BC-DESIGN: A Biochemistry-Aware Framework for Highly Accurate Inverse Protein Folding

Xiangru Tang [*1] **Xinwu Ye** [*1] **Fang Wu** [*2] **Daniel Shao** [3] **Dong Xu** [4] **Mark Gerstein** [1 3 5 6 7]

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, they typically represent biochemical properties as discrete features associated with individual residues. Here, we present BC-DESIGN, a framework that represents biochemical properties as continuous distributions across protein surfaces and interiors. Through contrastive learning, our model learns to encode essential biochemical information within structure embeddings, enabling sequence prediction using only structural input during inference—maintaining compatibility with real-world applications while leveraging biochemical awareness. BC-DESIGN achieves 88% sequence recovery versus state-of-the-art methods' 67% (a 21% absolute improvement) and reduces perplexity from 2.4 to 1.5 (39.5% relative improvement) on the CATH 4.2 benchmark. Notably, our model exhibits robust generalization across diverse protein characteristics, performing consistently well on proteins of varying sizes (50-500 residues), structural complexity (measured by contact order), and all major CATH fold classes. Through ablation studies, we demonstrate the complementary contributions of structural and biochemical information to this performance. Overall, BC-DESIGN establishes a new paradigm for integrating multimodal protein information, opening new avenues for computational protein engineering and drug discovery.

[*]Equal contribution [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]Department of EECS, Christopher S. Bond Life Sciences Center, University of Missouri, US [5]Program in Computational Biology & Bioinformatics, Yale University, US [6]Department of Molecular Biophysics & Biochemistry, Yale University, US [7]Department of Statistics & Data Science, Yale University, US. Correspondence to: Mark Gerstein <pi@gersteinlab.org>.

Proceedings of the Workshop on Generative AI for Biology at the 42nd International Conference on Machine Learning, Vancouver, Canada. PMLR 267, 2025.

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 (Gao et al., 2024; 2022a; McPartlon & Xu, 2023; Hsu et al., 2022; Maus et al., 2023; Krishna et al., 2024; Jing et al., 2021). This inverse problem is crucial for advancing our understanding of protein structure-function relationships and holds immense potential for therapeutic development (Jumper et al., 2021; Koehler Leman et al., 2023). 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 (Shanker et al., 2023; Dreyer et al., 2023; Cutting et al., 2024; Notin et al., 2024). 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 (Li et al., 2014; O'Connell et al., 2018), which integrate basic structural features like torsion angles and backbone angles. Subsequently, CNN-based models, including ProDCoNN and DenseCPD (Zhang et al., 2020; Qi & Zhang, 2020), 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 (Dauparas et al., 2022), PiFold (Gao et al., 2022a), and ESM-IF1 (Hsu et al., 2022) achieving remarkable results. While these structure-based approaches have shown progress, they do not explicitly account for the biochemical properties that fundamentally govern protein folding and stability. Previous attempts to incorporate

such properties in other protein-related tasks (Renaud et al., 2021; Lei et al., 2021; Mitra et al., 2024) have typically represented these features as discrete properties at the atomic or residue level. These discrete 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, existing methods lack effective ways to represent biochemical features as continuous distributions across protein surfaces, limiting their ability to capture the spatial arrangement of these critical features.

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 develop a methodology that effectively utilizes biochemical information during training while maintaining compatibility with real-world applications that require structure-only inputs*. This second challenge is particularly important for ensuring fair comparisons with existing methods and practical deployment in protein design workflows.

To address these challenges, we present BC-DESIGN, a biochemistry-aware framework for highly accurate inverse protein folding. During training, our approach represents biochemical properties as continuous distributions throughout the protein structure using constructed point clouds that sample both protein surfaces and internal spaces. A key innovation in our methodology is the use of contrastive learning to develop biochemistry-aware structural embeddings. By employing global and local contrastive learning between structural and biochemical features during training, our model learns to encode essential biochemical information within the structural representations themselves.

This contrastive learning strategy creates a critical bridge between the training and inference phases of our model. During training, the model leverages both structural information and biochemical feature distributions to learn rich, biochemistry-aware embeddings. However, at inference time—when designing sequences for novel protein structures—our model requires only the backbone structure as input, consistent with other structure-only methods and real-world protein design workflows. The structure embeddings, having implicitly captured biochemical information during training, can effectively guide sequence prediction without requiring explicit biochemical features as input. This approach offers the best of both worlds: it harnesses the power of biochemical awareness during training while maintaining the practical simplicity of structure-only input during application. In parallel to this contrastive learning framework, a STRUCT-ENCODER processes residue-level structural information through a hierarchical graph transformer, and a BC-FUSION module enables the framework to make biochemically informed sequence predictions.

The effectiveness of our biochemistry-aware design framework is validated through experiments on the CATH 4.2 dataset (Dawson et al., 2017). With our approach of incorporating biochemical features derived from protein structures, BC-DESIGN achieves high sequence prediction accuracy, with sequence recovery (the percentage of correctly predicted amino acids) of 88.37% and perplexity (a measure of prediction uncertainty, with lower values indicating more confident predictions) of 1.47. While this represents a substantial improvement over structure-only methods (such as SPDesign with 67.05% sequence recovery and 2.43 perplexity), we note that this performance gain stems from our novel incorporation of biochemical property distributions in the prediction process. Through detailed analysis, we demonstrate that this biochemistry-aware approach enables consistently accurate sequence prediction across proteins of different sizes (50-500 residues) and various CATH fold classes, highlighting the importance of biochemical context in protein sequence design.

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.

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 (Song et al.), which demonstrated these two properties to be the most informative for protein sequence design. Our feature construction process consists of four key stages, which are detailed in Sec. B in the appendix.

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. Specifically, 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 special aggregator nodes. The augmented node set is defined as $\mathcal{V}' = \mathcal{V} \cup \mathcal{V}_l \cup \{\mu\}$, where $\mathcal{V}_l = \{\nu_1, \nu_2, \ldots, \nu_{|\mathcal{V}_l|}\}$ represents a set of local aggregator nodes that summarize structural information from specific regions of the protein, and μ is a global aggregator node that captures protein-wide features. In our work, we use $|\mathcal{V}_l| = 8$ local aggregator nodes.

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}. This creates local receptive fields where each aggregator ν_i collects and summarizes information from a specific spatial region of the protein. Meanwhile, the global aggregator μ is connected to all nodes in $\mathcal{V}$ to capture protein-level features without spatial restrictions. 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: $H^{l+1}_{\mathcal{V}'} = \text{StrucGraphTrans}(H^l_{\mathcal{V}'}, W_{\mathcal{E}'}, GPE)$. 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

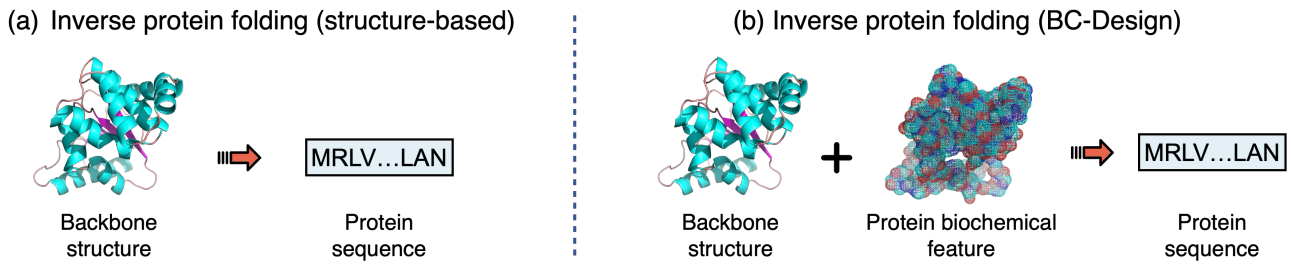

Figure 1. Comparison of traditional IPF and our biochemistry-aware approach. (a) Traditional IPF workflow inputs only the protein backbone structure and predicts the amino acid sequence. (b) During training, our approach augments the input with biochemical feature distributions (hydrophobicity and charge) alongside the backbone structure, enabling the model to learn relationships between structure and biochemical properties. Through contrastive learning, our model develops biochemistry-aware structural embeddings that capture essential physicochemical information. During inference, the model can perform sequence design using only structural inputs while benefiting from biochemical awareness, making it comparable to other structure-only methods.

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. This contrastive learning framework serves a crucial purpose beyond merely enhancing model performance - it enables our model to learn biochemistry-aware structural representations that can function independently during inference.

Primary Sequence Prediction Objective The 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 num-

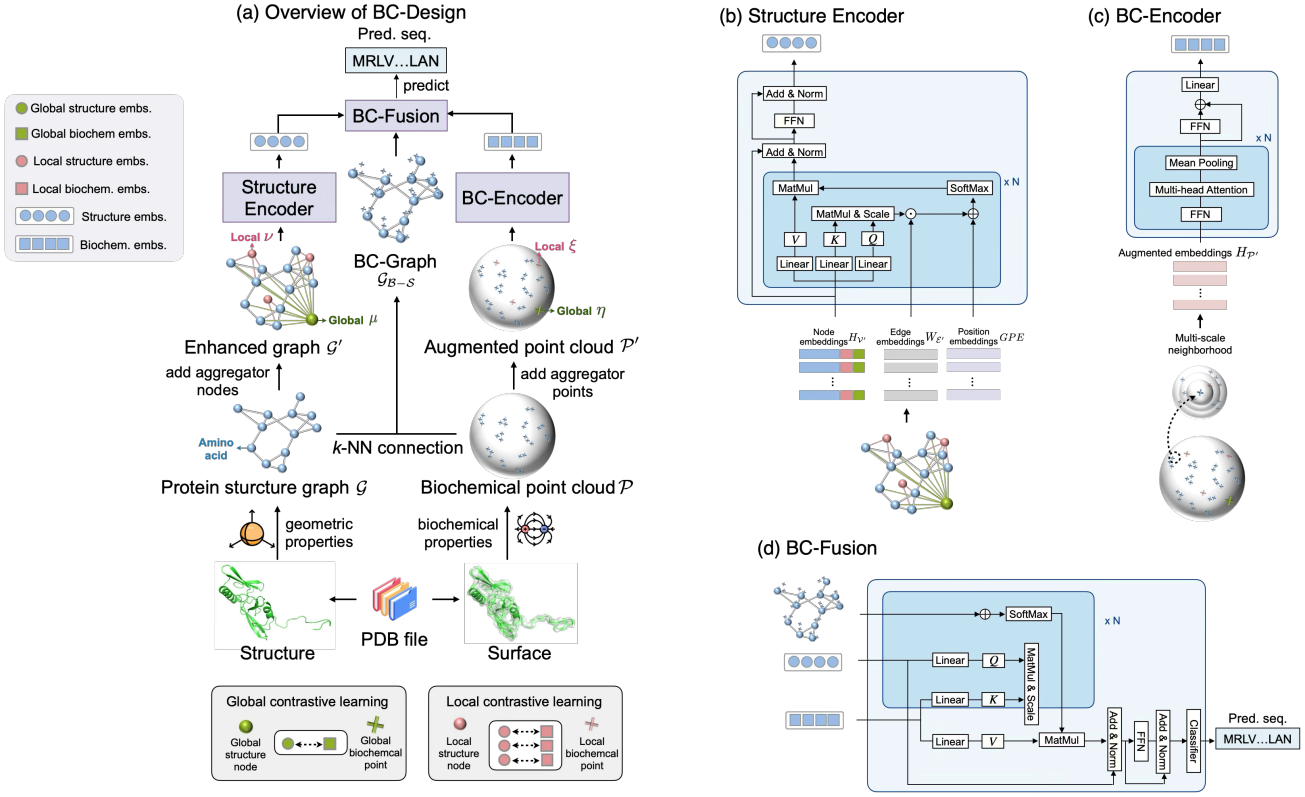

Figure 2. Overview of our proposed BC-DESIGN framework. **(a)** Architecture and workflow: The framework takes protein backbone structures (PDB files) as input and predicts amino acid sequences (MRLV...LAN) as output. The structure is represented as a graph $\mathcal{G}$ and enhanced with aggregator nodes (green: global structure μ, pink: local structure ν) to form $\mathcal{G}'$. Similarly, biochemical features (hydrophobicity and charge) are represented as a point cloud $\mathcal{P}$ augmented with aggregator points (green: global biochemical η, pink: local biochemical ξ) to form $\mathcal{P}'$. The STRUCT-ENCODER processes $\mathcal{G}'$ to produce structure embeddings (blue circles), while the BC-ENCODER processes $\mathcal{P}'$ to generate biochemical embeddings (blue squares). These embeddings are integrated through BC-FUSION using the bipartite BC-GRAPH ($\mathcal{G}_{B-S}$) to predict sequences. During training, contrastive learning aligns global structure embeddings with global biochemical embeddings (left box), and local structure embeddings with local biochemical embeddings (right box), enabling the model to function with only structural input during inference. **(b)** The STRUCT-ENCODER utilizes Structure Graph Transformers with attention mechanisms to process node embeddings (H'_V), edge embeddings (W'_E), and positional embeddings (GPE). **(c)** The BC-ENCODER employs multi-scale neighborhood sampling and multi-head attention to capture biochemical property distributions, producing augmented embeddings $H_{P'}$. **(d)** BC-FUSION combines structure and biochemical embeddings through masked Transformer decoder layers guided by the BC-GRAPH, ultimately predicting the amino acid sequence through feedforward networks and a softmax layer.

ber 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 (Sohn, 2016). $\tilde{\mathbf{h}}_\mu$ 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:

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

Combined Loss Function 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.

Training-Inference Consistency for Real-World Applications By teaching structural embeddings to implicitly encode biochemical information through contrastive learning, our approach creates a bridge between training and inference that addresses real-world application needs. During training, both structural and biochemical features are utilized to establish these relationships. However, at inference time—particularly for novel protein design scenarios where no native sequence exists—only the protein backbone structure is required as input.

The structure embeddings, having learned to capture essential biochemical information during training, can effectively guide sequence prediction without explicit biochemical features. This approach creates a fair comparison with existing structure-only methods, as our model operates under the same constraints during evaluation and practical application. When evaluating on validation and test sets, we follow this inference protocol, using only structural information as input without biochemical features, ensuring methodological consistency with other inverse folding techniques.

This strategy aligns with real-world protein design workflows where designers specify a target structure and seek compatible sequences. By encoding biochemical awareness within structural representations themselves, our model effectively utilizes the physicochemical context that governs protein folding while maintaining the practical simplicity of structure-only input during application.

3. Results: 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 (Orengo et al., 1997). 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 (Ingraham et al., 2019), 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 (Jing et al., 2020), StructGNN (Jing et al., 2020), GraphTrans (Ingraham et al., 2019), GCA (Tan et al., 2022)), recent advanced architectures (ProteinMPNN (Dauparas et al., 2022), PiFold (Gao et al., 2022a), AlphaDesign (Gao et al., 2022b), SPIN-CGNN (Zhang et al., 2023), GRADE-IF (Yi et al., 2024), DIPRoT (He et al., 2024), SPDesign (Wang et al., 2024), ProRefiner+ESM-IF1 (Zhou et al., 2023)), and language model-based methods (ESM-IF1 (Hsu et al., 2022), LM-Design (Zheng et al., 2023)). We also compare with methods incorporating external knowledge or multi-modal learning (Knowledge-Design (Gao et al., 2023), MMDesign (Zheng & Li, 2024), VFN-IFE (Mao et al., 2023)). For structure-level assessment, we evaluate against leading approaches, including ESM-Design (Verkuil et al., 2022), AlphaFold-Design (AF-Design) (Wang et al., 2022), PiFold, GraphTrans, GVP, ByProt (Lin et al., 2022), 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 (Dawson et al., 2017): sequence recovery, perplexity, and native sequence similarity recovery (nssr) (Löffler et al., 2017). Sequence recovery quantifies the model's accuracy in reproducing native amino acid sequences by calculating the percentage of positions where the predicted amino acid matches the native one. Perplexity, defined as $2^{-\frac{1}{L}\sum_{i=1}^{L}\log_2 p(a_i^*)}$ where $p(a_i^*)$ is the predicted probability of the native amino acid a_i^* at position i in a sequence of length L, measures the model's uncertainty in its predictions. Intuitively, perplexity represents the effective number of amino acids the model considers at each position—a perplexity of 2.0 indicates that the model is as uncertain as if it were randomly choosing between two equally likely amino acids at each position. The ideal perplexity is 1.0 (perfect certainty), while higher values (up to 20 for completely random predictions across

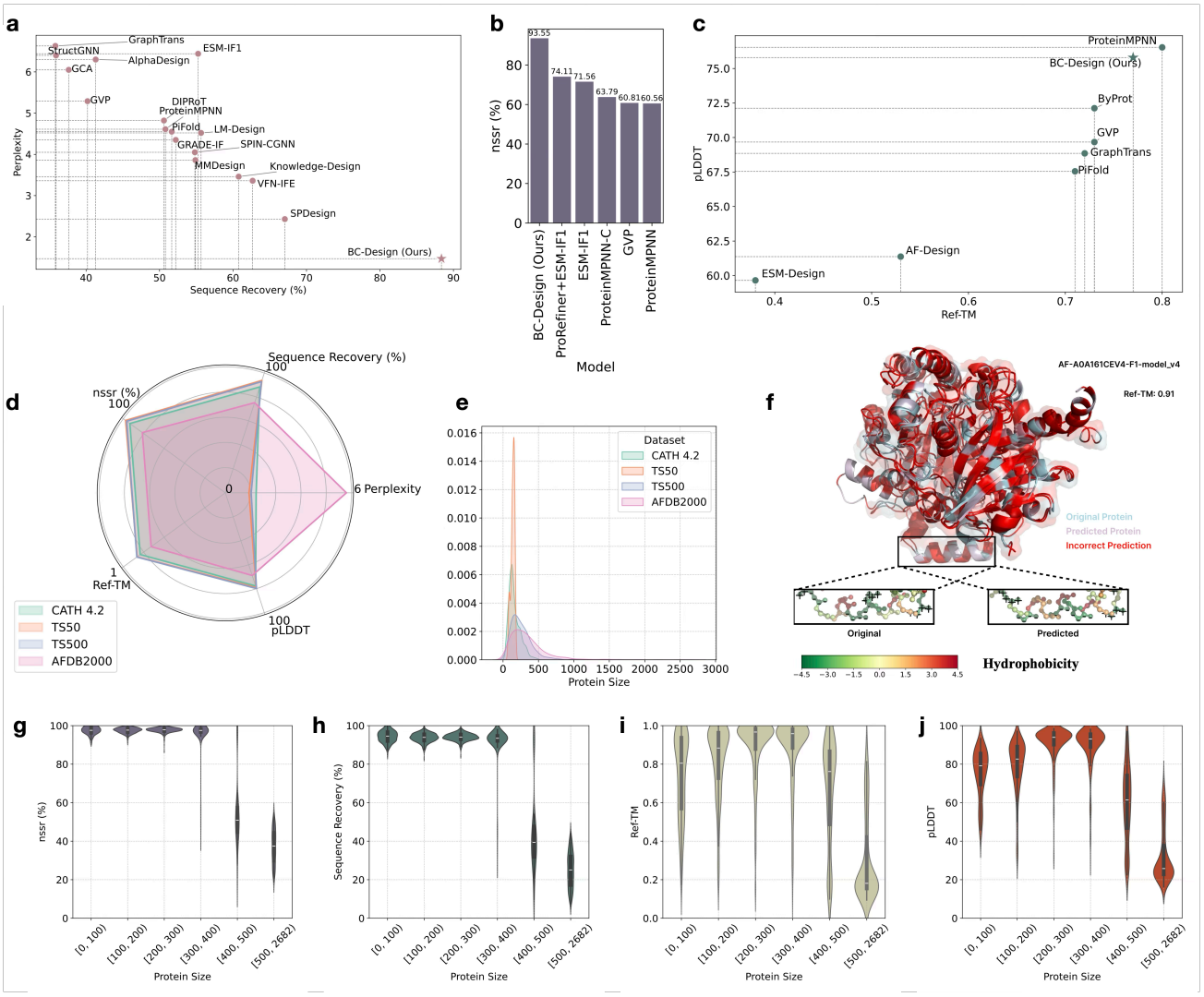

Figure 3. **(a)** Sequence recovery and perplexity of predictions by BC-DESIGN and baseline models on the CATH 4.2 test set. This plot demonstrates BC-DESIGN's superior performance with 88% sequence recovery versus the next best model (SPDesign) at 67%. An inverse correlation between sequence recovery and perplexity is evident across all models, with BC-DESIGN achieving optimal values in both metrics. **(b)** Nssr of predicted sequences by BC-DESIGN and baseline models on the CATH 4.2 test set. BC-DESIGN achieves 93.55% nssr, substantially outperforming ProteinMPNN-C (71.56%), ESM-IF1 (74.11%), and ProRefiner+ESM-IF1 (71.80%). This indicates BC-DESIGN not only recovers exact amino acids but also produces biochemically similar substitutions when needed, suggesting better functional preservation. **(c)** Ref-TM and pLDDT metrics for predicted structures (folded by ESMFold) on the filtered CATH 4.2 test set. BC-DESIGN maintains high structural quality comparable to ProteinMPNN despite having significantly better sequence accuracy. Models like ESM-Design and AF-Design show notably poorer structural metrics, highlighting BC-DESIGN's advantage in preserving structural integrity while improving sequence accuracy. **(d)** Performance metrics across multiple datasets. BC-DESIGN shows consistent results on CATH 4.2, TS50, and TS500, demonstrating strong generalization. Performance on AFDB2000 shows a modest decrease in sequence metrics while maintaining good structural quality, highlighting the model's robustness across dataset distributions. **(e)** Protein size distribution comparison across datasets. AFDB2000 contains significantly more proteins exceeding 500 amino acids (up to 3000 residues) compared to CATH 4.2, TS50, and TS500, explaining the performance variation observed in (d) and demonstrating BC-DESIGN's ability to handle proteins outside its training distribution. **(f)** Case study of protein AF-A0A161CEV4-F1-model_v4 (622 amino acids) showing strong structural recovery (Ref-TM score of 0.91) despite exceeding typical training sizes. The hydrophobicity patterns closely match between original and predicted structures, demonstrating BC-DESIGN's ability to capture essential biochemical determinants of protein folding even for larger proteins. **(g-j)** Performance stratification by protein size in AFDB2000. Sequence metrics (g,h) remain strong for proteins ¡400 amino acids but decline for larger proteins. Structural metrics (i,j) peak for proteins in the [200-300) range before gradually decreasing. Notably, proteins in [0,100) and [400,500) ranges show similar median Ref-TM values, indicating that while exact sequence recovery becomes challenging for larger proteins, BC-DESIGN maintains structural accuracy by capturing key sequence-structure relationships.

all amino acids) indicate increasing uncertainty. Perplexity is dimensionless and serves as a more sensitive measure of prediction quality than sequence recovery alone, as it accounts for the model's confidence in its predictions even when they don't exactly match the native sequence. The nssr metric extends beyond exact matches by considering residue similarity based on BLOSUM62 substitution scores (Henikoff & Henikoff, 1992), counting positions as correctly predicted if the model generates an amino acid that is biochemically similar to the native one, thereby providing a more functionally relevant evaluation of sequence design quality. 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 (Lin et al., 2023) for structure prediction due to its comparable accuracy to AlphaFold 2 (Jumper et al., 2021) with improved computational efficiency. Following (Wang et al., 2023), 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 (Wang et al., 2023) for structural similarity to native conformations, and pLDDT (Jumper et al., 2021) 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. pLDDT (predicted Local Distance Difference Test) is a confidence score introduced in AlphaFold2 that measures the expected accuracy of predicted structural elements on a per-residue basis, ranging from 0 to 100. Higher pLDDT values indicate greater confidence in the structural prediction—scores above 70 generally correspond to high-confidence regions with reliable atomic positions, while values below 50 indicate regions of low confidence. The high average pLDDT of 75.79 achieved by our method suggests that the proteins designed by BC-DESIGN fold into well-defined structures with high confidence, indicating not just sequence recovery but functional structural formation. These results 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 (Wang et al., 2023), which demonstrates consistent performance rankings across different structure prediction models, including ESMFold, AlphaFold 2, and OmegaFold (Wu et al., 2022).

4. Conclusion

By representing biochemical properties as continuous distributions and integrating them with protein structures through contrastive learning, our approach achieves exceptional accuracy in protein sequence design, significantly outperforming current state-of-the-art methods with an 88.37% sequence recovery rate. Our comprehensive evaluation demonstrates robust performance across diverse protein characteristics and structural classes.

While our current implementation uses biochemical features derived from ground-truth sequences during training, the contrastive learning framework enables our model to function effectively at inference time using only structural inputs. This represents a significant advancement in protein design methodology: our approach harnesses biochemical awareness during training while maintaining compatibility with real-world application scenarios where only structure is available. The ability of our structural embeddings to implicitly capture biochemical information demonstrates that protein design can benefit from multi-modal learning without sacrificing practical applicability.

Future work could further extend this paradigm by exploring user-specified biochemical property distributions, completely decoupling feature generation from sequence recovery. This would enable targeted design of proteins with specific biochemical characteristics while maintaining the same backbone structure. Additionally, our model shows decreased performance for larger proteins (> 400 residues), suggesting room for improvement in modeling long-range interactions. Extending our approach to handle multi-chain protein complexes and incorporating additional biochemical properties beyond hydrophobicity and charge represent promising directions for future research.

The success of BC-Design establishes a new paradigm for integrating multiple modalities of protein information through contrastive learning. By bridging the gap between information-rich training and structure-only inference, our approach opens new avenues for computational protein engineering and drug discovery that combine deep biochemical understanding with practical design workflows.

Impact Statement

This paper presents work whose goal is to advance the field of Machine Learning. There are many potential societal consequences of our work, none which we feel must be specifically highlighted here.

References

Cutting, D., Dreyer, F. A., Errington, D., Schneider, C., and Deane, C. M. De novo antibody design with se (3) diffusion. *arXiv preprint arXiv:2405.07622*, 2024.

Dauparas, J., Anishchenko, I., Bennett, N., Bai, H., Ragotte, R. J., Milles, L. F., Wicky, B. I., Courbet, A., de Haas, R. J., Bethel, N., et al. Robust deep learning–based protein sequence design using proteinmpnn. *Science*, 378 (6615):49–56, 2022.

Dawson, N. L., Lewis, T. E., Das, S., Lees, J. G., Lee, D., Ashford, P., Orengo, C. A., and Sillitoe, I. Cath: an expanded resource to predict protein function through structure and sequence. *Nucleic acids research*, 45(D1): D289–D295, 2017.

Dreyer, F. A., Cutting, D., Schneider, C., Kenlay, H., and Deane, C. M. Inverse folding for antibody sequence design using deep learning. *arXiv preprint arXiv:2310.19513*, 2023.

Gao, Z., Tan, C., Chacón, P., and Li, S. Z. Pifold: Toward effective and efficient protein inverse folding. *arXiv preprint arXiv:2209.12643*, 2022a.

Gao, Z., Tan, C., and Li, S. Z. Alphadesign: A graph protein design method and benchmark on alphafolddb. *arXiv preprint arXiv:2202.01079*, 2022b.

Gao, Z., Tan, C., and Li, S. Z. Knowledge-design: Pushing the limit of protein design via knowledge refinement. *arXiv preprint arXiv:2305.15151*, 2023.

Gao, Z., Tan, C., Zhang, Y., Chen, X., Wu, L., and Li, S. Z. Proteininvbench: Benchmarking protein inverse folding on diverse tasks, models, and metrics. *Advances in Neural Information Processing Systems*, 36, 2024.

He, J., Wu, W., and Wang, X. Diprot: A deep learning based interactive toolkit for efficient and effective protein design. *Synthetic and Systems Biotechnology*, 9(2):217–222, 2024.

Henikoff, S. and Henikoff, J. G. Amino acid substitution matrices from protein blocks. *Proceedings of the National Academy of Sciences*, 89(22):10915–10919, 1992.

Hsu, C., Verkuil, R., Liu, J., Lin, Z., Hie, B., Sercu, T., Lerer, A., and Rives, A. Learning inverse folding from millions of predicted structures. In *International conference on machine learning*, pp. 8946–8970. PMLR, 2022.

Ingraham, J., Garg, V., Barzilay, R., and Jaakkola, T. Generative models for graph-based protein design. *Advances in neural information processing systems*, 32, 2019.

Jing, B., Eismann, S., Suriana, P., Townshend, R. J., and Dror, R. Learning from protein structure with geometric vector perceptrons. *arXiv preprint arXiv:2009.01411*, 2020.

Jing, B., Eismann, S., Soni, P. N., and Dror, R. O. Equivariant graph neural networks for 3d macromolecular structure. *arXiv preprint arXiv:2106.03843*, 2021.

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.

Koehler Leman, J., Szczerbiak, P., Renfrew, P. D., Gligorijevic, V., Berenberg, D., Vatanen, T., Taylor, B. C., Chandler, C., Janssen, S., Pataki, A., et al. Sequence-structure-function relationships in the microbial protein universe. *Nature communications*, 14(1):2351, 2023.

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):eadl2528, 2024.

Lei, Y., Li, S., Liu, Z., Wan, F., Tian, T., Li, S., Zhao, D., and Zeng, J. A deep-learning framework for multi-level peptide–protein interaction prediction. *Nature communications*, 12(1):5465, 2021.

Li, Z., Yang, Y., Faraggi, E., Zhan, J., and Zhou, Y. 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.

Lin, Z., Akin, H., Rao, R., Hie, B., Zhu, Z., Lu, W., dos Santos Costa, A., Fazel-Zarandi, M., Sercu, T., Candido, S., et al. Language models of protein sequences at the scale of evolution enable accurate structure prediction. *BioRxiv*, 2022:500902, 2022.

Lin, Z., Akin, H., Rao, R., Hie, B., Zhu, Z., Lu, W., Smetanin, N., Verkuil, R., Kabeli, O., Shmueli, Y., et al. Evolutionary-scale prediction of atomic-level protein structure with a language model. *Science*, 379(6637): 1123–1130, 2023.

Löffler, P., Schmitz, S., Hupfeld, E., Sterner, R., and Merkl, R. Rosetta: Msf: a modular framework for multi-state computational protein design. *PLoS computational biology*, 13(6):e1005600, 2017.

Mao, W., Zhu, M., Sun, Z., Shen, S., Wu, L. Y., Chen, H., and Shen, C. De novo protein design using geometric vector field networks. *arXiv preprint arXiv:2310.11802*, 2023.

Maus, N., Zeng, Y., Anderson, D. A., Maffettone, P., Solomon, A., Greenside, P., Bastani, O., and Gardner, J. R. Inverse protein folding using deep bayesian optimization. *arXiv preprint arXiv:2305.18089*, 2023.

McPartlon, M. and Xu, J. 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.

Mitra, R., Li, J., Sagendorf, J. M., Jiang, Y., Cohen, A. S., Chiu, T.-P., Glasscock, C. J., and Rohs, R. Geometric deep learning of protein–dna binding specificity. *Nature Methods*, pp. 1–10, 2024.

Notin, P., Rollins, N., Gal, Y., Sander, C., and Marks, D. Machine learning for functional protein design. *Nature biotechnology*, 42(2):216–228, 2024.

O'Connell, J., Li, Z., Hanson, J., Heffernan, R., Lyons, J., Paliwal, K., Dehzangi, A., Yang, Y., and Zhou, Y. Spin2: Predicting sequence profiles from protein structures using deep neural networks. *Proteins: Structure, Function, and Bioinformatics*, 86(6):629–633, 2018.

Orengo, C. A., Michie, A. D., Jones, S., Jones, D. T., Swindells, M. B., and Thornton, J. M. Cath–a hierarchic classification of protein domain structures. *Structure*, 5(8):1093–1109, 1997.

Qi, Y. and Zhang, J. Z. 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.

Renaud, N., Geng, C., Georgievska, S., Ambrosetti, F., Ridder, L., Marzella, D. F., Réau, M. F., Bonvin, A. M., and Xue, L. C. Deeprank: a deep learning framework for data mining 3d protein-protein interfaces. *Nature communications*, 12(1):7068, 2021.

Shanker, V. R., Bruun, T. U., Hie, B. L., and Kim, P. S. Inverse folding of protein complexes with a structure-informed language model enables unsupervised antibody evolution. *bioRxiv*, 2023.

Sohn, K. Improved deep metric learning with multi-class n-pair loss objective. *Advances in neural information processing systems*, 29, 2016.

Song, Z., Huang, T., Li, L., and Jin, W. Surfpro: Functional protein design based on continuous surface. In *Forty-first International Conference on Machine Learning*.

Sundararajan, M., Taly, A., and Yan, Q. Axiomatic attribution for deep networks. In *International conference on machine learning*, pp. 3319–3328. PMLR, 2017.

Tan, C., Gao, Z., Xia, J., Hu, B., and Li, S. Z. Generative de novo protein design with global context. *arXiv preprint arXiv:2204.10673*, 2022.

Varadi, M., Bertoni, D., Magana, P., Paramval, U., Pidruchna, I., Radhakrishnan, M., Tsenkov, M., Nair, S., Mirdita, M., Yeo, J., 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.

Verkuil, R., Kabeli, O., Du, Y., Wicky, B. I., Milles, L. F., Dauparas, J., Baker, D., Ovchinnikov, S., Sercu, T., and Rives, A. Language models generalize beyond natural proteins. *BioRxiv*, pp. 2022–12, 2022.

Wang, C., Zhong, B., Zhang, Z., Chaudhary, N., Misra, S., and Tang, J. Pdb-struct: A comprehensive benchmark for structure-based protein design. *arXiv preprint arXiv:2312.00080*, 2023.

Wang, H., Liu, D., Zhao, K., Wang, Y., and Zhang, G. Spdesign: protein sequence designer based on structural sequence profile using ultrafast shape recognition. *Briefings in Bioinformatics*, 25(3):bbae146, 2024.

Wang, J., Lisanza, S., Juergens, D., Tischer, D., Watson, J. L., Castro, K. M., Ragotte, R., Saragovi, A., Milles, L. F., Baek, M., et al. Scaffolding protein functional sites using deep learning. *Science*, 377(6604):387–394, 2022.

Wu, R., Ding, F., Wang, R., Shen, R., Zhang, X., Luo, S., Su, C., Wu, Z., Xie, Q., Berger, B., et al. High-resolution de novo structure prediction from primary sequence. *BioRxiv*, pp. 2022–07, 2022.

Yi, K., Zhou, B., Shen, Y., Liò, P., and Wang, Y. Graph denoising diffusion for inverse protein folding. *Advances in Neural Information Processing Systems*, 36, 2024.

Zhang, X., Yin, H., Ling, F., Zhan, J., and Zhou, Y. Spincgnn: Improved fixed backbone protein design with contact map-based graph construction and contact graph neural network. *PLoS Computational Biology*, 19(12): e1011330, 2023.

Zhang, Y., Chen, Y., Wang, C., Lo, C.-C., Liu, X., Wei, W., and Zhang, J. Prodconn-protein design using a convolutional neural network. *Biophysical Journal*, 118(3): 43a–44a, 2020.

Zheng, J. and Li, S. Z. Progressive multi-modality learning for inverse protein folding. In *2024 IEEE International Conference on Multimedia and Expo (ICME)*, pp. 1–6. IEEE, 2024.

Zheng, Z., Deng, Y., Xue, D., Zhou, Y., Ye, F., and Gu, Q. Structure-informed language models are protein designers. In *International conference on machine learning*, pp. 42317–42338. PMLR, 2023.

Zhou, X., Chen, G., Ye, J., Wang, E., Zhang, J., Mao, C., Li, Z., Hao, J., Huang, X., Tang, J., et al. Prorefiner: an entropy-based refining strategy for inverse protein folding with global graph attention. *Nature Communications*, 14 (1):7434, 2023.

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. Detailed Biochemical Feature Construction

(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:** Once the unified point cloud is constructed, we need to associate relevant biochemical properties with each point to create a meaningful representation of the protein's chemical environment. For each point in the unified cloud $P_i \in \mathcal{P}$, we determine its nearest residue by computing distances to all C_α atoms in the protein structure. Specifically, we identify the residue r_j whose C_α position $\mathbf{c}_j$ minimizes the Euclidean distance $\|\mathbf{x}_i - \mathbf{c}_j\|_2$ to point $\mathbf{x}_i$. We then transfer the biochemical properties of this nearest residue to the point. Each point is decorated with two key biochemical features: hydrophobicity $h(\mathbf{x}_i)$ and charge $c(\mathbf{x}_i)$ values derived from standardized amino acid property tables (detailed in Sec. A in the appendix). For example, if a point is nearest to an isoleucine residue, it would be assigned a high hydrophobicity value of 4.5 and a neutral charge value of 0, whereas a point nearest to an arginine residue would receive a highly hydrophilic value of -4.5 and a positive charge value of +1. This assignment process creates a continuous representation of biochemical properties across both the protein surface and interior, where points in spatial proximity to similar residues will exhibit similar property distributions. 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 the hydrophobicity and charge values. This approach transforms discrete residue-level biochemical properties into a continuous field distributed throughout the protein's three-dimensional structure, providing a more natural representation of the biochemical environment as it exists in the folded protein state.

C. 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 (O'Connell et al., 2018). (ii) **TS500**. Similar to TS50, the TS500 dataset (O'Connell et al., 2018) contains 500 proteins. (iii) **AFDB2000**. The AlphaFold Database (AFDB) (Varadi et al., 2024) 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.

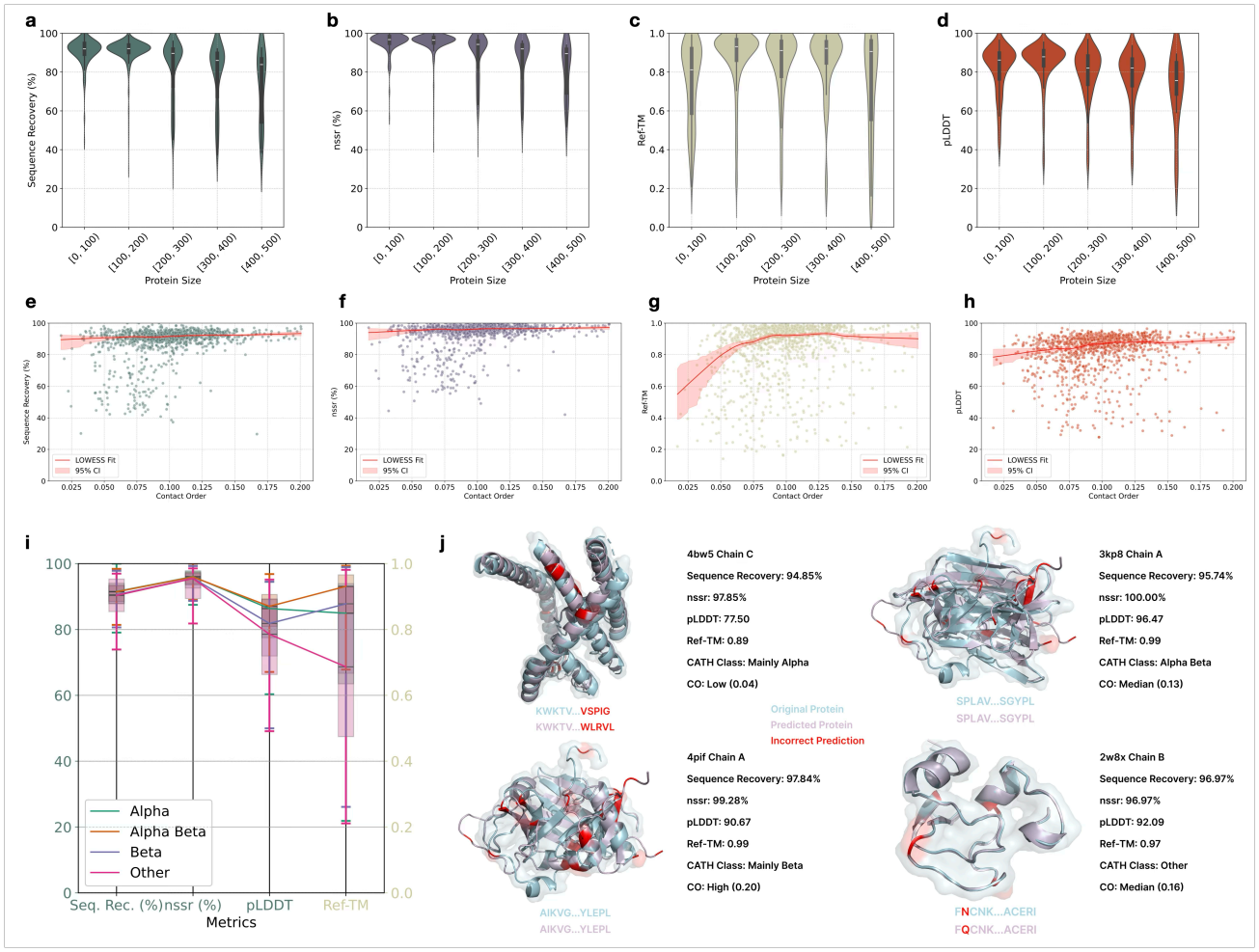

Figure 4. **(a-d)** Sequence and structure recovery of BC-DESIGN predictions for proteins of different sizes in the CATH 4.2 test set. These violin plots show excellent performance across all protein size ranges, with median sequence recovery $> 80\%$, nssr $> 85\%$, Ref-TM > 0.8, and pLDDT > 0.75. While sequence recovery slightly decreases for larger proteins, structural metrics remain strong even for the [400, 500) range, suggesting the model prioritizes maintaining backbone structure over exact sequence matching as complexity increases. **(e-h)** Relationship between recovery metrics and structural complexity (measured by contact order). Counterintuitively, all metrics improve as contact order increases, showing higher median values and reduced variability. This indicates proteins with more long-range interactions provide more favorable conditions for accurate prediction, likely because these complex topologies offer stronger biochemical and spatial constraints that guide sequence prediction. LOWESS fit lines confirm this positive correlation across all metrics. **(i)** Performance across different CATH Classes. While sequence metrics remain consistently high across all structural classes with minimal variation, structural metrics show significant differences. Alpha Beta proteins achieve the highest median Ref-TM scores, while Beta and Other classes exhibit lower structural metrics and higher variability, suggesting mixed alpha-beta elements present favorable prediction features, while beta-rich and structurally diverse proteins pose greater challenges. **(j)** Case studies of four proteins with different CATH Classes and structural complexity. Each example demonstrates robust performance across diverse structural contexts: a Mainly Alpha protein (94.85% recovery, Ref-TM 0.89, low contact order), an Alpha Beta protein (100% nssr, median contact order), a Mainly Beta protein (97.84% recovery despite high contact order), and an Other class protein (96.9% recovery, Ref-TM 0.97, median contact order). The displayed segments highlight both accurate predictions and occasional mismatches.

D. 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% 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 (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-

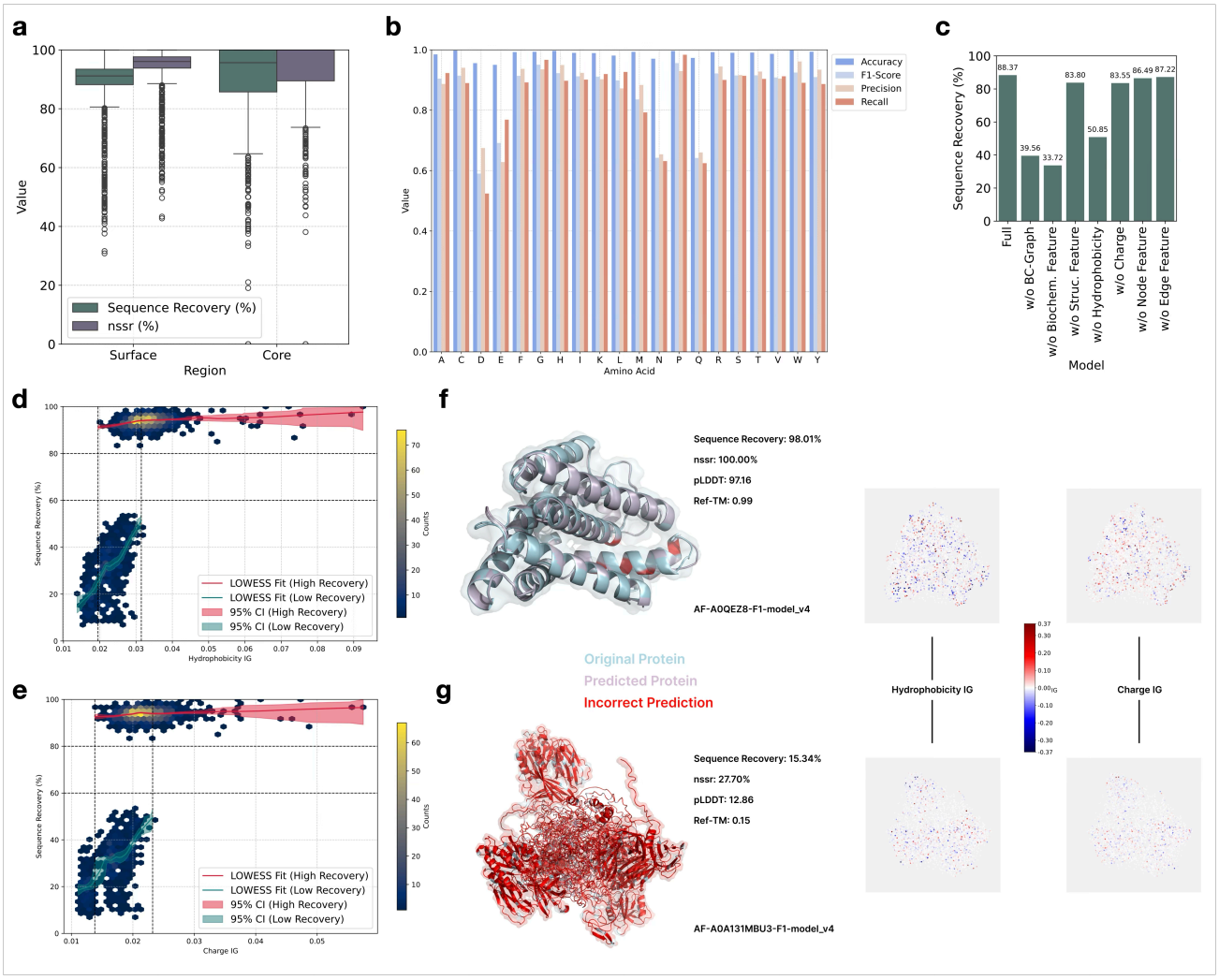

Figure 5. **(a)** Sequence recovery and nssr of BC-DESIGN predictions for surface versus core residues. Core residues show significantly higher prediction accuracy than surface residues, indicating more effective capture of conserved features critical for structural integrity in protein interiors. **(b)** Residue-level evaluation across amino acid types. Despite consistently high accuracy for all residues, negatively charged (D, E) and polar amide (N, Q) amino acids exhibit lower precision and recall, likely due to biochemical similarities and frequent occurrence in flexible surface regions. **(c)** Ablation study results quantifying each component's contribution. Biochemical features show the greatest impact (≈54% drop when removed), followed by the BC-Graph module (≈48% drop when replaced), while structure features provide a smaller but significant improvement (4.57%). **(d, e)** Relationship between sequence recovery and integrated gradient (IG) values for hydrophobicity and charge. Two distinct clusters emerge: low recovery (¡60%) with low IG values and high recovery ($> 80\%$) with higher IG values. LOWESS fits reveal steep positive correlations in the low recovery region but diminishing returns at higher IG values. **(f, g)** Contrasting case studies: a well-predicted protein (98.01% recovery, Ref-TM 0.99) versus a poorly predicted one (15.34% recovery, Ref-TM 0.15). IG distribution heatmaps demonstrate successful predictions correlate with higher magnitude biochemical feature attribution, while failed predictions show weaker attribution patterns.

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.

E. 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. The performance results for sequence recovery on the CATH 4.2 test set are summarized in Fig. 5 (c).

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.

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) (Sundararajan et al., 2017) to quantify the impact of hydrophobicity and charge on sequence prediction for each protein of AFDB2000.

Integrated gradients (IG) is an attribution method that measures feature importance by accumulating gradients along a straight-line path from a baseline (typically zero features) to the actual input. Formally, the integrated gradient for a feature i is defined as:

$$IG_i(x) = (x_i - x_i') \times \int_{\alpha=0}^{1} \frac{\partial F(x' + \alpha \times (x - x'))}{\partial x_i} d\alpha \tag{1}$$

where F is our model, x is the input, x' is the baseline, and the integral captures the gradients along the path from x' to x.

Specifically, for each biochemical feature (hydrophobicity and charge), 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 to obtain a global importance score. 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) reveal a clear pattern: proteins with higher sequence recovery rates consistently exhibit higher IG values for both hydrophobicity and charge features. This indicates that when our model successfully recovers sequences (e.g., with recovery rates above 80%), the biochemical features contribute significantly to the prediction, as evidenced by their higher IG scores. Conversely, proteins with low sequence recovery (below 60%) consistently show lower IG values, suggesting that the model fails to effectively leverage biochemical information in these cases. This stark contrast demonstrates that the effective utilization of biochemical features is a key determinant of successful sequence prediction in our model. The consistent correlation between IG values and prediction success provides strong evidence that our approach of modeling biochemical properties as continuous distributions on protein surfaces is instrumental to the model's performance.

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. This means that, in challenging cases where sequence recovery is low, biochemical features like hydrophobicity and charge play a critical role, and their enhancement can significantly boost model performance. 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.

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.

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.

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.

F. 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:

F.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.

F.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.

G. 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.

H. Future Work

Building on our success, several promising research directions could further enhance the approach:

H.1. Experimental Validation through Iterative Design Cycles

Unlike purely language model-based approaches, implementing a Trust Region-based optimization approach with experimental feedback loops could significantly improve real-world applicability. This would involve designing initial peptides, obtaining experimental effectiveness data, and refining designs through multiple iterations to achieve optimal performance against specific targets like fungi. The iterative nature of this approach allows for continuous improvement based on real experimental outcomes, potentially leading to more biologically relevant designs than those created through computational methods alone.

H.2. Design with Constraints for Functional Modification

A valuable extension would be developing the ability to modify proteins for new functions while maintaining structural integrity. This could be achieved by incorporating constraints for catalytic triads or active sites that must be preserved during the design process. Such an approach would enable the transformation of structural proteins into enzymes by ensuring that critical amino acids necessary for the desired function remain in place. Furthermore, by introducing biases based on known enzyme properties, the model could generate designs that not only satisfy structural requirements but also possess the biochemical characteristics essential for the target function. This constrained design paradigm would address real-world use cases where scientists begin with a protein of known structure but aim to engineer variants with enhanced or novel functionalities.

H.3. User-specified Biochemical Property Distributions

Currently, BC-DESIGN derives biochemical features from ground-truth sequences during evaluation, which may not fully reflect real-world design scenarios. Future work could explore methods for protein design with user-specified biochemical property distributions, effectively decoupling feature generation from sequence recovery. This advancement would enable designers to directly control the physicochemical environment of the protein, specifying desired hydrophobicity and charge distributions without relying on existing sequences. Such capability would significantly expand the creative possibilities for protein engineers, allowing them to explore novel sequence spaces that satisfy specific biochemical criteria while maintaining structural integrity.

H.4. Enhanced Modeling for Larger Proteins

The current approach shows decreased performance for proteins exceeding 400 residues, suggesting room for improvement in modeling long-range interactions. Addressing this limitation would require architectural modifications to better capture dependencies between distant residues in the protein structure. Additionally, extending the framework to handle multi-chain protein complexes would considerably broaden its application scope, enabling the design of more complex biological systems such as antibody-antigen interfaces, enzyme-substrate interactions, and protein-protein interaction networks. These enhancements would make BC-DESIGN applicable to a wider range of biologically relevant targets.

H.5. Incorporating Additional Biochemical Properties

Expanding beyond hydrophobicity and charge to include properties like hydrogen bonding potential, steric requirements, and conformational preferences could further enrich the model's understanding of physicochemical constraints. These additional properties play crucial roles in determining protein folding, stability, and function, and their inclusion would provide a more comprehensive representation of the factors governing protein behavior. By capturing these nuanced biochemical features, future versions of BC-DESIGN could achieve even greater accuracy in sequence prediction while generating proteins with more precisely tuned functional characteristics.

By addressing these areas, BC-DESIGN could evolve into an even more powerful tool for computational protein engineering and therapeutic development, bridging the gap between computational design and experimental validation through a more iterative and constraint-aware approach.

I. Comparison Between GVP-GNN and BC-DESIGN Approaches

GVP-GNN (Jing et al., 2020) and BC-DESIGN represent two significant advancements in learning from protein structure, yet they differ in several fundamental aspects. GVP-GNN primarily focuses on augmenting graph neural networks with geometric vector perceptrons to perform geometric reasoning while maintaining equivariance. It operates on both scalar and vector features, enabling direct representation of 3D information throughout graph propagation without reducing such information to scalars that may not fully capture complex geometry. The model learns to encode the 3D geometry through vector representations that transform appropriately under spatial rotations, creating a global coordinate system across the structure.

BC-DESIGN, in contrast, introduces an approach that explicitly represents biochemical properties as continuous distributions throughout the protein structure. Rather than encoding biochemical features as discrete properties of individual residues, it models hydrophobicity and charge as decorations on randomly sampled points across both exterior surfaces and internally bound regions. This provides a more natural way to capture the spatial distribution of biochemical properties as they exist in folded protein states, moving beyond the discrete residue-level representation prevalent in traditional models.

The architectural design of these two approaches also differs significantly. GVP-GNN employs relatively simple geometric vector perceptrons as its core computational unit, where vector channels directly encode geometric features. BC-DESIGN implements a more complex architecture comprising a STRUCT-ENCODER that processes residue-level structural information through a hierarchical graph transformer, a BC-ENCODER that handles biochemical features, and a BC-FUSION module that integrates structure and biochemistry through a bipartite graph structure. This multi-component design allows BC-DESIGN to process and fuse multiple information modalities.

A key innovation in BC-DESIGN not present in GVP-GNN is its use of contrastive learning to bridge training and inference phases. During training, BC-DESIGN leverages both structural information and biochemical feature distributions to learn rich, biochemistry-aware embeddings. However, at inference time, it requires only the backbone structure as input, with the structural embeddings having implicitly captured biochemical information. This creates a model that harnesses biochemical awareness during training while maintaining practical simplicity during application.

Performance-wise, BC-DESIGN demonstrates significant improvements over existing methods, including GVP-GNN, particularly in inverse protein folding tasks. BC-DESIGN achieves 88% sequence recovery compared to state-of-the-art methods' 67%, representing a 21% absolute improvement, and reduces perplexity from 2.4 to 1.47. These improvements stem from the model's ability to capture and leverage biochemical context in protein sequence design, highlighting the importance of representing biochemical properties as continuous distributions rather than discrete features.

Both approaches maintain rotation invariance in their scalar outputs and equivariance in their vector outputs with respect to 3D transformations, an essential property for learning from protein structure. However, BC-DESIGN's biochemistry-aware approach appears to provide a more comprehensive framework for capturing the physical and chemical principles that govern protein folding and stability, leading to its superior performance on inverse protein folding tasks.

J. Discussion on Atom-Based Sampling for Biochemical Feature Construction

The current BC-DESIGN approach represents biochemical features through point clouds that sample both protein surfaces and internal spaces, demonstrating significant improvements in inverse protein folding. However, the uniform sampling

strategy within a bounding box may not optimally capture the complex biochemical environment within proteins. Here, we explore the potential benefits and challenges of implementing an atom-based sampling approach as an alternative to the current methodology.

Atom-based sampling would leverage the precise atomic coordinates from the protein structure itself, rather than relying on geometric approximations. This approach could provide a more biologically accurate representation of the biochemical environment by directly sampling points based on actual atomic positions. Such sampling would inherently capture the non-uniform distribution of biochemical properties throughout the protein volume, potentially improving the model's ability to learn subtle patterns that influence amino acid preferences at specific positions. Furthermore, by correlating sampling density with functionally important regions or atoms, the model could develop enhanced sensitivity to critical areas that determine protein function.

A key advantage of atom-based sampling would be its ability to better represent the discrete nature of protein biochemistry while maintaining our continuous distribution paradigm. Different atom types contribute distinctly to the overall biochemical environment—polar atoms create hydrophilic regions, while carbon-rich areas form hydrophobic clusters. By basing our sampling on these atomic coordinates and types, we could more accurately represent the gradients and boundaries between different biochemical zones within the protein structure. This could be particularly beneficial for capturing features like binding pockets, catalytic sites, or stabilizing hydrophobic cores.

Implementation of atom-based sampling would require careful consideration of several factors. First, proteins contain varying numbers of atoms, necessitating adaptive sampling techniques to maintain consistent representation sizes across different structures. Second, a weighting scheme would need to be developed to determine sampling density around different atom types, potentially giving preference to side chain atoms that more strongly influence biochemical properties. Finally, computational efficiency must be considered, as the increased complexity of atom-based sampling could impact training and inference times.

Despite these challenges, atom-based sampling presents a promising direction for enhancing the biochemical awareness of our model. The potential improvements in capturing atomic-level biochemical environments could further refine sequence predictions, particularly for functionally specialized regions where precise biochemical conditions are essential for protein activity. Future work will explore hybrid approaches that combine the computational efficiency of uniform sampling with the biochemical accuracy of atom-based methods, potentially leading to even more accurate and biologically relevant protein designs.