distances highlighted by dashed lines are the CVs whose mean forces were used as $\mathbf{F}_0$. Panel (b) shows the FES of the distance between two Cα atoms in the two cystines within the model molecule, with the red dashed line indicating the value observed in the PDB structure. Panel (c)–(e) illustrate the effects of the biasing force on the sample distribution, presented by projecting the samples into the two-dimensional feature space spanned by end-to-end distance and AlphaRMSD. Panels (c), (d), and (e) use samples generated by the original Str2Str model, DiffPIE, and a metadynamics simulation, respectively. The background FES in panel (e) was generated by mapping configurations from the metadynamics simulation onto a two-dimensional Uniform Manifold Approximation and Projection (UMAP) feature space[59]. The PDB structure (red dot) and the samples generated by DiffPIE (yellow dots) are projected onto the same UMAP features. Representative structures selected from the simulation (bue) and from DiffPIE (red) are compared. Panel (f) shows 10 ns of unbiased MD relaxation simulations starting from randomly selected structures within low-energy clusters generated by DiffPIE. Initial structures (highlighted by star marks) and their trajectories (dots) are projected onto a FES defined by end-to-end distance and the number of backbone hydrogen bonds.

if the generated structure is not optimal, it can be quickly relaxed to a stable conformation. The key reason for stabilizing this conformation is the formation of a β-turn structure, which is correctly formed in the generated structure. We also examine the diversity of the generated structures in Fig. 2(f). The background FES of the UMAP CVs was generated after unbiasing the metadynamics simulation[59]. We map the generated structures onto the same UMAP CVs as the scatter points. It clearly demonstrates that DiffPIE is capable of generating structures that cover most of the stable conformations. The PDB structure, denoted by the red point, corresponds to a metastable conformation. Since the PDB structure was resolved for the Keap1–P3-F complex, it indicates that the protein–peptide interaction is the main driving force stabilizing the P3-F PDB structure. Even though the PDB structure is not the most stable conformation in the metadynamics simulation, DiffPIE successfully generates similar structures, highlighting a possible application of DiffPIE in predicting cyclic peptide drug binding[60,61].

2.3 Sampling Peptide Structures on Gold Surface

Gold nanoparticles have been widely used in biochemistry, with applications such as surface plasmon-enhanced spectroscopy[62,63] and drug delivery[64]. Although protein denaturation is well known to occur on naked gold surfaces, predicting the resulting denatured protein structures remains a challenge[65–68]. In this section, we demonstrate how to use DiffPIE to predict possible structures of proteins that bind to a gold surface. The system we study is Aβ_{16-22} on the Au(111) surface (Fig. 3(a)), which has been widely used as a model system to study the environmental impacts on protein fibrillation[69]. Additionally, the small size of this system allows us to exhaustively sample all possible conformations with metadynamics to benchmark DiffPIE. The biasing force $F_0(\mathbf{x})$ in this example is the thermodynamic force associated with binding Aβ_{16-22} to the gold surface. The force is approximated by the contributions of each residue. We perform metadynamics simulations of a dipeptide molecule for each type of amino acid in Aβ_{16-22} to generate a free energy surface (FES) based on three distances, d_1, d_2, and d_3, as shown in

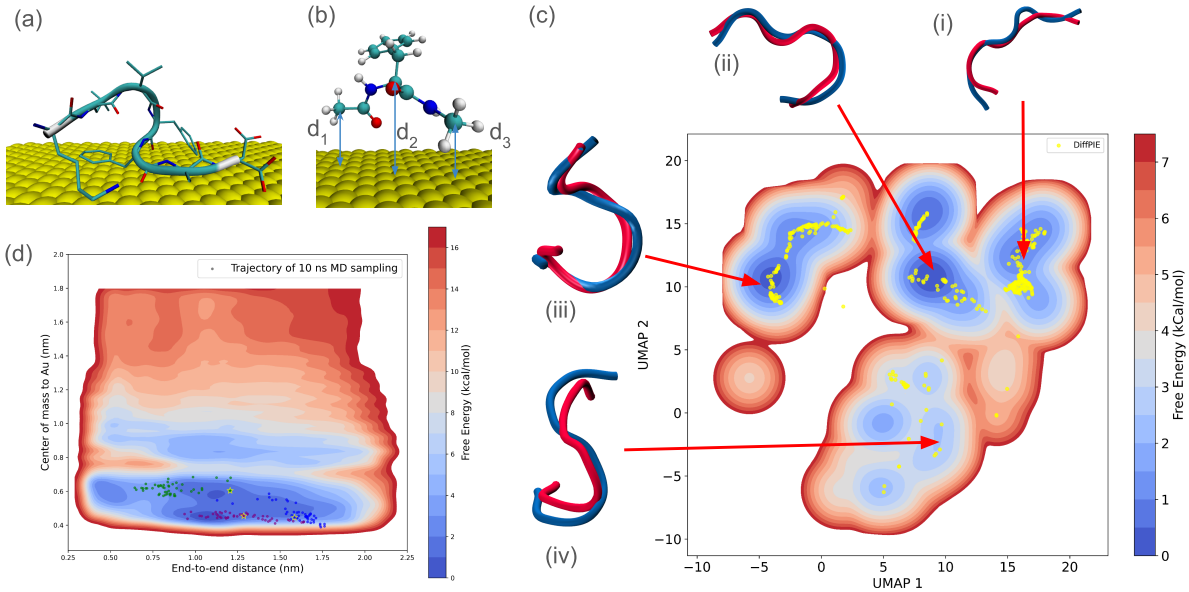

Figure 3. Panel (a) presents the system of $A\beta_{16-22}$ on the Au(111) surface. Panel (b) demonstrates a model dipeptide molecule (PHE in this case) binding on the gold surface to construct the biasing force $\mathbf{F}_0$. The binding free energy along three CVs, d_1, d_2, and d_3, was evaluated by a metadynamics simulation. Here, d_1, d_2, and d_3 are defined as the distances from the gold surface to the methyl carbon of the acetyl cap, the α-carbon, and the methyl carbon of the N-methylamide cap, respectively. Panel (c) shows the comparison between metadynamics simulations and DiffPIE. The background FES was generated by projecting the metadynamics simulation trajectory onto UMAP features. The DiffPIE samples are projected onto the same UMAP feature space as yellow dots. We further present the top-down view of representative simulated structures (blue) from low free energy clusters and similar DiffPIE samples (red), with the gold surface oriented parallel to the plane of the paper. Panel (d) shows short MD relaxations of three structures selected randomly from DiffPIE structure clusters. The starting structures (stars) and their relaxation trajectories (dots) are mapped onto the FES of the $A\beta_{16-22}$ end-to-end distance and the $A\beta_{16-22}$ center-to-gold surface distance.

Fig. 3(b). For residue i in Aβ_{16-22}, we generate the force on its Cα, while the Cα–surface distances of the residue $(i-1)$ and residue $(i+1)$ match d_1 and d_3, respectively. The design of $F_0(\mathbf{x})$ highlights the importance of the orientation of the residue, as reflected by the different values of d_1 and d_3. Unlike the previous example, $\mathbf{F}_0$ in this case cannot be expressed as the gradient of an energy function, highlighting the flexibility of designing $\mathbf{F}_0$. Similarly to the previous example, we project the structures generated by DiffPIE onto UMAP features trained from the benchmark metadynamics simulation. As shown in Fig. 3(c), the clusters of structures generated by DiffPIE overlap well with low-free-energy regions on the FES from metadynamics, suggesting that the populations of structures generated by DiffPIE are qualitatively consistent with conformation stability. DiffPIE is also capable of finding structures similar to representative conformations from metadynamics simulations. We emphasize that while some configurations are highly flexible and DiffPIE can only recover structures with similar overall shapes, for conformations that are energetically stabilized with low free energy, DiffPIE can generate structures that nearly overlap with the simulated structures. Again, generating structures with DiffPIE only took less than one hour, while metadynamics simulations may take days to complete. Finally, we tested the relaxation of generated structures by randomly selecting three structures as starting points for short MD simulations. Both the starting structures and the relaxation trajectories are presented in Fig. 3(d). For some starting structures, the relaxation leads to the same final conformation, suggesting that representative structures of stable conformations are included in our predictions. Meanwhile, high-free-energy structures quickly relax, indicating that even short MD simulations can significantly aid structure refinement. Importantly, all structures remained adsorbed on the gold surface without detaching, suggesting that DiffPIE successfully avoided the unpreferred adsorption modes.

3 Discussion and Conclusion

The examples in Sec.2.2 and 2.3 demonstrate that DiffPIE is capable of generating important structures for protein molecules with PEI. We highlight that DiffPIE does not require any further training or fine-tuning of the foundation diffusion model. In addition, DiffPIE does not require an exact PEI model for protein structure generation. In both examples, our objective was to

achieve a trade-off between the accuracy of PEI models and the computational cost of building them. Thus, approximations were used in both cases. In the first example, the presence of the gold surface disrupts the solvation shell of residues near the surface, leading to changes in the solvation free energy that are non-additive. In the second example, we ignore the non-bonded interactions between the organic linker and residues other than cysteines. The generated structures suggest that such approximations do not significantly affect the downgrade quality of structures generated by DiffPIE. However, we want to remind that the protocol of modeling biasing force may change for different PEI types. The generated structures can also be used as starting points for MD simulations or enhanced sampling simulations[70,71]. As suggested in recent studies, the introduction of enhanced sampling simulations starting from multiple important conformations predicted by machine learning models can significantly improve sampling efficiency[72,73]. We expect that DiffPIE can be combined with enhanced sampling methods or other free-energy calculation approaches to further improve structural exploration and thermodynamic analysis.

Ideally, we hope that DiffPIE will generate structures following the Boltzmann distribution of a protein molecule with PEI. However, in practice, it is limited by several theoretical and technical challenges. First, generating structures that follow the Boltzmann distribution of a protein requires that the foundation diffusion model is able to generate protein structures following the Boltzmann distribution without PEI. The foundation diffusion model used in this study, Str2Str, is known to be biased toward generating more helix-like structures. As shown in Sec. 2.2, the bias force in DiffPIE can help correct this bias, but cannot fully eliminate it. Some recent studies have proposed methods to correct Str2Str[45] or to develop new foundation diffusion models capable of generating protein ensembles without PEI[74,75]. We expect that improved foundation diffusion models will also enhance the performance of DiffPIE. Second, DiffPIE is limited by the accuracy of the PEI models. Therefore, we do not expect the current version of DiffPIE to generate structures that fully follow the Boltzmann distribution of proteins with PEI. Improving the quality of the DiffPIE-generated ensemble remains an important goal for future work.

Another challenge in DiffPIE is how to efficiently identify stable structures. The current solution is to screen structures using short MD simulations, which provides the most reliable, but also the most computationally expensive, screening approach. Other methods, such as energy minimization in vacuum or with an implicit solvent, work well for the cyclic peptide example but perform poorly for the protein-on-gold example. Alternative energy scoring functions for proteins, such as the Rosetta force field, are typically tuned for proteins without PEI and are not directly applicable to either example without careful reparameterization. Therefore, developing efficient and accurate screening methods for proteins with PEI remains an important area for future research.

In summary, we have developed DiffPIE, a framework that guides a foundation diffusion model with physics-based interactions to generate protein structures with PEI. This framework does not require additional training and can be adapted to various types of PEI. In addition to the examples discussed in this work, DiffPIE has the potential to greatly broaden the applications of protein foundation diffusion models, including protein–ligand binding, proteins at interfaces, and proteins in external fields.

4 Methods

Our proposed method is general for any score-based diffusion model (SBDM). A SBDM perturbs data distribution to noise prior with a diffusion process over a unit time by a linear stochastic differential equation (SDE):

$$d\mathbf{x} = f(\mathbf{x},t)dt + g(t)d\mathbf{w},\ t \in [0,1], \tag{4}$$

where $f(\mathbf{x},t), g(t)$ are user-defined drift and diffusion functions of the SDE and $\mathbf{w}$ denotes a standard Wiener process. With a carefully designed SDE, the marginal probability of $\mathbf{x}$ at diffusion time t, $P_t(\mathbf{x})$, changes from the data distribution $P(\mathbf{x})$ to approximately a simple Gaussian distribution. In this paper, we use an SDE with drift term $f(\mathbf{x},t) = 0$ and diffusion function $g(t) = \sqrt{\mathrm{d}\sigma^2(t)/\mathrm{d}t}$, where $\sigma(t)$ represents the noise scale. For any diffusion process in Eq.(4), it has a corresponding reverse-time SDE[49,76] :

$$d\mathbf{x} = [f(\mathbf{x},t) - g^2(t)\nabla_{\mathbf{z}} \log P_t(\mathbf{x})]dt + g(t)d\bar{\mathbf{w}}, \tag{5}$$

with $\bar{\mathbf{w}}$ a standard Wiener process in the reverse-time. The trajectories of the reverse SDE have the same marginal densities as those of the forward SDE. Thus, the reverse-time SDE can gradually convert noise to data. The SBDM parameterizes the time-dependent score function $\nabla_{\mathbf{x}} \log P_t(\mathbf{x})$ in reverse SDE with a neural network $\mathbf{s}_\theta(\mathbf{x}(t),t)$.

Suppose that the SBDM baseline aims to predict protein structures that follow the Boltzmann distribution, i.e. $P(\mathbf{x}) \approx p^{\mathrm{NPIE}}\mathrm{B}(\mathbf{x}) \propto \exp\{-\beta U(\mathbf{x})\}$ where $U(\mathbf{x})$ is a protein energy function that is similar to a "statistical potential". $\beta = 1/k_{\mathrm{B}}T$ where k_{B} is the Boltzmann constant and T is temperature. PEI can be written as an extra energy term $U_{\mathrm{ext}}(\mathbf{x})$. The Boltzmann distribution of protein structures PEI is $p_{\mathrm{B}}^{\mathrm{PIE}}(\mathbf{x}) \propto \exp\{-\beta(U(\mathbf{x}) + U_{\mathrm{ext}}(\mathbf{x}))\}$ It has been proved that sampling $P_{\mathrm{bias}}(\mathbf{x}) \approx p_{\mathrm{B}}^{\mathrm{PIE}}(\mathbf{x})$ requires a reverse diffusion process with time-dependent marginal distribution satisfying: $P_{\mathrm{bias},t}(\mathbf{x}(t)) \propto P_t(\mathbf{x}(t))\exp\{-\beta U_t(\mathbf{x}(t))\}$

where U_t is a time-dependent energy function defined as[45]

$$U_t(\mathbf{x}(t)) = -k_{\mathrm{B}}T \log \mathbb{E}_{p(\mathbf{x}(0)|\mathbf{x}(t))}\left(e^{-\beta U_{\mathrm{ext}}(\mathbf{x}(0))}\right). \tag{6}$$

$p(\mathbf{x}(0)|\mathbf{x}(t))$ in Eq.(6) is the probability of $\mathbf{x}(0)$ conditional on the location of $\mathbf{x}(t)$ in the unmodified revert diffusion process. To further simplify the expression of U_t, we adopted an approximation used in the manifold constraint algorithm for diffusion model:

$$p(\mathbf{x}(0)|\mathbf{x}(t)) \approx \delta(\mathbf{x}(0) - \hat{\mathbf{x}}(0)), \tag{7}$$

where $\delta(\cdot)$ is the Dirac-delta function and $\hat{\mathbf{x}}(0) = \mathbb{E}(\mathbf{x}(0)|\mathbf{x}(t))$. In certain types of diffusion models such as DDPM, $\hat{\mathbf{x}}_0$ is an explicit function of $\mathbf{x}(t)$ and t: $\hat{\mathbf{x}}_0 = \mathbf{h}(\mathbf{x}(t),t)$[48]. With the approximated $p(\mathbf{x}(0)|\mathbf{x}(t)$, Eq.(6) can be simplified as $U_t(\mathbf{x}(t)) \approx U_{\mathrm{ext}}(\mathbf{h}(\mathbf{x}(t),t))$. The approximation of using Eq.(7) works better when t is close to 0. Therefore, care must be taken at the initial stage of backward diffusion. A scaling factor is typically used to stabilize the initial reverse diffusion. With $U_{\mathrm{ext}}(\hat{\mathbf{x}}(0))$, we can calculate force $\mathbf{F}_0 = -\nabla U_{\mathrm{ext}}$ and we can push the force pack to the diffusion model: $\mathbf{F}(\mathbf{x}(t),t) = \mathbf{F}_0(\mathbf{h}(\mathbf{x}(t),t))$. Although using U_{ext} provides a clear picture about how to connect DiffPIE to the Boltzmann distribution of protein with PEI. We can relax the requirement of using an interaction potential U_{ext} for the structure generation task. This provides a more flexible tool for users to design $\mathbf{F}_0$, such as $\mathbf{F}_0$ representing non-equilibrium perturbations from the environment.

In this work, two examples of $\mathbf{F}0$ are used: one with an explicitly defined U_{ext}, and another that directly uses $\mathbf{F}_0$ without defining U_{ext}. In the example of the P3-F cyclic peptide, we use the FES of six distances (Fig. 2(b)) as U_{ext}. This six-dimensional potential was fitted using a logarithmic function of a mixture of Gaussians. We emphasize that the selected collective variables (CVs) provide joint information on both the distance and orientation between the two cystines. In the second example, $\mathbf{F}_0$ is designed as follows. First, we use metadynamics to generate a binding FES, A_i, for a dipeptide molecule that represents the residue i of the protein on the gold surface. A_i is a function of three distances, presented in Fig. 3(b), and can be viewed as a function of $\mathbf{y}_1$, $\mathbf{y}_2$, and $\mathbf{y}3$, representing the Cα atom ($\mathbf{y}_2$) and the two methyl carbons ($\mathbf{y}_1$ and $\mathbf{y}_3$) from the two capping groups. We define the biasing force on the i-th C$_\alpha$, $\mathbf{F}_0^{(i)}$, as $\mathbf{F}_0^{(i)} = -\nabla_{\mathbf{y}_1} A_i|_{\mathbf{y}_1=\mathbf{x}_{i-1},\mathbf{y}_2=\mathbf{x}_i,\mathbf{y}_3=\mathbf{x}_{i+1}}$ where $\mathbf{x}_{i-1}$, $\mathbf{x}_i$, and $\mathbf{x}_{i+1}$ are the coordinates of the $(i-1)$-th, i-th, and $(i+1)$-th Cα atoms, respectively. Please refer to the Supporting Information for details on the generation of $\mathbf{F}_0$.

Str2Str generates protein structures by composing translational and rotational operations on a fixed local frame for each residue. Biasing forces are applied only to the translational degrees of freedom, i.e. the Cartesian coordinates of Cα, rather than to the orientation degrees of freedom. This is because Tweedie's formula is not directly applicable to orientation spaces (i.e. the $SO(3)$ group)[55], and applying forces back onto the tangent space of $SO(3)$ in the diffusion model may cause numerical instability[45]. In this study, we empirically set $\lambda = 2.0$, and other λ values have also been tested. Choosing a different λ does not change the main conclusions of this paper, although a smaller λ generates more structures that follow the original Str2Str ensemble. Generating side chain conformations consistent with both the backbone structure and PEI is crucial, as interactions between side chains and environmental species significantly impact PEI strength, such as in protein binding affinity to material surfaces. Since local frames in Str2Str only determine the positions of backbone atoms, side chain backmapping is achieved using the packing software FASPR, as recommended by Str2Str[77]. However, directly using this module is not sufficient to generate side chains for proteins under PEI conditions. For example, the side chains generated by FASPR may clash with the material surfaces. Furthermore, in the first example, a proper organic linker structure must also be generated, which is beyond the capabilities of FASPR. To address these challenges, we designed an algorithm for generating missing structural components (e.g. side chains and linkers) from a backbone structure A generated by DiffPIE, by matching appropriate structures obtained from the MD simulations used to generate $\mathbf{F}_0$. In the first example, the two P3-F cystine side chains and the organic linker structures were reconstructed by aligning the P3-F cystine backbone atoms in the structure A with those in the MD simulations. In the second example, the side chain of each residue in structure A that is close to the gold surface is reconstructed by selecting the dipeptide configurations from the MD simulations based on two criteria: (1) the dipeptide backbone atoms must align with the local frame of the corresponding residue in structure A, and (2) the displacement of the target residue in structure A relative to the gold surface must match that observed in the MD simulations. In both examples, FASPR was used to generate any remaining side chains not reconstructed by this matching procedure.

References

1. Lovelock, S. L. *et al.* The road to fully programmable protein catalysis. *Nature* **606**, 49–58 (2022).
2. Pawson, T. & Nash, P. Protein–protein interactions define specificity in signal transduction. *Genes & development* **14**, 1027–1047 (2000).
3. Rosenberg, A. S. Effects of protein aggregates: an immunologic perspective. *The AAPS journal* **8**, E501–E507 (2006).

4. Rothman, J. E. Mechanisms of intracellular protein transport. *Nature* **372**, 55–63 (1994).
5. Sousa, S. F., Fernandes, P. A. & Ramos, M. J. Protein–ligand docking: current status and future challenges. *Proteins: structure, function, bioinformatics* **65**, 15–26 (2006).
6. Jones, S. & Thornton, J. M. Principles of protein-protein interactions. *Proc. Natl. Acad. Sci.* **93**, 13–20 (1996).
7. Steitz, T. A. Structural studies of protein–nucleic acid interaction: the sources of sequence-specific binding. *Q. reviews biophysics* **23**, 205–280 (1990).
8. Harter, R. D. & Stotzky, G. Formation of clay-protein complexes. *Soil Sci. Soc. Am. J.* **35**, 383–389 (1971).
9. Liu, J. & Peng, Q. Protein-gold nanoparticle interactions and their possible impact on biomedical applications. *Acta biomaterialia* **55**, 13–27 (2017).
10. Pantelejevs, T. *et al.* A recombinant approach for stapled peptide discovery yields inhibitors of the rad51 recombinase. *Chem. Sci.* **14**, 13915–13923 (2023).
11. Cheng, K.-Y. *et al.* The role of the phospho-cdk2/cyclin a recruitment site in substrate recognition. *J. Biol. Chem.* **281**, 23167–23179 (2006).
12. Wu, Y., Ng, D. Y., Kuan, S. L. & Weil, T. Protein–polymer therapeutics: a macromolecular perspective. *Biomater. science* **3**, 214–230 (2015).
13. Chen, H., Yuan, L., Song, W., Wu, Z. & Li, D. Biocompatible polymer materials: role of protein–surface interactions. *Prog. Polym. Sci.* **33**, 1059–1087 (2008).
14. Fry, D. C. Protein–protein interactions as targets for small molecule drug discovery. *Pept. Sci. Orig. Res. on Biomol.* **84**, 535–552 (2006).
15. Jao, D., Xue, Y., Medina, J. & Hu, X. Protein-based drug-delivery materials. *Materials* **10**, 517 (2017).
16. Drenth, J. *Principles of protein X-ray crystallography* (Springer Science & Business Media, 2007).
17. Mittermaier, A. K. & Kay, L. E. Observing biological dynamics at atomic resolution using nmr. *Trends biochemical sciences* **34**, 601–611 (2009).
18. Cheng, Y. Single-particle cryo-em—how did it get here and where will it go. *Science* **361**, 876–880 (2018).
19. Kobe, B. *et al.* Crystallography and protein–protein interactions: biological interfaces and crystal contacts. *Biochem. Soc. Transactions* **36**, 1438–1441 (2008).
20. Yip, K. M., Fischer, N., Paknia, E., Chari, A. & Stark, H. Atomic-resolution protein structure determination by cryo-em. *Nature* **587**, 157–161 (2020).
21. Bonilla, S. L. & Kieft, J. S. The promise of cryo-em to explore rna structural dynamics. *J. Mol. Biol.* **434**, 167802 (2022). Regulatory Mechanisms through RNA Conformational Switching and Dynamics.
22. Clore, G. M. & Gronenborn, A. M. Determining the structures of large proteins and protein complexes by nmr. *Trends Biotechnol.* **16**, 22–34 (1998).
23. Renaud, J.-P. *et al.* Cryo-em in drug discovery: achievements, limitations and prospects. *Nat. Rev. Drug Discov.* **17**, 471–492 (2018).
24. Kessler, H. Conformation and biological activity of cyclic peptides. *Angewandte Chemie Int. Ed. Engl.* **21**, 512–523 (1982).
25. Bernadó, P., Mylonas, E., Petoukhov, M. V., Blackledge, M. & Svergun, D. I. Structural characterization of flexible proteins using small-angle x-ray scattering. *J. Am. Chem. Soc.* **129**, 5656–5664 (2007).
26. Adochitei, A. & Drochioiu, G. Rapid characterization of peptide secondary structure by ft-ir spectroscopy. *Revue Roumaine de Chimie* **56**, 783–791 (2011).
27. Yang, J. T., Wu, C.-S. C. & Martinez, H. M. [11] calculation of protein conformation from circular dichroism. In *Enzyme Structure Part K*, vol. 130 of *Methods in Enzymology*, 208–269 (Academic Press, 1986).
28. Majewski, M. *et al.* Machine learning coarse-grained potentials of protein thermodynamics. *Nat. Commun.* **14**, 5739 (2023).
29. Pan, A. C. *et al.* Atomic-level characterization of protein–protein association. *Proc. Natl. Acad. Sci.* **116**, 4244–4249 (2019).

30. Ishraaq, R. & Das, S. All-atom molecular dynamics simulations of polymer and polyelectrolyte brushes. *Chem. Commun.* **60**, 6093–6129 (2024).

31. Souza, P. C. *et al.* Protein–ligand binding with the coarse-grained martini model. *Nat. communications* **11**, 3714 (2020).

32. Dandekar, B. R. & Mondal, J. Capturing protein–ligand recognition pathways in coarse-grained simulation. *The J. Phys. Chem. Lett.* **11**, 5302–5311 (2020).

33. Baaden, M. & Marrink, S. J. Coarse-grain modelling of protein–protein interactions. *Curr. opinion structural biology* **23**, 878–886 (2013).

34. Vilaseca, P., Dawson, K. A. & Franzese, G. Understanding and modulating the competitive surface-adsorption of proteins through coarse-grained molecular dynamics simulations. *Soft Matter* **9**, 6978–6985 (2013).

35. Shi, R., Qian, H.-J. & Lu, Z.-Y. Coarse-grained molecular dynamics simulation of polymers: Structures and dynamics. *Wiley Interdiscip. Rev. Comput. Mol. Sci.* **13**, e1683 (2023).

36. Noid, W. G. Perspective: Coarse-grained models for biomolecular systems. *The J. chemical physics* **139** (2013).

37. Jin, J. & Voth, G. A. Statistical mechanical design principles for coarse-grained interactions across different conformational free energy surfaces. *The J. Phys. Chem. Lett.* **14**, 1354–1362 (2023).

38. Jumper, J. *et al.* Highly accurate protein structure prediction with alphafold. *Nature* **596**, 583–589 (2021).

39. Ruff, K. M. & Pappu, R. V. Alphafold and implications for intrinsically disordered proteins. *J. molecular biology* **433**, 167208 (2021).

40. Beal, D. J. Esm 2.0: State of the art and future potential of experience sampling methods in organizational research. *Annu. Rev. Organ. Psychol. Organ. Behav.* **2**, 383–407 (2015).

41. Lu, J., Zhong, B., Zhang, Z. & Tang, J. Str2str: A score-based framework for zero-shot protein conformation sampling. *arXiv preprint arXiv:2306.03117* (2023).

42. Hayes, T. *et al.* Simulating 500 million years of evolution with a language model. *Science* eads0018 (2025).

43. Abramson, J. *et al.* Accurate structure prediction of biomolecular interactions with alphafold 3. *Nature* **630**, 493–500 (2024).

44. Yu, Z., Liu, Y., Lin, G., Jiang, W. & Chen, M. Esmadam: a plug-and-play all-purpose protein ensemble generator. *bioRxiv* (2025).

45. Wang, Y. *et al.* Protein conformation generation via force-guided se (3) diffusion models. *arXiv preprint ARXIV.2403.14088* (2024).

46. Hwang, M., Lee, Z., Kim, K.-S., Kim, K. & Yang, E. Hybridlinker: Topology-guided posterior sampling for enhanced diversity and validity in 3d molecular linker generation. *arXiv preprint arXiv:2502.17349* (2025).

47. Kulytė, P. *et al.* Improving antibody design with force-guided sampling in diffusion models. *arXiv preprint arXiv:2406.05832* (2024).

48. Ho, J., Jain, A. & Abbeel, P. Denoising diffusion probabilistic models. *NeurIPS* **33**, 6840–6851 (2020).

49. Song, Y. *et al.* Score-based generative modeling through stochastic differential equations. *arXiv preprint arXiv:2011.13456* (2020).

50. Lamtyugina, A., Behera, A. K., Nandy, A., Floyd, C. & Vaikuntanathan, S. Score-based generative diffusion with" active" correlated noise sources. *arXiv preprint arXiv:2411.07233* (2024).

51. Liu, Y. *et al.* Exendiff: An experiment-guided diffusion model for protein conformational ensemble generation. *bioRxiv* 2024–10 (2024).

52. Cao, L. *et al.* Design of protein-binding proteins from the target structure alone. *Nature* **605**, 551–560 (2022).

53. Corso, G., Stärk, H., Jing, B., Barzilay, R. & Jaakkola, T. Diffdock: Diffusion steps, twists, and turns for molecular docking. *ICLR* (2023).

54. Chung, H., Sim, B., Ryu, D. & Ye, J. C. Improving diffusion models for inverse problems using manifold constraints. *Adv. Neural Inf. Process. Syst.* **35**, 25683–25696 (2022).

55. Chung, H., Kim, J., Mccann, M. T., Klasky, M. L. & Ye, J. C. Diffusion posterior sampling for general noisy inverse problems. *arXiv preprint arXiv:2209.14687* (2022).

56. Iegre, J. *et al.* A cell-active cyclic peptide targeting the nrf2/keap1 protein–protein interaction. *Chem. Sci.* **14**, 10800–10805 (2023).

57. Barducci, A., Bussi, G. & Parrinello, M. Well-tempered metadynamics: a smoothly converging and tunable free-energy method. *Phys. Rev. Lett.* **100**, 020603 (2008).

58. Pietrucci, F. & Laio, A. A collective variable for the efficient exploration of protein beta-sheet structures: Application to sh3 and gb1. *J. Chem. Theory Comput.* **5**, 2197–2201 (2009).

59. McInnes, L., Healy, J. & Melville, J. Umap: Uniform manifold approximation and projection for dimension reduction. *arXiv preprint arXiv:1802.03426* (2018).

60. Zhang, C. *et al.* De novo design of cyclic peptide binders based on fragment docking and assembling. *J. Chem. Inf. Model.* (2025).

61. Hart, K. M., Ho, C. M., Dutta, S., Gross, M. L. & Bowman, G. R. Modelling proteins' hidden conformations to predict antibiotic resistance. *Nat. communications* **7**, 12965 (2016).

62. Amendola, V., Pilot, R., Frasconi, M., Maragò, O. M. & Iatì, M. A. Surface plasmon resonance in gold nanoparticles: a review. *J. physics: Condens. matter* **29**, 203002 (2017).

63. Dreaden, E. C., Alkilany, A. M., Huang, X., Murphy, C. J. & El-Sayed, M. A. The golden age: gold nanoparticles for biomedicine. *Chem. Soc. Rev.* **41**, 2740–2779 (2012).

64. Daraee, H. *et al.* Application of gold nanoparticles in biomedical and drug delivery. *Artif. cells, nanomedicine, biotechnology* **44**, 410–422 (2016).

65. Jahan Sajib, M. S., Sarker, P., Wei, Y., Tao, X. & Wei, T. Protein corona on gold nanoparticles studied with coarse-grained simulations. *Langmuir* **36**, 13356–13363 (2020).

66. Teichroeb, J., Forrest, J. & Jones, L. Size-dependent denaturing kinetics of bovine serum albumin adsorbed onto gold nanospheres. *The Eur. Phys. J. E* **26**, 411–415 (2008).

67. Wang, A., Vangala, K., Vo, T., Zhang, D. & Fitzkee, N. C. A three-step model for protein–gold nanoparticle adsorption. *The J. Phys. Chem. C* **118**, 8134–8142 (2014).

68. Deka, J., Paul, A. & Chattopadhyay, A. Estimating conformation content of a protein using citrate-stabilized au nanoparticles. *Nanoscale* **2**, 1405–1412 (2010).

69. Bellucci, L. *et al.* The interaction with gold suppresses fiber-like conformations of the amyloid β (16–22) peptide. *Nanoscale* **8**, 8737–8748 (2016).

70. Ribeiro, J. M. L., Bravo, P., Wang, Y. & Tiwary, P. Reweighted autoencoded variational bayes for enhanced sampling (rave). *J. Chem. Phys.* **149**, 072301 (2018).

71. Vani, B. P., Aranganathan, A., Wang, D. & Tiwary, P. Alphafold2-rave: From sequence to boltzmann ranking. *J. chemical theory computation* **19**, 4351–4354 (2023).

72. Vani, B. P., Aranganathan, A. & Tiwary, P. Exploring kinase asp-phe-gly (dfg) loop conformational stability with alphafold2-rave. *J. chemical information modeling* **64**, 2789–2797 (2023).

73. Lamim Ribeiro, J. M. & Tiwary, P. Toward achieving efficient and accurate ligand-protein unbinding with deep learning and molecular dynamics through rave. *J. chemical theory computation* **15**, 708–719 (2018).

74. Watson, J. L. *et al.* De novo design of protein structure and function with rfdiffusion. *Nature* **620**, 1089–1100 (2023).

75. Wu, K. E. *et al.* Protein structure generation via folding diffusion. *Nat. communications* **15**, 1059 (2024).

76. Anderson, B. D. Reverse-time diffusion equation models. *Stoch. Process. Their Appl.* **12**, 313–326 (1982).

77. Huang, X., Pearce, R. & Zhang, Y. Faspr: an open-source tool for fast and accurate protein side-chain packing. *Bioinformatics* **36**, 3758–3765 (2020).

Acknowledgements

Y.W. and M.C. gratefully acknowledge the support of the American Chemical Society Petroleum Research Fund (Grand Number 67307).

Author contributions statement

Y.W. and M.C. designed the research, Y.W. conducted the calculations, Y.W. and M.C. analyzed the results. All authors reviewed the manuscript.