CONFIDE: Hallucination Assessment for Reliable Biomolecular Structure Prediction and Design

Zijun Gao[1†], Mutian He[3†], Shijia Sun[3], Hanqun Cao[1], Jingjie Zhang[1], Zihao Luo[4], Xiaorui Wang[2], Xiaojun Yao[3], Chang-Yu Hsieh[2*], Chunbin Gu[1*], Pheng Ann Heng[1]

[1]Department of Computer Science and Engineering, The Chinese University of Hong Kong, Hong Kong.
[2]College of Pharmaceutical Sciences, Zhejiang University, Hangzhou, 310027, Zhejiang Province, China.
[3]Faculty of Applied Sciences, Macao Polytechnic University, Macao.
[4]School of Mechanical and Electrical Engineering, University of Electronic Science and Technology of China, Chengdu, 611731, Sichuan Province, China.

*Corresponding author(s). E-mail(s): kimhsieh@zju.edu.cn; guchunbin200888@gmail.com ;
†These authors contributed equally to this work.

Abstract

Reliable evaluation of protein structure predictions remains challenging, as metrics like pLDDT capture energetic stability but often miss subtle errors such as atomic clashes or conformational traps reflecting topological frustration within the protein-folding energy landscape. We present CODE (Chain of Diffusion Embeddings), a self-evaluating metric empirically found to quantify topological frustration directly from the latent diffusion embeddings of the AlphaFold3 series of structure predictors in a fully unsupervised manner. Integrating this with pLDDT, we propose CONFIDE, a unified evaluation framework that combines energetic and topological perspectives to improve the reliability of AlphaFold3 and related models. CODE strongly correlates with protein folding rates driven by topological frustration, achieving a correlation of 0.82 compared to pLDDT's 0.33 (a relative improvement of 148%). CONFIDE significantly enhances the reliability of quality evaluation in molecular glue structure prediction benchmarks, achieving a Spearman correlation of 0.73 with RMSD, compared to pLDDT's correlation of 0.42, a relative improvement of 73.8%. Beyond quality assessment,

our approach applies to diverse drug-design tasks, including all-atom binder design, enzymatic active-site mapping, mutation-induced binding-affinity prediction, nucleic acid aptamer screening, and flexible protein modeling. By combining data-driven embeddings with theoretical insight, CODE and CONFIDE outperform existing metrics across a wide range of biomolecular systems, offering robust and versatile tools to refine structure predictions, advance structural biology, and accelerate drug discovery.

Keywords: Protein Folding, Topological Frustration, Energy Landscape, AlphaFold3

1 Introduction

The advent of deep learning in structural biology, epitomized by AlphaFold3 (AF3), has transformed our ability to predict the three-dimensional structures of biomolecules including proteins, nucleic acids, and their complexes [1]. AF3 often achieves near-experimental accuracy across a broad range of targets, accelerating progress in structural biology and drug development. Yet, despite these advances, the internal mechanisms by which AF3 generates its predictions, and the reliability of its self-assessed confidence, remain incompletely understood.

AF3 reports a predicted Local Distance Difference Test (pLDDT) score to quantify structural confidence. While pLDDT generally correlates with model accuracy, it fails in critical cases involving complex molecular assemblies or atypical conformations. Similar to other large-scale generative models, AF3 can produce hallucination, i.e. structures with deceptively high confidence but substantial deviations from experimentally determined conformations. This high-confidence/low-accuracy paradox undermines downstream applications, particularly in drug discovery, where misleading structural models may misdirect experimental resources. A deeper understanding of AF3's generative process is therefore crucial for mitigating hallucinations and improving reliability, especially in data-scarce scenarios.

Classical interpretations of AlphaFold draw on protein-folding energy landscape theory. AlphaFold2 (AF2) was hypothesized to learn an energy function, using coevolutionary information from multiple sequence alignments (MSAs) to identify approximate global minima, followed by refinement through its structure module [2]. While this framework helps explain how AF2 resolves energetic frustration, it does not account for the additional constraints imposed by protein topology. Protein folding also reflects topological frustration, which arises from chain connectivity, conformational entropy losses, and the need to traverse rugged folding funnels without becoming trapped in local minima [3–5]. These topological constraints profoundly shape folding pathways and increase structural complexity, yet these intricate details of folding process are not well captured by existing data-driven methods.

Building upon these insights, we investigate whether the generative process of the AlphaFold3 series of structure predictors can be understood more comprehensively through the lens of energy landscape theory. Here, we introduce a theoretical framework for interrogating the predictions of the AlphaFold3 series of structure predictors

by quantifying topological frustration. Inspired by chain-of-thought reasoning in large language models (LLMs), we formalize the progressive latent representations within the diffusion-based structure module of the AlphaFold3 series of structure predictors as the Chain of Diffusion Embeddings (CODE) [6, 7]. We hypothesize that these progressively refined embeddings capture long-range inter-residue relationships reflective of backbone constraints and topological complexity [8]. As mentioned above, these subtle structural constraints cannot be easily recognized by more localized metrics such as pLDDT. To validate this, we conducted proof-of-concept experiments demonstrating that CODE strongly correlates with protein folding rates driven by topological frustration, achieving a correlation of 0.82 compared to pLDDT's 0.33 (a relative improvement of 148%).

Recognizing the complementarity between CODE and pLDDT, we develop CONFIDE (CONFIdence-integrated Diffusion Embedding), a self-evaluating metric that unifies energetic (pLDDT) and topological (CODE) perspectives to provide a more robust assessment of structural predictions. Across diverse benchmarks—including ternary complexes, flexible and rigid proteins, and other challenging datasets—CONFIDE achieves superior performance compared to pLDDT, with particularly notable improvements in molecular-glue complexes (31% Spearman correlation enhancement). Furthermore, in downstream applications, CONFIDE operates in a fully unsupervised manner to guide binder design, enzyme catalytic site identification, mutation-induced affinity prediction, and inhibitor/nucleic acid aptamer screening, achieving excellent results across all tasks. Notably, CONFIDE improves the success rate of binder design for IAI by 13%. Additionally, CONFIDE accurately detects affinity changes from resistance mutations in the BTK protein against Fenebrutinib, with a Spearman correlation of 0.83, far exceeding pLDDT's negligible 0.03.

By integrating embedding-based reasoning with theoretical insights into protein folding, CODE and CONFIDE provide a unified, interpretable framework for detecting structural hallucinations and improving model reliability. Addressing a key limitation in diffusion-based 3D generative modeling—widely used in structural biology, materials science, and engineering [9]—this framework universally applies to computational models with progressive structural encoding, ensuring robust evaluation of topological rationality across diverse domains.

2 Results

2.1 Overview

Inspired by the topological frustration theory in protein folding, we reformulated the diffusion embedding trajectories of the AF3 series structure predictors into CODE, a metric that captures topological frustration overlooked by the conventional confidence metric pLDDT. We propose CONFIDE, a unified framework integrating topological frustration and pLDDT-represented energetic frustration to comprehensively characterize the protein folding energy landscape (see Figure 1 for a schematic of the framework). To establish CODE's significant correlation with topological frustration, we conducted proof-of-concept experiments on single-domain proteins with minimal energetic frustration. These experiments demonstrated a strong correlation between

CODE and topological frustration, a weak correlation between pLDDT and topological frustration, and weak coupling between CODE and pLDDT, providing information gain-based interpretability for CONFIDE's robustness over individual metrics.

We categorized the applications of CODE and CONFIDE into two classes: hallucination detection and downstream applications. For hallucination detection, we evaluated ternary complexes, flexible and rigid proteins, and benchmark datasets (PDB test and CASP15). Comprehensive assessments of classification performance and correlation underscored CONFIDE's superior performance, with improvements explained from a topological frustration perspective based on atomic clash rates. In downstream applications, CODE and CONFIDE were applied to de novo binder design, enzyme active site identification, drug resistance mutation prediction, and virtual screening. These applications highlight the broad potential of CODE and CONFIDE as unsupervised self-evaluation tools, offering biologically reasonable interpretations (e.g., steric hindrance explaining drug resistance mutations) and enhanced design performance (e.g., generating more stable binder structures).

2.2 Topological Frustration-Mediated Interplay between CODE and Protein Folding Kinetics

Experimental evidence, together with comparisons to theoretical models, shows that proteins are robust folding entities: they can reach their native conformation under a wide range of conditions and preserve that fold even after multiple mutations in their amino acid sequences [10–13]. For small (single-domain) proteins, experiments further reveal that evolution has selected sequences whose energetic frustration on the free-energy landscape is so low that sensitivity to any particular mutation is the exception rather than the rule. Two observable consequences follow: (i) these proteins fold very rapidly on the microsecond-to-second time scale under typical laboratory conditions and (ii) the structural details of their folding mechanisms are governed mainly by what is commonly called topological frustration [3–5].

Topological frustration effectively describes the roughness of the folding landscape that arises from the interplay between chain connectivity and energetic bias toward the native state [14]. More precisely, this ruggedness results from a heterogeneous loss of conformational entropy associated with the formation of partially folded structures on the free-energy landscape. For proteins in which the heterogeneity of conformational entropy far exceeds that of energy, the principal folding pathways on the free-energy surface are tightly constrained and shaped by the protein's topology [15, 16].

To more intuitively explain how topological frustration affects the protein folding process and ultimately reflects on the folding rate, we provide three specific protein cases spanning different orders of magnitude of folding rates which is shown in Figure 2a. AcP has mild mixing entropy early in the folding process so the protein may sample a large volume of configuration space. A large bottleneck region slows the folding of this protein around the transition-state barrier. Diffuse structure flickers between conformations before the transition-state region, then the formation must become more ordered. Unfolded AcP may sample a large region of the configuration space available to it before it manages to pass through the narrow bottleneck that leads to the native state. These effects cause AcP to be one of the slowest two-state proteins.

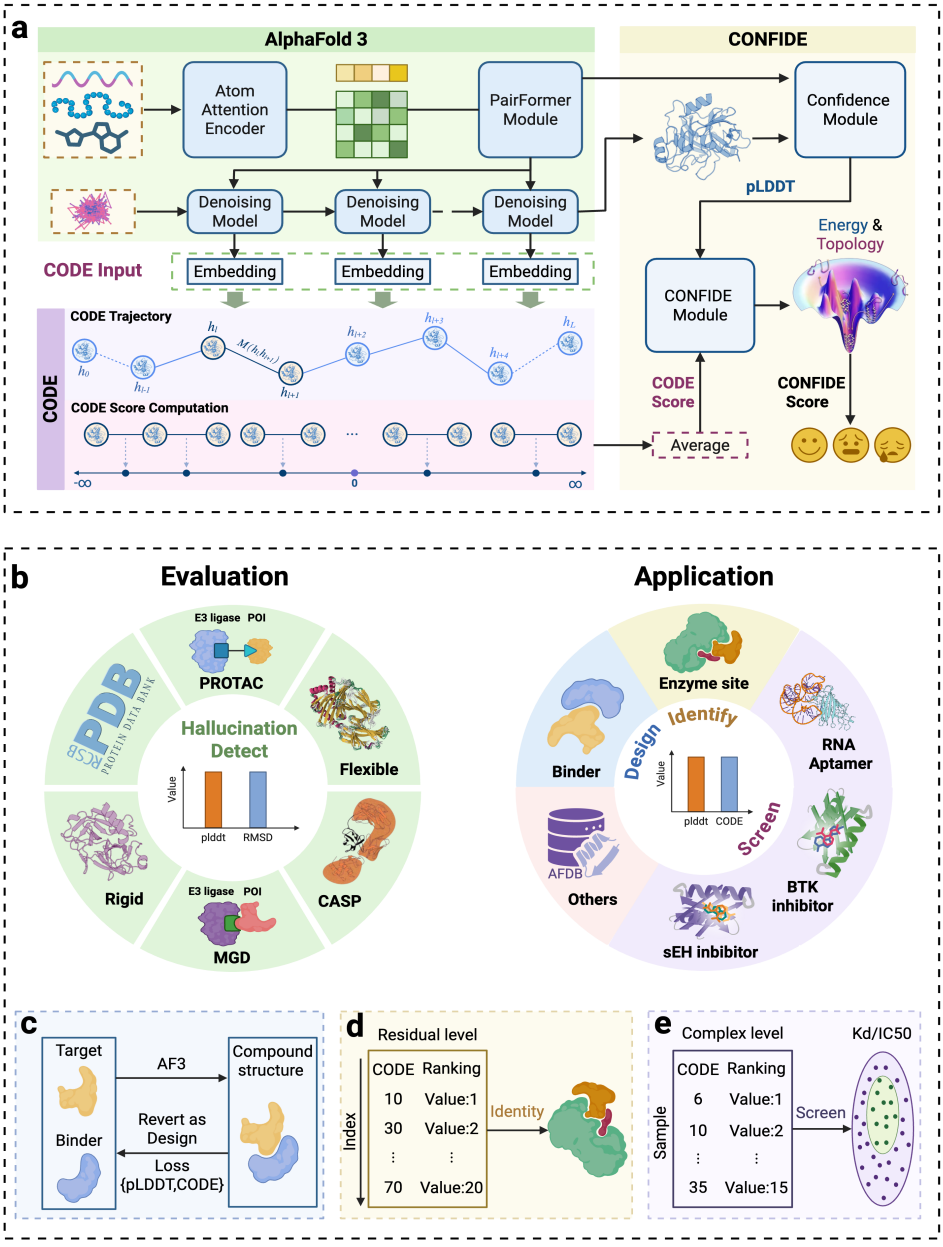


Fig. 1 Overview of CODE and CONFIDE. (a) The computational process of the CODE module and the CONFIDE module. First, the sequence of any biological molecule is input to the Alphafold3 series model to obtain the PairFormer embedding. Then, the embedding guides the noise to generate the corresponding structure. We construct the embedding change trajectory in the Denoising Module and calculate the CODE score to characterize topological frustration. The generated structure is used as the input of the Confidence Module to obtain the confidence score to represent energy frustration. Finally, the CONFIDE Module comprehensively considers the two frustrations and analyzes to obtain a more comprehensive CONFIDE score. (b) Shows the application scenarios of CODE and CONFIDE. In the hallucination detection of structure prediction, we verified 6 challenging datasets covering almost all biological molecules. We also expanded the evaluation to five application scenarios covering three types of tasks: binder design, enzyme site detection, and complex screening. (c) Shows the workflow of CODE application to hallucination-based binder design. (d) Shows the schematic diagram of CODE for residue level site detection. (e) Shows the schematic diagram of CODE/CONFIDE for virtual screening

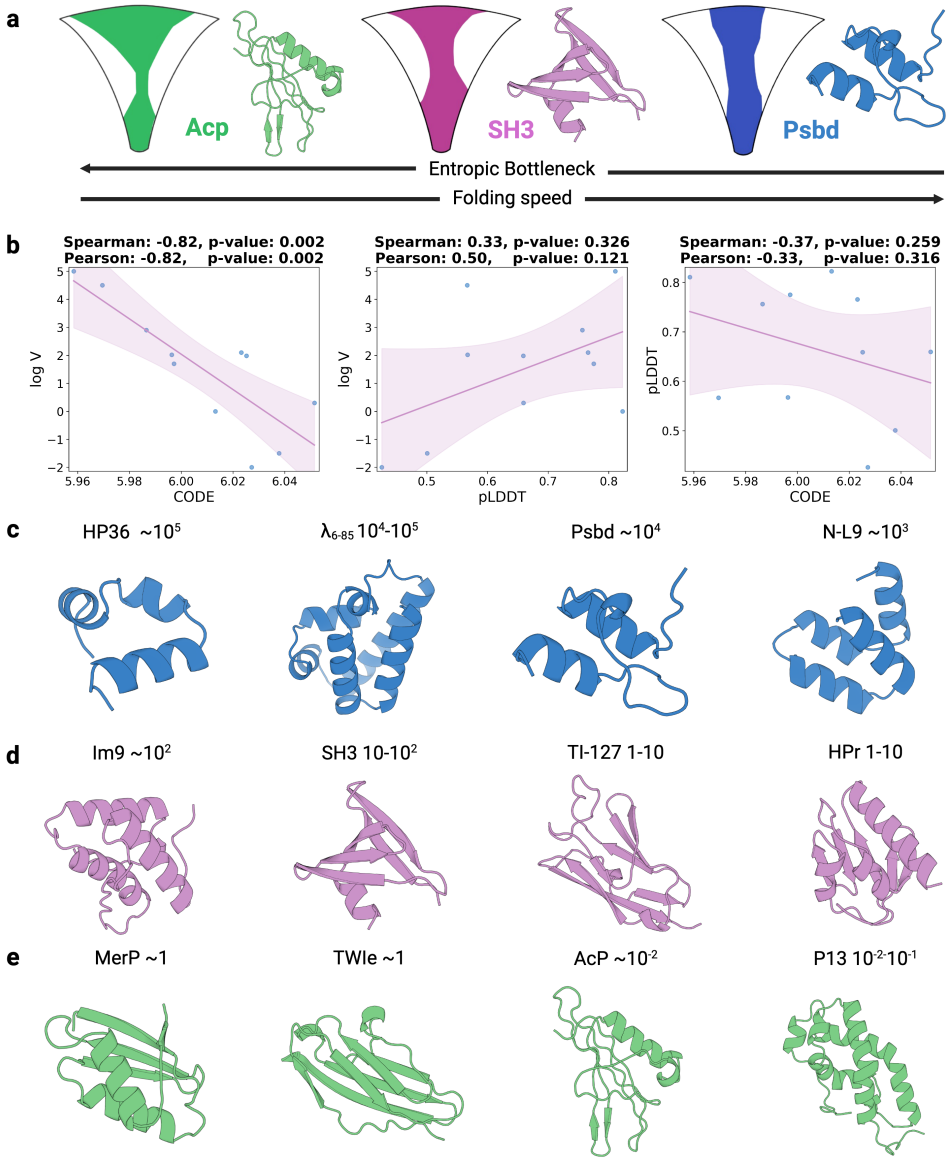


Fig. 2 CODE implicitly encodes the protein folding dynamics mediated by topological frustration. (a) Three proteins with different orders of magnitude of folding rates are shown. AcP has mild mixing entropy early in the folding process. A large bottleneck region slows the folding of this protein around the transition-state barrier. SH3 is a slower folding protein and shows little mild mixing entropy initially, and then a small bump appears in the route measure curve. Psbd is a very fast folding protein and shows almost constant mixing entropy measure throughout folding. (b) Correlation analysis among CODE, pLDDT and log V with linear regression fits and 95% confidence intervals. The first figure shows a strong Spearman correlation between CODE and log V. The second one shows a weak Spearman correlation between pLDDT and log V. The third one shows that there is only a weak correlation of -0.37 between CODE and pLDDT, indicating that they depict the energy landscape of protein folding from different perspective. (c) Protein structures with fast folding rates, showing their names and specific folding rates. (d) Protein structures with intermediate folding rates, showing their names and specific folding rates. (e) Protein structures with slower folding rates, showing their names and specific folding rates.

SH3 is a slower folding protein and shows little mixing entropy initially (similar to AcP), and then a small bump appears in the entropy measure curve. This indicates that at that time some portion of the protein is more likely to be structured relative to the rest of the protein. These contacts may need to be in place for SH3 to continue folding. After this short region is crossed, the rest of the curve is quite uniform and low. This protein never has its configuration space strongly reduced as it searches for its native state. The overall shape of the curve is similar to AcP, although the entropic bottleneck is smaller for SH3. Psbd is a very fast folding protein and shows almost constant mixing entropy measure throughout folding. The folding process from the unfolded state to the folded state is accompanied by a very slow decrease in the width of the sampled configuration space.

We conducted proof of concept tests of the link between CODE and folding-kinetics framework on a database of protein models that are completely free of energetic frustration [14, 17]. Proteins were selected whose experimentally measured folding rates span more than six orders of magnitude and whose overall fold topologies and secondary-structure compositions are diverse. The final data set consists of 16 two-state, single-domain proteins ranging from 36 to 115 residues in length. Structural information and detailed information on folding rates of selected proteins can be found in Supplementary Information S1.1. For detailed calculation methods of pLDDT and CODE, please refer to Section Methods 3 and Supplementary Information S4. Finally, we calculate the correlation coefficients of pLDDT and CODE with respect to the natural logarithm of the folding rate.

As shown in Figure 2b, CODE and the protein folding rate log V have a strong spearman correlation of -0.818, with a p-value of 0.002, while pLDDT has only a weak spearman correlation of 0.327, with a p-value of 0.326. This result suggests that CODE likely reflects topological frustration during protein folding, supported by a strong Spearman correlation with a statistically significant p-value. Protein folding is a process from a high-entropy unfolded state to a low-entropy native state, involving multiple intermediate states. Topological frustration may cause proteins to fall into local energy minima in the folding funnel, prolonging the folding time. High values of CODE may correspond to complex folding funnels containing more or deeper local traps. pLDDT, on the other hand, reflects more the energy stability of static structures rather than the dynamic process of folding dynamics. The folding rate depends not only on the stability of the final structure, but also on the folding path, transition states, and topological frustration. The weak correlation of pLDDT suggests that it may not be able to effectively capture these dynamic factors. Furthermore, as shown in the right side of Figure 2b, the low correlation of -0.372 between CODE and pLDDT with a p-value of 0.259 suggests that they are a set of features that tend to be weakly orthogonal and can be used together to provide a more comprehensive analysis of the folding process, where CODE captures topological frustration and pLDDT provides energy stability information.

2.3 Mitigating Hallucination Across Diverse Systems

Modern structure predictors have essentially reached experimental resolution for single-chain proteins, yet they still display a striking limitation in biologically realistic settings–the phenomenon we term "structural hallucination" [18, 19]. We define a hallucinated structure as one for which the model's internal confidence (pLDDT) is high, while the all-atom RMSD to the corresponding experimental structure remains unacceptably large. This high-confidence/low-accuracy mismatch can mis-rank drug-discovery campaigns, diverting resources toward spurious targets or lead compounds [20, 21]. We introduce CODE and CONFIDE, an evaluator that operates without external labels and is therefore naturally suited to data-poor or entirely novel protein families. To probe its application breadth, we benchmark CODE across three challenging regimes: (1) PROTAC and molecular-glue ternary complexes, which demand simultaneous assessment of the E3 ligase, the protein of interest and the small molecule, and in which current networks struggle with the dynamic nature of these multi-component assemblies, often overestimating small-molecule interfaces due to insufficient training data on such complex systems; (2) flexible proteins with intrinsically disordered regions or multiple conformational states that present challenges as current methods are optimized for single, well-folded structures and cannot adequately capture the ensemble nature of these dynamic systems and (3) a general benchmark comprising PDB Test and CASP15 targets, spanning membrane, nucleic acid binding and large multimeric assemblies in which data bias precipitates hallucination of unseen structures. We also evaluated the performance on rigid proteins. The detailed experimental analysis will be provided in the Supplementary Information S2.2 together with the results of PDB Test and CASP15 in the Supplementary Information S2.3.

CODE adds an intrinsic quality evaluation layer to any structure predictor, filtering hallucinated structures without additional supervision. Coupled with high-throughput experimental validation in the future, CODE is poised to (i) accelerate PROTAC/molecular-glue optimization by removing erroneous conformers early in the design–make–test loop and enabling construction of large, trustworthy datasets, and (ii) enhance the next generation of folding networks by serving as a self-distillation weight or adapter-gating signal that supplies real-time uncertainty feedback. Together, these advances offer a systematic remedy for structural hallucination and provide a more reliable structural foundation for AI-driven drug discovery.

2.3.1 Performance in Ternary Complex Structure Prediction

Targeted protein degradation (TPD) is a revolutionary therapeutic strategy that not only inhibits or activates proteins, but directly eliminates pathogenic proteins [22, 23]. At present, PROTAC is developing rapidly as one of the main methods of TPD, with more than 20 candidate drugs entering clinical trials and nearly 60 in preclinical development. PROTAC uses the ubiquitin-proteasome system to achieve targeted protein degradation. They promote the formation of a ternary complex between E3 ligase and target protein (POI), resulting in ubiquitination and subsequent degradation of the target protein [24]. Molecular glues are another important class of protein degraders

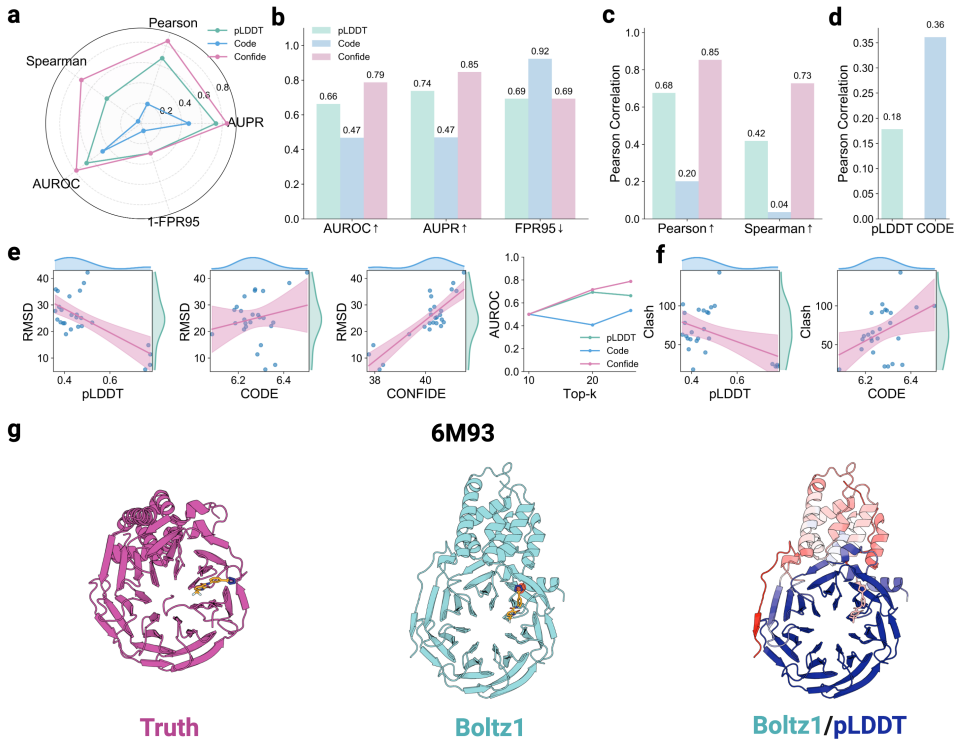


Fig. 3 Performance analysis on ternary complex structure prediction. (a) Radar chart of the five evaluation metrics of three scores (pLDDT, CODE and CONFIDE). (b) Histogram of classification evaluation metrics of three scores. (c) Histogram of correlation evaluation metrics of three scores. (d) Spearman correlation between the number of atomic conflicts and pLDDT/CODE. The ternary complex structure predicted by Boltz1 has physically abnormal local conformations, such as steric conflicts caused by atoms being less than 1.5 angstroms apart in space. CODE can capture the high topological frustration caused by these conformations, thus significantly improving the performance of both systems compared to pLDDT. (e) Scatter plots of the fit and distribution of pLDDT, CODE and CONFIDE with RMSD, including linear regression lines with 95% confidence intervals, as well as AUROC for different classification thresholds. (f) Scatter plots of the fit and distribution of pLDDT and CODE with the number of atom clashes, including linear regression lines with 95% confidence intervals. (g) The true structure, predicted structure, and predicted structure colored by pLDDT of 6M93 in MGD are shown from left to right. The red balls in the middle figure indicate atomic conflicts.

that promote protein degradation by directly binding to and stabilizing the interaction between E3 ligase and substrate protein [25, 26]. Unlike PROTAC, molecular glues are usually small molecule compounds that do not require a linker structure and can induce proteins that do not interact with each other to form stable degradation complexes. Well-known examples include immunomodulatory drugs such as lenalidomide, which can recruit specific substrate proteins to the CRBN E3 ligase complex for degradation [27].

With the continuous increase in TPD targets and mechanisms, it is essential to develop computational tools to optimize compound design. Current computational

methods include molecular dynamics simulation, molecular docking, and machine learning models [28–32]. Whether it is PROTAC or molecular glue system, the structure prediction of ternary complexes faces great challenges. First, these complexes involve complex interactions between multiple protein components, and their binding modes and conformational changes are difficult to accurately model. Second, the formation of ternary complexes is often a dynamic and transient process, and traditional static structure prediction methods are difficult to capture this time-dependent molecular behavior. In addition, different E3 ligases have different structural characteristics and binding preferences, and the diversity of target proteins further increases the complexity of prediction. For PROTAC, the length, flexibility and chemical properties of the linker significantly affect the stability and geometric configuration of the ternary complex. For molecular glue, it is necessary to accurately predict how small molecules bind to E3 ligases and substrate proteins at the same time and stabilize the interaction interface between them. Most importantly, there is currently a lack of sufficient experimental structural data to train and verify computational models, which makes data-driven structure prediction methods face serious data scarcity problems.

Alphafold3 series structure prediction models have obvious "hallucination" phenomenon in this system. We systematically evaluated the correlation of pLDDT, CODE and CONFIDE with RMSD on PROTACFOLD [33]. The PROTACFOLD dataset is derived from 48 known PROTAC-mediated complexes and 33 molecular glue-mediated complexes in the PDB. Detailed dataset information can be found in the Supplementary Information S1.2. For detailed calculation methods of pLDDT and CODE, please refer to Section Methods 3 and Supplementary Information S4. Finally, we screened all combinations of integer coefficients from -10 to 10 to weight CODE and pLDDT to obtain CONFIDE, and calculated the correlation coefficient under the optimal combination.

In Figure 3a, we first presented an overall comparison of the performance of the three scores in classification and correlation tasks, observing that CONFIDE demonstrates a significant advantage over pLDDT or CODE alone, achieving comprehensive superiority. In the classification task, CONFIDE improved the AUROC by 0.13 and the AUPR by 0.11 compared to pLDDT which is shown in Figure 3b. In the correlation analysis, CONFIDE achieved a Pearson correlation of 0.85, a 25% improvement over pLDDT, and a Spearman correlation of 0.73, a 74% improvement compared to pLDDT's 0.42 which is shown in Figure 3c. In Figure 3e, we displayed the distribution of the three scores across all samples along with their fitted curves, as well as the trend of AUROC changes with different top K selections.

Furthermore, we deeply elucidate the underlying mechanisms for CONFIDE's significant advantages from the theoretical perspective of topological frustration. We observed that in the predicted structures of Boltz1, there were numerous physical anomalies, namely atomic clashes. We define an atomic clash as a spatial distance between atoms of less than 1.5 Å. Such folding leads to a significant increase in topological frustration, which is not reflected in pLDDT scores. To illustrate this phenomenon, we presented a representative molecular glue as a case study (PDB ID: 6M93) in Figure 3g. The leftmost image shows the true crystal structure. The middle image depicts the Boltz1-predicted structure, with atomic clash regions marked by red

spheres. The rightmost image colors the predicted structure based on pLDDT scores, with blue indicating high confidence and red indicating low confidence. To some extent, pLDDT can reflect the quality of structural folding, particularly by assigning notably low scores to disordered or loop regions, indicating that pLDDT can capture this type of energy frustration. However, pLDDT is ineffective in detecting atomic clashes. In contrast, atomic clashes significantly increase topological frustration, enabling CODE to capture such structural prediction inaccuracies. We statistically quantified CODE's advantages. Figure 3d shows the Pearson correlation coefficients of pLDDT and CODE with the number of atomic clashes, where CODE achieved a correlation of 0.36, a 100% improvement over pLDDT's 0.18. Figure 3f illustrates the specific distribution and fitted curves. These significant improvements demonstrate that CODE captures critical complementary information related to topological frustration, providing substantial benefits for evaluating the structural quality of ternary or even higher-order complexes.

2.3.2 Assessment of Flexible Proteins

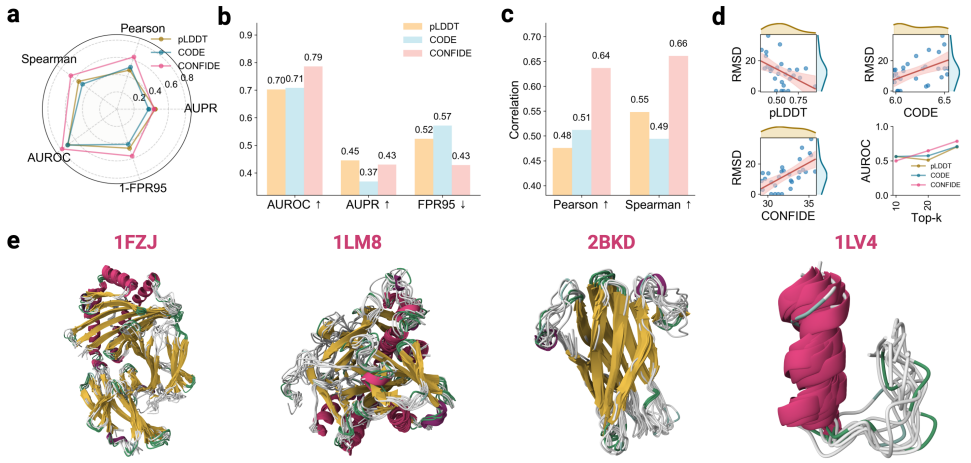


Fig. 4 Performance analysis on flexible protein structure prediction. (a) Radar chart of the five evaluation metrics of three scores (pLDDT, CODE and CONFIDE) on flexible protein. (b) Histogram of classification evaluation metrics of three scores on flexible protein. (c) Histogram of correlation evaluation metrics of three scores on flexible protein. (d) Scatter plots of the fit and distribution of three scores on flexible protein with RMSD, featuring linear regression fits with 95% confidence intervals, as well as AUROC for different classification thresholds. (e) The structures of four flexible proteins efficiently simulated using CABS-flex. The structure is presented in the form of an ensemble of all-atom structures, with different colors representing different secondary structures: yellow for sheet, green for turn, red for helix, and gray for coil.

The flexible disordered region gives the same protein the ability to respond quickly to the environment, achieve multifunctional regulation, and phase separation [34–36]. The difficulty of flexible proteins lies in their essential conformational heterogeneity and dynamic exchange from milliseconds to seconds. Traditional "single structure"

output fails, and sampling must span long time scales and explicitly consider conditions such as pH, ions, and chaperone proteins [37]. At the same time, they are limited by the high cost, noise, and lack of unified evaluation indicators of high-throughput data such as NMR, FRET, and SAXS.

To evaluate the ability of CODE and CONFIDE to detect implicit hallucinations in flexible conformation prediction tasks, we screened 30 disordered proteins with liquid-liquid phase separation from PDB as flexible protein datasets. Detailed dataset information can be found in the Supplementary Information S1.3. For detailed calculation methods of pLDDT and CODE, please refer to Section Methods 3 and Supplementary Information S4. Finally, we compare the performance of confidence score, CODE and CONFIDE in terms of pearson and spearman correlation coefficient. The results are shown in Figure 4.

In Figure 4a, we first presented an overall comparison of the performance of the three scores in classification and correlation tasks, observing that CONFIDE demonstrates a significant advantage over pLDDT or CODE alone, achieving comprehensive superiority. In the classification task, CONFIDE improved the AUROC by 0.09 and reduced the FPR95 by 0.09 compared to pLDDT which is shown in Figure 4b. In terms of Pearson correlation, CONFIDE achieved 0.64, a 33.3% improvement over pLDDT's 0.48. For Spearman correlation, CONFIDE reached 0.66, a 20% improvement compared to pLDDT's 0.55 which is shown in Figure 4c. In Figure 4d, we displayed the distribution of the three scores across all samples along with their fitted curves, as well as the trend of AUROC changes with different top K selections. Figure 4e shows four flexible proteins efficiently simulated using CABS-flex [38]. The structure is presented in the form of an ensemble of all-atom structures, with different colors representing different secondary structures: yellow for sheet, green for turn, red for helix, and gray for coil.

2.4 Diverse Applications in Structural Biology and Drug Discovery

Our application is organized into three levels, capturing the broad scope of modern structure-guided research. (i) At the design level, we integrate CODE with a hallucination-based sequence design model [39, 40] to generate small molecule binders, demonstrating that CODE can steer de novo protein design toward topologically coherent and functionally competent folds. (ii) At the site prediction level, CODE operates zero-shot to identify catalytic and active residues in enzymes, enabling rapid functional annotation of structural data [41]. (iii) At the complex prediction level, CODE ranks protein-ligand and protein-RNA complexes, identifying drug resistance in different BTK protein mutations [42], quantifying inhibitor affinities in congeneric sEH series [43] and screening GFP-binding RNA aptamers [44]. These applications guide hit-to-lead optimization and anticipatory resistance management. Spanning enzymes, kinases, metabolites, drug-like compounds, and structured nucleic acids, CODE's chemistry-agnostic nature, lack of dependence on retraining backbone models, and mechanistically interpretable scoring unify energetic and topological insights.

This breadth—from binder design to functional annotation and resistance prediction—positions CODE as a versatile and high-impact tool for structural biology and drug discovery.

2.4.1 Improving Binder Design Capabilities

The design of protein binders that can specifically interact with molecular targets to mediate catalytic activity, molecular recognition, structural support, and protein-biomolecule interactions represents a critical goal in synthetic biology and therapeutic development [45–47]. Deep learning-based computational methods have made great progress in recent years with the improving protein design success rate. One popular method involves fine-tuning structure prediction models into diffusion models to generate novel structures, such as RfDiffusion [48] and RfDiffusionAA [49]. Another method does not require additional training and directly uses the structure prediction model back propagation to iteratively optimize the sequence until convergence. The back-propagation method can model sequence and structure simultaneously, and has the unique ability to predict the structure of ligands and binders during the design process, thereby being able to explore the induced fit effect during binding, which is currently unavailable to other computational methods [40]. More importantly, this method allows for custom loss function optimization without retraining the model, which provides the possibility for flexible programmability of binder design. Naturally, since the CODE we proposed reflects topological frustration to a certain extent, it can provide strong guidance information for the topological conformation design near the binding site, so it is very suitable for improving the quality of binder design.

To illustrate that CODE can be used as a guide for binder design, we follow BoltzDesign1 [40], which implements iterative sequence design by inverting the all-atom prediction model Boltz-1 [50], and directly optimizes the probability distribution of pairing features through the predicted distribution graph. Specifically, BoltzDesign1 combines the confidence module with the Pairformer to achieve joint optimization based on distance map loss and confidence scores. The confidence module represents the fit between structures sampled from the diffusion module and the pairwise feature probability distributions. In certain scenarios, such as protein-DNA/RNA interactions where pairwise features are ambiguous, additional sampling within the diffusion module may better assess the accessible structural space. To enable loss backpropagation from the confidence module to the inputs, BoltzDesign1 modified the process to allow gradient flow between the Pairformer and confidence module. Addtionally, we calculate CODE when calling the structure diffusion module as an additional loss function to guide binder design compared with BoltzDesign1. As a simple verification, we follow the setting in BoltzDesign1: use the small molecules tested in RfDiffusionAA as a benchmark to design protein binders. The workflow and calculation steps are as follows:

$$\hat{s}_{trunk}, \hat{z}_{trunk} = \text{Pairformer}(s_{trunk}, z_{trunk}, S_{\text{input}}) \tag{1}$$

$$xyz_{atom} = \text{Structure Module}(\hat{s}_{trunk}, \hat{z}_{trunk}).stop\_gradient \tag{2}$$

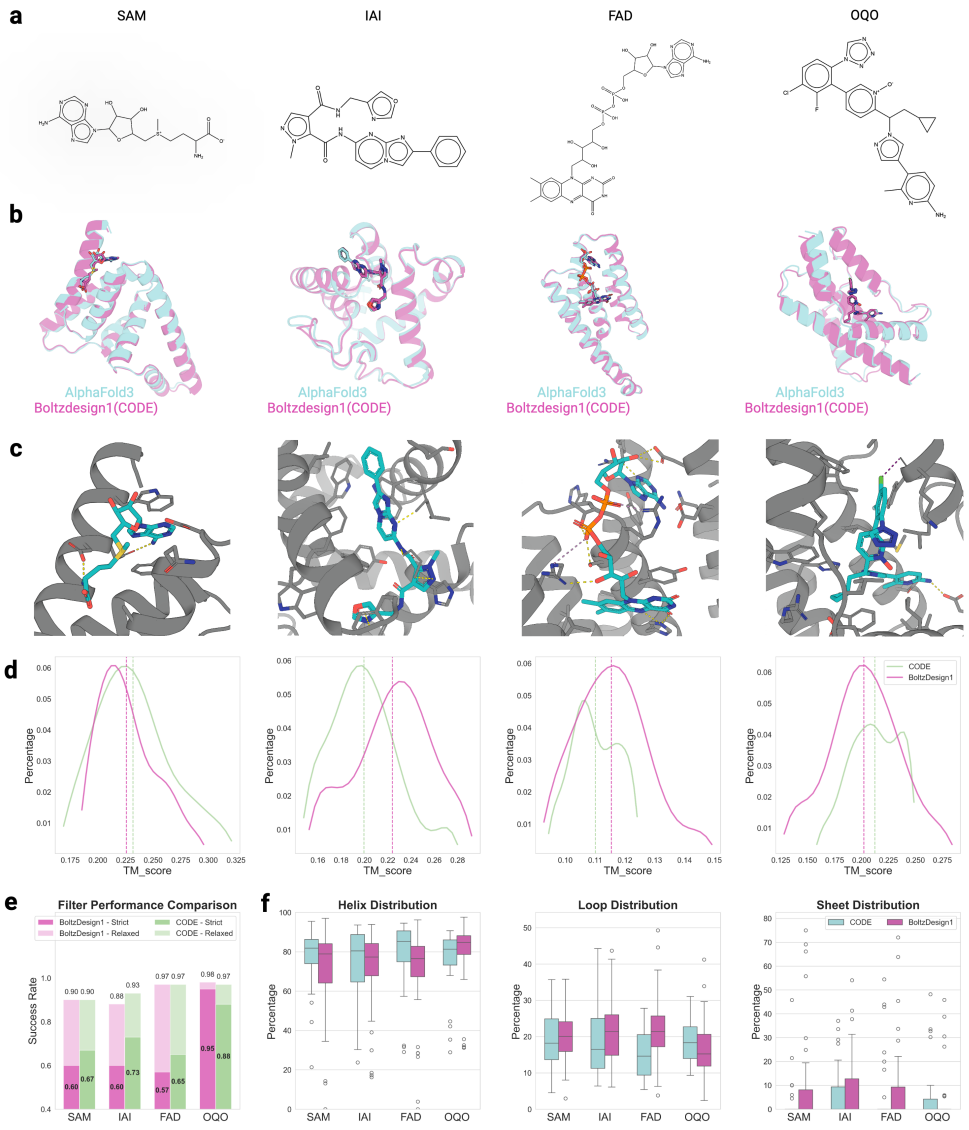


Fig. 5 CODE guides BoltzDesign1 to generate diverse small molecule binders with high designability and AF3 success rates. (a) Chemical structures of four target small molecules used in BoltzDesign1: SAM, IAI, FAD, and OQO. (b) Representative examples of designed binders for these four target molecules, with purple structures showing CODE-optimized designs and green structures showing AlphaFold3 predictions based on the designed sequences. The high structural consistency demonstrates CODE's ability to successfully design binders for small molecule ligands. (c) Zoomed-in view of protein-ligand binding pockets, with yellow dashed lines indicating hydrogen bonds between ligand and protein, and purple dashed lines representing salt bridges. CODE-designed binders form structurally compact and energetically favorable pockets with their targets. (d) Distribution of pairwise TMscores between CODE-optimized binders (green)/original BoltzDesign1 binders (purple) and wild-type proteins. (e) Bar plots compare AF3 success rates between CODE-optimized and original BoltzDesign1 approaches under two filtering criteria. To evaluate intrinsic design capabilities, BoltzDesign1 output structures were used directly without LigandMPNN redesign. (f) Box plots showing secondary structure distribution in CODE-optimized binders. From left to right: percentages of helix, loop/coil, and beta sheet content. Default BoltzDesign1 helix loss weight parameters were applied (minimum -0.3, maximum 0).The target molecules SAM, IAI, FAD, and OQO correspond to the wild-type PDB structures 7C7M, 5SDV, 7BKC, and 7V11, respectively.

$$\text{Confidence score} = \text{Confidence Module}(\hat{s}_{trunk}, \hat{z}_{trunk}, xyz_{atom}, S_{\text{input}}) \quad (3)$$

where s_{trunk} represents the single-sequence representation, z_{trunk} denotes the pairwise representation, and S_{input} is the initial sequence input. xyz_{atom} corresponds to the denoised 3D structure coordinates. Here, $\hat{s}_{trunk}$ and $\hat{z}_{trunk}$ are the outputs from the Pairformer, where gradients are allowed to flow through to the confidence module.

CODE guides BoltzDesign1 to achieve higher in silico success rates

After performing hallucination design through Boltz1, BoltzDesign1 also calls LigandMPNN [51] to redesign the preliminary designed sequence. However, we only want to explore the basic design capabilities without considering the unfair information introduced by LigandMPNN. So we disabled all post-sequence design. BoltzDesign1 presented in silico success rates of binder design for four small molecules: IAI, FAD, SAM and OQO (Figure 5a). For each ligand, we used the default settings of BoltzDesign1 and the configuration with CODE loss to design 30 structures. Then Alphafold3 [1] was used to predict the complex structure. We evaluate these results based on structure confidence scores, specifically the pLDDT from AlphaFold3 for the entire complex and the ipAE for the binding interface confidence between the binder and the small molecule. The calculation of the success rate is divided into two modes. In strict mode, complex $pLDDT > 70$ and $ipAE < 10$ are required, and in relaxed mode, complex $pLDDT > 70$ and $ipAE < 15$ are required [52–54].

As shown in Figure 5b, CODE-based design binders have strong structural consistency between models. For four targets, these designed binders optimized based on CODE loss well show the interactions between proteins and ligands, including hydrogen bonds, hydrophobic interactions and $\pi - \pi$ interactions (Figure 5c). On the first three targets, the design samples with CODE loss restrictions all showed better AF3 success rates than the original BoltzDesign1. Specifically, on SAM, we achieved a strict success rate of 67%, a relative improvement of 11.6% compared to BoltzDesign1's 60%. On IAI, we reached a strict success rate of 73%, a 21.6% relative improvement over BoltzDesign1's 60%. On FAD, we attained a strict success rate of 65%, a 14% relative improvement compared to BoltzDesign1's 57%. In terms of relaxed success rate, our model performed on par with the baseline model, with a 5% relative improvement on IAI. On OQO, due to the clearer binding mode, both methods achieved good comparable results.

These significant improvements demonstrate that the topological frustration information introduced by CODE effectively captures interactions between biological molecules, providing novel insights for binder design that BoltzDesign1 overlooks by ignoring key structural topological constraints.

CODE preserves the diversity of generated samples

To evaluate whether the introduction of topological frustration loss would reduce the diversity of generated structures, we compared with the original Boltzdesign1. For each target molecule, we calculated the TMscore of the pairwise generated binder structure

and the wild type, without any post-processing filtering. We show the distribution of TMscore in Figure 5d. In general, the introduction of topological frustration CODE will emphasize the local structure maintained by the binder design, but will not harm the diversity of generated structures, and even achieve better diversity for some targets. For the targets 5SDV and 7BKC, the average TMscore of the binder designed with CODE loss is 0.199 and 0.110, respectively, while the TMscore without CODE loss is 0.224 and 0.115. The addition of CODE loss enhances diversity in designs for these two targets, outperforming BoltzDesign1, which produces more similar samples. For targets 7C7M and 7V11, although the average TMscore of the binder designed with CODE loss is 0.232 and 0.212, respectively, and the TMscore without CODE loss is 0,226 and 0.202, they are almost consistent, with no obvious changes in diversity.

We also evaluated the distribution of secondary structure in the designed binders in Figure 5f. Among the four binder designs, an increase in the helix proportion was found to be a favorable design factor. In the binder designs for SAM, IAI, and FAD, we observed a higher percentage of helix while reducing the content of the loop region. In contrast, in OQO, the trend was the opposite, and this result showed consistency with the design success rate. No clear pattern was observed in the sheet distribution; however, increasing the percentage of sheet could stabilize the overall structure of the binder. In the future, a corresponding loss term could be introduced to enhance its content.

Topological frustration refers to energetically conflicted regions in protein folding that exhibit high conformational flexibility. In binder design, targeting these dynamic sites on protein surfaces can enhance induced fit or conformational selection mechanisms. Due to their functional significance, topologically frustrated regions are often ideal drug targets. By leveraging local frustration energy, CODE optimizes binder design to improve affinity for target proteins.

2.4.2 Enzyme Catalytic Active Sites Identification

Accurate identification of enzyme active sites remains a fundamental challenge in biochemistry, despite its critical importance for understanding metabolic processes and advancing drug development. While DNA sequencing has generated millions of enzyme sequences, fewer than 0.7% have properly annotated active sites due to the complexity of predicting these functional regions [55]. This bottleneck stems from the need to understand specific enzyme-substrate interactions, reaction mechanisms and the ability to distinguish between different site types. The scarcity of high-quality training data makes direct training of predictive methods particularly challenging [41, 56–58]. However, CODE uses topological frustration theory in protein folding to gain universal structural insights, providing a practical solution for data-scarce scenarios without requiring specific training.

We implemented the residue-level CODE algorithm to prioritize amino acid rankings. The rationale is rooted in evolutionary biophysics: residues critical for function experience strong purifying selection, mutate infrequently, and therefore are more likely to occupy structurally stable regions with minimal topological frustration [59]. CODE is expected to reach its minimum values at these functionally critical locations.

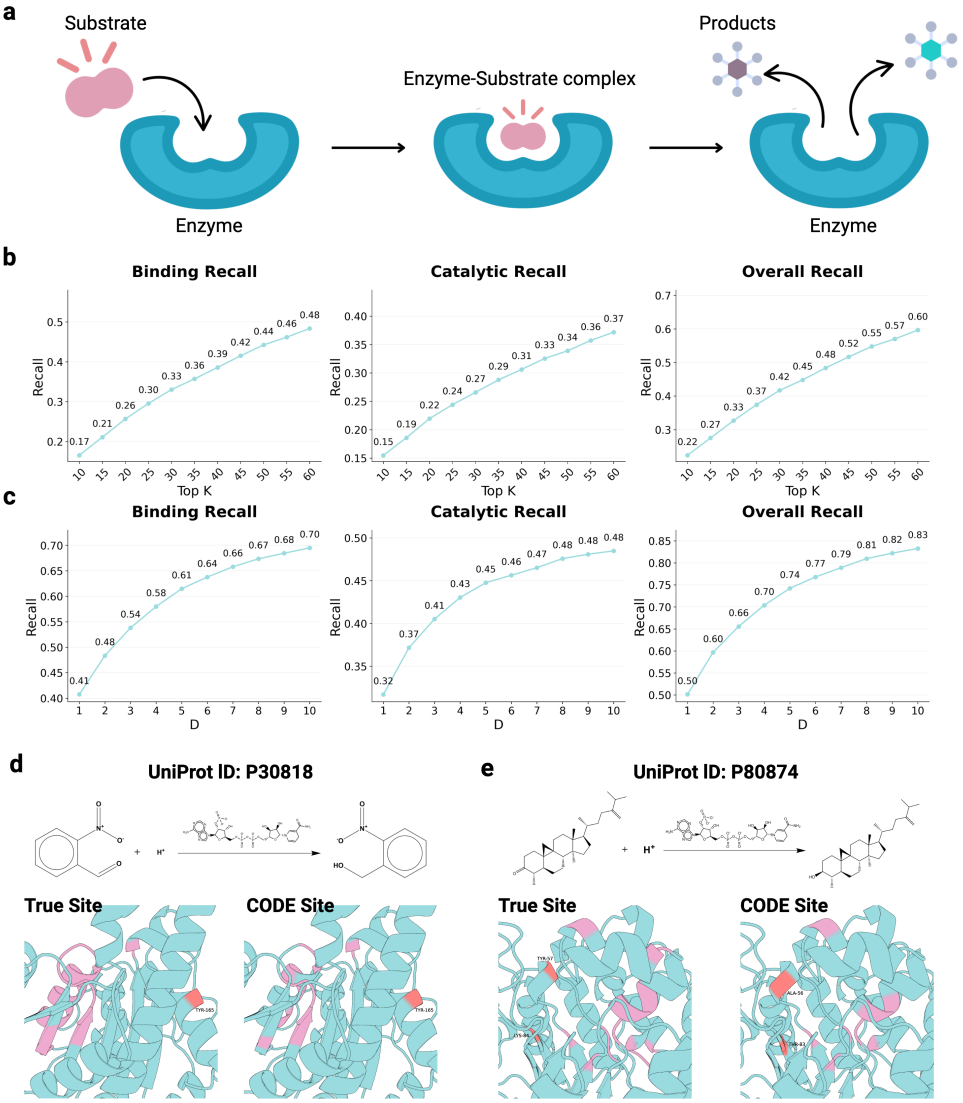


Fig. 6 Residue level CODE ranking can reveal the catalytic active site and binding site of the enzyme. (a) Schematic diagram of the process of enzyme catalyzing substrate reaction. (b) Under the condition $D = 2$, the CODE values are sorted in ascending order, with the top K selected as recognition sites. The recall rates for binding sites, catalytic sites, and overall sites are calculated and presented. (c) After selecting the top 60 values and sorting them in ascending order as recognition sites, recall rates for binding sites, catalytic sites, and overall sites are calculated across varying tolerance values D. (d-e) Chemical reaction equations catalyzed by enzymes in two case studies (UniProt ID: P30818 and P80874) are presented. The actual enzyme catalytic sites (highlighted in orange), binding sites (highlighted in pink), and CODE-based sorting recognition results are compared. CODE demonstrates a significant recall rate, serving as an effective tool for identifying key enzyme sites.

Specifically, we adopted the test set released by EasIFA [41], comprising 853 non-redundant enzymes with residue-level annotations for ligand-binding sites, catalytic sites and others. For every enzyme, we computed CODE score s_i for residue i and ranked residues in ascending order. CODE was applied in a fully zero-shot manner without fitting model parameters to the dataset. We introduced the Relaxed Overlap Rate (ROR), a novel metric to quantify CODE's detection accuracy by comparing ranked predictions to curated functional annotations with adjustable positional tolerance. Let $\mathcal{A}$ denote the intervals of annotated functional residues, $\mathcal{P}$ the intervals covered by the top-k CODE-ranked residues, and D the allowed sequence-index tolerance. Then

$$\mathcal{P}_D = \{\, i \,|\, \exists\, j \in \mathcal{P},\ |i - j| \le D\} \tag{4}$$

$$\mathrm{ROR} = \frac{|\mathcal{A} \cap \mathcal{P}_D|}{|\mathcal{A}|} \tag{5}$$

We first analyzed the impact of different Top K selections on identification Recall, which is shown in Figure 6b. We observed that as K increased, Recall continuously improved, but the rate of increase gradually slowed. In Top 60, we achieved a Binding Site Recall of 48%, a Catalytic Site Recall of 37%, and an overall Recall of 60%. Given that the amino acid length of these enzymes is generally above 500, achieving a 60% Recall by unsupervised selection of 10% of the sites without additional training is a highly promising result. Then we fixed the Top 60 selection and observed the changes in Recall with varying tolerance levels, which is shown in Figure 6c. We noted that as the tolerance (D) increased, Recall continuously improved, but the rate of increase gradually slowed. At D=10, we achieved a Binding Site Recall of 70%, a Catalytic Site Recall of 48%, and an overall Recall of 83%. Since CODE is not an algorithm specifically designed for enzyme site identification and topological frustration is influenced by surrounding amino acids, CODE achieves high identification accuracy within a reasonable error tolerance range. Finally, we selected two representative enzymes (UniProt ID: P30818 and P80874) in Figure 6d and 6e for detailed, visually intuitive case studies.

Catalytic residues are often responsible for proton relay, nucleophilic attack, or transition-state stabilization, but the chemistry can proceed only if these side chains are held in a geometrically precise and electrostatically pre-organized constellation. Evolution therefore embeds them in a highly cooperative network of contacts that dampens local strain and maximizes structural rigidity. In the CODE formalism, such cooperativity is reflected by low CODE scores, i.e., minimal topological frustration. Strikingly, we observe similarly depressed CODE values for ligand-binding pockets located adjacent to the catalytic center. This convergence toward regions of low frustration provides a coherent biophysical rationale for the frequent spatial overlap of binding and catalytic motifs. Because CODE requires only a structure prediction model and no task-specific training, it offers a scalable route to annotate the millions of orphan enzyme sequences currently lacking functional labels. By coupling CODE-based priors with sparse experimental data, one can envisage active-site mapping at proteome scale, accelerating enzyme engineering and structure-guided drug discovery.

2.4.3 Drug-Resistant Mutants Prediction

Drug resistance represents one of the most formidable obstacles in modern cancer therapeutics, particularly affecting the clinical efficacy of targeted kinase inhibitors that have transformed treatment paradigms for numerous malignancies. Bruton's tyrosine kinase (BTK) exemplifies this challenge, serving as a critical therapeutic target in B-cell malignancies including chronic lymphocytic leukemia and mantle cell lymphoma [42]. While BTK inhibitors such as pirtobrutinib have demonstrated remarkable clinical success, the inevitable emergence of resistance mutations necessitates sophisticated approaches to predict and overcome therapeutic resistance before it manifests clinically [60].

The molecular basis of kinase inhibitor resistance typically involves structural perturbations that alter the binding landscape of the ATP-binding pocket or allosteric regulatory sites. These mutations can range from subtle amino acid substitutions that modestly reduce inhibitor affinity to dramatic structural rearrangements that completely abrogate drug binding. Understanding how specific mutations translate to quantitative changes in binding affinity is crucial for developing next-generation inhibitors and optimizing treatment strategies. However, experimental characterization of all possible resistance mutations is impractical, highlighting the urgent need for computational approaches that can accurately predict the impact of mutations on inhibitor binding.

We evaluated the performance of CODE and CONFIDE in predicting drug resistance effects using well-characterized BTK mutations known to reduce the efficacy of pirtobrutinib, a clinically approved non-covalent BTK inhibitor. Our analysis extended beyond pirtobrutinib to include three additional non-covalent BTK inhibitors: ARQ-531, Vecabrutinib and Fenebrutinib, allowing us to examine whether CODE could accurately capture resistance patterns across multiple compounds targeting the same kinase. Binding affinities between various compounds and both wild-type and mutant forms of purified BTK protein were quantified through surface plasmon resonance measurements, yielding dissociation equilibrium constants (K_D). These experimental data were sourced from earlier research [42]. We inputted the wild-type or mutant protein sequences along with ligands into Boltz-1 to generate predicted structures, from which we extracted confidence scores and combined them with topological frustration values (CODE) to calculate CONFIDE scores. Finally, we computed spearman and pearson correlation coefficients between confidence scores, CODE, and CONFIDE against K_D values. In this task, since we are more concerned about the ranking ability of the scores, we mainly consider the performance of the spearman correlation coefficient.

The correlation between different scores and K_d is shown in Figure 7b. For drugs Vecabrutinib and Fenebrutinib, CODE takes the lead, plddt provides additional information, and CONFIDE is the most comprehensive. For drug Vecabrutinib, the absolute Spearman correlation of pLDDT was 0.66, while CODE improved this to 0.71 (a 7.6% relative increase over pLDDT). CONFIDE achieved an absolute correlation of 0.89, which represents a 25.4% relative improvement over CODE and a 34.8% relative increase compared to pLDDT. Similarly, for drug Fenebrutinib, pLDDT achieved the

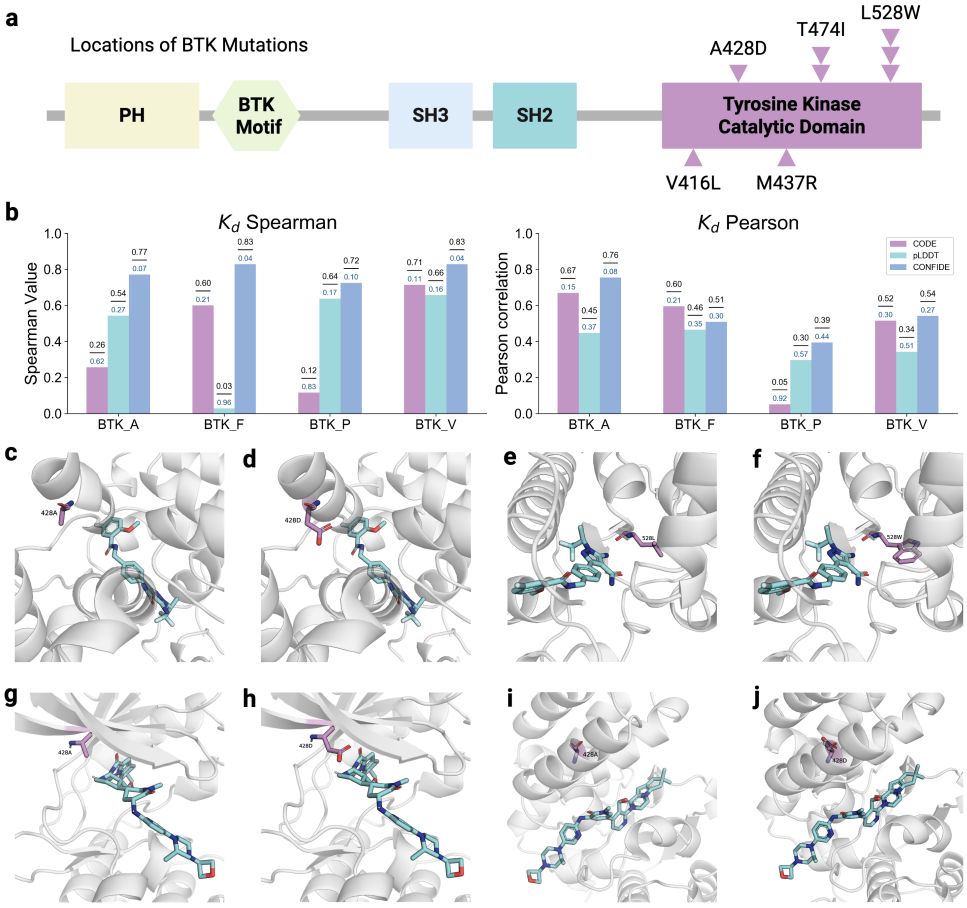


Fig. 7 CODE exhibits high effectiveness in predicting mutation-induced drug resistance effects. We focus on the relevance of five BTK protein mutations to changes in affinity (expressed as K_d) for four non-covalent inhibitors (Pirtobrutinib as BTK_P, ARQ-531 as BTK_A, Vecabrutinib as BTK_V and Fenebrutinib as BTK_F). (a) shows BTK mutations outside the BTK C481 residue found in patients with resistance to pirtobrutinib. Each individual occurrence of a mutation is depicted as an arrowhead. PH denotes pleckstrin homology domain, and SH Src homology domain. (b) Spearman and Pearson correlation coefficients are presented for CODE, pLDDT, and CONFIDE in relation to K_d, with each bar displaying the correlation coefficient value (in black) and the p-value (in blue), separated by a hyphen. The predicted structures that conform to the real mechanism when there is no hallucination are shown in (c-f). (c)(e) show the reasonable interaction between the wild-type protein and Pirtobrutinib and appropriate pocket size in 428A and 528L. (d) shows A428D mutation which introduces a side chain that significantly blocks the kinase pocket entrance, causing substantial steric hindrance with the drug molecule and eliminating binding affinity. (f) shows L528W mutation which increases steric hindrance but causes less severe conflict than A428D, allowing weak binding affinity at 100 μM. Figures (g-j) illustrate that pLDDT fails to accurately reflect structural interpretations of affinity changes in the presence of hallucinated predictions. (g) shows the reasonable interaction between the wild-type protein and Fenebrutinib and appropriate pocket. (h) shows significant increase in steric hindrance for the A428D mutation structure obtained through homology modeling, which is consistent with the experimental observation that no binding has been detected. (i) shows predicted structure by Boltz1 between the wild-type protein and Fenebrutinib. (j) shows the predicted A428D mutation structure by Boltz1, exhibiting no significant steric hindrance, which fails to explain the lack of binding and results in a very low correlation between pLDDT and K_d.

lowest absolute correlation at 0.03, while CODE increased this to 0.60 (a 1900% relative improvement). CONFIDE again performed best, achieving an absolute correlation of 0.83, which is 38.3% higher than CODE and 2758% better than pLDDT.

However, for drugs Pirtobrutinib and ARQ-531, the trend of pLDDT and CODE reversed. Specifically, for drug Pirtobrutinib, CODE achieved an absolute spearman correlation of 0.12, pLDDT improved this to 0.64, while CONFIDE reached 0.73, representing a further relative improvement of 14% over pLDDT and a total relative improvement of 508% compared to CODE. Similarly, for drug ARQ-531, CODE showed an absolute correlation of 0.26, while pLDDT increased to 0.54. CONFIDE again outperformed both, achieving 0.77, which is 42.6% relative higher than pLDDT and an overall relative 196% increase compared to CODE. These results demonstrate that CONFIDE consistently integrates the strengths of both topological and energy frustration, yielding the highest predictive accuracy for all these four drugs.

In addition to its strong sorting ability, CONFIDE also basically maintains the leading Pearson correlation that reflects the linear correlation. In drugs ARQ-531, Fenebrutinib, and Vecabrutinib, CODE takes the lead, and pLDDT serves as auxiliary information. In drug ARQ-531, CODE's spearman correlation reached 0.67, pLDDT was 0.45, and CONFIDE surpassed both to reach 0.76, achieving a relative improvement of 13.4% and 68.9%, respectively. In drug Fenebrutinib, CODE had the best spearman correlation, reaching 0.60, pLDDT was 0.47, and CONFIDE was 0.51. Because of the consideration of stronger sorting ability, CONFIDE made certain sacrifices in linear relationships. In drug Vecabrutinib, CODE's spearman correlation reached 0.52, pLDDT was 0.34, and CONFIDE surpassed both to reach 0.54, achieving a relative improvement of 58.8% compared to pLDDT. However, in drug Pirtobrutinib, the Pearson correlation coefficients of pLDDT and CODE are 0.30 and 0.05, respectively. CONFIDE is still able to integrate the information of the two to obtain the best result of 0.39, achieving a relative improvement of 33.3% and 700%.

These results suggest that while CONFIDE consistently outperforms the other methods, the relative contributions of topological and energy frustration vary depending on the drug. For drugs Pirtobrutinib and ARQ-531, energy frustration plays a more significant role, as reflected in the greater improvement of pLDDT over CODE. In contrast, for drugs Vecabrutinib and Fenebrutinib, topological frustration appears to dominate, as CODE outperforms pLDDT and provides a solid foundation for CONFIDE's superior performance. This highlights the context-dependent nature of these methods and emphasizes the value of combining both metrics in CONFIDE for robust predictive power across diverse drug classes.

In-depth Case Studies

We conduct a more in-depth case analysis based on the structural biology perspective of steric hindrance. We will analyze the two situations of whether hallucination occurs or not: in the absence of hallucinations, we investigated the fundamental mechanism underlying pLDDT's functionality. When hallucination occurs, CODE provides additional information to help understand the predicted structure more comprehensively, so as to avoid being misled by hallucination. According to previous studies[42, 61, 62], drug-resistant mutations mainly play a role through steric hindrance - the L528W

mutation inside the binding pocket and the A428D mutation at the entrance will introduce a large side chain to produce spatial constraints, thereby physically hindering the optimal positioning of the inhibitor.

In Figure 7c-f, we show the predicted structure that conforms to the real mechanism when there is no hallucination. Figure 7c shows the interaction between the wild-type protein and Pirtobrutinib. Due to the reasonable ligand conformation and appropriate pocket size, the wild type has a strong affinity. However, when the A428D mutation occurs, that is, the structure shown in Figure 7d, the side chain will significantly block the entrance of the kinase pocket, resulting in a large steric hindrance between the protein and the drug molecule, so no binding affinity is observed. Similarly, Figure 7e shows the interaction between the wild-type protein and Pirtobrutinib, and Figure 7f shows the L528W mutation. Although the L528W mutation also increases steric hindrance, it does not cause such a severe conflict as A428D. Therefore, L528W can also be observed to bind weakly at 100 μM. The reason why plddt can show a relative strong Pearson correlation of 0.64 under this drug is that the predicted structure shown in Figures a-d can correctly reflect the above mechanism.

However, when hallucination occurs, plddt cannot be served as a good judgment standard. Figure 7g shows the interaction between the wild-type protein and the Fenebrutinib drug. We used homology modeling to obtain the A428D mutation structure in Figure 7h. In reality, this mutation will produce large steric hindrance, which is consistent with the experimental observation that no binding has been detected. However, Figures 7i-j show the protein-drug binding structure before and after the mutation predicted by Boltz1. As can be seen, no significant increase in steric hindrance is observed in the predicted structure, so when plddt evaluates the structure, it is likely that it will not recognize that this is an unreasonable conformation, thus giving a misleadingly high confidence. However, CODE can effectively capture such spatial steric effects, accurately reflecting binding affinity and aligning closely with experimental results.

CODE reflects topological frustration, providing valuable mechanistic insights for comprehensive analysis, complementing the structural predictions observed in our BTK resistance studies. Higher topological frustration values may indicate a more severe loss of binding ability, while a moderate decrease in frustration values may correspond to a moderate increase in affinity. Regions with minimal topological frustration in wild-type BTK represent evolutionarily conserved structural elements that are critical for inhibitor recognition, explaining why mutations such as C481S within these low-frustration regions are able to maintain high binding affinity by retaining essential protein-ligand contacts. The differential correlation patterns observed between different inhibitors - Vecabrutinib and Fenbutinib show clear structure-activity relationships, while Pirotinib and ARQ-531 do not - may reflect different reliance on specific frustration landscapes. Most importantly, the case of A428D on Fenebrutinib, which exhibits a high confidence score despite a significant loss in experimental affinity, highlights how topological frustration analysis can capture subtle local destabilizing effects that are not detectable from static structural predictions alone, leading to a more nuanced understanding of how mutations may disrupt binding through mechanisms of steric interference.

2.4.4 Inhibitor Affinity Prediction

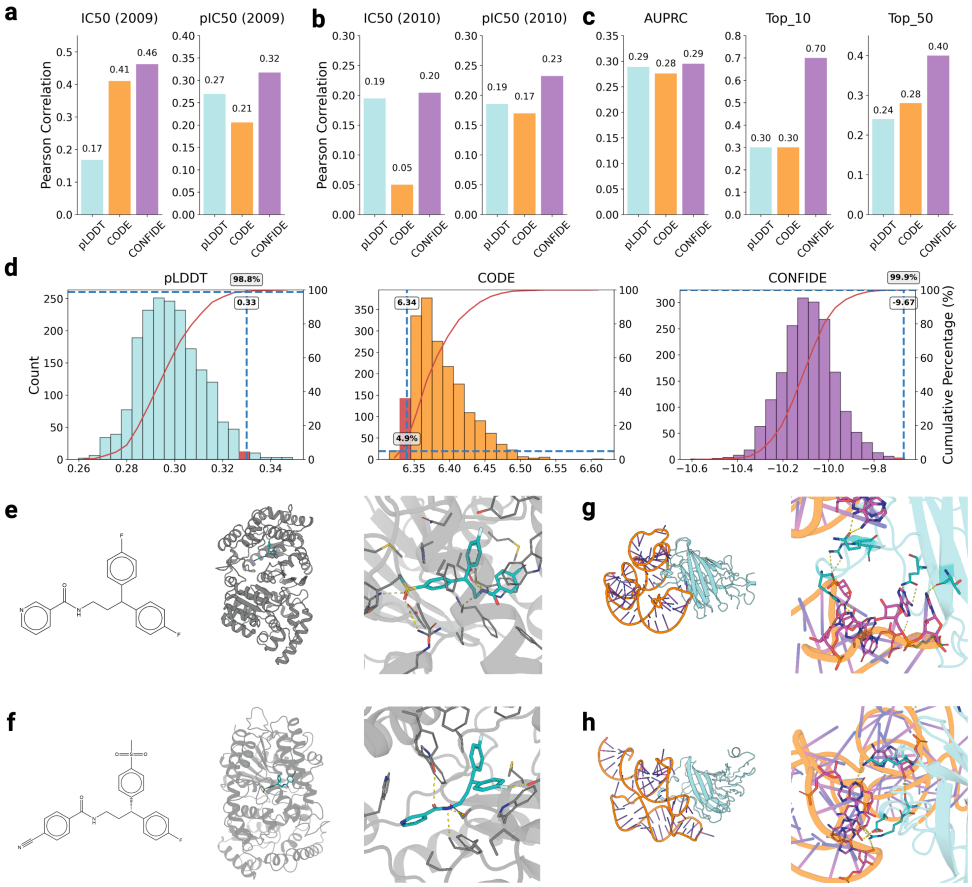


Fig. 8 CODE is an unsupervised virtual screening tool. (a-b) show the Pearson correlation of CODE, pLDDT, and CONFIDE with activity data (IC_{50} and pIC_{50}) for the sEH2009 and sEH2010 datasets, respectively. (c) presents the AUPRC classification metrics for CODE, pLDDT, and CONFIDE in RNA aptamer screening, alongside the hit rates for the top 10 and top 50 candidates. (d) illustrates the distribution of pLDDT, CODE, and CONFIDE scores on the RNA aptamer dataset, with the red highlighted region indicating the positions of these scores corresponding to a top 10 hit aptamer in the distribution. (e-f) show the binding interfaces of two screened sEH inhibitors with their corresponding molecules and proteins. (g-h) show the binding interfaces of two screened RNA aptamers.

Soluble epoxide hydrolase (sEH) represents a compelling therapeutic target for cardiovascular, inflammatory, and neurological disorders due to its central role in regulating bioactive lipid mediators [63, 64]. This enzyme catalyzes the hydrolysis of epoxyeicosatrienoic acids (EETs), which possess anti-inflammatory, vasodilatory, and cardioprotective properties. By inhibiting sEH, these beneficial epoxide metabolites

are preserved, offering a therapeutic strategy for conditions ranging from hypertension to chronic pain. The development of potent and selective sEH inhibitors has therefore attracted significant attention in medicinal chemistry, necessitating sophisticated approaches to identify compounds with optimal binding characteristics [43].

The challenge of inhibitor screening and optimization for sEH exemplifies a broader problem in structure-based drug design: predicting how subtle molecular modifications translate to significant changes in binding affinity. Traditional high-throughput screening approaches, while comprehensive, are resource-intensive and may miss nuanced structure-activity relationships that emerge from the dynamic nature of protein-ligand interactions. This limitation underscores the need for computational methods that can effectively discriminate between compounds with varying affinities within related chemical series.

We inputted all protein-ligand pairs into Boltz-1 for structure prediction and obtained the confidence scores as energy frustration metrics. By combining the confidence score and the topological frustration-aware CODE, we calculate CONFIDE. Finally, we computed Pearson and Spearman correlation coefficients between the CODE, CONFIDE and confidence scores and both IC50 and pIC50 values. The experimental results are shown in Figure 8.

Our correlation analysis of the two datasets shows that CONFIDE's Pearson correlation consistently outperforms the pLDDT and CODE metrics alone. In the 2009 dataset, CODE showed a stronger Pearson correlation on IC50 (0.41 vs. 0.17 for pLDDT, a relative improvement of 142%), while CONFIDE surpassed both to 0.46, a relative improvement of 170% over pLDDT. On pIC50, pLDDT showed a better Pearson correlation (0.27 vs. 0.21 for CODE), while CONFIDE surpassed both to 0.32, an improvement of 18.5% over pLDDT. This pattern changed in the 2010 dataset. For IC50, pLDDT has a higher Pearson correlation coefficient (0.195 vs. CODE's 0.050), and CONFIDE still maintains the lead at 0.204, a relative improvement of 4.6% compared to pLDDT. For pIC50, pLDDT also has a higher Pearson correlation coefficient than CODE (0.185 vs. CODE's 0.170). CONFIDE reached 0.232, a 25.4% improvement over the second highest pLDDT. These results clearly show that CONFIDE is the most superior metric, significantly improving the performance of single energy or topological metrics in all comparisons, and also highlighting the complementary advantages of pLDDT and CODE when prioritizing linear relationships.

CODE offers a novel understanding of binding affinity prediction. In the context of protein-ligand complexes, regions of low topological frustration typically correspond to structurally conserved binding sites that have evolved to accommodate specific molecular interactions with high fidelity. These minimally frustrated regions often correlate with critical binding determinants where even minor structural perturbations can significantly impact affinity. Conversely, areas of higher frustration may represent more adaptable binding environments that can tolerate diverse ligand modifications without dramatic affinity changes. By mapping topological frustration patterns within sEH-inhibitor complexes, we can potentially identify which molecular features are most crucial for maintaining high-affinity interactions, thereby enhancing our ability to predict and rank compound affinities based on their interactions with these frustration-sensitive regions.

2.4.5 Virtual Screening of RNA Aptamers

RNA aptamers represent an important class of therapeutic and diagnostic molecules that combine the specificity of antibodies with the versatility and synthetic accessibility of nucleic acids [65, 66]. These short, structured RNA sequences fold into unique three-dimensional conformations that enable highly specific binding to target proteins, cells, or other biomolecules. Green fluorescent protein (GFP) serves as an excellent model target for aptamer development due to its well-characterized structure, widespread use as a research tool, and potential applications in biosensing and cellular imaging. The identification of high-affinity GFP-binding aptamers from large candidate pools represents both a significant technical challenge [67, 68].

Traditional experimental aptamer selection through SELEX (Systematic Evolution of Ligands by EXponential enrichment) is inherently time-consuming and resource-intensive, often requiring multiple rounds of selection and amplification to identify optimal binders from libraries containing up to 10^{15} unique sequences [69, 70]. Virtual screening approaches offer a valuable alternative, enabling the computational evaluation of vast aptamer libraries to prioritize candidates for experimental validation [44, 71]. However, the success of such approaches critically depends on accurate prediction of binding affinities and the ability to distinguish high-affinity aptamers from weak or non-specific binders within structurally diverse RNA populations.

We sought to identify high-affinity GFP-binding aptamers from an extensive candidate library. Our analysis utilized a curated dataset of GFPapt mutants by prior work [44], originally derived from the earlier study [72]. The dataset encompasses aptamers exhibiting a broad spectrum of binding affinities, with dissociation constants (K_d) ranging from 0 nM to 125 nM. For classification purposes, aptamers demonstrating K_d values below 10 nM were designated as high-affinity binders and classified as positive hits for subsequent analysis. We inputted all protein-RNA pairs into Boltz-1 for structure prediction, from which we extracted confidence scores and combined them with topological frustration values (CODE) to calculate CONFIDE scores. Finally, we used these three scores to compute three classification metrics: AUPRC, Precision@10, and Precision@50.

In Figure 8c, we reported the results for the three scores. We observed that CONFIDE demonstrates a significant advantage in screening for the highest-affinity aptamers. For the top 10 screening, CONFIDE achieved a hit rate of 70%, a 133% improvement compared to pLDDT's 30%. For the top 50 screening, CONFIDE reached a hit rate of 40%, surpassing pLDDT's 24% and CODE's 28%, representing a 66.6% improvement over pLDDT. To further analyze the properties of the screened candidate aptamers, we examined the distribution of pLDDT, CODE, and CONFIDE scores in the candidate library, as shown in Figure 8d. We highlighted representative aptamers in the top 10 with red, noting that their pLDDT values were distributed in the higher range, while CODE scores were lower, ultimately leading to higher CONFIDE score rankings, resulting in their selection.

Regions of minimal topological frustration in GFP likely correspond to evolutionarily conserved structural elements that form stable, high-fidelity binding interfaces. Aptamers that interact primarily with these low-frustration regions may exhibit

enhanced binding stability and specificity, as these areas represent energetically favorable interaction sites that are less tolerant of suboptimal contacts. Conversely, aptamer binding modes that engage highly frustrated regions of GFP may result in weaker or more promiscuous interactions. By incorporating topological frustration mapping into virtual screening workflows, we can potentially enhance our ability to identify aptamers that form thermodynamically stable complexes with GFP, while simultaneously gaining mechanistic insights into the structural determinants of RNA-protein recognition that could inform rational aptamer design strategies.

3 Methods

3.1 Overview

The current AlphaFold3 confidence metrics, especially pLDDT, often give the hallucination of high confidence but low prediction accuracy when faced with many difficult structure prediction scenarios. To solve this problem, we proposed CODE and COFIDE. Extensive research suggests that protein folding occurs in a complex energy landscape and is mainly affected by two different types of frustration: energy frustration and topological frustration [3]. Although evolutionary selection minimizes the energy frustration of natural proteins, topological frustration remains another important intrinsic property that determines folding pathways and dynamics. CODE extracts the embedding of each layer in the diffusion transformer of the diffusion module in alphafold3, and analyzes the changing trajectory of the embedding in high-dimensional space to reflect the model's own estimation of topological frustration information. CONFIDE is built on the basis of both CODE and pLDDT (or confidence score), providing the most comprehensive modeling perspective on the folding landscape. In the following sections, we will introduce each score one by one and their contribution to understanding the quality assessment of predicted structures.

3.2 Chain of Diffusion Embedding

To characterize topological frustration in AlphaFold3, we draw inspiration from the Chain-of-Embedding (COE) paradigm developed for transformer-based language models [6]. Research suggests that transformer-based models progressively encode increasingly complex semantic information across different layers, with hidden state evolution representing a physical correspondence to incremental reasoning [8, 73–75]. Thus, we hypothesize that AlphaFold3's diffusion transformer, through its Chain of Diffusion Embeddings (CODE), progressively models inter-residue relationships that directly correspond to the backbone constraints and energy biases central to topological frustration. Each transformer block refines the representation of molecular connectivity patterns, with the cumulative trajectory encoding the complexity of topological constraints present in the target structure. This connection is theoretically motivated by the physical correspondence between folding pathway complexity and representational refinement difficulty. The rate of change in trajectory magnitude reflects energy landscape ruggedness from a topological perspective: smaller magnitude changes correspond to smoother transitions through conformational space, while

larger changes indicate more constrained, difficult-to-navigate pathways. This provides a complementary assessment to energetic frustration metrics, enabling more comprehensive evaluation of structural plausibility.

Building upon the established connection between latent embedding trajectories and topological frustration, we then formalize the Chain of Diffusion Embedding (CODE) under Alphafold3's diffusion transformer f. Given that the diffusion transformer consists of L hidden layers, we can partition f into a series of ordered submodules:

$$f = f_{head} \circ f_l \circ \cdots \circ f_1 \circ f_0 \tag{6}$$

In Eq.6, $f_{\text{head}} : \mathbb{R}^d \to \mathbb{R}^3$ is the final coordinate prediction layer, $f_0 : \mathbb{R}^d \to \mathbb{R}^d$, is the diffusion conditioning module that receives the embedding from the noise coordinates and pairformer output. f_l $(1 \leq l \leq L) : \mathbb{R}^d \to \mathbb{R}^d$ is the intermediate transformer block. Here d is the embedding dimensions. Given a sequence $\mathbf{x}$ with length N as input to f, when AlphaFold3 generates its ultimate structural prediction at the final denoising timestep, we denote the i-th residue's hidden representation at layer l as $\mathbf{z}_l^i$. Following the definitions in earlier research [76, 77], we define the average embedding at layer l as:

$$\mathbf{h}_l^{diff} = \frac{1}{N} \sum_{i=1}^{N} \mathbf{z}_l^i \tag{7}$$

which represents the l-th global structural representation in the diffusion process. Then, the CODE is expressed as a progressive chain $\mathcal{H}$ of all structural hidden states formalized as follows:

$$\mathcal{H}^{diff} = \underbrace{\mathbf{h}_0^{diff}}_{\text{Condition}} \to \underbrace{\mathbf{h}_1^{diff} \to \cdots \to \mathbf{h}_l^{diff} \to \cdots \to \mathbf{h}_{L-1}^{diff}}_{\text{Intermediate Hidden States}} \to \underbrace{\mathbf{h}_L^{diff}}_{\text{Output State}} \tag{8}$$

To quantify the trajectory of diffusion embedding changes, we need to extract basic geometric information from the chain. The magnitude of change is obviously a direct feature of the path, which is used to reflect the changes in global structural information during the wandering process [78, 79]. We define this basic feature as the L2-norm of magnitude change between any two adjacent embedding pair:

$$M^{\text{diff}}(h_l, h_{l+1}) = \|h_{l+1}^{\text{diff}} - h_l^{\text{diff}}\|_2 \tag{9}$$

CODE needs to consider the magnitude change characteristics of the entire change trajectory. Specifically, it sums the magnitude of each adjacent embedded change and takes the average:

$$\text{CODE}(\mathcal{H}^{diff}) = \frac{1}{L} \sum_{l=0}^{L} \frac{M^{\text{diff}}(h_l, h_{l+1})}{Z^{\text{diff}}} \tag{10}$$

$$Z^{\text{diff}} = \|h_L^{\text{diff}} - h_0^{\text{diff}}\|_2 \tag{11}$$

To mitigate the potential sample bias, range scaling factor Z^{diff} is introduced for the following reasons: If the input and output of one sample are far apart in the latent space, its trajectory naturally has a longer wandering distance. By setting scaling

factors, we convert the absolute magnitude changes of each adjacent state pair into relative changes, specifically, the changes of (h_l, h_{l+1}) is relative to the changes of the input-output states (h_0, h_L), thereby avoiding measurement noise caused by inherent differences between samples.

3.3 Confidence Score Computing

The prior work [2] pointed out that Alphafold has learned a certain folding energy function. They hypothesized that AlphaFold's folding mechanism relies on an initial conformation estimate from the MSA and recycling mechanisms, enabling rapid localization of the energy landscape's local minimum. In the second step, the structural module iteratively scores the initial conformation to yield a reliable predicted structure. Thus, we believe the AlphaFold partially captures energy frustration by predicting multiple confidence metrics learned during training, accounting for overall structural rationality.

In this paper, we mainly focus on two of them: predicted local distance difference (pLDDT) and predicted template modeling score (pTM). The pLDDT provides a per-atom confidence estimate. For atom a, pLDDT is calculated as:

$$pLDDT_a = \frac{1}{|R|} \sum_{m \in R} \frac{1}{4} \sum_{c \in \{0.5,1,2,4\}} [|d_{am}^{\text{pred}} - d_{am}^{\text{true}}| < c \text{ Å}] \tag{12}$$

where d_{am}^{pred} and d_{am}^{true} are the predicted and true distances between atoms l and m, respectively. The set R contains atoms satisfying some criteria which can be found in earlier research [1] in detail.

The pTM provides a global confidence assessment by approximating the TM-score through predicted alignment errors. The method introduces a pairwise error matrix e_{ij}, which captures the positional error of the C_α atom of residue j when predicted and true structures are aligned using the backbone frame of residue i. The error distribution is discretized into 64 bins covering the range from 0 to 31.5 Å with 0.5 Å bin width. The pairwise error e_{ij} is computed as a linear projection of the pair representation z_{ij} followed by softmax normalization. Using the predicted error matrix, pTM is approximated as:

$$\text{pTM} = \max_i \frac{1}{N_{\text{res}}} \sum_j \mathbb{E}[t(e_{ij})] \tag{13}$$

where the expectation is taken over the probability distribution defined by e_{ij}, N_{res} represents the residue number and t represents the TM-score transformation function. Details can be found in the prior study [80].

Aggregating both pLDDT and pTM, Boltz1 proposed Confidence Score to rank predicted structures:

$$\text{Confidence Score} = 0.8 \times \text{pLDDT} + 0.2 \times \text{pTM} \tag{14}$$

The Confidence Score has a range of [0,1], where higher values indicate higher confidence.

3.4 Reweighting CODE and Confidence Score

Different biological systems require distinct weighting of energetic and topological frustrations due to their inherent structural and functional characteristics. The relative importance of energy frustration and topology frustration depends on the folding pathway complexity: fast-folding proteins with simple topologies may be primarily limited by energetic conflicts, while proteins with complex structures or those requiring chaperone assistance face substantial topological barriers. Therefore, we adaptively weight two components based on the specific structural class, folding mechanism, and environmental context of the target system.

To address this problem, we proposed CONFIDE (CONFIdence integrated chain of Diffusion Embedding), which dynamically integrates Confidence Score and CODE to comprehensively consider folding conformations from the perspectives of energy frustration and topological frustration. We formalize the dynamic weight optimization process for CONFIDE as follows. Given a structure prediction task, let $\mathbf{s}_{\text{code}} \in \mathbb{R}$ denote the CODE and $\mathbf{s}_{\text{conf}} \in \mathbb{R}$ represent the Confidence Scores. The CONFIDE score is computed as a weighted linear combination:

$$\text{CONFIDE}(\mathbf{s}_{\text{code}}, \mathbf{s}_{\text{conf}}; w_1, w_2) = w_1 \cdot \mathbf{s}_{\text{code}} + w_2 \cdot \mathbf{s}_{\text{conf}} \tag{15}$$

To establish optimal weighting that maximizes predictive power for structural accuracy, we perform a systematic grid search over the discrete weight space $\mathcal{W} = \{-10, -9, \ldots, 9, 10\}^2$. For each weight configuration $(w_1, w_2) \in \mathcal{W}$, we evaluate the Spearman rank correlation between the composite CONFIDE scores and RMSD values across the dataset $\mathcal{D}$:

$$\rho_{w_1,w_2} = \text{Spearman}\left(\{\text{CONFIDE}(\mathbf{s}^{(i)}_{\text{code}}, \mathbf{s}^{(i)}_{\text{conf}}; w_1, w_2)\}_{i\in\mathcal{D}}, \{\text{RMSD}^{(i)}\}_{i\in\mathcal{D}}\right) \tag{16}$$

The optimal weight configuration is determined through:

$$(w_1^*, w_2^*) = \arg \max_{(w_1,w_2)\in\mathcal{W}} \rho_{w_1,w_2} \tag{17}$$

This adaptive weighting mechanism enables CONFIDE to dynamically balance the complementary information from topology frustration and energy frustration, automatically discovering the optimal linear combination that best captures the relationship.

4 Discussion

We introduce CODE, a latent-space trajectory–based metric, and CONFIDE, an integrated score that jointly quantify topological and energetic frustrations underlying unreliable AlphaFold3 predictions. By reframing diffusion embeddings as structural "reasoning paths," CODE captures folding pathway constraints overlooked by conventional confidence metrics such as pLDDT. Our investigations highlight CODE and CONFIDE's ability to markedly enhance the reliability of data-driven biomolecular structure prediction.

As unsupervised, plug-and-play tools, CODE and CONFIDE demonstrate versatility across biochemical contexts. CONFIDE consistently outperforms pLDDT, showing substantial gains in hallucination detection, flexible protein modeling, and RNA aptamer screening, while CODE-guided optimization improves binder design by increasing success rates and enhancing conserved secondary-structure content without sacrificing diversity. In predictive tasks such as affinity estimation and drug-resistance profiling, CONFIDE achieves strong correlations with experimental measures (e.g., up to 0.89 in BTK inhibitor resistance prediction), underscoring its value for robust and interpretable structural modeling across applications central to drug discovery.

By revealing biophysical constraints embedded within deep structure predictors, CODE and CONFIDE provide a scalable, mechanistically interpretable toolkit with broad applications. Paired with AlphaFold3, they can generate high-quality structural datasets across biomolecular modalities, fueling advances in diffusion-based predictors, inverse-folding models, multi-conformational design, and cryptic-pocket identification. Ultimately, we envision CODE and CONFIDE as a new paradigm for unsupervised structural evaluation—accelerating progress in computational structural biology, deepening our understanding of protein folding, and driving innovations in AI-guided therapeutic discovery.

Declarations

Author Contributions Statement

Z.G. and C.G. conceived the idea, developed the theoretical formalism, constructed the model and conducted experiments, and wrote the manuscript; M.H., S.S. and X.W. prepared the data; H.C. helped write and revise the manuscript; J.Z. investigated the Drug resistance prediction; Z.L. verified the analytical methods; X.Y. revised manuscript and provided computing resources; C.H., C.G. and P.H. supervised the project and revised the manuscript. All authors discussed the results and contributed to the final manuscript.

Competing Interests Statement

The authors declare no competing interests.

Data and Code Availability

This study utilizes 13 datasets. The topological frustration dataset is accessible from https://pubs.acs.org/doi/10.1021/ja049510%2B. The ternary complex structure prediction dataset is available from https://github.com/NilsDunlop/PROTACFold. The flexible and rigid protein datasets can be found at https://www.nature.com/articles/s41592-023-01831-0#data-availability and https://www.nature.com/articles/s41467-023-36699-3#data-availability. The PDB test and CASP 15 datasets is accessible from https://drive.google.com/file/d/1JvHlYUMINOaqPTunI9wBYrfYniKgVmxf/view. For data resources of binder design, please refer to https://github.com/yehlincho/BoltzDesign1. The enzyme site dataset is available from https://www.nature.com/articles/

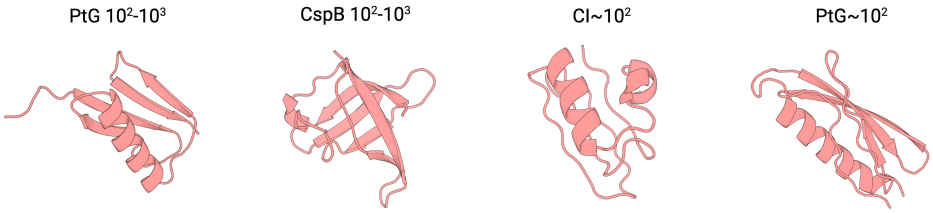
s41467-024-51511-6#data-availability. For the drug-resistant mutation dataset, please refer to https://www.nejm.org/doi/full/10.1056/NEJMoa2114110 for detailed PDB ID. For affinity data of sEH 2009 and 2010 versions, please refer to https://doi.org/10.1101/2025.04.07.647682. The RNA aptamer dataset is accessible from https://github.com/Graph-and-Geometric-Learning/Frame-Averaging-Transformer/tree/main/dataset/aptamer/raw. All code for structure prediction and the anlysis of CODE and CONFIDE can be accessed at https://github.com/zjgao02/CONFIDE.git.

S1 Dataset Description

S1.1 Datasets on topological frustration

We utilized a database of energetically unfrustrated single domain proteins for our analysis [14]. This database represents the extrapolation of the minimal (energetic) frustration principle to the limit of completely unfrustrated protein-like chains. By studying the folding landscape in this energetically unfrustrated protein world, we were able to concentrate on features determined solely by backbone topology (i.e., configuration entropy). The entropic roughness across different proteins in this dataset provided a clean explanation for the different folding scenarios experimentally detected: bottlenecks in configuration entropy were identified along the folding routes of slow folding proteins, whereas the smooth entropy landscapes associated with the fastest folders left more room for energetic perturbations (sequence dependence) to shape the minimal free energy pathways to the native state.

The proteins in the database were chosen to cover a range of experimentally determined folding rates spanning over six orders of magnitude, while exhibiting diverse overall folding topologies and secondary structure compositions. We analyzed sixteen two-state, single-domain proteins, with lengths ranging from 36 to 115 residues. Table 1 provides a summary of the structural details and folding rates of the selected proteins. Figure 1 shows the predicted structures of the remaining four proteins in the main text.



Supplementary Fig. 1 The predicted structures of the remaining four proteins in the main text.

S1.2 Datasets on ternary complex structure

We utilized the dataset curated 62 PROTAC-mediated ternary and binary complexes from the PDB [81]. Each structure contains experimentally resolved coordinates for the

Supplementary Table 1 Summary of Most Relevant Structural and Folding Features for All Two-State Folding Proteins Used in the proof-of-concept experiments on the correlation between topological frustration and CODE.

Protein	L	Structural Information	k_f (s^{-1})
HP36	36	All helical; smallest naturally occurring, independently folding protein domain	$\sim 10^5$
λ_{6-85}	80	Five-helix bundle	$10^4 - 10^5$
Psbd	43	Very small three-helix bundle	$\sim 10^4$
N-L9	56	Three-stranded antiparallel β-sheet sandwiched between two helices	$\sim 10^3$
CspB	67	Small β-barrel	$10^2 - 10^3$
PtG	56	Four-stranded β-sheet spanned by an α-helix (similar to PtL)	$10^2 - 10^3$
CI2	64	Six-stranded β-sheet packed against an α-helix	$\sim 10^2$
PtL	61	Four-stranded β-sheet spanned by an α-helix (similar to PtG)	$\sim 10^2$
Im9	86	Four-helix bundle	$\sim 10^2$
SH3	57	Two antiparallel β-sheets orthogonally packed	$10 - 10^2$
TI-I27	89	Two antiparallel β-sheets packed against each other (Greek key topology)	$1 - 10$
HPr	85	Three α-helices packed against a four-stranded antiparallel β-sheet	$1 - 10$
MerP	72	Antiparallel four-stranded β-sheet, with two helices packed on one side	~ 1
TWIg	93	Two antiparallel β-sheets packed against each other (Greek key topology)	~ 1
AcP	98	Two antiparallel α-helices packed against a five-stranded β-sheet	$\sim 10^{-2}$
P13	115	Filled β-barrel	$10^{-2} - 10^{-1}$

complete PROTAC molecule–including warhead, linker, and E3 ligase binding moiety–as well as at least one of either the protein of interest (POI) or the E3 ubiquitin ligase. This collection represents all publicly available, fully crystallized PROTAC ternary and binary complexes as of May 2024. Structures with partially resolved PROTACs, such as those capturing solely the warhead component, were identified but excluded from the dataset.

The completeness of PROTAC crystallization was validated through systematic review of each structure's associated publication, enhanced by computational analysis tools. To analyze the complete ternary complex interaction, we finally retained 48 samples with complete POI, E3 and ligand. Similarly, we also cleaned out 33 molecular glue samples. These samples were finally processed into YAML files and input into Boltz1 for structure prediction without MSA. In order to better characterize ligands, we used CCD format input for ligands.

S1.3 Datasets on flexible and rigid protein structure

Flexible proteins are disordered proteins from liquid-liquid phase separation, obtained from the CD-CODE platform. This platform provides a wide range of protein sequences related to biomolecular condensates, including 223 Diver LLPS protein sequences. We randomly selected 29 of these sequences as a validation dataset. Rigid

proteins were obtained from the rigid and MOAD PDB data sets of negative samples constructed using pocketminer [37]. For details, see Supplementary Data 1 in [37].

S1.4 Datasets on PDB test and CASP15

We evaluate the performance of the model on two benchmarks: the diverse test set of recent PDB structures and CASP15, the last community-wide protein structure prediction competition where for the first time RNA and ligand structures were also evaluated [82, 83]. These benchmarks feature a diverse range of structures, including protein complexes, nucleic acids, and small molecules, making them ideal for evaluating models like Boltz-1 that predict the structure of various biomolecules.

For PDB test, We followed the dataset processing strategy of the Boltz1 benchmark, and the final test set size was 593. The detailed processing process can be found in section 2.2 in [50].

For CASP15, the benchmark processing methodology from Boltz-1 was adopted to extract all competing targets using the following filtering criteria: (1) targets that were not withdrawn from the competition, (2) targets with associated PDB identifiers to obtain ground truth crystal structures, (3) targets where the stoichiometry information matched the number of provided chains, and (4) targets with total residue counts below 2000. This filtering process resulted in 76 structures. For the test dataset, structures containing covalently bound ligands were excluded since the current version of the Chai-1 public repository does not support configuration of these interactions. Additionally, for both datasets, structures that exceeded memory limitations or failed due to other technical issues on A100 80GB GPUs were removed from consideration. Following these preprocessing and filtering steps, the final evaluation datasets comprised 66 structures for CASP15 and 541 structures for the test set, ensuring computational feasibility and methodological consistency across all evaluated prediction methods.

S1.5 Datasets on enzyme site identification

We used the test dataset proposed in [41] for enzyme site identification. The test dataset contains 894 data, including the enzyme catalytic reaction equation, enzyme commission number, PDB ID, amino acid sequence, site interval and label. For the label of the site interval, 0 represents the binding site, 1 represents the catalytic site, and 2 represents others.

S1.6 Datasets on BTK mutants prediction

[42] provided the source data, which contained experimentally measured affinity data for four BTKs with their five mutants, and the wild type. The dissociation equilibrium constants (K_d) of compounds to purified wild-type or mutant BTK protein are determined with the use of surface plasmon resonance technology. We treat 'no binding detected' data as 100 μM for the correlation analysis following [84].

S1.7 Datasets on sEH inhibitor affinity prediction

Following [84], we selected two well-characterized series of soluble epoxide hydrolase (sEH) inhibitors from the CHEMBL database, comprising compounds with a wide range of experimental IC50 and pIC50 values (pIC50 equals the negative $\log_{10}$IC50) but sharing common scaffolds. We utilized a dataset containing protein-ligand pairs with two versions from 2009 and 2010, comprising 25 and 50 pairs respectively.

S1.8 Datasets on RNA aptamers screening

Our analysis utilized a curated dataset of GFPapt mutants by [44], originally derived from the work of [72]. The dataset encompasses aptamers exhibiting a broad spectrum of binding affinities, with dissociation constants (K_d) ranging from 0 nM to 125 nM. For classification purposes, aptamers demonstrating K_d values below 10 nM were designated as high-affinity binders and classified as positive hits for subsequent analysis.

S2 Additional Results Analysis

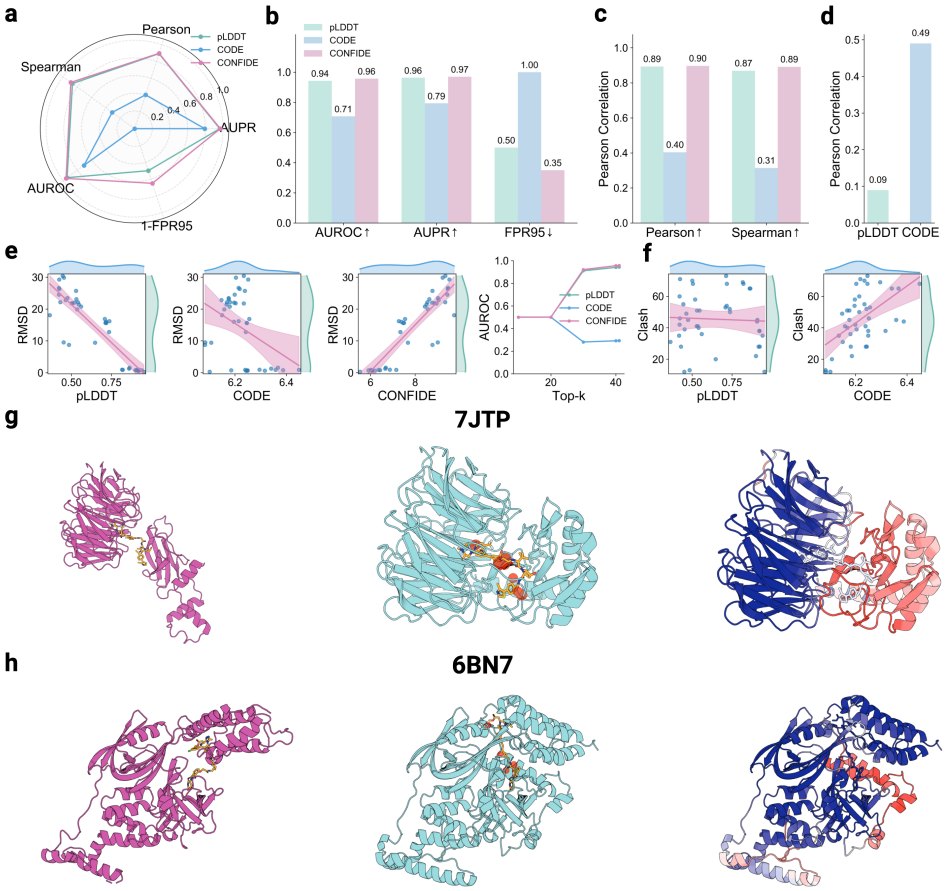
S2.1 Performance in PROTAC Structure Prediction

In Figure 2a, we first presented an overall comparison of the performance of the three scores in classification and correlation tasks, observing that CONFIDE demonstrates an advantage over pLDDT or CODE alone. In the classification task, CONFIDE improved the AUROC by 0.02 and the FPR95 by 0.15 compared to pLDDT which is shown in Figure 2b. In the Pearson correlation analysis, CONFIDE reached 0.9, which is 0.50 higher than using CODE alone and 0.01 higher than pLDDT. In the Spearman correlation coefficient, CONFIDE reached 0.89, which is 0.58 higher than using CODE alone and 0.02 higher than pLDDT, which is shown in Figure 2c. In Figure 2e, we displayed the distribution of the three scores across all samples along with their fitted curves, as well as the trend of AUROC changes with different top K selections.

As in our in-depth analysis of molecular glue in the main text, we also demonstrated the powerful ability of CODE to recognize topological frustration caused by atomic conflicts in the PROTAC complex. Here we presented two representative PROTAC as a case study (PDB ID: 7JTP and 6BN7) in Figure 2g and Figure 2h. The leftmost image shows the true crystal structure. The middle image depicts the Boltz1-predicted structure, with atomic clash regions marked by red spheres. The rightmost image colors the predicted structure based on pLDDT scores, with blue indicating high confidence and red indicating low confidence. It was observed that pLDDT can, to some extent, reflect the quality of structural folding, particularly in assigning notably low scores to disordered or loop regions, indicating that pLDDT can capture this type of energy frustration. However, pLDDT is ineffective in detecting atomic clashes. In contrast, atomic clashes significantly increase topological frustration, enabling CODE to capture such structural prediction inaccuracies.

In addition, we statistically quantified CODE's advantages. Figure 2d shows the Pearson correlation coefficients of pLDDT and CODE with the number of atomic clashes, where CODE achieved a correlation of 0.49, a 444% improvement

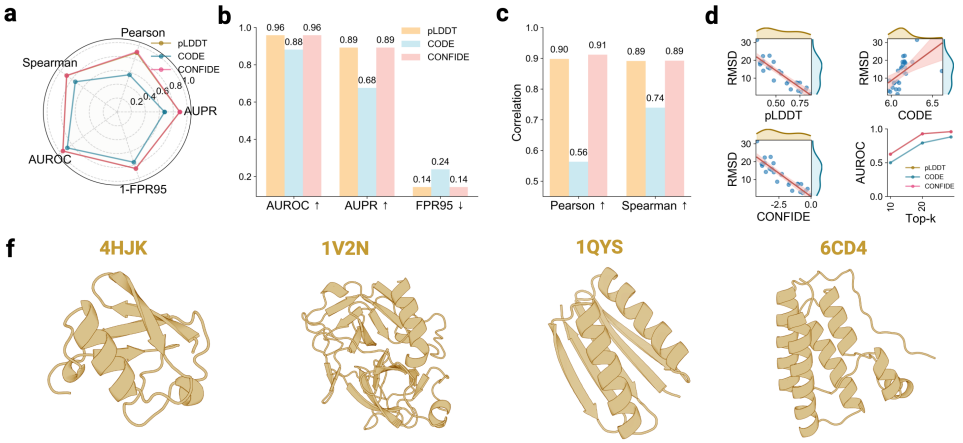
over pLDDT's 0.09. Figure 2f illustrates the specific distribution and fitted curves. These significant improvements demonstrate that CODE captures critical complementary information related to topological frustration, providing substantial benefits for evaluating the structural quality of ternary or even higher-order complexes.



Supplementary Fig. 2 Performance analysis on PROTAC structure prediction. (a) Radar chart of the five evaluation metrics of pLDDT, CODE and CONFIDE. (b) Histogram of classification evaluation metrics of pLDDT, CODE and CONFIDE. (c) Histogram of correlation evaluation metrics of pLDDT, CODE and CONFIDE. (d) Spearman correlation between the number of atomic conflicts and pLDDT/CODE. (e) Scatter plots showing the fit and distribution of pLDDT, CODE, and CONFIDE against RMSD, including linear regression fits and 95% confidence intervals, as well as AUROC for different classification thresholds. (f) Scatter plots of the fit and distribution of pLDDT and CODE with RMSD, incorporating linear regression fits and 95% confidence intervals. (g-h) The true structure, predicted structure, and predicted structure colored by pLDDT of 7JTP and 6BN7 in PROTAC are shown from left to right. The red balls in the middle figure indicate atomic conflicts.

S2.2 Performance in Rigid Structure Prediction

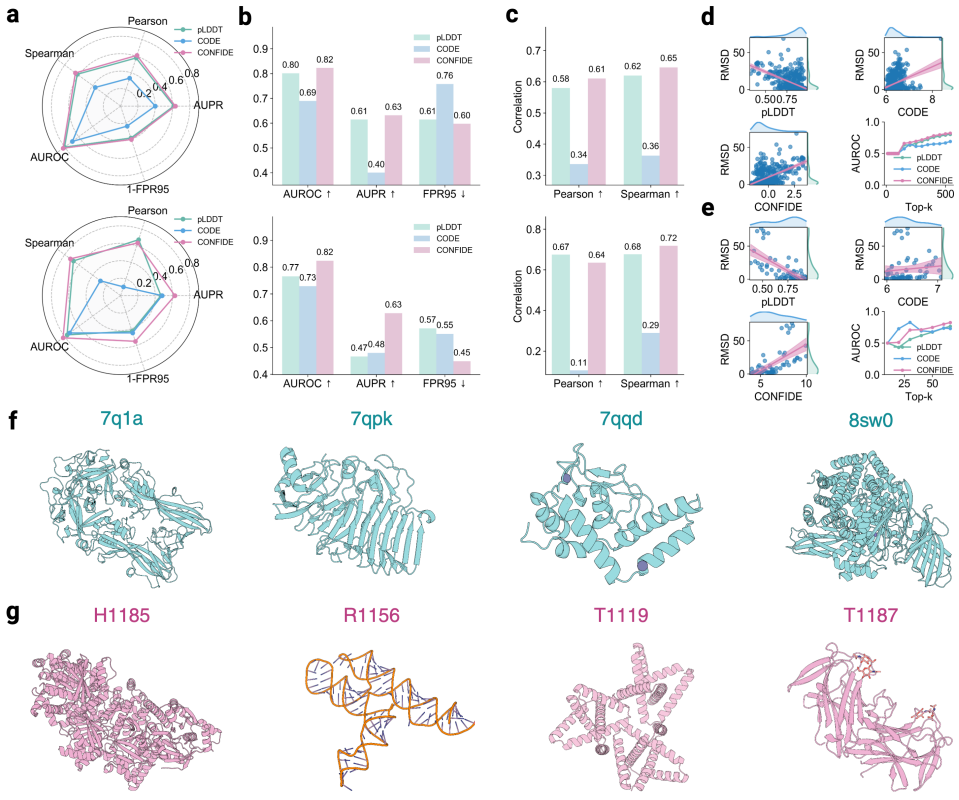
In Figure 3a, we first presented an overall comparison of the performance of the three scores in classification and correlation tasks. Since the structure prediction of rigid proteins is relatively simple and Boltz1 predicts well, pLDDT alone can achieve a strong correlation of 0.9. Therefore, all indicators of CONFIDE are almost on par with pLDDT and are close to saturation. Only the Pearson correlation is higher than pLDDT alone by 0.01. In Figure 3d, we displayed the distribution of the three scores across all samples along with their fitted curves, as well as the trend of AUROC changes with different top K selections. Figure 3e shows four rigid proteins.



Supplementary Fig. 3 Performance analysis on rigid protein structure prediction. (a) Radar chart of the five evaluation metrics of three scores (pLDDT, CODE and CONFIDE) on rigid protein. (b) Histogram of classification evaluation metrics of three scores on rigid protein. (c) Histogram of correlation evaluation metrics of three scores on rigid protein. (d) Scatter plots of the fit and distribution of three scores on flexible proteins with RMSD, including linear regression fits and 95% confidence intervals, as well as AUROC for different classification thresholds. (e) The structures of four rigid proteins.

S2.3 Comprehensive Benchmark Evaluation on CASP15 and PDBTEST Datasets

Not only do CODE and CONFIDE perform amazingly well on challenging datasets, to assess their general capabilities, we also evaluate performance on two benchmark datasets: a diverse test set containing recent PDB structures (details are provided in the supplementary material) [85]; and CASP15 (the community-wide protein structure prediction competition), which for the first time evaluated structures of RNA and ligands [82, 86]. Both benchmark datasets contain a very diverse set of structures, including protein complexes, nucleic acids, and small molecules, making them excellent testbeds for evaluating models that can predict the structure of arbitrary biomolecules, such as Boltz-1.



Supplementary Fig. 4 Performance analysis on two recognized structure prediction benchmark PDB-test and CASP15. (a) Radar chart of the five evaluation metrics of three scores (pLDDT, CODE and CONFIDE) on two datasets. (b) Histogram of classification evaluation metrics of three scores on two datasets. (c) Histogram of correlation evaluation metrics of three scores on two datasets. (d-e) Scatter plots of the fit and distribution of three scores on the two datasets with RMSD, including linear regression fits and 95% confidence intervals, as well as AUROC for different classification thresholds. (f) Four representative structures predicted by Boltz1 on PDB-test. (g) Four representative structures predicted by Boltz1 on CASP15.

In Figure 4a, we presented an overall comparison of the performance of the three scores on the PDB test and CASP 15 datasets in classification and correlation tasks. In Figure 4b and Figure 4c, for the PDB test dataset, in the classification task, CONFIDE improved the AUROC by 0.02 and the AUPR by 0.02 compared to pLDDT. In terms of Pearson correlation, CONFIDE achieved 0.61, surpassing pLDDT by 0.03. For Spearman correlation, CONFIDE reached 0.65, exceeding pLDDT by 0.03. For the CASP 15 dataset, in the classification task, CONFIDE achieved an AUROC of 0.82, a 0.05 improvement over pLDDT's 0.77. In AUPR, CONFIDE reached 0.63, a 34% improvement compared to pLDDT's 0.47. In FPR95, CONFIDE achieved 0.45, a 21% improvement over pLDDT's 0.57. In Pearson correlation, CONFIDE and pLDDT were nearly equivalent, possibly due to CODE's limited understanding of nucleic acid and

small molecule folding. In Spearman correlation, CONFIDE reached 0.72, surpassing pLDDT by 0.04. In Figure 4d, we displayed the distribution of the three scores across all samples along with their fitted curves, as well as the trend of AUROC changes with different top K selections. In Figures In Figure 4f and Figure 4g, we presented representative structures from the PDB test and CASP 15 datasets, respectively.

S3 Evaluation Metric Description

To comprehensively evaluate the performance of pLDDT, CODE, and CONFIDE in different application scenarios, we use six metrics. Each metric provides a unique perspective on the classification or correlation analysis results, ensuring a comprehensive and balanced analysis of different use cases.

- **Area Under the Receiver Operating Characteristic Curve (AUROC)**

$$\text{AUROC} = \int_0^1 \text{TPR}(\text{FPR}^{-1}(t))\,dt = \int_0^1 \frac{TP}{TP+FN}\,d\left(\frac{FP}{FP+TN}\right),$$

  where TPR is the true positive rate and FPR is the false positive rate. AUROC measures the area under the ROC curve, providing a single scalar value representing the model's discriminative ability across all classification thresholds.
- **Area Under the Precision-Recall Curve (AUPR)**

$$\text{AUPR} = \int_0^1 \text{Precision}(\text{Recall}^{-1}(r))\,dr = \int_0^1 \frac{TP}{TP+FP}\,d\left(\frac{TP}{TP+FN}\right).$$

  AUPR quantifies the area under the precision-recall curve, which is particularly useful for imbalanced datasets as it focuses on the positive class performance and is less influenced by the large number of true negatives.
- **False Positive Rate at 95% True Positive Rate (FPR95)**

$$\text{FPR95} = \frac{FP}{FP+TN}\Big|_{\text{TPR}=0.95},$$

  where the threshold is set such that the true positive rate equals 95%. FPR95 measures the false positive rate when the model achieves 95% sensitivity, providing insight into the trade-off between sensitivity and specificity at high recall levels.
- **Pearson Correlation Coefficient**

$$\text{Pearson} = \frac{\sum_{i=1}^{n}(x_i-\bar{x})(y_i-\bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i-\bar{x})^2}\sqrt{\sum_{i=1}^{n}(y_i-\bar{y})^2}},$$

  where x_i and y_i are the predicted and actual values, and $\bar{x}$ and $\bar{y}$ are their respective means. The Pearson correlation measures the linear relationship between predicted and actual values, ranging from -1 to 1.

- **Spearman Rank Correlation Coefficient**

$$\text{Spearman} \;=\; 1 - \frac{6\sum_{i=1}^{n} d_i^2}{n(n^2-1)},$$

where d_i is the difference between the ranks of corresponding predicted and actual values, and n is the number of observations. Spearman correlation assesses the monotonic relationship between variables, making it robust to outliers and non-linear relationships.

- **Area Under the Precision-Recall Curve (AUPRC)**

$$\text{AUPRC} \;=\; \int_0^1 \text{Precision}(\text{Recall}^{-1}(r))\, dr \;=\; \int_0^1 \frac{TP}{TP+FP}\, d\left(\frac{TP}{TP+FN}\right).$$

AUPRC is identical to AUPR and measures the area under the precision-recall curve, providing a comprehensive evaluation of model performance especially for imbalanced datasets where positive instances are rare.

- **Recall**

$$\text{Recall} \;=\; \frac{TP}{TP+FN}.$$

Also known as sensitivity or true positive rate, recall measures the proportion of actual positive instances that are correctly identified by the model. High recall indicates the model's ability to detect most positive cases, which is crucial in applications where missing positive instances has serious consequences.

S4 Structural Prediction and Calculation Details

Due to copyright restrictions on alphafold3, our structural inference model uses the open-source Boltz1 model. To adhere to the theoretical assumptions in the paper and to facilitate large-scale structural inference, we employed an MSA-free inference mode. After obtaining the output of Boltz1, we used the confidence score in the output JSON file to reflect energy frustration. The confidence score is calculated as $0.8 * complex_{plddt} + 0.2 * iptm$ (ptm for single chains). CODE is calculated in the hidden layers of Boltz1, and the embeddings of each layer are saved during forward inference, with a dimension of 768. After obtaining the CODE and confidence score, we selected the optimal combination coefficient based on the Spearman coefficient, or we could simply add them together. All experiments were conducted on four NVIDIA A800 80G GPUs.

References

[1] Abramson, J., Adler, J., Dunger, J., Evans, R., Green, T., Pritzel, A., Ronneberger, O., Willmore, L., Ballard, A.J., Bambrick, J., *et al.*: Accurate structure prediction of biomolecular interactions with alphafold 3. Nature **630**(8016), 493–500 (2024)

[2] Roney, J.P., Ovchinnikov, S.: State-of-the-art estimation of protein model accuracy using alphafold. Physical Review Letters **129**(23), 238101 (2022)

[3] Brockwell, D.J., Smith, D.A., Radford, S.E.: Protein folding mechanisms: new methods and emerging ideas. Current opinion in structural biology **10**(1), 16–25 (2000)

[4] Grantcharova, V., Alm, E.J., Baker, D., Horwich, A.L.: Mechanisms of protein folding. Current opinion in structural biology **11**(1), 70–82 (2001)

[5] Gunasekaran, K., Eyles, S.J., Hagler, A.T., Gierasch, L.M.: Keeping it in the family: folding studies of related proteins. Current opinion in structural biology **11**(1), 83–93 (2001)

[6] Wang, Y., Zhang, P., Yang, B., Wong, D.F., Wang, R.: Latent space chain-of-embedding enables output-free llm self-evaluation. arXiv preprint arXiv:2410.13640 (2024)

[7] Wei, J., Wang, X., Schuurmans, D., Bosma, M., Xia, F., Chi, E., Le, Q.V., Zhou, D., *et al.*: Chain-of-thought prompting elicits reasoning in large language models. Advances in neural information processing systems **35**, 24824–24837 (2022)

[8] Peters, M.E., Neumann, M., Zettlemoyer, L., Yih, W.-t.: Dissecting contextual word embeddings: Architecture and representation. arXiv preprint arXiv:1808.08949 (2018)

[9] Zeni, C., Pinsler, R., Zügner, D., Fowler, A., Horton, M., Fu, X., Shysheya, S., Crabbé, J., Sun, L., Smith, J., et al.: Mattergen: a generative model for inorganic materials design. arXiv preprint arXiv:2312.03687 (2023)

[10] Itzhaki, L.S., Otzen, D.E., Fersht, A.R.: The structure of the transition state for folding of chymotrypsin inhibitor 2 analysed by protein engineering methods: evidence for a nucleation-condensation mechanism for protein folding. Journal of molecular biology **254**(2), 260–288 (1995)

[11] Bryngelson, J.D., Onuchic, J.N., Socci, N.D., Wolynes, P.G.: Funnels, pathways, and the energy landscape of protein folding: a synthesis. Proteins: Structure, Function, and Bioinformatics **21**(3), 167–195 (1995)

[12] Kuhlman, B., Baker, D.: Native protein sequences are close to optimal for their structures. Proceedings of the National Academy of Sciences **97**(19), 10383–10388 (2000)

[13] Bryngelson, J.D., Wolynes, P.G.: Spin glasses and the statistical mechanics of protein folding. Proceedings of the National Academy of sciences **84**(21), 7524–7528 (1987)

[14] Chavez, L.L., Onuchic, J.N., Clementi, C.: Quantifying the roughness on the free energy landscape: entropic bottlenecks and protein folding rates. Journal of the American Chemical Society **126**(27), 8426–8432 (2004)

[15] Plotkin, S.S., Onuchic, J.N.: Structural and energetic heterogeneity in protein folding. i. theory. The Journal of chemical physics **116**(12), 5263–5283 (2002)

[16] Plotkin, S.S., Onuchic, J.N.: Investigation of routes and funnels in protein folding by free energy functional methods. Proceedings of the National Academy of Sciences **97**(12), 6509–6514 (2000)

[17] Plaxco, K.W., Simons, K.T., Ruczinski, I., Baker, D.: Topology, stability, sequence, and length: defining the determinants of two-state protein folding kinetics. Biochemistry **39**(37), 11177–11183 (2000)

[18] Orr, B., Crilly, S.E., Akpinaroglu, D., Zhu, E., Keiser, M.J., Kortemme, T.: An improved model for prediction of de novo designed proteins with diverse geometries. bioRxiv, 2025–06 (2025)

[19] Rathkopf, C.: Hallucination, reliability, and the role of generative ai in science. arXiv preprint arXiv:2504.08526 (2025)

[20] Hong, X., Gao, B., Jia, Y., Zhu, W., Chen, Q., Tian, X., Zhong, Z., Wang, J., Lan, Y.: How good is alphafold3 at ranking drug binding affinities? bioRxiv, 2025–05 (2025)

[21] Desai, D., Kantliwala, S.V., Vybhavi, J., Ravi, R., Patel, H., Patel, J., RAVI, R.: Review of alphafold 3: transformative advances in drug design and therapeutics. Cureus **16**(7) (2024)

[22] Zhao, L., Zhao, J., Zhong, K., Tong, A., Jia, D.: Targeted protein degradation: mechanisms, strategies and application. Signal transduction and targeted therapy **7**(1), 113 (2022)

[23] Schapira, M., Calabrese, M.F., Bullock, A.N., Crews, C.M.: Targeted protein degradation: expanding the toolbox. Nature reviews Drug discovery **18**(12), 949–963 (2019)

[24] Békés, M., Langley, D.R., Crews, C.M.: Protac targeted protein degraders: the past is prologue. Nature reviews Drug discovery **21**(3), 181–200 (2022)

[25] Toriki, E.S., Papatzimas, J.W., Nishikawa, K., Dovala, D., Frank, A.O., Hesse, M.J., Dankova, D., Song, J.-G., Bruce-Smythe, M., Struble, H., *et al.*: Rational chemical design of molecular glue degraders. ACS central science **9**(5), 915–926 (2023)

[26] Dewey, J.A., Delalande, C., Azizi, S.-A., Lu, V., Antonopoulos, D., Babnigg, G.:

Molecular glue discovery: current and future approaches. Journal of medicinal chemistry **66**(14), 9278–9296 (2023)

[27] Jan, M., Sperling, A.S., Ebert, B.L.: Cancer therapies based on targeted protein degradation—lessons learned with lenalidomide. Nature Reviews Clinical Oncology **18**(7), 401–417 (2021)

[28] Xie, L., Xie, L.: Elucidation of genome-wide understudied proteins targeted by protac-induced degradation using interpretable machine learning. PLOS Computational Biology **19**(8), 1010974 (2023)

[29] Zhang, W., Roy Burman, S.S., Chen, J., Donovan, K.A., Cao, Y., Shu, C., Zhang, B., Zeng, Z., Gu, S., Zhang, Y., *et al.*: Machine learning modeling of protein-intrinsic features predicts tractability of targeted protein degradation. Genomics, Proteomics & Bioinformatics **20**(5), 882–898 (2022)

[30] García Jiménez, D., Rossi Sebastiano, M., Vallaro, M., Mileo, V., Pizzirani, D., Moretti, E., Ermondi, G., Caron, G.: Designing soluble protacs: strategies and preliminary guidelines. Journal of Medicinal Chemistry **65**(19), 12639–12649 (2022)

[31] Nori, D., Coley, C.W., Mercado, R.: De novo protac design using graph-based deep generative models. arXiv preprint arXiv:2211.02660 (2022)

[32] Chen, Z., Gu, C., Tan, S., Wang, X., Li, Y., He, M., Lu, R., Sun, S., Hsieh, C.-Y., Yao, X., et al.: Interpretable protac degradation prediction with structure-informed deep ternary attention framework. bioRxiv, 2024–11 (2024)

[33] Erazo, F., Dunlop, N., Jalalypour, F., Mercado, R.: Enhancing protac ternary complex prediction with ligand information in alphafold 3 (2025)

[34] Karshikoff, A., Nilsson, L., Ladenstein, R.: Rigidity versus flexibility: the dilemma of understanding protein thermal stability. The FEBS journal **282**(20), 3899–3917 (2015)

[35] Schlessinger, A., Rost, B.: Protein flexibility and rigidity predicted from sequence. Proteins: Structure, Function, and Bioinformatics **61**(1), 115–126 (2005)

[36] Schlessinger, A., Yachdav, G., Rost, B.: Profbval: predict flexible and rigid residues in proteins. Bioinformatics **22**(7), 891–893 (2006)

[37] Meller, A., Ward, M.D., Borowsky, J.H., Lotthammer, J.M., Kshirsagar, M., Oviedo, F., Ferres, J.L., Bowman, G.: Predicting the locations of cryptic pockets from single protein structures using the pocketminer graph neural network. Biophysical journal **122**(3), 445 (2023)

[38] Wróblewski, K., Zalewski, M., Kuriata, A., Kmiecik, S.: Cabs-flex 3.0: an online

tool for simulating protein structural flexibility and peptide modeling. Nucleic Acids Research, 412 (2025)

[39] Pacesa, M., Nickel, L., Schellhaas, C., Schmidt, J., Pyatova, E., Kissling, L., Barendse, P., Choudhury, J., Kapoor, S., Alcaraz-Serna, A., et al.: Bindcraft: one-shot design of functional protein binders. bioRxiv, 2024–09 (2024)

[40] Cho, Y., Pacesa, M., Zhang, Z., Correia, B.E., Ovchinnikov, S.: Boltzdesign1: Inverting all-atom structure prediction model for generalized biomolecular binder design. bioRxiv, 2025–04 (2025)

[41] Wang, X., Yin, X., Jiang, D., Zhao, H., Wu, Z., Zhang, O., Wang, J., Li, Y., Deng, Y., Liu, H., *et al.*: Multi-modal deep learning enables efficient and accurate annotation of enzymatic active sites. Nature Communications **15**(1), 7348 (2024)

[42] Wang, E., Mi, X., Thompson, M.C., Montoya, S., Notti, R.Q., Afaghani, J., Durham, B.H., Penson, A., Witkowski, M.T., Lu, S.X., *et al.*: Mechanisms of resistance to noncovalent bruton's tyrosine kinase inhibitors. New England Journal of Medicine **386**(8), 735–743 (2022)

[43] Shen, H.C., Hammock, B.D.: Discovery of inhibitors of soluble epoxide hydrolase: a target with multiple potential therapeutic indications. Journal of medicinal chemistry **55**(5), 1789–1808 (2012)

[44] Huang, T., Song, Z., Ying, R., Jin, W.: Protein-nucleic acid complex modeling with frame averaging transformer. arXiv preprint arXiv:2406.09586 (2024)

[45] Yang, Z., Zeng, X., Zhao, Y., Chen, R.: Alphafold2 and its applications in the fields of biology and medicine. Signal Transduction and Targeted Therapy **8**(1), 115 (2023)

[46] Watkins, A.M., Arora, P.S.: Structure-based inhibition of protein–protein interactions. European journal of medicinal chemistry **94**, 480–488 (2015)

[47] Lu, H., Zhou, Q., He, J., Jiang, Z., Peng, C., Tong, R., Shi, J.: Recent advances in the development of protein–protein interactions modulators: mechanisms and clinical trials. Signal transduction and targeted therapy **5**(1), 213 (2020)

[48] Watson, J.L., Juergens, D., Bennett, N.R., Trippe, B.L., Yim, J., Eisenach, H.E., Ahern, W., Borst, A.J., Ragotte, R.J., Milles, L.F., *et al.*: De novo design of protein structure and function with rfdiffusion. Nature **620**(7976), 1089–1100 (2023)

[49] Krishna, R., Wang, J., Ahern, W., Sturmfels, P., Venkatesh, P., Kalvet, I., Lee, G.R., Morey-Burrows, F.S., Anishchenko, I., Humphreys, I.R., *et al.*: Generalized biomolecular modeling and design with rosettafold all-atom. Science **384**(6693), 2528 (2024)

[50] Wohlwend, J., Corso, G., Passaro, S., Reveiz, M., Leidal, K., Swiderski, W., Portnoi, T., Chinn, I., Silterra, J., Jaakkola, T., et al.: Boltz-1: Democratizing biomolecular interaction modeling. bioRxiv, 2024–11 (2024)

[51] Dauparas, J., Lee, G.R., Pecoraro, R., An, L., Anishchenko, I., Glasscock, C., Baker, D.: Atomic context-conditioned protein sequence design using ligandmpnn. Nature Methods, 1–7 (2025)

[52] Bennett, N.R., Coventry, B., Goreshnik, I., Huang, B., Allen, A., Vafeados, D., Peng, Y.P., Dauparas, J., Baek, M., Stewart, L., *et al.*: Improving de novo protein binder design with deep learning. Nature Communications **14**(1), 2625 (2023)

[53] Harteveld, Z., Van Hall-Beauvais, A., Morozova, I., Southern, J., Goverde, C., Georgeon, S., Rosset, S., Defferrard, M., Loukas, A., Vandergheynst, P., *et al.*: Exploring "dark-matter" protein folds using deep learning. Cell systems **15**(10), 898–910 (2024)

[54] Jendrusch, M.A., Yang, A.L., Cacace, E., Bobonis, J., Voogdt, C.G., Kaspar, S., Schweimer, K., Perez-Borrajero, C., Lapouge, K., Scheurich, J., et al.: Alphadesign: A de novo protein design framework based on alphafold. Molecular Systems Biology, 1–24 (2025)

[55] Uniprot: the universal protein knowledgebase in 2023. Nucleic acids research **51**(D1), 523–531 (2023)

[56] Petrova, N.V., Wu, C.H.: Prediction of catalytic residues using support vector machine with selected protein sequence and structural properties. BMC bioinformatics **7**, 1–12 (2006)

[57] Jones, D.T.: Protein secondary structure prediction based on position-specific scoring matrices. Journal of molecular biology **292**(2), 195–202 (1999)

[58] Teukam, Y.G.N., Dassi, L.K., Manica, M., Probst, D., Schwaller, P., Laino, T.: Language models can identify enzymatic binding sites in protein sequences. Computational and Structural Biotechnology Journal **23**, 1929–1937 (2024)

[59] Todd, A.E., Orengo, C.A., Thornton, J.M.: Plasticity of enzyme active sites. Trends in biochemical sciences **27**(8), 419–426 (2002)

[60] Joseph, R.E., Wales, T.E., Jayne, S., Britton, R.G., Fulton, D.B., Engen, J.R., Dyer, M.J., Andreotti, A.H.: Impact of the clinically approved btk inhibitors on the conformation of full-length btk and analysis of the development of btk resistance mutations in chronic lymphocytic leukemia. eLife **13**, 95488 (2024)

[61] Wong, R.L., Choi, M.Y., Wang, H.-Y., Kipps, T.J.: Mutation in bruton tyrosine kinase (btk) a428d confers resistance to btk-degrader therapy in chronic lymphocytic leukemia. Leukemia **38**(8), 1818–1821 (2024)

[62] Mato, A.R., Woyach, J.A., Brown, J.R., Ghia, P., Patel, K., Eyre, T.A., Munir, T., Lech-Maranda, E., Lamanna, N., Tam, C.S., *et al.*: Pirtobrutinib after a covalent btk inhibitor in chronic lymphocytic leukemia. New England Journal of Medicine **389**(1), 33–44 (2023)

[63] Rose, T.E., Morisseau, C., Liu, J.-Y., Inceoglu, B., Jones, P.D., Sanborn, J.R., Hammock, B.D.: 1-aryl-3-(1-acylpiperidin-4-yl) urea inhibitors of human and murine soluble epoxide hydrolase: Structure- activity relationships, pharmacokinetics, and reduction of inflammatory pain. Journal of medicinal chemistry **53**(19), 7067–7075 (2010)

[64] Eldrup, A.B., Soleymanzadeh, F., Taylor, S.J., Muegge, I., Farrow, N.A., Joseph, D., McKellop, K., Man, C.C., Kukulka, A., De Lombaert, S.: Structure-based optimization of arylamides as inhibitors of soluble epoxide hydrolase. Journal of medicinal chemistry **52**(19), 5880–5895 (2009)

[65] Brody, E.N., Gold, L.: Aptamers as therapeutic and diagnostic agents. Reviews in Molecular Biotechnology **74**(1), 5–13 (2000)

[66] Coutinho, M.F., Matos, L., Santos, J.I., Alves, S.: Rna therapeutics: how far have we gone? The mRNA Metabolism in Human Disease, 133–177 (2019)

[67] Morsch, F., Umasankar, I.L., Moreta, L.S., Latawa, P., Lange, D.B., Wengel, J., Konjen, H., Code, C.: Aptabert: Predicting aptamer binding interactions. bioRxiv, 2023–11 (2023)

[68] Chandola, C., Neerathilingam, M.: Aptamers for targeted delivery: Current challenges and future. Role of novel drug delivery vehicles in nanobiomedicine, 1 (2020)

[69] Ellington, A.D., Szostak, J.W.: In vitro selection of rna molecules that bind specific ligands. nature **346**(6287), 818–822 (1990)

[70] Stoltenburg, R., Reinemann, C., Strehlitz, B.: Selex—a (r) evolutionary method to generate high-affinity nucleic acid ligands. Biomolecular engineering **24**(4), 381–403 (2007)

[71] Nori, D., Jin, W.: Rnaflow: Rna structure & sequence design via inverse folding-based flow matching. arXiv preprint arXiv:2405.18768 (2024)

[72] Shui, B., Ozer, A., Zipfel, W., Sahu, N., Singh, A., Lis, J.T., Shi, H., Kotlikoff, M.I.: Rna aptamers that functionally interact with green fluorescent protein and its derivatives. Nucleic acids research **40**(5), 39–39 (2012)

[73] Tenney, I., Das, D., Pavlick, E.: Bert rediscovers the classical nlp pipeline. arXiv preprint arXiv:1905.05950 (2019)

[74] Jawahar, G., Sagot, B., Seddah, D.: What does bert learn about the structure of language? In: ACL 2019-57th Annual Meeting of the Association for Computational Linguistics (2019)

[75] Chen, J., Yang, Z., Yang, D.: Mixtext: Linguistically-informed interpolation of hidden space for semi-supervised text classification. arXiv preprint arXiv:2004.12239 (2020)

[76] Ren, J., Luo, J., Zhao, Y., Krishna, K., Saleh, M., Lakshminarayanan, B., Liu, P.J.: Out-of-distribution detection and selective generation for conditional language models. arXiv preprint arXiv:2209.15558 (2022)

[77] Wang, Y., Zhang, P., Yang, B., Wong, D., Zhang, Z., Wang, R.: Embedding trajectory for out-of-distribution detection in mathematical reasoning. Advances in Neural Information Processing Systems **37**, 42965–42999 (2024)

[78] Helland-Hansen, W., Hampson, G.: Trajectory analysis: concepts and applications. Basin Research **21**(5), 454–483 (2009)

[79] Rintoul, M.D., Wilson, A.T.: Trajectory analysis via a geometric feature space approach. Statistical Analysis and Data Mining: The ASA Data Science Journal **8**(5-6), 287–301 (2015)

[80] Jumper, J., Evans, R., Pritzel, A., Green, T., Figurnov, M., Ronneberger, O., Tunyasuvunakool, K., Bates, R., Žídek, A., Potapenko, A., *et al.*: Highly accurate protein structure prediction with alphafold. nature **596**(7873), 583–589 (2021)

[81] Dunlop, N., Erazo, F., Jalalypour, F., Mercado, R.: Predicting protac-mediated ternary complexes with alphafold3 and boltz-1 (2025)

[82] Das, R., Kretsch, R.C., Simpkin, A.J., Mulvaney, T., Pham, P., Rangan, R., Bu, F., Keegan, R.M., Topf, M., Rigden, D.J., *et al.*: Assessment of three-dimensional rna structure prediction in casp15. Proteins: Structure, Function, and Bioinformatics **91**(12), 1747–1770 (2023)

[83] Robin, X., Studer, G., Durairaj, J., Eberhardt, J., Schwede, T., Walters, W.P.: Assessment of protein–ligand complexes in casp15. Proteins: Structure, Function, and Bioinformatics **91**(12), 1811–1821 (2023)

[84] Zheng, H., Lin, H., Alade, A.A., Chen, J., Monroy, E.Y., Zhang, M., Wang, J.: Alphafold3 in drug discovery: A comprehensive assessment of capabilities, limitations, and applications. bioRxiv, 2025–04 (2025)

[85] Burley, S.K., Berman, H.M., Kleywegt, G.J., Markley, J.L., Nakamura, H., Velankar, S.: Protein data bank (pdb): the single global macromolecular structure archive. Protein crystallography: methods and protocols, 627–641 (2017)

[86] Elofsson, A.: Progress at protein structure prediction, as seen in casp15. Current opinion in structural biology **80**, 102594 (2023)