BC-DESIGN: A Biochemistry-Aware Framework for High-Precision Inverse Protein Folding

Xiangru Tang[†1], Xinwu Ye[†1], Fang Wu[†2], Daniel Shao[3], Yin Fang[4], Siming Chen[5], Dong Xu[6], and Mark Gerstein[1,3,7,8,9*]

[1] Department of Computer Science, Yale University, US
[2] Computer Science Department, Stanford University, US
[3] Department of Biomedical Informatics & Data Science, Yale University, US
[4] National Center for Biotechnology Information, National Institutes of Health, US
[5] School of Data Science, Fudan University, China
[6] Department of Electrical Engineering and Computer Science, University of Missouri, US
[7] Program in Computational Biology & Bioinformatics, Yale University, US
[8] Department of Molecular Biophysics & Biochemistry, Yale University, US
[9] Department of Statistics & Data Science, Yale University, US

Abstract. Inverse protein folding, which aims to design amino acid sequences for desired protein structures, is fundamental to protein engineering and therapeutic development. While recent deep-learning approaches have made remarkable progress in addressing this challenge, they typically represent biochemical properties as discrete features associated with individual residues. Here, we present BC-DESIGN, an approach that explicitly represents these properties as decorations on randomly sampled points on exterior surfaces and within internally bound regions representing the complete molecular extent of the protein. This provides a more natural way to capture the spatial distribution of properties. We demonstrate that BC-DESIGN significantly outperforms all current methods, improving sequence recovery from 67% to **88.37%** over the state-of-the-art methods (a **21.32%** absolute improvement) and reducing perplexity from 2.4 to **1.47** (a **39.51%** relative improvement) on the CATH 4.2 benchmark. Notably, our model exhibits robust generalization across diverse protein characteristics, achieving consistently high performance on proteins of varying sizes (50-500 residues), structural complexity (measured by contact order), and all major CATH fold classes. Through ablation tests, we compare the relative contribution of both structure encoding information and the encoded property information, and we show that both substantially contribute equally to this strong performance. Overall, this opens new avenues for computational protein engineering and drug discovery.

Keywords: Protein Design · Deep learning · Structural Bioinformatics · Biochemistry Encoding.

[†] Equal first authorship.
[*] Corresponding Author. Email: pi@gersteinlab.org

1 Introduction

Inverse protein folding, which aims to determine amino acid sequences that will adopt a specified three-dimensional structure, represents a fundamental challenge in computational biology [1–7]. This inverse problem is crucial for advancing our understanding of protein structure-function relationships and holds immense potential for therapeutic development [8, 9]. While nature has evolved proteins over billions of years to perform diverse functions, the ability to rationally design protein sequences that fold into predetermined structures would enable the creation of novel enzymes, therapeutic antibodies, and biomaterials with tailored properties [10–13]. The successful solution to this inverse problem could accelerate protein engineering by circumventing the traditional trial-and-error approaches, potentially leading to breakthroughs in drug development, vaccine design, and sustainable catalyst creation.

Traditional methods for inverse folding typically rely on different network architectures to model protein structures. Early methods primarily employed MLP-based frameworks such as SPIN and SPIN2 [14, 15], which integrate basic structural features like torsion angles and backbone angles. Subsequently, CNN-based models, including ProDCoNN and DenseCPD [16, 17], were proposed to extract higher-dimensional features from distance matrices and atomic distributions. More recently, the approaches for IPF primarily focus on optimizing network architectures, with methods like ProteinMPNN [18], PiFold [2], and ESM-IF1 [4] achieving remarkable results. On the other hand, while some previous models in other fields [19–21] have explored various ways to incorporate biochemical properties, they typically represent these features as discrete properties at the atomic or residue level. These representations, however, do not fully capture the continuous nature of biochemical properties across protein surfaces. Physicochemical properties such as hydrophobicity and charge distributions are inherently continuous features that define both the protein's spatial configuration and its interaction patterns. Despite numerous efforts to model these properties, existing methods lack effective ways to represent biochemical features as continuous decorations on protein surfaces, limiting their ability to capture the spatial distribution of these critical features. Therefore, developing a protein design method that can naturally encode continuous biochemical properties remains an important challenge.

Building upon these observations, we identify two major challenges in incorporating biochemical properties into IPF: (1) *how to construct appropriate representations of biochemical features as continuous decorations on protein surfaces, rather than discrete properties at individual residues*, and (2) *how to effectively model the relationship between these continuous biochemical features and residue-level structural information for accurate sequence prediction*. While structural information is naturally represented through graph-based encodings of residue relationships, the representation of continuous biochemical properties requires careful consideration of their spatial distribution across protein surfaces.

To address these challenges, we present BC-DESIGN, a biochemistry-aware framework for high-precision inverse protein folding. At its core, our approach represents biochemical properties as continuous distributions throughout the protein structure, including both surface and internal regions, rather than as discrete residue-level features. This biochemical encoding is achieved through a proposed BC-ENCODER that captures the spatial distribution of properties; here, we use hydrophobicity and charge at multiple scales, using constructed point clouds that sample both protein surfaces and internal spaces. In parallel, a STRUCT-ENCODER processes residue-level structural information through a hierarchical graph transformer. A BC-FUSION module then enables the framework to make biochemically informed sequence predictions by modeling the relationship between continuous biochemical features and discrete structural information. The effectiveness of this biochemistry-aware design is validated through our experiments on the CATH 4.2 dataset, where BC-DESIGN achieves unprecedented precision, significantly outperforming all IPF methods, including current state-of-the-art SPDesign by improving sequence recovery from 67.05% to 88.37% (a 21.32% absolute improvement) and reducing perplexity from 2.43 to 1.47 (a 39.51% relative improvement). We demonstrate that our biochemistry-aware approach enables consistently high-precision sequence prediction through detailed analysis across proteins of different sizes (50-500 residues) and various CATH fold classes.

2 Methodology

2.1 Input Representation

2.1.1 Structural Representation A protein is a three-dimensional macromolecule containing one or multiple polypeptide chains, each chain formed by amino acid residues. Given a protein structure, we represent it as a graph $\mathcal{G}(\mathcal{V}, \mathcal{E}, \mathcal{F}_\mathcal{V}, \mathcal{F}_\mathcal{E})$ according to the featurizer in PiFold, where each node in the node-set $\mathcal{V}$ represents a residue centered at its C_α atom position. The edge set $\mathcal{E}$ encodes spatial relationships between residues by connecting each node to its k nearest neighbors ($k = 30$ in our work), forming a k-NN graph that captures local interactions between nearby residues in the protein structure.

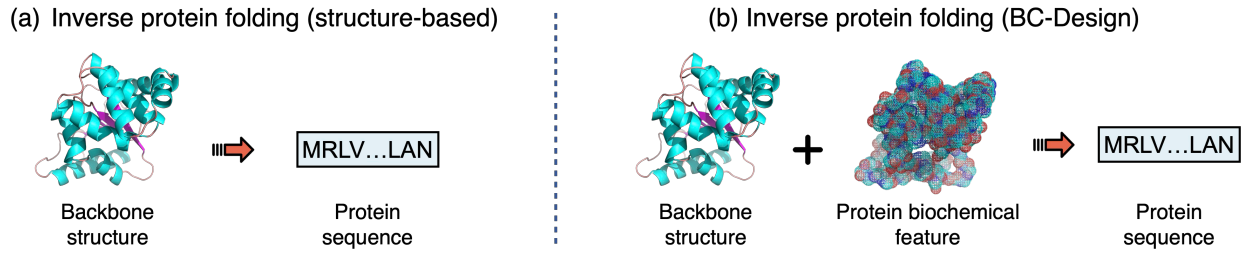

Fig. 1: Comparison of traditional IPF and our approach. (a) Traditional IPF workflow inputs only the protein backbone structure and predicts the amino acid sequence. (b) Our approach augments the input with biochemical feature distributions (hydrophobicity and charge) in addition to the backbone structure. The biochemical features are represented as continuous point clouds in 3D space.

To achieve translational and rotational invariance of our structural representation, we construct a local coordinate system $Q = [\mathbf{x}_Q, \mathbf{y}_Q, \mathbf{x}_Q \times \mathbf{y}_Q]$ for each residue based on its backbone atoms (C_α, C, N, and O). Each feature vector in the node feature set $\mathcal{F}_\mathcal{V} = \{\mathbf{f}_{v_1}, \mathbf{f}_{v_2}, \ldots, \mathbf{f}_{v_{|\mathcal{V}|}}\}$ encodes essential geometric properties of each residue: **Distances**: Distances between key backbone atoms (C_α, C, N, and O). **Angles**: Three bond angles ($N_i - C_{\alpha i} - C_i$, $C_{i-1} - N_i - C_{\alpha i}$, and $C_{\alpha i} - C_i - N_{i+1}$) and three dihedral angles ($N_i - C_{\alpha i}$, $C_i - C_{\alpha i}$, and $C_i - N_{i+1}$) that define the local backbone conformation. **Orientations**: Direction vectors from C, N, and O atoms to the central C_α.

Each feature in the edge feature set $\mathcal{F}_\mathcal{E} = \{\mathbf{f}_{e_1}, \mathbf{f}_{e_2}, \ldots, \mathbf{f}_{e_{|\mathcal{E}|}}\}$ captures the relative spatial arrangements between connected residues: **Inter-residue Distances**: Distances between backbone atoms of connected residues. **Inter-residue Rotations**: Quaternion representation $\psi(Q_i^T Q_j)$ of the relative rotation between local coordinate frames of residues i and j, where $\psi : \mathbb{R}^{3\times3} \to \mathbb{R}^4$ is a quaternion encoding function. **Inter-residue Orientations**: Direction vectors from the backbone atoms of residue j to the C_α atom of residue i.

This graph representation captures the local backbone geometry and spatial arrangements essential for protein structure characterization. Combining these carefully chosen geometric features ensures our model can learn from conformational preferences and structural interactions.

2.1.2 Biochemical Feature Construction In protein structures, biochemical properties are crucial in determining protein folding and function. These properties form distributions throughout the protein structure, which can be theoretically represented as a function $\phi : \mathbb{R}^3 \to \mathbb{R}^d$ that maps any spatial coordinate $\mathbf{x}_i \in \mathbb{R}^3$ to its corresponding biochemical features. To approximate this continuous distribution in a computationally tractable manner, we develop a discretized feature representation through point sampling and feature assignment, capturing surface and internal distributions essential for protein sequence design.

Among various biochemical properties, we specifically choose hydrophobicity and charge to construct our biochemical features, according to SurfPro [22], which demonstrated these two properties to be the most informative for protein sequence design. Our feature construction process consists of four key stages:

(i) **Surface Representation:** We first generate a simplified-protein surface representation using MSMS (based just on alpha carbon spheres), which constructs a triangulated mesh based on the solvent-accessible surface area (SASA) model. The vertices of this mesh form an initial surface point cloud $\mathcal{P}_S$, which undergoes Gaussian smoothing followed by octree-based compression to ensure high-quality and uniform point distribution along the protein surface. (ii) **Internal Space Sampling:** To capture the biochemical environment within the protein core, we construct a bounding box encompassing the protein structure and uniformly sample 5000 points within this volume. We retain only those points that fall within the protein interior, as determined by a Delaunay triangulation (and associated Voronoi construction) of $\mathcal{P}_S$, forming the internal point cloud $\mathcal{P}_I$. (iii) **Point Cloud Integration:** The surface ($\mathcal{P}_S$) and internal ($\mathcal{P}_I$) point clouds are merged and subsampled to create a unified representation of 5000 points that captures both surface and internal biochemical environments. (iv) **Biochemical Feature Assignment:** For each point in the unified cloud, we assign biochemical features based on their spatial relationship to the root alpha carbon of each residue. Specifically, each point is assigned the biochemical properties (hydrophobicity and charge values, detailed in Appendix A) of its nearest residue, determined by C_α positions. The resulting attributed point cloud $\mathcal{P} = \{P_1, P_2, ..., P_{5000}\}$ consists of points $P_i = \{\mathbf{x}_i, \phi(\mathbf{x}_i)\}$, where $\mathbf{x}_i \in \mathbb{R}^3$ represents spatial coordinates and $\phi(\mathbf{x}_i) = (h(\mathbf{x}_i), c(\mathbf{x}_i))$ encodes hydrophobicity and charge values.

2.2 Model Architecture

2.2.1 STRUCT-ENCODER To effectively capture both local and global structural information, we augment the protein structure graph with auxiliary nodes that serve as dedicated feature aggregators at different spatial scales. Specif-

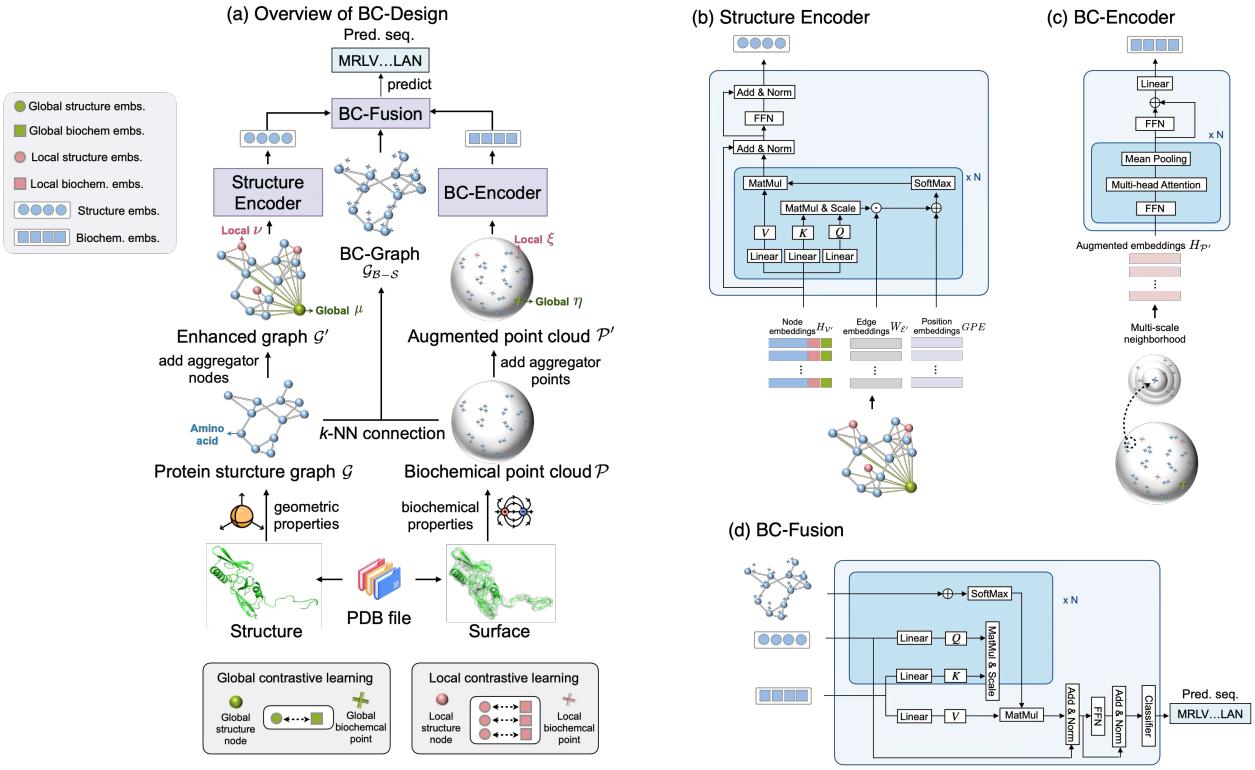

Fig. 2: Overview of our proposed BC-DESIGN. **(a)** Illustration of the overall model architecture and model training. **(b)** The STRUCT-ENCODER, showing Structure Graph Transformers with local and global aggregator nodes and integration of heat kernel embeddings to capture structural information from the structure graph. **(c)** The BC-ENCODER is characterized by multi-scale feature aggregation and the integration of biochemical aggregator points. **(d)** BC-FUSION uses the BC-GRAPH for masked Transformer decoder layers.

ically, we introduce two types of aggregator nodes: (1) a global aggregator that summarizes features of the entire protein structure and (2) several local aggregators that capture structural patterns within specific spatial regions. This hierarchical aggregation scheme is inspired by the [CLS] (classification) token mechanism in Transformer models, where special tokens accumulate and summarize information from input sequences.

Formally, we augment the original structure graph $\mathcal{G}$ into an enhanced graph $\mathcal{G}'(\mathcal{V}', \mathcal{E}', \mathcal{F}_\mathcal{V}, \mathcal{F}_\mathcal{E})$ by adding these aggregator nodes. The augmented node set is defined as $\mathcal{V}' = \mathcal{V} \cup \mathcal{V}_l \cup \{\mu\}$, where $\mathcal{V}_l = \{\nu_1, \nu_2, \ldots, \nu_{|\mathcal{V}_l|}\}$ contains the local aggregator nodes, and μ is the global aggregator node. In our work, $|\mathcal{V}_l|$ is set to 8. To enable effective feature aggregation, we construct additional edges $\mathcal{E}_c$ to connect these aggregator nodes to the protein structure: we randomly select $|\mathcal{V}_l|$ centers $\{v_{k_1}, v_{k_2}, \ldots, v_{k_{|\mathcal{V}_l|}}\}$ from the original nodes, and each local aggregator ν_i is connected to all nodes within a sphere region $\mathcal{R}_i$ of radius r_L centered at its corresponding v_{k_i}. The global aggregator μ is connected to all nodes in V to capture protein-level features. The complete edge set is thus $\mathcal{E}' = \mathcal{E} \cup \mathcal{E}_c$.

After constructing $\mathcal{G}'$, the node feature matrix $F_\mathcal{V} = \left[\mathbf{f}_{v_1}, \mathbf{f}_{v_2}, \ldots, \mathbf{f}_{v_{|\mathcal{V}|}}\right]^\top \in \mathbb{R}^{|\mathcal{V}| \times d_v}$ and edge feature tensor $F_\mathcal{E} \in \mathbb{R}^{|\mathcal{V}| \times |\mathcal{V}| \times d_e}$ are processed by feed-forward networks (FFNs): $H_\mathcal{V} = \text{FFN}_\mathcal{V}(F_\mathcal{V}) \in \mathbb{R}^{|\mathcal{V}| \times d_\mathcal{V}}$ and $W_\mathcal{E} = \text{FFN}_\mathcal{E}(F_\mathcal{E}) \in \mathbb{R}^{|\mathcal{V}| \times |\mathcal{V}|}$, where $F_\mathcal{E}[i, j, :] = \mathbf{f}_{(v_i, v_j)}$ if $(v_i, v_j) \in \mathcal{E}$ and 0 otherwise. With structure aggregator node embeddings initialized with learnable parameters $H_{\mathcal{V}_l \cup \{\mu\}} \in \mathbb{R}^{(|\mathcal{V}_l|+1) \times d_\mathcal{V}}$, the node embeddings are concatenated as $H_{\mathcal{V}'} = [H_\mathcal{V}; H_{\mathcal{V}_c}]$. $W_\mathcal{E}$ is extended by adding $|\mathcal{V}_l| + 1$ additional rows and columns: For each position corresponding to an edge in $\mathcal{E}_c$, the value is set to 1, and 0 otherwise, resulting in the weight matrix $W_{\mathcal{E}'} \in \mathbb{R}^{|V'| \times |V'|}$. We also apply a heat kernel to the graph Laplacian L to construct graph positional encoding: $H(t) = \exp(-tL)$ is obtained with diffusion time $t = 1$ and extended with $|\mathcal{V}_l| + 1$ zero rows and columns to form the final $GPE \in \mathbb{R}^{|\mathcal{V}'| \times |\mathcal{V}'|}$.

The STRUCT-ENCODER then processes $H_{\mathcal{V}'}$, $W_{\mathcal{E}'}$, and GPE as input using a specialized Structure Graph Transformer: $H_{\mathcal{V}'}^{l+1} = \text{StrucGraphTrans}(H_{\mathcal{V}'}^l, W_{\mathcal{E}'}, GPE)$. For each Structure Graph Transformer Layer, the input embeddings $H_{\mathcal{V}'}$ are linearly projected into query (Q), key (K), and value (V) matrices for each attention head. The attention score matrix S is computed using scaled dot-product attention, $\frac{QK^T}{\sqrt{d_k}}$. Afterwards, S is modified by $W_{\mathcal{E}'}$ and GPE: $S' = S \odot W_{\mathcal{E}'} + GPE$. The modified attention scores S' are passed through softmax to obtain attention weights,

which are used to compute the attention output for each attention head. The outputs from all heads are then concatenated and passed through a final linear projection to obtain the multi-head attention output. A residual connection is employed around each layer, followed by layer normalization. Afterward, an FFN is applied, which is also surrounded by a residual connection.

2.2.2 The BC-ENCODER Similar to the introduction of aggregator nodes in Sec. 2.2.1, the local biochemical aggregator points, $\mathcal{P}_l = \{\xi_1, \xi_2, \ldots, \xi_{|\mathcal{V}_l|}\}$, and a global biochemical aggregator point, η, are added to the point cloud $\mathcal{P}$ together for each protein. Specifically, we define a local biochemical aggregator point ξ_i in the STRUCT-ENCODER for each spherical region $\mathcal{R}_i \subset \mathbb{R}^3$. Denote the augmented point cloud as $\mathcal{P}' = \mathcal{P} \cup \mathcal{P}_l \cup \{\eta\}$. Similar to the structure aggregator nodes, the embeddings of the biochemical aggregator points are initialized as learnable parameters $H_{\mathcal{P}_l}$ and H_η. For each point $P_i \in \mathcal{P}$, their features $\phi(\mathbf{x}_i)$ are passed through an FFN, resulting in a set of embeddings $H_{\mathcal{P}}$. We then define the augmented embeddings as $H_{\mathcal{P}'} = [H_{\mathcal{P}}; H_{\mathcal{P}_l}; H_\eta]$.

The encoder processes the augmented point cloud using a multi-scale approach. For each point $P_i \in \mathcal{P}$, we define its multi-scale neighborhood, $\mathcal{N}_r(P_i)$, which consists of all points within a radius r and the aggregator nodes ξ_k and η: $\mathcal{N}_r(P_i) = \{P_j \mid \|\mathbf{x}_i - \mathbf{x}_j\| \leq r\} \cup \{\xi_k \mid P_i \in \mathcal{R}_k\} \cup \{\eta\}$. This neighborhood defines the local context around point P_i at different spatial scales, where each radius r captures a different level of detail. Meanwhile, for each local aggregator nodes ξ_k and global node η, their neighborhoods are defined to contain themselves only: $\mathcal{N}_r(\xi_k) = \{\xi_k\}$ and $\mathcal{N}_r(\eta) = \{\eta\}$.

Next, feature aggregation is performed using multi-head attention (MHA) applied to the embeddings of the point P_i' and its neighborhood $\mathcal{N}_r(P_i')$. The result of the MHA operation is pooled using mean pooling to obtain the aggregated feature for each scale: $\mathbf{z}_i^r = \text{MeanPool}(\text{MHA}(H_{\mathcal{P}'}[\mathcal{N}_r(P_i')]))$. This operation is repeated for multiple radii, creating a set of aggregated features, each corresponding to a different scale. To combine these multi-scale features, we concatenate the aggregated features from different radii: $\mathbf{z}_i = [(\mathbf{z}_i^{r_1})^T, (\mathbf{z}_i^{r_2})^T, \ldots, (\mathbf{z}_i^{r_m})^T]^T$. In our work, the number of scales, m, is set to 4. Finally, the fused feature $\mathbf{z}_i$ is passed through an FFN surrounded by a residual connection and followed by a linear transformation, yielding the final output: $\tilde{\mathbf{z}}_i = (\text{FFN}(\mathbf{z}_i) + \mathbf{z}_i)W$.

2.2.3 BC-FUSION Decoder and Residue Classification BC-FUSION serves as a specially designed decoder to integrate structural and biochemical information for amino acid sequence prediction in three steps:

First, it establishes spatial correspondences between structural and biochemical features through a bipartite graph structure (BC-GRAPH). Specifically, the BC-GRAPH $\mathcal{G}_{\mathcal{B}-\mathcal{S}}(\mathcal{V}_\mathcal{B}, \mathcal{E}_\mathcal{B})$ connects each residue node in $\mathcal{V}$ to its $k_\mathcal{B}$ nearest points in the biochemical point cloud $\mathcal{P}$, where $k_\mathcal{B} = \max(1, \left\lfloor \frac{1400}{|\mathcal{V}|} \right\rfloor)$ adapts to protein size. Second, it performs feature fusion using masked transformer decoder layers, where the attention mechanism is guided by the BC-GRAPH's adjacency matrix. This ensures that each residue position attends only to its relevant biochemical features during decoding. Finally, for each residue position, the fused features are transformed through an FFN and softmax layer to predict probabilities over the 20 standard amino acids: $\hat{Y} = \text{softmax}(\text{FFN}(H_\mathcal{V}'))$, where $H_\mathcal{V}'$ represents the fused embeddings at each residue position. This hierarchical decoding process enables the model to leverage both structural context and spatially aligned biochemical properties when predicting amino acid identities.

2.3 Contrastive Learning and Training Objectives

The model is trained with a primary objective of accurate sequence prediction, supplemented by global and local contrastive learning to enhance the fusion of structural and biochemical features. The primary sequence prediction objective is formulated as a cross-entropy loss: $\mathcal{L}_{\text{CE}} = -\frac{1}{N}\sum_{i=1}^{N}\sum_{c=1}^{20} y_{i,c}\log(\hat{y}_{i,c})$, where N is the number of residues, $y_{i,c}$ is the ground truth one-hot encoding for residue i and amino acid class c, and $\hat{y}_{i,c}$ is the predicted probability for residue i being amino acid c.

Global Contrastive Learning (GCL) For each protein structure, we form positive pairs using its global structure aggregator node and global biochemical aggregator point. To construct negative pairs, we maintain two continuously updated queues, $\mathcal{Q}_\mathcal{S}$ and $\mathcal{Q}_\mathcal{B}$, which store global structure embeddings and global biochemical embeddings, respectively, from the last $K_\mathcal{B}$ proteins (set to 64 in this paper) during training. The global contrastive loss is defined as:

$$\mathcal{L}_{\text{GCL}} = \frac{1}{2}\left(\text{NT-Xent}(\tilde{\mathbf{z}}_\eta, \tilde{\mathbf{h}}_\mu, \mathcal{Q}_\mathcal{S}) + \text{NT-Xent}(\tilde{\mathbf{h}}_\mu, \tilde{\mathbf{z}}_\eta, \mathcal{Q}_\mathcal{B})\right),$$

where NT-Xent is the normalized temperature-scaled cross entropy (NT-Xent) loss function, a contrastive loss used to maximize agreement between similar pairs of embeddings and push dissimilar pairs apart [23]. $\tilde{\mathbf{h}}_\mu$ and $\tilde{\mathbf{z}}_\eta$ denote the global structure and global biochemical embeddings, respectively.

Local Contrastive Learning (LCL) At the local level, the non-corresponding node-point pairs within the same protein are treated as negative pairs. Since biochemical properties like hydrophobicity and charge are inherent to amino acids and less variable than structural conformations, we encourage the local structure embeddings to learn and reflect essential biochemical information. To adjust the local structure embeddings to better align with the corresponding local biochemical embeddings and preserve the local biochemical embeddings as stable representations of biochemical properties, we design asymmetric contrastive learning by only using local structure embeddings as anchors in the NT-Xent losses. $\tilde{\mathbf{h}}_{\nu_i}$ and $\tilde{\mathbf{z}}_{\xi_i}$ are the local structure and local biochemical embeddings, respectively, the local contrastive loss is defined as:

$$\mathcal{L}_{\text{LCL}} = \frac{1}{|\mathcal{V}_l|} \sum_{i=1}^{|\mathcal{V}_l|} \text{NT-Xent}\left(\tilde{\mathbf{h}}_{\nu_i}, \tilde{\mathbf{z}}_{\xi_i}, \{\tilde{\mathbf{z}}_{\xi_1}, \tilde{\mathbf{z}}_{\xi_2}, \ldots, \tilde{\mathbf{z}}_{\xi_{i-1}}, \tilde{\mathbf{z}}_{\xi_{i+1}}, \ldots, \tilde{\mathbf{z}}_{\xi_{|\mathcal{P}_l|}}\}\right).$$

Finally, the model is trained using a combined loss function: $\mathcal{L} = \mathcal{L}_{\text{CE}} + \lambda_1 \mathcal{L}_{\text{GCL}} + \lambda_2 \mathcal{L}_{\text{LCL}}$, where $\mathcal{L}_{\text{CE}}$ is the cross-entropy loss for sequence prediction, and $\lambda_1 = \lambda_2 = 1$ are the weights for the global and local contrastive losses. This combined objective ensures that the model learns to predict amino acid sequences while maintaining consistency between structural and biochemical representations at both global and local scales.

3 Results

3.1 Cross Validation on CATH 4.2 Dataset

To assess the effectiveness of BC-DESIGN in generating protein sequences from backbone structures and biochemical feature distributions, we evaluate its performance against recent state-of-the-art methods on the **CATH 4.2** dataset [24]. This comprehensive dataset encompasses complete protein chains up to 500 residues in length, with structures organized at 40% non-redundancy based on their CATH classification (Class, Architecture, Topology, Homologous). Following the protocol established in previous work [25], we adopt an identical data split comprising **18,024** training, **608** validation, and **1,120** test proteins, with no CAT overlap across sets.

To comprehensively evaluate BC-DESIGN's performance, we conduct both sequence-level and structure-level validations. At the sequence level, we benchmark against a wide spectrum of state-of-the-art methods, including traditional graph-based approaches (GVP [26], StructGNN [26], GraphTrans [25], GCA [27]), recent advanced architectures (ProteinMPNN [18], PiFold [2], AlphaDesign [28], SPIN-CGNN [29], GRADE-IF [30], DIPRoT [31], SPDesign [32], ProRefiner+ESM-IF1 [33]), and language model-based methods (ESM-IF1 [4], LM-Design [34]). We also compare with methods incorporating external knowledge or multi-modal learning (Knowledge-Design [35], MMDesign [36], VFN-IFE [37]). For structure-level assessment, we evaluate against leading approaches, including ESM-Design [38], AlphaFold-Design (AF-Design) [39], PiFold, GraphTrans, GVP, ByProt [40], and ProteinMPNN.

Sequence-level validation. To evaluate BC-DESIGN's performance in protein design, we employ three complementary metrics on the CATH 4.2 test set: sequence recovery, perplexity, and native sequence similarity recovery (nssr) [41]. Sequence recovery quantifies the model's accuracy in reproducing native amino acid sequences, while perplexity measures the alignment between predicted and actual amino acid probabilities at each position. The nssr metric extends beyond exact matches by considering residue similarity based on BLOSUM62 scores [42], where positive-scoring residue pairs are counted as similar. Among these metrics, sequence recovery and perplexity are most widely adopted in the field, and we assess BC-DESIGN in comparison to most of the baselines. We benchmark BC-DESIGN against leading approaches, including ProRefiner+ESM-IF1, ESM-IF1, GVP, ProteinMPNN, and ProteinMPNN-C, specifically for the nssr metric. As shown in Fig. 3 (a, b), BC-DESIGN significantly outperforms existing methods, achieving **88.37%** sequence recovery, **1.47** perplexity, and **93.55%** nssr score. These results demonstrate that our integrated approach of modeling both backbone structure and biochemical features substantially enhances protein design accuracy across diverse proteins in the CATH 4.2 dataset.

Structure-level validation. While sequence-level metrics provide valuable insights, they cannot fully capture the practical effectiveness of protein design models. Given that protein function critically depends on structure, we perform a structural evaluation of our designs. We utilize ESMFold [43] for structure prediction due to its comparable accuracy to AlphaFold 2 [8] with improved computational efficiency. Following [44], we use a curated subset of **82** proteins from the CATH 4.2 test set by selecting one protein from each CATH family randomly. The structural quality is assessed using two complementary metrics: Ref-TM [44] for structural similarity to native conformations, and pLDDT [8] for prediction confidence. As shown in Fig. 3 (c), BC-DESIGN achieves superior structural accuracy compared to most baseline methods, with Ref-TM of **0.77** and pLDDT of **75.79**, second only to ProteinMPNN. These results

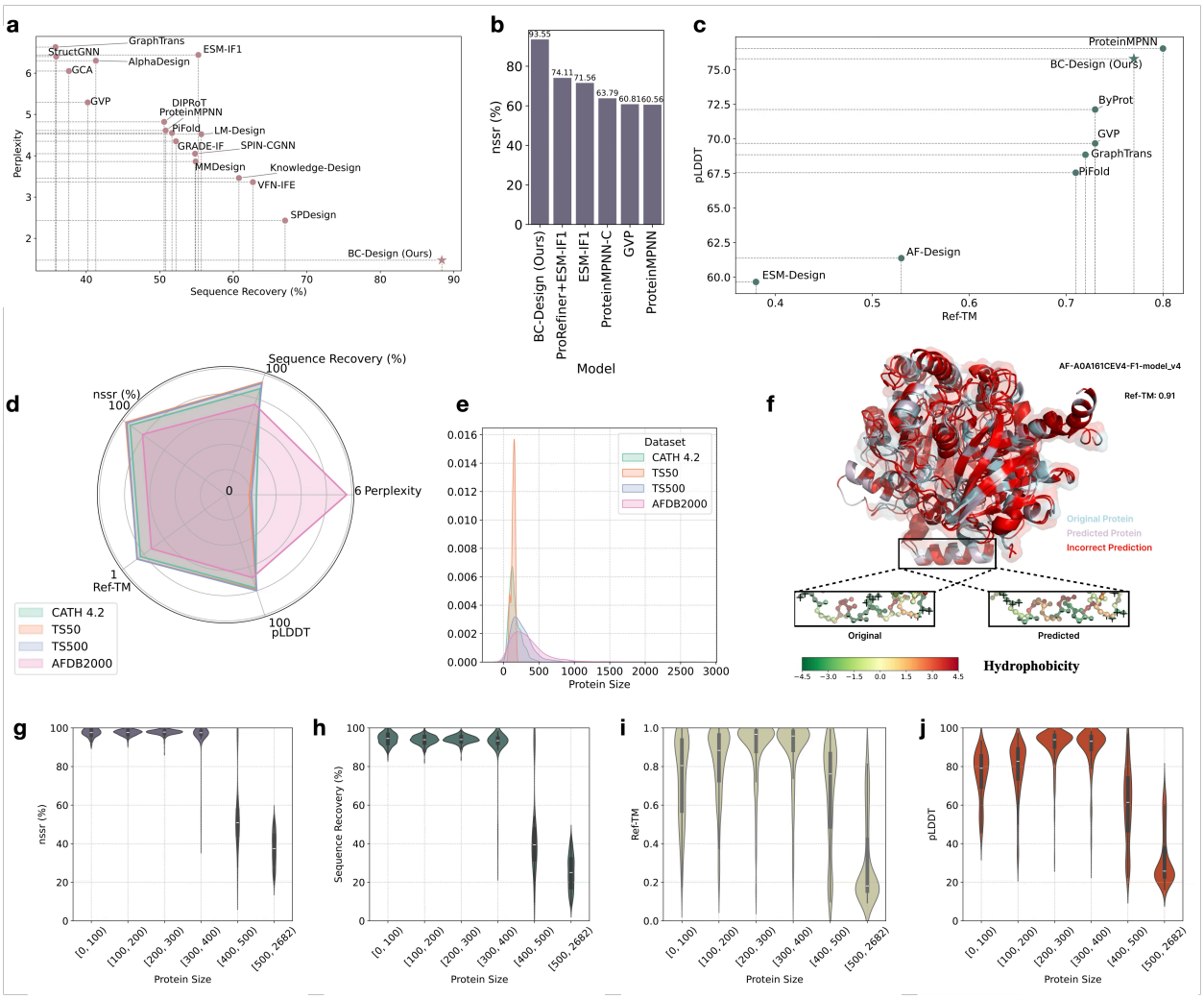

Fig. 3: **(a)** Sequence recovery and perplexity of predictions by BC-DESIGN and baseline models on the CATH 4.2 test set. **(b)** Nssr of predicted sequences by BC-DESIGN and baseline models on the CATH 4.2 test set. **(c)** Ref-TM and pLDDT of predicted structures (folded by ESMFold) for sequences predicted by BC-DESIGN and baseline models on the filtered CATH 4.2 test set. **(d)** Perplexity, sequence recovery, nssr, Ref-TM, and pLDDT for BC-DESIGN predictions (or their structures folded by ESMFold) on the CATH 4.2 test set, TS50, TS500, and AFDB2000. **(e)** Protein size distribution in the CATH 4.2, TS50, TS500, and AFDB2000 datasets, showing that AFDB2000 contains a larger proportion of long sequences. **(f)** Visualization of original and predicted sequences of a protein with a long sequence from AFDB2000 showing strong structural recovery. **(g-j)** Sequence and structure recovery of BC-DESIGN predictions for proteins of different sizes in AFDB2000.

suggest that BC-Design effectively captures the complex relationships between sequences, structures, and biochemical features, rather than merely optimizing sequence similarity. The reliability of these results is reinforced by [44], which demonstrates consistent performance rankings across different structure prediction models, including ESMFold, AlphaFold 2, and OmegaFold [45].

3.2 Generalization

To assess the generalizability of BC-DESIGN, we conduct further experiments on the following three datasets in addition to the CATH 4.2 test set: (i)**TS50**. TS50 is a benchmark set of 50 protein chains proposed by [15]. (ii) **TS500**. Similar to TS50, the TS500 dataset [15] contains 500 proteins. (iii) **AFDB2000**. The AlphaFold Database (AFDB) [46] is a public database providing AF-predicted 3D protein structures. To assess the model's ability to generalize to new structures, we select the first 2,000 proteins from Swiss-Prot with AF-predicted structures (ordered alphabetically by their accession numbers, which do not reflect any biological classification or functional information), resulting in an unbiased dataset AFDB2000. AFDB2000 contains no overlapping structures with other datasets.

The experimental results are presented in Fig. 3 (d). On both TS50 and TS500 datasets, BC-DESIGN consistently outperforms or achieves comparable results to the CATH 4.2 test set across all metrics at both the sequence and structure levels. On the AFDB2000 dataset, while BC-DESIGN still performs well, it shows inferior results compared to the other test datasets. Specifically, sequence recovery, nssr, Ref-TM, and pLDDT are approximately 0.1 lower on AFDB2000 than on the CATH 4.2 test set, and the perplexity value is higher. To investigate the reasons behind BC-DESIGN's reduced performance on AFDB2000, we analyzed the protein size distributions across the four test datasets, as depicted in Fig. 3 (e). It is observed that AFDB2000 contains a significantly larger proportion of large proteins compared to the other datasets.

To explore this further, we divided the AFDB2000 dataset into groups based on protein sequence length: [0, 100), [100, 200), [200, 300), [300, 400), [400, 500), and [500, $+\infty$). For each group, we evaluated four key metrics: sequence recovery, nssr, Ref-TM, and pLDDT. As shown in Fig. 3 (g, h), at the sequence level, BC-DESIGN performs well for proteins smaller than 400 amino acids, achieving high sequence recovery and nssr values comparable to those observed in the other test datasets. However, for proteins in larger size divisions, these values decline significantly. At the structure level (Fig. 3 (i, j)), the metrics initially increase with protein size but then decrease, with BC-DESIGN showing its best Ref-TM and pLDDT values for proteins in the [200, 300) and [300, 400) length ranges. Interestingly, the median Ref-TM values for the [0, 100) and [400, 500) divisions are quite similar. This suggests that, while the model struggles with larger proteins in terms of sequence recovery, it still manages to achieve comparable structural recovery and foldability. Despite this, BC-DESIGN is able to achieve good performance on large proteins from a structural perspective, even though their sizes fall outside the model's training data distribution. For instance, some proteins in the largest size division (over 500 amino acids) exhibit high (> 0.8) Ref-TM values. As an illustrative example, we evaluate BC-DESIGN on protein AF-A0A161CEV4-F1-model_v4, which has a length of 622 amino acids (Fig. 3 (f)). BC-DESIGN achieves a Ref-TM score of 0.91, demonstrating that the model can generalize to out-of-distribution proteins in terms of size when assessed structurally. In summary, BC-DESIGN performs best on small to medium-sized proteins (less than 400 amino acids) across datasets with different distributions and can generalize to predict sequences with appreciable structural accuracy and foldability even for larger proteins.

3.3 Stratified Validation

As demonstrated, BC-DESIGN's performance might vary with protein size. To further analyze its strengths and weaknesses across different protein types, we evaluate BC-DESIGN on subsets of the CATH 4.2 test set using various stratifications. This allows us to assess the model's performance across diverse protein characteristics.

Protein size. We evaluate BC-DESIGN on subsets of the CATH 4.2 test set, where the dataset was divided based on protein size, illustrated in Fig. 4 (a-d). BC-DESIGN demonstrated excellent performance across all groups, with median sequence recovery $>$ **80%**, nssr $>$ **85%**, Ref-TM $>$ **0.8**, and pLDDT $>$ **0.75**. Overall, BC-DESIGN performs relatively better on shorter sequences at the sequence level, likely due to the simpler sequence-to-structure relationships in shorter proteins. However, at the structure level, BC-DESIGN achieves comparable median Ref-TM scores for proteins even in the [400, 500) length range, similar to shorter sequences. This could be because the model prioritizes structural consistency over exact sequence recovery, ensuring that the predicted sequences maintain the backbone structure's fold. Additionally, longer proteins may have more sequence variability in regions that do not significantly impact the overall fold, which may explain why sequence recovery declines but structural metrics remain strong.

Structural complexity. We investigate the relationship between BC-DESIGN's performance and structural complexity of proteins. To quantify structural complexity, we used contact order, a metric that represents the normalized average sequence distance between residues that are in contact within the protein structure. As shown in Fig. 4 (e-h), four metrics all show higher median values and reduced variability as contact order increases in terms of the overall trend. This suggests that proteins with higher structural complexity, as indicated by higher contact order, provide more favorable conditions for accurate prediction by the model. The consistent observation across different metrics implies that the model was better able to handle the biochemical and spatial complexity when the folding topology involved more long-range interactions, leading to more reliable predictions. Nevertheless, the variations in model performance across different levels of structural complexity were not very substantial. As shown in Fig. 4 (j) with cases, BC-DESIGN demonstrates consistently high sequence recovery and structural recovery for proteins representing low, median, and high contact order. This indicates that the model is robust across a spectrum of structural complexities.

CATH class. To further examine the influence of structural properties, we divide the CATH 4.2 test set according to the CATH classification system. CATH categorizes protein domains into four main classes based on their overall secondary structure content: 'Mainly Alpha,' 'Mainly Beta,' 'Alpha Beta,' and 'Few Secondary Structures.' Approximately 90%

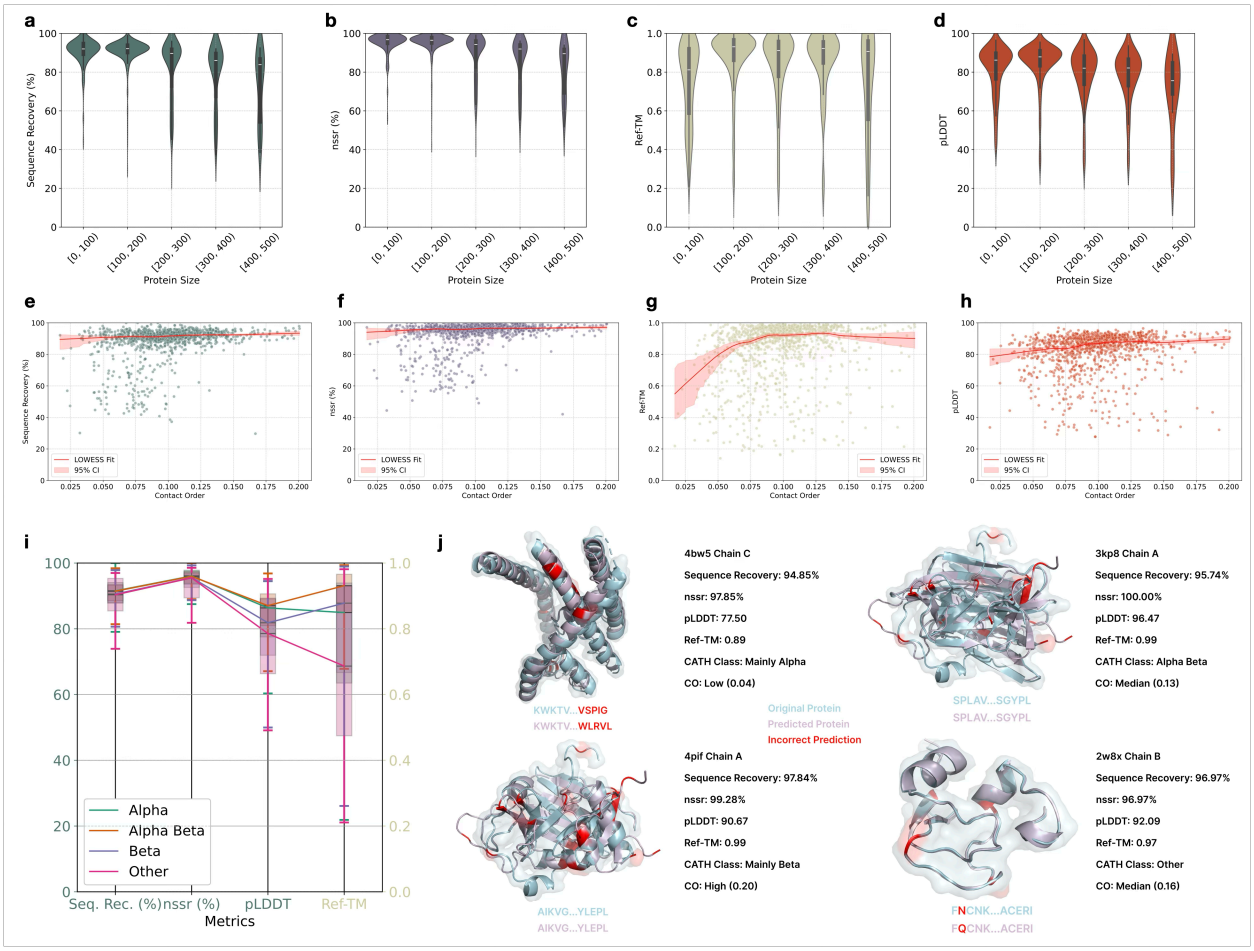

Fig. 4: **(a-d)** Sequence and structure recovery of BC-DESIGN predictions for proteins of different sizes in the CATH 4.2 test set. **(e-h)** Relationship between sequence and structure recovery metrics of BC-DESIGN predictions and structural complexity (measured by contact order) on the CATH 4.2 test set. **(i)** Sequence and structure recovery of BC-DESIGN predictions for proteins from different CATH Classes in the CATH 4.2 test set. **(j)** Visualization of original and predicted sequences for four proteins with different CATH Classes and structural complexity in the CATH 4.2 test set.

of protein domains are classified into these four categories. For those that do not fit into any of these classes, we denote them as 'Other' in this paper. Additionally, we merge the 'Few Secondary Structures' class into the 'Other' category. Specifically, each protein is classified according to its largest domain's assigned class. As shown in Fig. 4 (i), BC-DESIGN achieves high median sequence recovery and nssr across all classes, with only slight differences among them. This indicates that the model is broadly effective at predicting the amino acid sequence regardless of the specific structural characteristics represented by the CATH classes. However, at the structural level, BC-DESIGN demonstrates variable median performance and variance across the classes, suggesting that the model's ability to accurately predict fine structural details might depend on the protein's structural characteristics. For the structure-level metrics, Kruskal-Wallis tests followed by Dunn's tests indicate significant differences in the median pLDDT of the 'Mainly Beta' class compared to the 'Mainly Alpha' and 'Alpha Beta' classes. Additionally, statistically significant differences were found in median Ref-TM scores, with the 'Alpha Beta' class achieving the highest median score compared to the other classes. This result suggests that the mixed alpha and beta structural elements may present favorable features for prediction by the model, potentially due to their stabilized folding patterns involving both types of secondary structures. Conversely, the lower scores for the 'Other' class indicate that irregular or less structured proteins are more challenging for the model to predict accurately. These findings indicate challenges posed by special topologies, particularly for beta-rich and structurally diverse proteins. Finally, illustrative cases representing the four classes are shown in Fig. 4 (j).

Protein region. Next, we investigate the accuracy of sequence prediction in different regions of the protein, specifically focusing on the core versus surface residues. By analyzing the performance separately for buried (core) and exposed

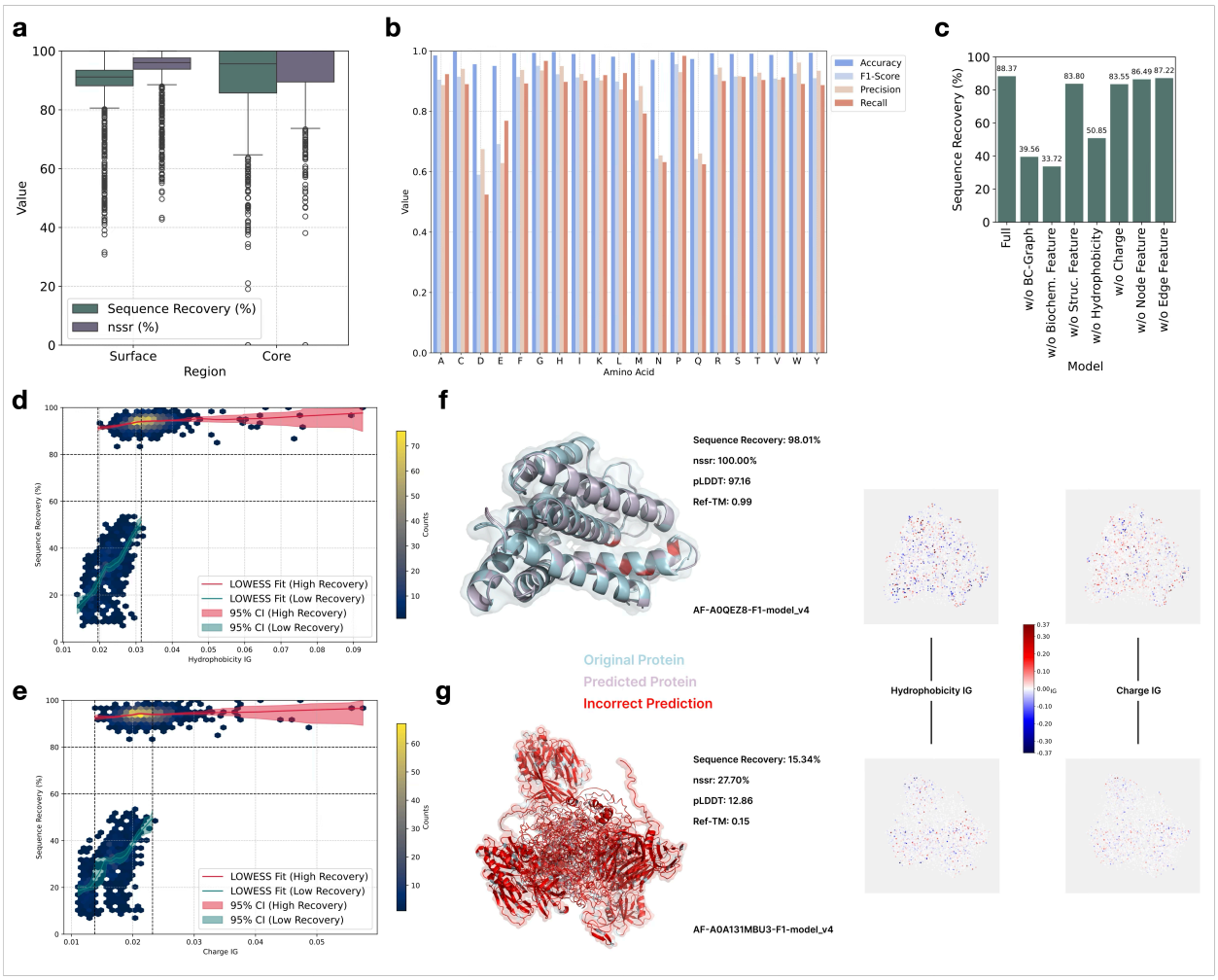

Fig. 5: **(a)** Sequence recovery and nssr of BC-DESIGN predictions for surface and core residues on the CATH 4.2 test set. **(b)** Residue-level evaluation of BC-DESIGN predictions on the CATH 4.2 test set. **(c)** Results of the ablation study. **(d, e)** Relationship between sequence recovery and IG values for hydrophobicity and charge features. **(f, g)** Case examples illustrating the relationship between BC-DESIGN performance and IG values of biochemical features.

(surface) regions, we find that while BC-DESIGN achieves strong results in both areas, the core residues—typically more conserved across homologous proteins—have significantly higher sequence recovery and nssr scores (Fig. 5 (a)). This indicates that our method is more effective in capturing these essential conserved features, which are critical for the structural integrity and biological function of the protein. These findings align with earlier studies on structure-only inverse folding, suggesting that the core regions, due to their conserved nature, are easier for the model to predict accurately, thereby potentially aiding in the retention of the protein's biological activity.

Amino acid type. In Fig. 5 (b), we examine each amino acid type using metrics including accuracy, precision, recall, and F1-score. The results demonstrate that accuracy remains consistently high across all amino acid types, highlighting the robustness of our method. However, for certain residues, including D (Aspartic acid), E (Glutamic acid), N (Asparagine), and Q (Glutamine), the precision, recall, and F1-score metrics are lower compared to other amino acids. From a biochemical perspective, D, E, N, and Q are all highly hydrophilic, and they share a similar degree of hydrophilicity. D and E are both acidic residues with highly similar negative charges at physiological pH, while N and Q are both uncharged. These similarities imply that D, E, N, and Q tend to exist in comparable biochemical environments within the protein, which contributes to the challenge of distinguishing them accurately. Structurally, these residues are frequently located in surface-exposed and flexible regions, such as loops, which are less ordered compared to core regions. These regions' inherent flexibility and variability introduce challenges for the model in predicting based on the structure features. This finding also explains the previous results that the model has inferior performance in surface regions compared to core regions.

3.4 Ablation Study and Analysis

We conducted an ablation study to investigate the impact of the BC-FUSION module and different input components. All models were trained under the same settings described in Sec. 3.1. The performance results for sequence recovery on the CATH 4.2 test set are summarized in Fig. 5 (c).

BC-DESIGN leverages both biochemical features and structure features. As demonstrated in the ablation study, biochemical features have a dominant effect on predicting amino acid sequences compared to structural features. To better understand this observation, we use integrated gradients (IG) [47] to quantify the impact of hydrophobicity and charge on sequence prediction for each protein of AFDB2000. Specifically, for each biochemical feature, we calculate the IG for each point in the point cloud representing the biochemical feature distribution with respect to the predicted probabilities of the original amino acids, resulting in two IG point clouds per protein. For each protein, we then calculate the average magnitude of IG values for each biochemical feature. Finally, we illustrate the relationship between these IG values and sequence recovery, as shown in Fig. 5 (d) and (e).

Fig. 5 d and e illustrate a significantly stratified relationship between the IG scores and sequence recovery. Specifically, we observe two distinct clusters: one with low sequence recovery and low IG values (approximately below 0.03 for hydrophobicity and 0.025 for charge), indicating that the model struggles when these features have a limited impact; and another cluster with very high sequence recovery and generally higher IG values, demonstrating improved prediction accuracy as these features become more influential. To gain more accurate insights, we identified two threshold effects for these biochemical features: the **recovery threshold effect** and the **IG threshold effect**.

The **recovery threshold effect** is characterized by the absence of cases with sequence recovery between 60% and 80%. Cases with sequence recovery below 60% and those with sequence recovery above 80% show distinct relationships between IG and recovery. Specifically, for predictions with sequence recovery below 60%, the IG values are consistently low, suggesting that insufficient contributions from hydrophobicity or charge hinder accurate sequence prediction. In contrast, for predictions with sequence recovery above 80%, the IG values are generally higher, indicating that strong contributions from these features facilitate accurate prediction. Further analysis helps solidify these observations: By dividing the data into clusters based on the recovery threshold, we conducted LOWESS fitting to explore the trends. As shown in Fig. 5 (d, e), for recovery values below 0.6, there is a steep, significant positive correlation between IG and sequence recovery, indicating that increasing IG has a substantial effect on improving sequence recovery. However, for recovery values above 0.8, although there is still a positive correlation between IG and recovery, it becomes much less significant. This suggests that once a sufficient level of recovery is achieved, additional IG improvements contribute less towards enhancing recovery, indicating a point of diminishing returns.

The observation highlights that both hydrophobicity and charge play critical roles in determining accurate sequence prediction. This further supports the earlier ablation findings that biochemical features are critical to BC-DESIGN's success. Importantly, the results imply that capturing sufficient information about these key biochemical properties is crucial for achieving high sequence recovery. Additionally, the model's performance is highly predictable based on the impact of biochemical features. To intuitively illustrate this observation, we visualize two cases, each consisting of the original and predicted proteins, along with their respective IG distributions for biochemical features (Fig. 5 (f, g)). Fig. 5 (f) shows a well-predicted case (AF-A0QEZ8-F1-model_v4), achieving 98.01% sequence recovery and a Ref-TM score of 0.99. In contrast, Fig. 5 (g) represents a poorly predicted case (AF-A0A131MBU3-F1-model_v4), with only 15.34% sequence recovery and a Ref-TM score of 0.15. Comparing their biochemical feature IG distributions, it is clear that the well-predicted case exhibits IGs with consistently larger magnitudes across both biochemical features.

4 Conclusion

By representing biochemical features as continuous distributions, our approach achieves unprecedented accuracy in protein sequence design, significantly outperforming current state-of-the-art methods with an 88.37% sequence recovery rate. Our evaluation demonstrates robust performance across diverse protein characteristics. The success of BC-DESIGN not only validates the importance of continuous biochemical feature representations but also establishes a new paradigm for integrating multiple modalities of protein information.

Bibliography

[1] Zhangyang Gao, Cheng Tan, Yijie Zhang, Xingran Chen, Lirong Wu, and Stan Z Li. Proteininvbench: Benchmarking protein inverse folding on diverse tasks, models, and metrics. Advances in Neural Information Processing Systems, 36, 2024.

[2] Zhangyang Gao, Cheng Tan, Pablo Chacón, and Stan Z Li. Pifold: Toward effective and efficient protein inverse folding. arXiv preprint arXiv:2209.12643, 2022.

[3] Matthew McPartlon and Jinbo Xu. An end-to-end deep learning method for protein side-chain packing and inverse folding. Proceedings of the National Academy of Sciences, 120(23):e2216438120, 2023.

[4] Chloe Hsu, Robert Verkuil, Jason Liu, Zeming Lin, Brian Hie, Tom Sercu, Adam Lerer, and Alexander Rives. Learning inverse folding from millions of predicted structures. In International conference on machine learning, pages 8946–8970. PMLR, 2022.

[5] Natalie Maus, Yimeng Zeng, Daniel Allen Anderson, Phillip Maffettone, Aaron Solomon, Peyton Greenside, Osbert Bastani, and Jacob R Gardner. Inverse protein folding using deep bayesian optimization. arXiv preprint arXiv:2305.18089, 2023.

[6] Rohith Krishna, Jue Wang, Woody Ahern, Pascal Sturmfels, Preetham Venkatesh, Indrek Kalvet, Gyu Rie Lee, Felix S Morey-Burrows, Ivan Anishchenko, Ian R Humphreys, et al. Generalized biomolecular modeling and design with rosettafold all-atom. Science, 384(6693):eadl2528, 2024.

[7] Bowen Jing, Stephan Eismann, Pratham N Soni, and Ron O Dror. Equivariant graph neural networks for 3d macromolecular structure. arXiv preprint arXiv:2106.03843, 2021.

[8] John Jumper, Richard Evans, Alexander Pritzel, Tim Green, Michael Figurnov, Olaf Ronneberger, Kathryn Tunyasuvunakool, Russ Bates, Augustin Žídek, Anna Potapenko, et al. Highly accurate protein structure prediction with alphafold. nature, 596(7873):583–589, 2021.

[9] Julia Koehler Leman, Pawel Szczerbiak, P Douglas Renfrew, Vladimir Gligorijevic, Daniel Berenberg, Tommi Vatanen, Bryn C Taylor, Chris Chandler, Stefan Janssen, Andras Pataki, et al. Sequence-structure-function relationships in the microbial protein universe. Nature communications, 14(1):2351, 2023.

[10] Varun R Shanker, Theodora UJ Bruun, Brian L Hie, and Peter S Kim. Inverse folding of protein complexes with a structure-informed language model enables unsupervised antibody evolution. bioRxiv, 2023.

[11] Frédéric A Dreyer, Daniel Cutting, Constantin Schneider, Henry Kenlay, and Charlotte M Deane. Inverse folding for antibody sequence design using deep learning. arXiv preprint arXiv:2310.19513, 2023.

[12] Daniel Cutting, Frédéric A Dreyer, David Errington, Constantin Schneider, and Charlotte M Deane. De novo antibody design with se (3) diffusion. arXiv preprint arXiv:2405.07622, 2024.

[13] Pascal Notin, Nathan Rollins, Yarin Gal, Chris Sander, and Debora Marks. Machine learning for functional protein design. Nature biotechnology, 42(2):216–228, 2024.

[14] Zhixiu Li, Yuedong Yang, Eshel Faraggi, Jian Zhan, and Yaoqi Zhou. Direct prediction of profiles of sequences compatible with a protein structure by neural networks with fragment-based local and energy-based nonlocal profiles. Proteins: Structure, Function, and Bioinformatics, 82(10):2565–2573, 2014.

[15] James O'Connell, Zhixiu Li, Jack Hanson, Rhys Heffernan, James Lyons, Kuldip Paliwal, Abdollah Dehzangi, Yuedong Yang, and Yaoqi Zhou. Spin2: Predicting sequence profiles from protein structures using deep neural networks. Proteins: Structure, Function, and Bioinformatics, 86(6):629–633, 2018.

[16] Yuan Zhang, Yang Chen, Chenran Wang, Chun-Chao Lo, Xiuwen Liu, Wu Wei, and Jinfeng Zhang. Prodconn-protein design using a convolutional neural network. Biophysical Journal, 118(3):43a–44a, 2020.

[17] Yifei Qi and John ZH Zhang. Densecpd: improving the accuracy of neural-network-based computational protein sequence design with densenet. Journal of chemical information and modeling, 60(3):1245–1252, 2020.

[18] Justas Dauparas, Ivan Anishchenko, Nathaniel Bennett, Hua Bai, Robert J Ragotte, Lukas F Milles, Basile IM Wicky, Alexis Courbet, Rob J de Haas, Neville Bethel, et al. Robust deep learning–based protein sequence design using proteinmpnn. Science, 378(6615):49–56, 2022.

[19] Nicolas Renaud, Cunliang Geng, Sonja Georgievska, Francesco Ambrosetti, Lars Ridder, Dario F Marzella, Manon F Réau, Alexandre MJJ Bonvin, and Li C Xue. Deeprank: a deep learning framework for data mining 3d protein-protein interfaces. Nature communications, 12(1):7068, 2021.

[20] Yipin Lei, Shuya Li, Ziyi Liu, Fangping Wan, Tingzhong Tian, Shao Li, Dan Zhao, and Jianyang Zeng. A deep-learning framework for multi-level peptide–protein interaction prediction. Nature communications, 12(1):5465, 2021.

[21] Raktim Mitra, Jinsen Li, Jared M Sagendorf, Yibei Jiang, Ari S Cohen, Tsu-Pei Chiu, Cameron J Glasscock, and Remo Rohs. Geometric deep learning of protein–dna binding specificity. Nature Methods, pages 1–10, 2024.

[22] Zhenqiao Song, Tinglin Huang, Lei Li, and Wengong Jin. Surfpro: Functional protein design based on continuous surface. In Forty-first International Conference on Machine Learning.

[23] Kihyuk Sohn. Improved deep metric learning with multi-class n-pair loss objective. Advances in neural information processing systems, 29, 2016.

[24] Christine A Orengo, Alex D Michie, Susan Jones, David T Jones, Mark B Swindells, and Janet M Thornton. Cath–a hierarchic classification of protein domain structures. Structure, 5(8):1093–1109, 1997.

[25] John Ingraham, Vikas Garg, Regina Barzilay, and Tommi Jaakkola. Generative models for graph-based protein design. Advances in neural information processing systems, 32, 2019.

[26] Bowen Jing, Stephan Eismann, Patricia Suriana, Raphael JL Townshend, and Ron Dror. Learning from protein structure with geometric vector perceptrons. arXiv preprint arXiv:2009.01411, 2020.

[27] Cheng Tan, Zhangyang Gao, Jun Xia, Bozhen Hu, and Stan Z Li. Generative de novo protein design with global context. arXiv preprint arXiv:2204.10673, 2022.

[28] Zhangyang Gao, Cheng Tan, and Stan Z Li. Alphadesign: A graph protein design method and benchmark on alphafolddb. arXiv preprint arXiv:2202.01079, 2022.

[29] Xing Zhang, Hongmei Yin, Fei Ling, Jian Zhan, and Yaoqi Zhou. Spin-cgnn: Improved fixed backbone protein design with contact map-based graph construction and contact graph neural network. PLoS Computational Biology, 19(12):e1011330, 2023.

[30] Kai Yi, Bingxin Zhou, Yiqing Shen, Pietro Liò, and Yuguang Wang. Graph denoising diffusion for inverse protein folding. Advances in Neural Information Processing Systems, 36, 2024.

[31] Jieling He, Wenxu Wu, and Xiaowo Wang. Diprot: A deep learning based interactive toolkit for efficient and effective protein design. Synthetic and Systems Biotechnology, 9(2):217–222, 2024.

[32] Hui Wang, Dong Liu, Kailong Zhao, Yajun Wang, and Guijun Zhang. Spdesign: protein sequence designer based on structural sequence profile using ultrafast shape recognition. Briefings in Bioinformatics, 25(3):bbae146, 2024.

[33] Xinyi Zhou, Guangyong Chen, Junjie Ye, Ercheng Wang, Jun Zhang, Cong Mao, Zhanwei Li, Jianye Hao, Xingxu Huang, Jin Tang, et al. Prorefiner: an entropy-based refining strategy for inverse protein folding with global graph attention. Nature Communications, 14(1):7434, 2023.

[34] Zaixiang Zheng, Yifan Deng, Dongyu Xue, Yi Zhou, Fei Ye, and Quanquan Gu. Structure-informed language models are protein designers. In International conference on machine learning, pages 42317–42338. PMLR, 2023.

[35] Zhangyang Gao, Cheng Tan, and Stan Z Li. Knowledge-design: Pushing the limit of protein design via knowledge refinement. arXiv preprint arXiv:2305.15151, 2023.

[36] Jiangbin Zheng and Stan Z Li. Progressive multi-modality learning for inverse protein folding. In 2024 IEEE International Conference on Multimedia and Expo (ICME), pages 1–6. IEEE, 2024.

[37] Weian Mao, Muzhi Zhu, Zheng Sun, Shuaike Shen, Lin Yuanbo Wu, Hao Chen, and Chunhua Shen. De novo protein design using geometric vector field networks. arXiv preprint arXiv:2310.11802, 2023.

[38] Robert Verkuil, Ori Kabeli, Yilun Du, Basile IM Wicky, Lukas F Milles, Justas Dauparas, David Baker, Sergey Ovchinnikov, Tom Sercu, and Alexander Rives. Language models generalize beyond natural proteins. BioRxiv, pages 2022–12, 2022.

[39] Jue Wang, Sidney Lisanza, David Juergens, Doug Tischer, Joseph L Watson, Karla M Castro, Robert Ragotte, Amijai Saragovi, Lukas F Milles, Minkyung Baek, et al. Scaffolding protein functional sites using deep learning. Science, 377(6604):387–394, 2022.

[40] Zeming Lin, Halil Akin, Roshan Rao, Brian Hie, Zhongkai Zhu, Wenting Lu, Allan dos Santos Costa, Maryam Fazel-Zarandi, Tom Sercu, Sal Candido, et al. Language models of protein sequences at the scale of evolution enable accurate structure prediction. BioRxiv, 2022:500902, 2022.

[41] Patrick Löffler, Samuel Schmitz, Enrico Hupfeld, Reinhard Sterner, and Rainer Merkl. Rosetta: Msf: a modular framework for multi-state computational protein design. PLoS computational biology, 13(6):e1005600, 2017.

[42] Steven Henikoff and Jorja G Henikoff. Amino acid substitution matrices from protein blocks. Proceedings of the National Academy of Sciences, 89(22):10915–10919, 1992.

[43] Zeming Lin, Halil Akin, Roshan Rao, Brian Hie, Zhongkai Zhu, Wenting Lu, Nikita Smetanin, Robert Verkuil, Ori Kabeli, Yaniv Shmueli, et al. Evolutionary-scale prediction of atomic-level protein structure with a language model. Science, 379(6637):1123–1130, 2023.

[44] Chuanrui Wang, Bozitao Zhong, Zuobai Zhang, Narendra Chaudhary, Sanchit Misra, and Jian Tang. Pdb-struct: A comprehensive benchmark for structure-based protein design. arXiv preprint arXiv:2312.00080, 2023.

[45] Ruidong Wu, Fan Ding, Rui Wang, Rui Shen, Xiwen Zhang, Shitong Luo, Chenpeng Su, Zuofan Wu, Qi Xie, Bonnie Berger, et al. High-resolution de novo structure prediction from primary sequence. BioRxiv, pages 2022–07, 2022.

[46] Mihaly Varadi, Damian Bertoni, Paulyna Magana, Urmila Paramval, Ivanna Pidruchna, Malarvizhi Radhakrishnan, Maxim Tsenkov, Sreenath Nair, Milot Mirdita, Jingi Yeo, et al. Alphafold protein structure database in 2024: providing structure coverage for over 214 million protein sequences. Nucleic acids research, 52(D1):D368–D375, 2024.

[47] Mukund Sundararajan, Ankur Taly, and Qiqi Yan. Axiomatic attribution for deep networks. In International conference on machine learning, pages 3319–3328. PMLR, 2017.

A Biochemical Features

Here, we present the biochemical properties of all 20 standard amino acids, specifically their hydrophobicity and charge values. These values form the foundation for generating biochemical feature distributions in our work. The hydrophobicity values are derived from the Kyte-Doolittle Hydrophobicity Scale obtained from the ImMunoGeneTics information system.

In this table, hydrophobicity values range from -4.5 to 4.5, where: Positive values indicate hydrophobic amino acids (e.g., Isoleucine: 4.5, Valine: 4.2, Leucine: 3.8). Negative values indicate hydrophilic amino acids (e.g., Arginine: -4.5, Lysine: -3.9).

The charge values are simplified to represent the ionic state at physiological pH: +1 for positively charged amino acids (Lysine, Arginine). -1 for negatively charged amino acids (Aspartic acid, Glutamic acid) 0 for neutral amino acids. 0.1 for Histidine due to its special properties at physiological pH.

These biochemical features play a crucial role in our BC-DESIGN framework, as they help capture the physical and chemical properties that determine protein folding and stability. By incorporating both hydrophobicity and charge distributions, our model can better understand and predict the amino acid sequences that would adopt a desired protein structure.

Amino Acid	**Hydrophobicity**	**Charge**
I (Isoleucine)	4.5	0
V (Valine)	4.2	0
L (Leucine)	3.8	0
F (Phenylalanine)	2.8	0
C (Cysteine)	2.5	0
M (Methionine)	1.9	0
A (Alanine)	1.8	0
W (Tryptophan)	-0.9	0
G (Glycine)	-0.4	0
T (Threonine)	-0.7	0
S (Serine)	-0.8	0
Y (Tyrosine)	-1.3	0
P (Proline)	-1.6	0
H (Histidine)	-3.2	0.1
N (Asparagine)	-3.5	0
D (Aspartic acid)	-3.5	-1
Q (Glutamine)	-3.5	0
E (Glutamic acid)	-3.5	-1
K (Lysine)	-3.9	1
R (Arginine)	-4.5	1

Table 1: Hydrophobicity and charge values of amino acids

B Fairness of Biochemical Feature Introduction

The introduction of biochemical features in our model raises potential concerns about information leakage, particularly whether these features could unfairly encode sequence information. Here we demonstrate the fairness of our approach from two key aspects:

B.1 Universality of Biochemical Properties

The biochemical features (hydrophobicity and charge) assigned to point clouds represent universal physicochemical properties of proteins rather than sequence-specific information. Multiple amino acids share similar biochemical characteristics, making it impossible to uniquely determine residue types based on these features alone:

1. Hydrophobicity groups:
   - Highly hydrophobic: Isoleucine (4.5), Valine (4.2), Leucine (3.8)
   - Moderately hydrophobic: Phenylalanine (2.8), Cysteine (2.5)
   - Highly hydrophilic: Lysine (-3.9), Arginine (-4.5)
   - Moderately hydrophilic: Asparagine (-3.5), Glutamine (-3.5), Aspartic acid (-3.5), Glutamic acid (-3.5)
2. Charge groups:
   - Positively charged ($+1$): Lysine, Arginine
   - Negatively charged (-1): Aspartic acid, Glutamic acid
   - Neutral (0): Majority of amino acids

This degeneracy in biochemical properties means that multiple amino acid sequences could potentially satisfy the same biochemical feature distribution, making it impossible to reconstruct the original sequence solely from these features.

B.2 Randomness in Point Cloud Generation

The point cloud generation process incorporates several layers of randomness that further prevent sequence information leakage:

1. Spatial sampling:
   - Points are uniformly sampled within the bounding box
   - The final point cloud is randomly subsampled to 5000 points
2. Feature assignment: The continuous and random nature of the point cloud representation blurs discrete residue positions

These characteristics ensure that while the biochemical features provide valuable information about the chemical environment necessary for protein folding, they do not encode or leak sequence information that would make the inverse folding task trivial. This maintains the fundamental challenge of the task while enriching the model's understanding of the physicochemical constraints that guide protein folding.

C Implementation Details

Our BC-Design architecture employs a hidden size of 256, featuring a STRUCT-ENCODER with 3 Structure Graph Transformer Layers and 8 attention heads, complemented by a BC-Encoder with 4 hierarchical levels and a 4-head MHA block. The fusion component incorporates 3 BC-Fusion blocks, each equipped with 8 attention heads. We train the model for 20 epochs on NVIDIA A100 GPUs using the AdamW optimizer and OneCycle learning rate scheduler, with a batch size of 4 and a learning rate of 0.0002. The model training requires a total of 6h 19m for 20 epochs on 4 NVIDIA A100 GPUs. During inference, evaluated on a single A100 GPU with a batch size of 1, the model achieves an average inference time of 0.0651 seconds per batch, with a standard deviation of 0.1450 seconds.

D A Linear Regression Analysis (Fig. 5 d & e)

As a further verification, we implement linear regression analysis to complement these findings by providing a quantitative perspective. For the high recovery group, the regression equation for hydrophobicity is $Recovery = 0.84IG + 0.91$, with an R^2 value (coefficient of determination) of 0.046, and a p-value of 3.67×10^{-16} for the slope, indicating a statistically significant but weak correlation. Similarly, for charge, the R^2 value is 0.014, with a p-value of 8.41×10^{-6} for the slope, also suggesting a weak relationship. These findings highlight the limited impact of biochemical features when recovery is already high. Conversely, for the low recovery group, the regression for hydrophobicity yields $Recovery = 17.37IG - 0.09$, with an R^2 value of 0.333 and a p-value of 1.55×10^{-52} for the slope, indicating a strong and significant correlation. For charge, the equation is $Recovery = 21.51IG - 0.06$, with an R^2 value of 0.311 and a p-value for the slope of 1.94×10^{-48}. These results demonstrate a much steeper and stronger relationship between IG and sequence recovery for the low-recovery group, with high statistical significance. This reinforces the notion that IG has a critical influence when sequence recovery is low, but its influence diminishes once a certain recovery threshold is crossed. Together, these results emphasize the importance of biochemical feature contributions in determining sequence recovery. For proteins in the low recovery cluster, increasing IG can substantially enhance sequence prediction accuracy. However, for proteins already achieving high recovery, further increasing IG yields only marginal benefits. This differential impact underlines the importance of focusing efforts on enhancing learning about biochemical feature distribution specifically for challenging cases, where the model has the most to gain from improved biochemical feature contributions.

E Discussion on IG results

The **IG threshold effect** shows that when IG values are below a certain threshold (approximately 0.02 for hydrophobicity and 0.013 for charge), sequence recovery consistently remains below 60%. Conversely, when IG values are above a higher threshold (approximately 0.032 for hydrophobicity and 0.025 for charge), sequence recovery is consistently very high (above 80%). However, there is also an intermediate region between these two thresholds where sequence recovery can vary significantly, indicating a transition zone with more variability in prediction accuracy. This suggests that beyond a certain level of contribution from hydrophobicity or charge, the additional impact becomes less pronounced for enhancing sequence recovery. The regions with low IG values correspond to more challenging proteins where biochemical features may not be well-defined or may exhibit higher flexibility, making them harder for the model to predict accurately.

F Discussion on Ablation Study

Key findings include: (i) Each input component and the BC-FUSION module makes significant contributions to the model's overall prediction accuracy. (ii) Biochemical features have the greatest effect on performance, resulting in an approximate **54%** gain. (iii) The BC-GRAPH-ablated model, which replaces the BC-FUSION module with a general Transformer Decoder, has approximately **48%** lower sequence recovery compared to the full model, despite using essentially the same input data. This reinforces the value of the dedicated fusion module for effectively integrating multi-modal inputs. (iv) In contrast, the sequence recovery improvement (**4.57%**) due to the addition of structure features is relatively minor but still significant. (v) Notably, combining features, such as edge and node features, provides greater performance improvements than the sum of their individual effects, highlighting the synergistic benefits of

feature interactions. Overall, these results underscore the critical role of each input component and the effectiveness of the STRUCT-ENCODER, BE-Encoder, and BC-FUSION module in leveraging structural and biochemical features, and, most importantly, combining them to achieve substantial performance gains.