RATE-CONSTRAINED GEOMETRY-AWARE PROTEIN INVERSE FOLDING WITH SYNCHRONIZED MASKING

Anonymous authors
Paper under double-blind review

ABSTRACT

Structure-conditioned protein design requires bridging discrete one-dimensional sequence representations and continuous three-dimensional geometry while preserving evolutionary and structural constraints. We propose GeoSeqAlign, a probabilistic inverse-folding framework that aligns sequence and structure through a variational information bottleneck (VIB). To prevent trivial reliance on local cross-modal correspondences, we introduce synchronized structure–sequence corruption, which masks both modalities at the same residue positions. Combined with a bottleneck that balances reconstruction fidelity against latent information capacity, this mechanism encourages the model to capture predictive sequence–structure relationships while reducing sensitivity to geometric perturbations. To integrate geometric constraints with evolutionary priors, GeoSeqAlign injects $SE(3)$-invariant pairwise distances and relative orientations as attention biases into a frozen protein language model, enabling geometry-aware sequence generation without coordinate-equivariant updates. Across CATH+ 4.3, TS50, and CASP15/TS45, GeoSeqAlign achieves the highest amino-acid recovery among the evaluated fixed-backbone baselines while tuning only 6.4% of its parameters. Controlled ablations support the complementary contributions of bottleneck regularization and synchronized corruption, with greater benefits under structural distribution shift and backbone perturbations. These results highlight information-constrained cross-modal learning as an effective approach to robust, parameter-efficient protein inverse folding. Source code is available at: https://anonymous.4open.science/r/GeoSeqAlign-17466.

1 INTRODUCTION

Proteins execute complex biological functions through an intricate interplay between discrete amino-acid sequences and folded three-dimensional structures. While large-scale protein language models (PLMs) have successfully distilled rich evolutionary regularities from unannotated sequence corpora (Rives et al., 2021; Madani et al., 2023; Ferruz et al., 2022), and state-of-the-art structure predictors have demonstrated the remarkable predictability of geometry from sequence (Jumper et al., 2021; Lin et al., 2023), protein design poses a fundamentally more constrained inverse problem. Given a target backbone geometry, the objective is to generate sequences compatible with that spatial topology while assessing their structural self-consistency.

This inverse-folding paradigm is non-trivial because sequences and structures inhabit vastly different representational spaces. A sequence is a one-dimensional symbolic string rich in evolutionary semantics, whereas a backbone is a continuous, $SE(3)$-symmetric geometric entity. Standard PLM attention, while highly expressive, natively lacks awareness of long-range 3D spatial contacts. Graph-based and coordinate-equivariant models encode geometry explicitly, while PLM-conditioned approaches additionally exploit pretrained sequence priors. Recent advancements such as ProteinMPNN(Dauparas et al., 2022), PiFold(Gao et al., 2023a), ESM-IF1(Hsu et al., 2022), and diffusion-based models (e.g., GRADE-IF(Yi et al., 2023), MapDiff(Bai et al., 2025)) have substantially elevated fixed-backbone design performance (Wang et al., 2024; Tang et al., 2024). A remaining question is how to regularize this conditioning so that performance transfers to unfamiliar or perturbed backbones. In paired denoising, independently masked modalities also permit a training-only bypass through unmasked evidence at the same residue position.

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

We study this question through an information bottleneck perspective: retain information useful for sequence prediction and auxiliary geometry reconstruction while penalizing the rate carried by a stochastic latent. Global coordinate invariance is imposed by the geometric representation; the KL penalty regularizes capacity rather than establishing invariance or identifying a unique sufficient statistic. We instantiate this framework with GeoSeqAlign and three design principles. First, synchronized corruption removes local sequence and geometric evidence at shared training positions, encouraging prediction from broader context. Second, invariant pairwise geometry is injected as scalar attention biases into a frozen PLM. Third, a VIB penalty controls latent rate. The individual ingredients build on established methods; our contribution is their integration and controlled evaluation in fixed-backbone inverse folding:

1. We combine a VIB-style rate–distortion objective with synchronized structure–sequence corruption, using a KL information-rate bound to regularize paired denoising.

2. We connect raw relative-frame descriptors to $SE(3)^N/SE(3)$ and inject invariant geometric biases through trainable adapters that condition a frozen PLM.

3. We evaluate recovery, robustness, diversity, structural overlap, and computational cost, with matched KL and independent-mask controls to quantify the effects of KL regularization and mask alignment.

2 Related Work

Protein Language Models. Transformer PLMs learn evolutionary and functional sequence regularities from large corpora (Rives et al., 2021; Madani et al., 2023; Ferruz et al., 2022). Sequence generation still requires structural validation. ESMFold showed that PLM representations support structure prediction (Lin et al., 2023), motivating explicit geometric conditioning of frozen sequence representations.

Inverse Folding. The core objective of inverse folding is mapping 3D coordinates to 1D sequences. Early approaches like Structured Transformer utilized graph-conditioned autoregression (Ingraham et al., 2019), while GVP-GNN incorporated scalar and vector features to capture structural motifs (Jing et al., 2021). ProteinMPNN remains a highly practical and robust baseline for standard fixed-backbone design (Dauparas et al., 2022). Scaling efforts, such as ESM-IF1, trained on millions of AlphaFold2-distilled structures (Hsu et al., 2022), and PiFold optimized graph architectures for one-shot efficiency (Gao et al., 2023a). More recently, diffusion-based methods like GRADE-IF and MapDiff have sought to explicitly model sequence uncertainty and increase diversity (Yi et al., 2023; Bai et al., 2025). GeoSeqAlign studies how KL regularization and synchronized corruption affect transfer and robustness in a geometry-conditioned PLM.

External Knowledge and Structure-aware Pretraining. Several state-of-the-art models inject external inductive biases to boost performance. SaProt constructs a hybrid vocabulary utilizing Foldseek structural tokens for joint pretraining (Su et al., 2024). SPDesign incorporates structural analogs and profile-based priors (Wang et al., 2024), whereas BC-Design embeds biochemical domain knowledge into the representation learning phase (Tang et al., 2024). While these approaches are highly effective, they often require auxiliary inputs beyond a bare atomic backbone. We report their evaluated fixed-backbone settings explicitly; matching targets and decoding does not equalize external pretraining resources.

Geometric Learning and Information Bottlenecks. Equivariant models such as SE(3)-Transformers and EGNNs support coordinate-valued prediction (Fuchs et al., 2020; Satorras et al., 2021). Invariant backbone features are also established in ProteinMPNN and PiFold. AlphaFold2 introduced pair representations and frame-aligned point error (FAPE), and AlphaFold3 further develops biomolecular token/pair modeling (Jumper et al., 2021; Abramson et al., 2024). We treat these geometric ingredients as prior art. Our regularization follows the information bottleneck and its variational approximation (Tishby et al., 1999; Alemi et al., 2017). Structure language models separate invariant structure encoding from latent generation (Lu et al., 2025); we use continuous invariant conditioning for fixed-backbone sequence design.

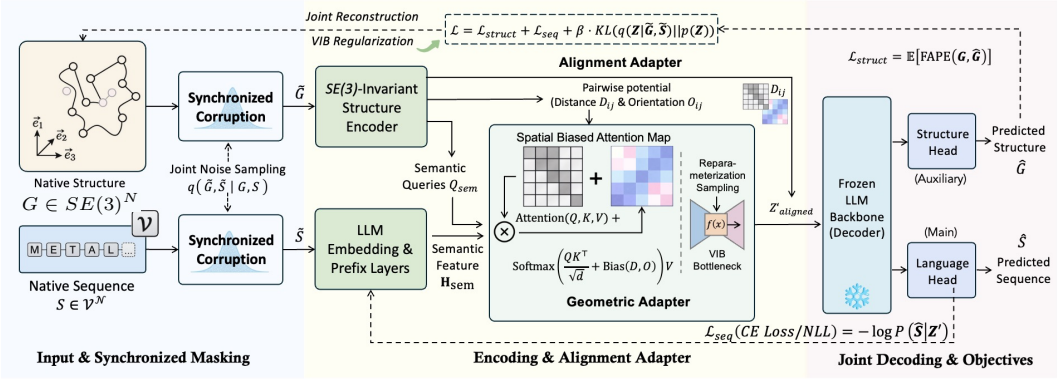


Figure 1: **Overview of GeoSeqAlign.** Synchronized corrupted views are aligned through invariant pairwise geometry, attention biases, and a stochastic VIB latent. Trainable prefixes/adapters condition a frozen PLM and an amino-acid output head. The framed-backbone quotient is $SE(3)^N/SE(3)$. Auxiliary geometry reconstruction is used only during training.

PLM-conditioned and Richer-context Design. LM-Design uses structural adapters and iterative refinement (Zheng et al., 2023); KW-Design refines predictions using pretrained knowledge (Gao et al., 2024); Bridge-IF uses structure-modulated Markov bridges (Zhu et al., 2024). We instead study a KL-regularized latent trained on synchronized corrupted views. ADFLIP uses all-atom structural contexts (Yi et al., 2025); CoFlow jointly models sequence and discrete structure tokens and also supports inverse folding (Yang et al., 2025). Faithful matched comparisons with these richer-context systems remain outside the present evaluation.

3 Method

3.1 Problem Formulation and Invariance

We consider paired data $(G, S) \sim p_{\text{data}}$, where $S = (s_1, \ldots, s_N) \in \mathcal{V}_{\text{aa}}^N$ and $\mathcal{V}_{\text{aa}}$ contains the 20 standard amino acids. A residue-indexed framed backbone is

$$G = ((x_i, R_i))_{i=1}^N, \qquad x_i \in \mathbb{R}^3, \quad R_i \in SO(3). \tag{1}$$

Writing $g_i = (R_i, x_i) \in SE(3)$ gives the ambient space $\mathcal{X} = SE(3)^N$. A single global motion $h = (R, t)$ acts diagonally: $h \cdot G = ((Rx_i + t, RR_i))_i$. We model $p(S \mid [G])$, where $[G] \in \mathcal{Q} = SE(3)^N/SE(3)$ removes global pose. Physically realizable backbones form a constrained subset of this ambient space.

3.2 Synchronized Structure–Sequence Corruption

During training, a shared stochastic state ω couples the corrupted views:

$$q(\widetilde{G}, \widetilde{S} \mid G, S) = \int q_{\text{geo}}(\widetilde{G} \mid G, \omega) q_{\text{seq}}(\widetilde{S} \mid S, \omega) p(\omega)\, d\omega. \tag{2}$$

Let $M_i \in \{0, 1\}$ denote visibility. With $\epsilon_i \sim \mathcal{N}(0, \sigma^2 I_3)$, the coordinate view is

$$\widetilde{x}_i = \begin{cases} x_i + \epsilon_i, & M_i = 1, \\ \perp, & M_i = 0, \end{cases} \tag{3}$$

$$\tilde{s}_i = \begin{cases} s_i, & M_i = 1, \\ \texttt{[MASK]}, & M_i = 0. \end{cases} \tag{4}$$

Here $\perp$ denotes missing geometry, not a coordinate at the origin. Frames and pair descriptors involving hidden geometric evidence are masked; distances are not computed to a fictitious mask coordinate. Available frames must transform equivariantly with their backbone atoms. Isotropic coordinate noise respects global rotation in distribution.

With independent masks, evidence hidden in one modality can remain visible at the same position in the other. Synchronization erases this direct local channel: for a fixed hidden set $\mathcal{M}$ and data-independent masks, $I(S_{\mathcal{M}}; \widetilde{G}_{\mathcal{M}} \mid M) = 0$. Appendix A.2.1 proves the statement and the corresponding context bound. Local geometry remains legitimate input to inverse folding; complete target geometry is available at standard inference.

3.3 $SE(3)$-INVARIANT GEOMETRIC ENCODING

The structural encoder maps available geometry and validity masks to scalar node and pair features, $H_{\text{geo}} \in \mathbb{R}^{N \times d_g}$ and $E_{\text{geo}} \in \mathbb{R}^{N \times N \times d_e}$. For valid pairs, the raw descriptors are

$$\begin{aligned} D_{ij} &= \|\widetilde{x}_j - \widetilde{x}_i\|_2, \qquad Q_{ij} = \widetilde{R}_i^\top \widetilde{R}_j, \\ u_{ij} &= \frac{\widetilde{R}_i^\top (\widetilde{x}_j - \widetilde{x}_i)}{D_{ij} + \varepsilon}, \qquad O_{ij} = [u_{ij}, \mathrm{vec}(Q_{ij})], \quad \varepsilon > 0. \end{aligned} \tag{5}$$

Each channel is unchanged by a common rigid motion. On a complete, valid framed backbone, the descriptors recover $g_i^{-1} g_j = (Q_{ij}, (D_{ij} + \varepsilon) u_{ij})$ exactly, including the numerical stabilizer. Equality of all relative transforms implies $g_i' = h g_i$ for a single h, so the raw descriptors separate global-pose orbits. Appendix A.1 proves the quotient correspondence.

3.4 GEOMETRY-AWARE ATTENTION

The discrete corrupted sequence is embedded into Euclidean semantic features $H_{\text{sem}} = \mathcal{E}_{\text{sem}}(\widetilde{S}) \in \mathbb{R}^{N \times d_s}$. To inject the invariant 3D topological constraints into the semantic sequence modeling, we compute a pairwise geometric bias for each attention head h:

$$B_{\text{geo}}^{(h)}(i,j) = \mathrm{MLP}_{\psi}^{(h)} \left(\gamma_D(D_{ij}) \oplus \gamma_O(O_{ij}) \oplus E_{\text{geo}}(i,j) \right), \tag{6}$$

where γ_D and γ_O act as radial and angular basis embeddings. The attention logits are then modulated deterministically:

$$A^{(h)} = \mathrm{Softmax}\left(\frac{Q^{(h)} {K^{(h)}}^\top}{\sqrt{d_k}} + B_{\text{geo}}^{(h)} + B_{\text{mask}} \right), \tag{7}$$

yielding the multi-head aligned representation Z_{align}. Here softmax acts row-wise over valid keys; B_{mask} handles padding/visibility independently of global pose. The scalar geometric bias supplies a topology-aware routing mechanism without coordinate-valued decoder updates.

3.5 VARIATIONAL INFORMATION BOTTLENECK AND DECODING

The aligned representation parameterizes a diagonal Gaussian stochastic encoder:

$$q_\phi(z \mid \widetilde{G}, \widetilde{S}) = \mathcal{N}\left(\mu_\phi(Z_{\text{align}}), \mathrm{diag}(\sigma_\phi^2(Z_{\text{align}})) \right). \tag{8}$$

Latent variables are drawn via $z = \mu_\phi + \sigma_\phi \odot \epsilon$, with $\epsilon \sim \mathcal{N}(0, I)$. A standard normal $p(z) = \mathcal{N}(0, I)$ is the reference distribution in the KL penalty, not a hard bound on z.

The frozen parameters of the pre-trained protein language model f_{PLM} are preserved to retain evolutionary semantics. We condition this backbone via trainable prefix adapters:

$$H_{\text{dec}} = f_{\text{PLM}}\left(\mathrm{Query}_N, \mathrm{Prefix}_\theta(z) \right). \tag{9}$$

The sequence head linearly projects the decoded states to amino-acid logits over $\mathcal{V}_{\text{aa}}$, parameterized by θ, yielding the sequence loss:

$$\mathcal{L}_{\text{seq}} = -\sum_{i=1}^{N} w_i \log p_\theta(s_i \mid z), \tag{10}$$

where $w_i \geq 0$ weights residue losses, emphasizing masked positions. At standard inference, the sequence view is entirely `[MASK]`, geometry is complete and unperturbed, and $z = \mu_\phi(G, \texttt{[MASK]}^N)$ is decoded with auxiliary geometry heads disabled. Greedy decoding gives single-sample AAR;

candidate diversity comes from categorical sampling at the specified temperature, not posterior sampling. For this Gaussian encoder, the mean is also the variational MAP estimate; we use that mode to fix the conditioning representation and remove additional latent Monte Carlo variability at test time. We retain latent sampling to optimize the VIB expectation during training, while conditional PLM sampling at the fixed mode controls output diversity; this is a plug-in approximation to the latent-marginal prediction (Appendix A.2.2).

3.6 Auxiliary Geometry Reconstruction

An auxiliary head $(\widehat{x}_i, \widehat{R}_i, \widehat{D}_{ij}, \widehat{O}_{ij}) = h_{\text{geo}}(H_{\text{dec}})$ encourages structural information retention. Following AlphaFold2 (Jumper et al., 2021), we use a robust frame-aligned reconstruction loss:

$$\mathcal{L}_{\text{FAPE}} = \frac{1}{N^2} \sum_{i,j} \rho\Big(\Big\| \widehat{R}_i^\top (\widehat{x}_j - \widehat{x}_i) - R_i^\top (x_j - x_i) \Big\|_2 \Big) . \tag{11}$$

Here ρ is the clipped Huber penalty used for reconstruction. Pair targets are computed from the *clean* backbone: $\mathcal{L}_{\text{pair}} = \mathbb{E}_{i,j}[\ell_D(\widehat{D}_{ij}, D_{ij}^{\text{clean}}) + \ell_O(\widehat{O}_{ij}, O_{ij}^{\text{clean}})]$, and $\mathcal{L}_{\text{geo}} = \lambda_{\text{FAPE}}\mathcal{L}_{\text{FAPE}} + \lambda_{\text{pair}}\mathcal{L}_{\text{pair}}$. Valid-residue masks restrict these sums where atoms or frames are missing. We constrain auxiliary rotations through six-dimensional Gram–Schmidt orthonormalization on nondegenerate inputs (Appendix A.3); hidden features remain Euclidean.

3.7 Rate–distortion Objective and Theoretical Properties

Let X be the corrupted encoder input (including validity masks), and let $\Phi(G) = (g_i^{-1} g_j)_{ij}$ be the clean relative-frame array encoded by Eq. (5). Then $Y = (S, \Phi(G))$ is the clean sequence and pose-invariant geometry. The information bottleneck motivation is

$$\mathcal{J}_{\text{IB}} = -I_q(Z;Y) + \beta I_q(Z;X). \tag{12}$$

We optimize a variational rate–distortion objective (Alemi et al., 2017; Bissiri et al., 2016):

$$\begin{aligned} \mathcal{L}(\phi,\theta) = \mathbb{E}_{X,Y}\big[& \mathbb{E}_{q_\phi(z|X)}[\lambda_{\text{seq}}\mathcal{L}_{\text{seq}} + \lambda_{\text{geo}}\mathcal{L}_{\text{geo}}] \\ & + \beta D_{\text{KL}}(q_\phi(z \mid X)\|p(z))\big]. \end{aligned} \tag{13}$$

The expected KL satisfies $R(\phi) = I_q(X;Z) + D_{\text{KL}}(q_\phi(z)\|p(z)) \geq I_q(X;Z)$. For each (X,Y), set $D_\theta = \lambda_{\text{seq}}\mathcal{L}_{\text{seq}} + \lambda_{\text{geo}}\mathcal{L}_{\text{geo}}$ and define $\mathcal{Z}_\beta = \int p(z) e^{-D_\theta/\beta}\, dz$, $\pi_\beta(z) \propto p(z) e^{-D_\theta/\beta}$ for $\beta > 0$. The exact identity

$$\mathbb{E}_q D_\theta + \beta D_{\text{KL}}(q\|p) = -\beta \log \mathcal{Z}_\beta + \beta D_{\text{KL}}(q\|\pi_\beta)$$

gives a generalized variational bound with an explicit approximation gap. It covers the weighted, clipped reconstruction losses in Eq. (13); a normalized emission model recovers the usual likelihood ELBO. Appendix A.0 derives both forms and their normalization conditions. The KL term also upper-bounds $I_q(X;Z)$, linking this variational objective to information-rate control.

4 Experiments

4.1 Experimental Setup

Tasks. We evaluate GeoSeqAlign† on structure-conditioned protein design. The auxiliary geometry head is used strictly as a training regularizer. Unless explicitly stated otherwise, structural metrics are computed by folding each designed sequence with the ESMFold oracle and comparing the predicted geometry against the clean target backbone.

Datasets. We benchmark on CATH+ 4.3(Sillitoe et al., 2021) for in-distribution inverse folding and TS50(Gao et al., 2023a) for zero-shot evaluation. For highly challenging CASP-style targets, we evaluate on the CASP15/TS45 domain subset using the evaluation framework of ProteinInvBench (Gao et al., 2023b). Chains are clustered via MMseqs2 at a 30% sequence identity threshold (Steinegger & Söding, 2017) to reduce sequence overlap. Structural overlap and external PLM pretraining overlap require separate consideration (Appendix F.2).

Table 1: **Structure-conditioned protein design on CATH+ 4.3.** All methods are evaluated on the same test targets with a unified ESMFold oracle. AAR and Best@8 are reported in percentages; PAE is in Å. Best results are **bolded**, and second-best are underlined.

Model	Train / Extra Data	Seq. Plausibility		Design Accuracy		Foldability / Self-Consistency			Lat. $N = 200$ ms ↓
		PPL ↓	Rep ↓	AAR (%) ↑	Best@8 (%) ↑	pLDDT ↑	PAE (Å) ↓	scTM ↑	
Struct. Transformer	CATH	6.852	0.151	38.214	41.532	71.245	8.512	0.724	45
GVP-GNN	CATH	6.521	0.124	40.325	43.118	73.512	7.823	0.751	35
ProteinMPNN	CATH	4.614	0.082	52.418	55.621	81.234	5.415	0.842	28
PiFold	CATH	4.552	0.071	53.142	56.415	81.921	5.218	0.854	**15**
ESM-IF1†	12M AF2 Structs.	4.125	0.063	54.215	57.824	83.512	4.814	0.862	210
SaProt†	External	3.951	0.052	56.518	59.215	84.821	4.512	0.881	230
GRADE-IF	CATH/Ext.	4.054	0.061	55.824	58.712	84.118	4.621	0.874	35
SPDesign	CATH/Templates	3.982	0.054	56.135	59.512	84.524	4.418	0.883	28
BC-Design	CATH/Biochem.	3.751	0.042	57.218	60.824	85.815	4.012	0.892	38
MapDiff	CATH	3.824	0.045	57.612	61.218	85.621	4.124	0.895	40
GeoSeqAlign† **(Ours)**	CATH+4.3	**3.652**	**0.031**	**58.145**	**61.832**	**86.215**	**3.912**	**0.904**	225

Baselines. We restrict our primary comparisons to structure-conditioned sequence generation models. Pure backbone generators (e.g., RFdiffusion (Watson et al., 2023)) are excluded as they do not natively model conditional sequence likelihoods on fixed coordinates. All methods are evaluated on identical test splits using a standardized sequence-decoding budget. Baselines use officially released checkpoints; disparities in external pretraining data are explicitly documented.

Implementation Details. For parameter-efficient adaptation, the underlying protein language model in GeoSeqAlign is strictly frozen. Only the structure encoder, geometry-aware attention adapter, VIB posterior, and mapping heads (totalling 45M parameters) are trained. We optimize via AdamW with a learning rate of 10^{-4}, weight decay of 0.01, and batch size of 128. To prevent early latent collapse, the VIB coefficient β is linearly annealed from 0 to 0.1 over the first $5{,}000$ steps. We use three seeds for the main stochastic evaluations and retain the $\beta = 0.1$ training schedule; we examine additional coefficients, including $\beta = 0.05$, in a separate sensitivity study (Appendix F.5).

4.2 Metrics

Sequence Plausibility. We report conditional residue perplexity (PPL) over $\mathcal{V}_{\text{aa}}$. Repetitiveness (Rep) measures within-sequence repeated n-grams; it is not a measure of cross-sample diversity. Separate uniqueness, identity, and clustering analyses are given in Appendix F.1.

Design Accuracy. Amino-acid recovery (AAR, %) is position-wise identity to the native sequence. Single-sample AAR uses greedy decoding; Best@8 reports the maximum recovery among eight sampled sequences. We use the same eight-candidate budget to measure the recovery attainable through a modest amount of sampling while keeping evaluation cost comparable. Best@8 complements greedy AAR as a native-sequence-aware diagnostic, rather than a deployable selection rule.

Foldability and Self-consistency. Each generated sequence is evaluated in silico via ESMFold(Lin et al., 2023). pLDDT and PAE are confidence estimates produced by the folding oracle (Jumper et al., 2021; Mariani et al., 2013); they are not measured errors against the target. Self-consistency TM-score (scTM) and RMSD (scRMSD) compare oracle-predicted structures with the clean target (Zhang & Skolnick, 2004; 2005). These are computational proxies, not experimental foldability measurements.

4.3 Main Results

As summarized in Table 1, GeoSeqAlign achieves the highest AAR and scTM in this comparison using the same targets and folding oracle. Table 2 adds LM-Design, KW-Design, and Bridge-IF. The clean CATH margin over KW-Design is only 0.3 percentage points at the displayed precision; differences are larger on CASP15/TS45 and under 2Å noise. In the noise-condition columns, perturbations are applied exclusively to inverse-folder inputs, and the reported values correspond to aggregate estimates. Sequence recovery and structural consistency are evaluated under an identical decoding budget. However, pretraining resources and model architectures vary across the compared methods.

Table 2: **Quantitative benchmarking and structural robustness analysis. Left:** Comparison against state-of-the-art structure-conditioned PLMs across datasets, metrics, and decoding latency. **Right:** Self-consistency TM-scores (↑) under escalating Gaussian coordinate noise ($\sigma \in [0, 2]$ Å).

	AAR (%) ↑			scTM ↑			Lat.
Model	**CATH**	**TS50**	**CASP**	**Clean**	**1 Å**	**2 Å**	**(ms) ↓**
ProteinMPNN	52.4	48.6	35.8	0.842	0.58	0.35	**28**
LM-Design	56.5	52.1	39.8	0.885	0.62	0.40	181
KW-Design	57.8	53.6	40.5	0.892	0.64	0.41	309
Bridge-IF	57.2	53.1	41.2	0.898	0.65	0.43	242
GeoSeqAlign	**58.1**	**54.5**	**42.4**	**0.904**	**0.68**	**0.48**	225

Model	$\sigma = 0.0$ Å	$\sigma = 0.5$ Å	$\sigma = 1.0$ Å	$\sigma = 2.0$ Å
ProteinMPNN	0.84 ± 0.01	0.72 ± 0.01	0.58 ± 0.02	0.35 ± 0.04
ESM-IF1	0.86 ± 0.01	0.75 ± 0.01	0.62 ± 0.02	0.38 ± 0.03
SaProt	0.88 ± 0.00	0.77 ± 0.01	0.65 ± 0.02	0.42 ± 0.02
SPDesign	0.88 ± 0.01	0.78 ± 0.01	0.63 ± 0.03	0.40 ± 0.04
BC-Design	$\underline{0.89} \pm 0.00$	$\underline{0.81} \pm 0.01$	$\underline{0.67} \pm 0.02$	$\underline{0.46} \pm 0.03$
MapDiff	$\underline{0.89} \pm 0.01$	0.80 ± 0.01	0.66 ± 0.02	0.45 ± 0.03
GeoSeqAlign	**0.90**±0.00	**0.82**±0.01	**0.68**±0.01	**0.48**±0.02

Table 3: **Zero-shot generalization.** TS50 and CASP15/TS45 targets are not used for fine-tuning. AAR is reported in percent; PAE is reported in Å.

Model	**TS50**				**CASP15/TS45**			
	AAR (%) ↑	pLDDT ↑	PAE (Å) ↓	scTM ↑	AAR (%) ↑	pLDDT ↑	PAE (Å) ↓	scTM ↑
ProteinMPNN	48.6± 0.381	80.5± 0.427	5.6± 0.142	0.82± 0.013	35.8± 0.485	68.2± 0.518	7.5± 0.183	0.64± 0.021
PiFold	49.2± 0.453	80.8± 0.472	5.4± 0.126	0.83± 0.014	36.5± 0.551	68.5± 0.584	7.3± 0.167	0.65± 0.019
ESM-IF1†	51.5± 0.329	82.1± 0.284	5.1± 0.092	0.84± 0.011	38.5± 0.376	70.8± 0.352	6.8± 0.174	0.68± 0.015
SaProt†	53.2± 0.264	83.5± 0.258	4.8± 0.117	0.86± 0.009	40.1± 0.421	72.1± 0.395	6.4± 0.123	0.70± 0.012
GRADE-IF	52.8± 0.412	83.0± 0.347	4.9± 0.138	0.85± 0.013	39.4± 0.463	71.5± 0.451	6.6± 0.228	0.69± 0.017
SPDesign	52.5± 0.375	83.2± 0.316	4.8± 0.129	0.85± 0.011	39.8± 0.472	71.8± 0.464	6.5± 0.191	0.69± 0.018
BC-Design	<u>54.0± 0.283</u>	<u>84.2± 0.271</u>	<u>4.5± 0.086</u>	<u>0.87± 0.008</u>	<u>41.8± 0.354</u>	<u>73.8± 0.372</u>	<u>6.1± 0.147</u>	<u>0.73± 0.016</u>
MapDiff	53.8± 0.297	84.0± 0.263	4.6± 0.115	<u>0.87± 0.012</u>	41.2± 0.388	73.2± 0.419	6.2± 0.162	0.72± 0.014
GeoSeqAlign† **(ours)**	**54.5± 0.174**	**84.8± 0.182**	**4.5± 0.063**	**0.88± 0.005**	**42.4± 0.261**	**74.5± 0.257**	**6.1± 0.098**	**0.74± 0.011**

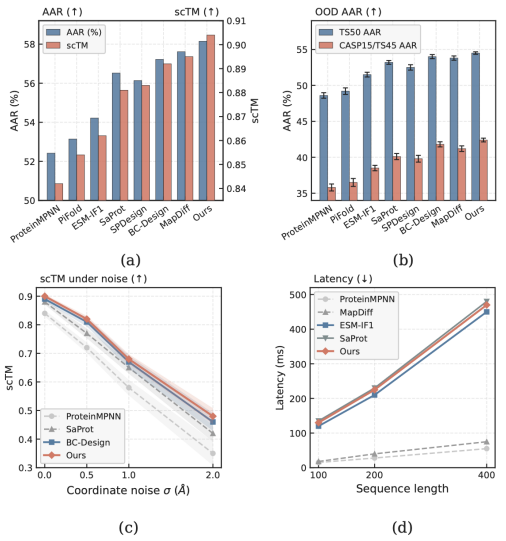
Zero-shot Generalization. Table 3 retains the zero-shot TS50 and CASP15/TS45 results. The observed CASP AAR of 42.4% exceeds the evaluated baselines, but dataset-level transfer does not by itself establish topology-disjoint generalization. A separate Foldseek-filtered evaluation retains 321/500 held-out targets and yields 53.2 AAR and 0.869 scTM, compared with 51.8 and 0.845 for KW-Design (Appendix F.2).

Table 4: **Robustness under severe structural corruption.** scTM (↑) scores of designed sequences evaluated after injecting controlled Gaussian noise into the inference coordinates.

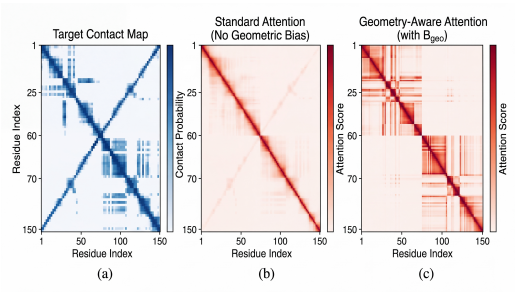
Model	$\sigma = 0.0$Å	$\sigma = 0.5$Å	$\sigma = 1.0$Å	$\sigma = 2.0$Å
ProteinMPNN	0.84± 0.01	0.72± 0.01	0.58± 0.02	0.35± 0.04
ESM-IF1	0.86± 0.01	0.75± 0.01	0.62± 0.02	0.38± 0.03
SaProt	0.88± 0.00	0.77± 0.01	0.65± 0.02	0.42± 0.02
SPDesign	0.88± 0.01	0.78± 0.01	0.63± 0.03	0.40± 0.04
BC-Design	<u>0.89± 0.00</u>	<u>0.81± 0.01</u>	<u>0.67± 0.02</u>	<u>0.46± 0.03</u>
MapDiff	<u>0.89± 0.01</u>	0.80± 0.01	0.66± 0.02	0.45± 0.03
GeoSeqAlign†	**0.90± 0.00**	**0.82± 0.01**	**0.68± 0.01**	**0.48± 0.02**

Robustness to Noisy Backbones. Table 4 perturbs only the inverse folder's input; oracle predictions are compared with pristine targets. At $\sigma = 2$Å, GeoSeqAlign obtains scTM 0.48 versus 0.45 for MapDiff and 0.46 for BC-Design. These controlled perturbations characterize robustness to this noise model. Appendix G illustrates the corresponding geometric views and separates visual annotations from the benchmark measurements.

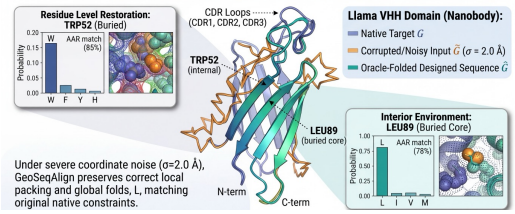
Ablation Study. Table 5 reports component ablations. Removing geometric bias produces the largest AAR decrease among these removals; this does not decompose absolute contributions. The added factorial analysis holds the geometric model and stochastic encoder fixed (Table 6). With



(a) Macro benchmark and latency trade-offs



(b) Geometry-aware attention routing



(c) Local packing under coordinate noise (PDB: 1G9E)

Figure 2: **Comprehensive empirical validation and mechanistic analysis.** **(a)** Global Pareto trade-offs across sequence recovery (CATH+ 4.3), zero-shot OOD generalization (TS50/CASP15), coordinate noise degradation ($\sigma \in [0, 2]$ Å), and batch-1 decoding latency. **(b)** Qualitative alignment between the target residue contact map and PLM self-attention weights before and after incorporating B_{geo}. **(c)** Structural and residue-level packing validation on a representative VHH domain (PDB: 1G9E) under severe coordinate noise ($\sigma = 2.0$ Å), demonstrating the preservation of core hydrophobic packing (TRP52, LEU89) and global tertiary foldability.

Table 5: **Component analysis on CATH+ 4.3.** We systematically disable key mechanisms to isolate their contribution to sequence recovery and structural fidelity.

Variant	AAR (%) ↑	**ΔAAR**	pLDDT ↑	PAE (Å) ↓	scTM ↑
Full GeoSeqAlign†	**58.1**± **0.16**	–	**86.2**± **0.15**	**3.9**± **0.07**	**0.90**± **0.00**
w/o KL regularization ($\beta = 0$)	56.8± 0.28	−1.3	85.3± 0.26	4.2± 0.11	0.88± 0.01
w/o synchronized corruption	56.2± 0.31	−1.9	84.8± 0.29	4.4± 0.13	0.87± 0.01
w/o geometry-aware bias B_{geo}	54.9± 0.38	−3.2	83.5± 0.35	4.9± 0.18	0.85± 0.02
Learned pos. bias instead of B_{geo}	55.4± 0.29	−2.7	84.1± 0.31	4.7± 0.14	0.86± 0.01
w/o auxiliary geometry loss	57.5± 0.19	−0.6	85.8± 0.24	4.1± 0.09	0.89± 0.00
Mean-only posterior, KL retained	57.0± 0.26	−1.1	85.4± 0.28	4.2± 0.12	0.88± 0.01

synchronized corruption fixed, KL regularization adds 1.3 CATH, 1.7 TS50, and 2.2 CASP AAR points and 0.06 scTM at 2Å. Independent masks give 40.6 CASP AAR and 0.43 scTM at 2Å, below synchronization's 42.4 and 0.48. The interaction contrasts are 0.2 CATH AAR points and 0.01 scTM at 2Å, consistent with largely additive effects of the two training components.

Table 6: **KL and corruption controls.** The PLM, geometric encoder, stochastic latent architecture, data, and optimization budget are held fixed; $\beta = 0$ removes the KL penalty, not stochastic encoding. Sequence corruption is retained in the sequence-only controls, while geometric corruption is removed. Independent masks use matched per-modality masking rates. AAR is in percent.

Variant	CATH	TS50	CASP	scTM@1Å	scTM@2Å
Full model	**58.1**	**54.5**	**42.4**	**0.68**	**0.48**
Synchronized, $\beta = 0$	56.8	52.8	40.2	0.63	0.42
Sequence-only corruption + KL	56.2	52.1	39.2	0.61	0.39
Sequence-only corruption, $\beta = 0$	55.1	50.8	37.2	0.56	0.34
Independent masks + KL	57.1	53.2	40.6	0.64	0.43

Table 7: **Computational footprint.** Pure sequence decoding latency (batch size 1, isolated A100-80GB) and resident parameter overhead across model architectures.

Model	Params (Train/Total)	Peak VRAM	Latency (ms) ↓		
			100	200	400
ProteinMPNN	1.9M / 1.9M	2.5 GB	15	28	55
SPDesign	15M / 15M	2.8 GB	14	28	52
MapDiff	75M / 75M	3.8 GB	18	40	75
ESM-IF1	142M / 142M	14.2 GB	120	210	450
SaProt	650M / 650M	15.1 GB	135	230	480
GeoSeqAlign†	45M / 700M	14.8 GB	130	225	470

Table 8: **Retrospective mutation evaluation.** Zero-shot variant fitness ranking on ProteinGym assays ($\Delta \log p$) and Rosetta thermodynamic relaxation proxies.

Model	ProteinGym		Fold	Rosetta
	Spearman ↑	NDCG@10 ↑	Pres. ↑	$\Delta\Delta G$ ↓
ProteinMPNN	0.42± 0.02	0.48± 0.02	85%	-2.5± 0.34
ESM-IF1	0.48± 0.02	0.55± 0.03	88%	-3.2± 0.21
SPDesign	0.50± 0.02	0.60± 0.02	90%	-3.7± 0.18
SaProt	0.51± 0.01	0.61± 0.02	91%	-3.8± 0.18
MapDiff	0.52± 0.02	0.63± 0.02	92%	-4.0± 0.19
BC-Design	0.53± 0.02	0.64± 0.03	91%	-4.1± 0.22
GeoSeqAlign†	**0.54**± 0.01	**0.65**± 0.02	**94%**	**-4.3**± 0.17

Efficiency and Functional Evaluation. Table 7 reports the computational footprint: 45M trainable parameters out of approximately 700M total (6.4%). At $N = 200$, decoding takes 225 ms versus 28 ms for ProteinMPNN and 15 ms for PiFold. Freezing weights reduces adaptation costs but does not remove their forward-pass cost. Full-trajectory FLOPs and recent PLM comparisons appear in Appendix F.4.

Table 8 reports retrospective ProteinGym variant ranking (Spearman 0.54, NDCG@10 0.65) and Rosetta energy proxies. These results do not establish prospective function or binding. Cross-sample evaluation on 100 backbones gives 99.2% exact uniqueness and 93.4% of samples above scTM 0.5, with 18.2 foldable clusters per 50 samples at 80% identity. Other methods yield more clusters at several thresholds; Appendix F.1 reports the complete quality–diversity comparison.

Mechanistic Validation and Attention Routing. Quantitative comparisons against recent structure-conditioned PLMs are consolidated in Table 2. While top-performing methods show competitive parity on clean in-distribution targets, GeoSeqAlign establishes pronounced advantages on zero-shot CASP15 domains (42.4% AAR) and under severe 2.0 Å coordinate noise (0.48 scTM). Figure 2 provides a unified empirical and mechanistic overview: Figure 2(a) highlights the global Pareto trade-offs across accuracy, generalization, noise degradation, and computational cost. Figure 2(b) inspects the internal attention modulation, confirming that the invariant geometric bias B_{geo} successfully redirects attention mass away from the trivial sequence diagonal toward off-diagonal tertiary contact regions. Finally, the molecular case study in Figure 2(c) demonstrates that even when backbone coordinates are severely perturbed ($\sigma = 2.0$ Å), the VIB regularizer filters out local coordinate noise, preserving the steric hydrophobic packaging of critical buried residues (TRP52, LEU89) and maintaining overall foldability.

5 Conclusion

We presented GeoSeqAlign, a fixed-backbone inverse-folding framework combining invariant geometric conditioning, a frozen PLM, synchronized corruption, and a KL-regularized stochastic latent. Raw relative-frame descriptors connect the implementation to the global-pose quotient, while matched ablations quantify robustness and transfer gains from KL regularization and synchronized corruption. Recovery, structural proxies, diversity, and compute reveal a useful but non-universal trade-off. Partner-aware conditioning, faithful richer-context baselines, and experimental validation remain future work.

AI USE STATEMENT

We used generative AI tools during manuscript revision for language editing, consistency checks, reference checking and readability improvements. We did not use generative AI tools to develop the research methodology, design experiments, interpret results, or formulate scientific claims. All AI-assisted edits were reviewed by the authors. We are responsible for the final text, mathematical claims, references, and empirical evidence.

REPRODUCIBILITY STATEMENT

Sections 3–4 describe the model and evaluation. Appendix A gives assumptions and proofs, Appendix B gives the training algorithm, and Appendices C and E describe metrics and the reported environment. Appendices F.1–F.5 give additional results and their evaluation scope. The checkpoint labels are descriptive rather than immutable source hashes.

ETHICS STATEMENT

Protein sequence design can support beneficial research, but computational compatibility does not establish biological safety, activity, or binding. Generated candidates require appropriate expert assessment and experimental validation. Broader impacts and the limits of the present fixed-backbone task are discussed in Appendix D.

REFERENCES

Josh Abramson, Jonas Adler, Jack Dunger, Richard Evans, Tim Green, Alexander Pritzel, Olaf Ronneberger, Lindsay Willmore, Andrew J. Ballard, Joshua Bambrick, et al. Accurate structure prediction of biomolecular interactions with AlphaFold 3. *Nature*, 630:493–500, 2024. doi: 10.1038/s41586-024-07487-w. URL https://www.nature.com/articles/s41586-024-07487-w.

Alexander A. Alemi, Ian Fischer, Joshua V. Dillon, and Kevin Murphy. Deep variational information bottleneck. In *International Conference on Learning Representations*, 2017. URL https://openreview.net/forum?id=HyxQzBceg.

Rebecca F. Alford, Andrew Leaver-Fay, Jeliazko R. Jeliazkov, Matthew J. O'Meara, Frank P. DiMaio, Hahnbeom Park, Maxim V. Shapovalov, P. Douglas Renfrew, Vikram K. Mulligan, Kalli Kappel, et al. The Rosetta all-atom energy function for macromolecular modeling and design. *Journal of Chemical Theory and Computation*, 13(6):3031–3048, 2017. doi: 10.1021/acs.jctc.7b00125. URL https://doi.org/10.1021/acs.jctc.7b00125.

Peizhen Bai, Filip Miljkovic, Xianyuan Liu, Leonardo De Maria, Rebecca Croasdale-Wood, Owen Rackham, and Haiping Lu. Mask-prior-guided denoising diffusion improves inverse protein folding. *Nature Machine Intelligence*, 2025. doi: 10.1038/s42256-025-01042-6. URL https://doi.org/10.1038/s42256-025-01042-6.

Pier Giovanni Bissiri, Chris C. Holmes, and Stephen G. Walker. A general framework for updating belief distributions. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 78 (5):1103–1130, 2016. doi: 10.1111/rssb.12158.

Justas Dauparas, Ivan Anishchenko, Nathaniel Bennett, Hua Bai, Robert J. Ragotte, Lucas F. Milles, Basile I. M. Wicky, Alexis Courbet, Rob J. de Haas, Neville Bethel, et al. Robust deep learning-based protein sequence design using ProteinMPNN. *Science*, 378(6615):49–56, 2022. doi: 10.1126/science.add2187. URL https://doi.org/10.1126/science.add2187.

Noelia Ferruz, Steffen Schmidt, and Birte Höcker. ProtGPT2 is a deep unsupervised language model for protein design. *Nature Communications*, 13:4348, 2022. doi: 10.1038/s41467-022-32007-7. URL https://doi.org/10.1038/s41467-022-32007-7.

Fabian B. Fuchs, Daniel E. Worrall, Volker Fischer, and Max Welling. SE(3)-transformers: 3d roto-translation equivariant attention networks. In *Advances in Neural Information Processing Systems*, volume 33, pp. 1970–1981, 2020. URL https://proceedings.neurips.cc/paper_files/paper/2020/hash/15231a7ce4ba789d13b722cc5c955834-Abstract.html.

Zhangyang Gao, Cheng Tan, and Stan Z. Li. PiFold: Toward effective and efficient protein inverse folding. In *International Conference on Learning Representations*, 2023a. URL https://openreview.net/forum?id=oMsN9TYwJ0j.

Zhangyang Gao, Cheng Tan, Yijie Zhang, Xingran Chen, Lirong Wu, and Stan Z. Li. ProteinInvBench: Benchmarking protein inverse folding on diverse tasks, models, and metrics. In *Advances in Neural Information Processing Systems, Datasets and Benchmarks Track*, 2023b. URL https://papers.neurips.cc/paper_files/paper/2023/hash/d73078d49799693792fb0f3f32c57fc8-Abstract-Datasets_and_Benchmarks.html.

Zhangyang Gao, Cheng Tan, Xingran Chen, Yijie Zhang, Jun Xia, Siyuan Li, and Stan Z. Li. KW-Design: Pushing the limit of protein design via knowledge refinement. In *International Conference on Learning Representations*, 2024. URL https://proceedings.iclr.cc/paper_files/paper/2024/hash/9410d94d47adfb07b41a0b226270f068-Abstract-Conference.html.

Chloe Hsu, Robert Verkuil, Jason Liu, Zeming Lin, Brian Hie, Tom Sercu, Adam Lerer, and Alexander Rives. Learning inverse folding from millions of predicted structures. In *Proceedings of the 39th International Conference on Machine Learning*, volume 162, pp. 8946–8970. PMLR, 2022. URL https://proceedings.mlr.press/v162/hsu22a.html.

John Ingraham, Vikas Garg, Regina Barzilay, and Tommi Jaakkola. Generative models for graph-based protein design. In *Advances in Neural Information Processing Systems*, volume 32, 2019. URL https://proceedings.neurips.cc/paper_files/paper/2019/hash/f3a4ff4839c56a5f460c88cce3666a2b-Abstract.html.

Bowen Jing, Stephan Eismann, Patricia Suriana, Raphael J. L. Townshend, and Ron O. Dror. Learning from protein structure with geometric vector perceptrons. In *International Conference on Learning Representations*, 2021. URL https://openreview.net/forum?id=1YLJDvSx6J4.

John Jumper, Richard Evans, Alexander Pritzel, Tim Green, Michael Figurnov, Olaf Ronneberger, Kathryn Tunyasuvunakool, Russ Bates, Augustin Žídek, Anna Potapenko, et al. Highly accurate protein structure prediction with AlphaFold. *Nature*, 596(7873):583–589, 2021. doi: 10.1038/s41586-021-03819-2. URL https://doi.org/10.1038/s41586-021-03819-2.

Zeming Lin, Halil Akin, Roshan Rao, Brian Hie, Zhongkai Zhu, Wenting Lu, Allan dos Santos Costa, Maryam Fazel-Zarandi, Tom Sercu, Sal Candido, et al. Evolutionary-scale prediction of atomic-level protein structure with a language model. *Science*, 379(6637):1123–1130, 2023. doi: 10.1126/science.ade2574. URL https://doi.org/10.1126/science.ade2574.

Jiarui Lu, Xiaoyin Chen, Stephen Z. Lu, Chence Shi, Hongyu Guo, Yoshua Bengio, and Jian Tang. Structure language models for protein conformation generation. In *International Conference on Learning Representations*, 2025. URL https://openreview.net/forum?id=OzUNDnpQyd.

Ali Madani, Ben Krause, Eric R. Greene, Subu Subramanian, Benjamin P. Mohr, James M. Holton, Jose Luis Olmos Jr., Caiming Xiong, Zachary Z. Sun, Richard Socher, et al. Large language models generate functional protein sequences across diverse families. *Nature Biotechnology*, 41(8):1099–1106, 2023. doi: 10.1038/s41587-022-01618-2. URL https://doi.org/10.1038/s41587-022-01618-2.

Valerio Mariani, Marco Biasini, Alessandro Barbato, and Torsten Schwede. lDDT: a local superposition-free score for comparing protein structures and models using distance difference tests. *Bioinformatics*, 29(21):2722–2728, 2013. doi: 10.1093/bioinformatics/btt473. URL https://doi.org/10.1093/bioinformatics/btt473.

Pascal Notin, Aaron Kollasch, Daniel Ritter, Lood van Niekerk, Steffanie Paul, Han Spinner, Nathan Rollins, Ada Shaw, Rose Orenbuch, Ruben Weitzman, Jonathan Frazer, Mafalda Dias, Dinko Franceschi, Yarin Gal, and Debora Marks. ProteinGym: Large-scale benchmarks for protein fitness prediction and design. In *Advances in Neural Information Processing Systems, Datasets and Benchmarks Track*, 2023. URL https://proceedings.neurips.cc/paper_files/paper/2023/hash/cac723e5ff29f65e3fcbb0739ae91bee-Abstract-Datasets_and_Benchmarks.html.

Alexander Rives, Joshua Meier, Tom Sercu, Siddharth Goyal, Zeming Lin, Jason Liu, Demi Guo, Myle Ott, C. Lawrence Zitnick, Jerry Ma, and Rob Fergus. Biological structure and function emerge from scaling unsupervised learning to 250 million protein sequences. *Proceedings of the National Academy of Sciences*, 118(15):e2016239118, 2021. doi: 10.1073/pnas.2016239118. URL https://doi.org/10.1073/pnas.2016239118.

Victor Garcia Satorras, Emiel Hoogeboom, and Max Welling. E(n) equivariant graph neural networks. In *Proceedings of the 38th International Conference on Machine Learning*, volume 139, pp. 9323–9332. PMLR, 2021. URL https://proceedings.mlr.press/v139/satorras21a.html.

Ian Sillitoe, Nicola Bordin, Natalie Dawson, Vaishali P. Waman, Paul Ashford, Harry M. Scholes, Camilla S. M. Pang, Laurel Woodridge, Clemens Rauer, Neeladri Sen, et al. CATH: increased structural coverage of functional space. *Nucleic Acids Research*, 49(D1):D266–D273, 2021. doi: 10.1093/nar/gkaa1079. URL https://academic.oup.com/nar/article/49/D1/D266/6006195.

Martin Steinegger and Johannes Söding. MMseqs2 enables sensitive protein sequence searching for the analysis of massive data sets. *Nature Biotechnology*, 35(11):1026–1028, 2017. doi: 10.1038/nbt.3988. URL https://doi.org/10.1038/nbt.3988.

Jin Su, Chenchen Han, Yuyang Zhou, Junjie Shan, Xibin Zhou, and Fajie Yuan. SaProt: Protein language modeling with structure-aware vocabulary. In *International Conference on Learning Representations*, 2024. URL https://proceedings.iclr.cc/paper_files/paper/2024/hash/1c42513b8895ab11fbbb5b7e8e6b6b02-Abstract-Conference.html.

Xiangru Tang, Xinwu Ye, Fang Wu, Daniel Shao, Yin Fang, Siming Chen, Dong Xu, and Mark Gerstein. BC-Design: A biochemistry-aware framework for high-precision inverse protein folding, 2024. URL https://www.biorxiv.org/content/10.1101/2024.10.28.620755v1.

Naftali Tishby, Fernando C. Pereira, and William Bialek. The information bottleneck method. In *Proceedings of the 37th Annual Allerton Conference on Communication, Control, and Computing*, pp. 368–377, 1999. URL https://arxiv.org/abs/physics/0004057.

Michel van Kempen, Stephanie S. Kim, Charlotte Tumescheit, Milot Mirdita, Jeongjae Lee, Cameron L. M. Gilchrist, Johannes Söding, and Martin Steinegger. Fast and accurate protein structure search with Foldseek. *Nature Biotechnology*, 42:243–246, 2024. doi: 10.1038/s41587-023-01773-0. URL https://www.nature.com/articles/s41587-023-01773-0.

Hui Wang, Dong Liu, Kailong Zhao, Yajun Wang, and Guijun Zhang. SPDesign: protein sequence designer based on structural sequence profile using ultrafast shape recognition. *Briefings in Bioinformatics*, 25(3):bbae146, 2024. doi: 10.1093/bib/bbae146. URL https://doi.org/10.1093/bib/bbae146.

Joseph L. Watson, David Juergens, Nathaniel R. Bennett, Brian L. Trippe, Jason Yim, Helen E. Eisenach, Woody Ahern, Andrew J. Borst, Robert J. Ragotte, Lukas F. Milles, et al. De novo design of protein structure and function with RFdiffusion. *Nature*, 620(7976):1089–1100, 2023. doi: 10.1038/s41586-023-06415-8. URL https://doi.org/10.1038/s41586-023-06415-8.

Sen Yang, Lingli Ju, Peng Cheng, Jianglin Zhou, Yamin Cai, and Dawei Feng. Co-design protein sequence and structure in discrete space via generative flow. *Bioinformatics*, 41(5):

btaf248, 2025. doi: 10.1093/bioinformatics/btaf248. URL https://academic.oup.com/bioinformatics/article/41/5/btaf248/8123382.

Kai Yi, Bingxin Zhou, Yiqing Shen, Pietro Liò, and Yu Guang Wang. Graph denoising diffusion for inverse protein folding. In *Advances in Neural Information Processing Systems*, volume 36, 2023. URL https://papers.neurips.cc/paper_files/paper/2023/hash/20888d00c5df685de2c09790040e0327-Abstract-Conference.html.

Kai Yi, Kiarash Jamali, and Sjors H. W. Scheres. All-atom inverse protein folding through discrete flow matching. In *Proceedings of the 42nd International Conference on Machine Learning*, volume 267, pp. 72244–72257. PMLR, 2025. URL https://proceedings.mlr.press/v267/yi25a.html.

Yang Zhang and Jeffrey Skolnick. Scoring function for automated assessment of protein structure template quality. *Proteins: Structure, Function, and Bioinformatics*, 57(4):702–710, 2004. doi: 10.1002/prot.20264. URL https://doi.org/10.1002/prot.20264.

Yang Zhang and Jeffrey Skolnick. TM-align: a protein structure alignment algorithm based on the TM-score. *Nucleic Acids Research*, 33(7):2302–2309, 2005. doi: 10.1093/nar/gki524. URL https://doi.org/10.1093/nar/gki524.

Zaixiang Zheng, Yifan Deng, Dongyu Xue, Yi Zhou, Fei Ye, and Quanquan Gu. Structure-informed language models are protein designers. In *Proceedings of the 40th International Conference on Machine Learning*, volume 202, pp. 42317–42338. PMLR, 2023. URL https://proceedings.mlr.press/v202/zheng23a.html.

Yi Zhou, Connelly Barnes, Jingwan Lu, Jimei Yang, and Hao Li. On the continuity of rotation representations in neural networks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 5745–5753, 2019. URL https://openaccess.thecvf.com/content_CVPR_2019/html/Zhou_On_the_Continuity_of_Rotation_Representations_in_Neural_Networks_CVPR_2019_paper.html.

Yiheng Zhu, Jialu Wu, Qiuyi Li, Jiahuan Yan, Mingze Yin, Wei Wu, Mingyang Li, Jieping Ye, Zheng Wang, and Jian Wu. Bridge-IF: Learning inverse protein folding with markov bridges. In *Advances in Neural Information Processing Systems*, volume 37, 2024. URL https://openreview.net/forum?id=Q8yfhrBBD8.

Appendices

A Additional theoretical details

Let X include corrupted views and validity masks, $Z \sim q_\phi(z \mid X)$, and $Y = (S, \Phi(G))$. Information quantities use $q(x, y, z) = p_{\text{data,corrupt}}(x, y)q_\phi(z \mid x)$ and the finite integrals stated below.

A.0 Multi-task variational formulation and emission models

We first derive a likelihood bound from an explicit joint model, then give the exact variational identity for the loss used in our implementation. This separates the choice of observation model from the variational argument. The same separation underlies the latent-variable formulation of structure language models (Lu et al., 2025).

A normalized conditional model. We use $U = (T_k^{-1}T_i)_{i\neq k} \in SE(3)^{N-1}$ as coordinates for the clean quotient geometry. Sequence observations use counting measure and U uses a product Haar measure. A conditional model with prior $r_\eta(z \mid X)$ and conditionally independent emissions has joint density

$$p_{\theta,\eta}(S, U, z \mid X) = r_\eta(z \mid X)p_{\theta_s}(S \mid z, X)p_{\theta_g}(U \mid z, X). \tag{14}$$

The independence is a modeling assumption given (z, X), not an independence property of protein sequences and structures. For any variational density $q = q_\phi(z \mid X)$ with the required support and finite terms, Jensen's inequality gives

$$\begin{aligned}\log p_{\theta,\eta}(S, U \mid X) &= \log \int q(z)\frac{r_\eta(z \mid X)p_{\theta_s}(S \mid z, X)p_{\theta_g}(U \mid z, X)}{q(z)}\, dz \\ &\geq \mathbb{E}_q \log p_{\theta_s}(S \mid z, X) + \mathbb{E}_q \log p_{\theta_g}(U \mid z, X) - D_{\text{KL}}(q\|r_\eta(\cdot \mid X)).\end{aligned} \tag{15}$$

The difference between the two sides is exactly $D_{\text{KL}}(q_\phi(z \mid X)\|p_{\theta,\eta}(z \mid S, U, X))$. Restricting the inference network to X preserves the inequality, although it may enlarge this gap. A conditional prior must remain inside its KL term. If both the prior and emissions ignore X, the generative model is unconditional even when its variational encoder takes X as input.

Normalized local emissions and loss scales. Let $y_{ij} = R_i^\top(x_j - x_i)$, $\widehat{y}_{ij} = \widehat{R}_i^\top(\widehat{x}_j - \widehat{x}_i)$, and $e_{ij} = \widehat{y}_{ij} - y_{ij}$. For an independently observed local vector in $\mathbb{R}^3$, the radial Laplace density is

$$p_b(e) = \frac{1}{8\pi b^3}\exp(-\|e\|_2/b), \qquad \int_{\mathbb{R}^3} e^{-\|e\|_2/b}\, de = 4\pi\int_0^\infty r^2 e^{-r/b}\, dr = 8\pi b^3. \tag{16}$$

Thus $-\log p_b(e) = \|e\|_2/b + \log(8\pi b^3)$; for a Gaussian residual it is $\|e\|_2^2/(2\sigma^2) + (3/2)\log(2\pi\sigma^2)$. An unclipped Huber kernel is integrable because its tail grows linearly. A clipped penalty $\rho \leq c$ gives $e^{-\rho} \geq e^{-c}$ and cannot normalize over all of $\mathbb{R}^3$. On a bounded observation domain Ω, the kernel instead has the finite normalizer

$$C(\widehat{y}) = \int_\Omega e^{-\lambda\rho(\|\widehat{y}-y\|)}\, dy,$$

which may depend on $\widehat{y}$ and must be included in the likelihood.

A rotation-valued emission can use the matrix Fisher density

$$p_F(R) = c(F)^{-1}\exp\{\text{tr}(F^\top R)\}, \qquad c(F) = \int_{SO(3)} \exp\{\text{tr}(F^\top R)\}\, d\mu(R), \tag{17}$$

where μ is normalized Haar measure. Compactness makes $c(F)$ finite for finite F; changing F also changes its normalizer. This is a possible rotation emission, not the local displacement loss in Eq. (11). Likewise, all pairwise clean displacements are constrained observations: multiplying independent residual densities in an ambient product space does not automatically yield a normalized density on the physical quotient. An anchor-coordinate density or a correctly normalized joint energy is needed for that interpretation.

These examples give precise scale meanings: the quadratic Gaussian weight is $1/(2\sigma^2)$ and the Laplace weight is $1/b$. Weighted categorical log scores require their own normalizer to define a powered probability model; generally $p(s \mid z)^\lambda / \sum_v p(v \mid z)^\lambda \neq p(s \mid z)^\lambda$. We therefore treat the coefficients of our multi-task distortion as loss scales and use the following exact generalized variational construction.

The implemented generalized variational bound. For fixed (X, Y), write $D_\theta(z; X, Y) = \lambda_{\text{seq}}\mathcal{L}_{\text{seq}} + \lambda_{\text{geo}}\mathcal{L}_{\text{geo}} \geq 0$. Let $p(z) = \mathcal{N}(0, I)$ be the normalized reference, $\beta > 0$, and assume D_θ is finite p-almost everywhere. Define

$$\mathcal{Z}_\beta(Y; X) = \int p(z) e^{-D_\theta(z;X,Y)/\beta}\, dz, \qquad \pi_\beta(z \mid X, Y) = \frac{p(z) e^{-D_\theta(z;X,Y)/\beta}}{\mathcal{Z}_\beta(Y; X)}. \tag{18}$$

Here $0 < \mathcal{Z}_\beta \leq 1$, so the loss defines a proper Gibbs distribution over z, including when the geometric loss is clipped. This is a loss-based variational construction (Bissiri et al., 2016).

Theorem A.1 (Generalized multi-task variational identity)**.** *For a density q with finite expected distortion and $D_{\text{KL}}(q\|p)$,*

$$\underbrace{\mathbb{E}_q D_\theta + \beta D_{\text{KL}}(q\|p)}_{\mathcal{F}_\beta(q;X,Y)} = -\beta \log \mathcal{Z}_\beta(Y; X) + \beta D_{\text{KL}}(q\|\pi_\beta). \tag{19}$$

Consequently $-\mathcal{F}_\beta/\beta \leq \log \mathcal{Z}_\beta$, with equality if and only if $q = \pi_\beta$ almost everywhere. Equation (13) is $\mathbb{E}_{X,Y}\mathcal{F}_\beta(q_\phi; X, Y)$.

Proof. Substituting Eq. (18) into the KL and expanding each term gives

$$\begin{aligned} D_{\text{KL}}(q\|\pi_\beta) &= \int q(z) \log \frac{q(z)\mathcal{Z}_\beta}{p(z)e^{-D_\theta(z)/\beta}}\, dz \\ &= D_{\text{KL}}(q\|p) + \beta^{-1}\mathbb{E}_q D_\theta + \log \mathcal{Z}_\beta. \end{aligned} \tag{20}$$

Rearranging proves the identity; non-negativity and the equality condition of KL prove the bound. □

The variational quantity here is the log partition score, which need not integrate to one over Y. If $D_\theta = -\log p_\theta(Y \mid z, X)$ for normalized emissions, $\beta = 1$, and the reference is the chosen generative prior, the same identity becomes the usual ELBO with all emission normalizers included. For our weighted clipped losses, Eq. (19) establishes the objective directly without changing its implementation. The zero-KL setting $\beta = 0$ is the distortion-only ablation; it is not covered by division by β. During positive-β annealing, the identity holds at each step with its current coefficient.

Connection to the information bottleneck. For normalized emissions and the joint defined above, the following two propositions recover the VIB rate and relevance bounds (Alemi et al., 2017). The rate identity remains valid for our stochastic encoder independently of the emission interpretation.

Proposition A.1 (Capacity control)**.** *Let $q_\phi(z) = \int q_\phi(z \mid \widetilde{X}) p(\widetilde{X})\, d\widetilde{X}$ be the aggregated posterior. Then, the expected Kullback-Leibler divergence satisfies:*

$$\mathbb{E}_{p(\widetilde{X})}\Big[D_{\text{KL}}(q_\phi(z \mid \widetilde{X}) \,\|\, p(z))\Big] = I(Z; \widetilde{X}) + D_{\text{KL}}(q_\phi(z) \,\|\, p(z)). \tag{21}$$

Proof. The proof follows directly by expanding the expected KL divergence and introducing the marginal $q_\phi(z)$:

$$\begin{aligned} \mathbb{E}_{p(\widetilde{X})}\left[\mathbb{E}_{q_\phi(z|\widetilde{X})}\left[\log \frac{q_\phi(z \mid \widetilde{X})}{p(z)}\right]\right] &= \mathbb{E}_{p(\widetilde{X}), q_\phi(z|\widetilde{X})}\left[\log \frac{q_\phi(z \mid \widetilde{X})}{q_\phi(z)} + \log \frac{q_\phi(z)}{p(z)}\right] \\ &= I(Z; \widetilde{X}) + D_{\text{KL}}(q_\phi(z) \,\|\, p(z)). \end{aligned} \tag{22}$$

□

Proposition A.2 (Reconstruction lower bound)**.** *For a normalized decoder $p_\theta(Y \mid Z)$ and finite relevant entropies, the clean-target mutual information satisfies:*

$$I_q(Z; Y) \geq \mathbb{E}_{q(Y,Z)}[\log p_\theta(Y \mid Z)] + H_q(Y). \tag{23}$$

Proof. By non-negativity of $\mathbb{E}_{q(Z)} D_{\mathrm{KL}}(q(Y \mid Z) \| p_\theta(Y \mid Z))$, the expected decoder log-likelihood is at most $\mathbb{E}_q \log q(Y \mid Z)$. Adding $H_q(Y)$ proves the bound. For continuous targets, densities and entropies are defined relative to the chosen base measure. The generalized bound in Appendix A.0 applies directly to the clipped reconstruction loss.

□

A.1 The quotient and maximal invariance

We consider complete, valid residue frames with fixed indexing and $N \geq 1$. We identify $g_i = (R_i, x_i)$ with the homogeneous matrix

$$T_i = \begin{pmatrix} R_i & x_i \\ 0 & 1 \end{pmatrix} \in SE(3), \qquad T = (T_1, \ldots, T_N) \in \mathcal{X} = SE(3)^N.$$

We remove a common world frame through $H \cdot T = (HT_1, \ldots, HT_N)$, $H \in SE(3)$, and write $\mathcal{Q} = \mathcal{X}/SE(3)$. Physical backbone constraints select a subset of this ambient configuration space.

Fix k and set $U_i = T_k^{-1} T_i$, with $U_k = I$. The global smooth change of variables

$$T \longmapsto \big(T_k, (U_i)_{i \neq k}\big) \tag{24}$$

has inverse $T_i = T_k U_i$. The group acts only on the first factor, so $\mathcal{Q}$ is diffeomorphic to $SE(3)^{N-1}$ and has dimension $6(N-1)$. These are ambient dimensions, not independent physical degrees of freedom for a constrained backbone.

We define $\Phi(T)_{ij} = T_i^{-1} T_j$ and let $\mathrm{Rel}(N)$ contain the arrays $A_{ij} \in SE(3)$ satisfying, for every i, j, k,

$$A_{ii} = I, \qquad A_{ji} = A_{ij}^{-1}, \qquad A_{ij} A_{jk} = A_{ik}. \tag{25}$$

Proposition A.3 (Quotient correspondence and maximal invariance)**.** *The relative-frame map obeys*

$$\begin{aligned} \Phi(H \cdot T) &= \Phi(T), \\ \Phi(T) = \Phi(T') &\iff \exists H \in SE(3) : T' = H \cdot T. \end{aligned} \tag{26}$$

It induces a bijection $\overline{\Phi} : \mathcal{Q} \to \mathrm{Rel}(N)$, which is a diffeomorphism under the submanifold structure given by the anchor coordinates.

Proof. We have $(HT_i)^{-1}(HT_j) = T_i^{-1} H^{-1} H T_j = T_i^{-1} T_j$. Conversely, equality of the two arrays implies $T_k^{-1} T_i = (T_k')^{-1} T_i'$ for every i. Thus $T_i' = H T_i$ with the same $H = T_k' T_k^{-1}$, proving maximality and injectivity on the quotient. For any consistent array A, set $T_k = I$ and $T_i = A_{ki}$. Its relations give $T_i^{-1} T_j = A_{ki}^{-1} A_{kj} = A_{ij}$, proving surjectivity. Group multiplication and inversion are smooth, and row k recovers the anchor coordinates, giving a smooth inverse. □

The cycle identities reduce an $N \times N$ array to $N - 1$ independent $SE(3)$ elements: $A_{ij} = A_{ki}^{-1} A_{kj}$. Thus $\dim \mathrm{Rel}(N) = 6(N-1)$ despite quadratic storage. This is a smooth, information-preserving coordinate description under a diagonal *left* action; it is not an adjoint action, nor a geodesic isometry without a separately specified metric. Physical bond and steric constraints further restrict the valid subset. Figure 3 illustrates this orbit correspondence.

A.2 Why Euclidean features preserve the required invariance

We distinguish geometric objects, numerical descriptors, and learned features. A relative frame is (Q_{ij}, t_{ij}), where $Q_{ij} = R_i^\top R_j$ and $t_{ij} = R_i^\top (x_j - x_i)$. Equation (5) gives

$$D_{ij} = \|t_{ij}\|_2, \qquad u_{ij} = \frac{t_{ij}}{D_{ij} + \varepsilon}, \qquad t_{ij} = (D_{ij} + \varepsilon) u_{ij}. \tag{27}$$

Thus the raw descriptors recover the complete relative transform, including at $D_{ij} = 0$ when the frames remain valid. Their equality is equivalent to equality of $\Phi(T)$, so they retain maximal invariance. Learned basis embeddings, MLPs, sparse selections, and the VIB latent need not retain injectivity.

Euclidean operations need not preserve a geometric manifold. The midpoint of $(1,0),(-1,0) \in S^1$ is $(0,0) \notin S^1$; likewise

$$I,\ \mathrm{diag}(-1,-1,1) \in SO(3), \qquad \tfrac{1}{2}\big(I + \mathrm{diag}(-1,-1,1)\big) = \mathrm{diag}(0,0,1) \notin SO(3).$$

We use invariant descriptor entries as scalar channels, without interpreting their averages or learned activations as rotations. Invariance concerns a function's response to transformed inputs; it does not require its features to lie in the input manifold.

Proposition A.4 (Invariance under Euclidean processing)**.** *Suppose all geometric inputs are functions of valid invariant descriptors and pose-independent masks, and the remaining inputs do not depend on global pose. Arbitrary deterministic Euclidean processing then preserves global-pose invariance. In particular,*

$$q_\phi(z \mid H \cdot \widetilde{G}, \widetilde{S}) = q_\phi(z \mid \widetilde{G}, \widetilde{S}),$$

and latent-marginal and posterior-mode sequence predictions are well-defined functions of $[G]$.

Proof. Let D collect descriptors and masks, so $D(H \cdot \widetilde{G}) = D(\widetilde{G})$. For any Euclidean map f, we have $f(D(H \cdot \widetilde{G})) = f(D(\widetilde{G}))$. We apply this identity successively to embeddings, scalar biases, row-wise attention, and posterior heads. The Gaussian means and variances agree, so the distributions agree. Integrating a shared sequence decoder against equal distributions preserves equality; evaluating it at their common mode does so as well. □

We use a Euclidean latent and a standard normal reference distribution, without latent charts, tangent coordinates, exponential/logarithmic maps, or a manifold-valued prior. Input construction enforces invariance; KL regularization controls rate. The proposition applies to transformed realized corrupted views, while isotropic corruption also respects the symmetry in distribution. Missing frames do not satisfy the completeness assumptions for maximality.

Stability of raw descriptors. Invariance, completeness on the quotient, and continuity under perturbations are distinct properties. The first two follow above. For the third, assume both original and perturbed frames are valid and write $\eta_i = \|x'_i - x_i\|$, $\omega_i = \|R'_i - R_i\|_2$. Direct subtraction gives

$$\begin{aligned} |D'_{ij} - D_{ij}| &\le \eta_i + \eta_j, \\ \|t'_{ij} - t_{ij}\| &\le \eta_i + \eta_j + \omega_i D_{ij}, \\ \|Q'_{ij} - Q_{ij}\|_F &\le \|R'_i - R_i\|_F + \|R'_j - R_j\|_F, \\ \|u'_{ij} - u_{ij}\| &\le \varepsilon^{-1}\|t'_{ij} - t_{ij}\|. \end{aligned} \tag{28}$$

The first bound is the reverse triangle inequality. For the second, expand $t'_{ij} - t_{ij} = R_i'^{\top}[(x'_j - x_j) - (x'_i - x_i)] + (R_i'^{\top} - R_i^{\top})(x_j - x_i)$. For the third, add and subtract $R_i'^{\top} R_j$ and use invariance of the Frobenius norm under orthogonal multiplication. Finally, the map $v \mapsto v/(\|v\| + \varepsilon)$ is globally $1/\varepsilon$-Lipschitz: its radial and tangential derivative magnitudes are $\varepsilon/(\|v\| + \varepsilon)^2$ and $1/(\|v\| + \varepsilon)$, with the derivative at zero given by I/ε. Integrating along a segment proves the last bound. The constants quantify continuity; they need not be small for small ε or large separations.

These bounds condition on valid frames. Stability of atom-to-frame construction additionally requires nondegenerate defining atoms, and stability of output orthonormalization requires the conditions of Proposition A.7. Neither raw continuity nor the quotient correspondence alone proves robustness of the learned predictor.

A.2.1 Bypass suppression through synchronized masking

We formalize a co-located shortcut as a direct observation channel at a residue whose sequence identity is hidden. We condition on a fixed trained encoder; its internal random seed is independent of the evaluated sample. Masks are data-independent, and an erased channel is replaced by a fixed symbol together with its visibility flag.

Independent masking leaves a local observation channel. Let V_i be a local geometric observation (possibly noisy), and let $M_i^{(S)}, M_i^{(G)}$ be mutually independent Bernoulli$(1-p)$ variables, independent of (S_i, V_i), with $0 < p < 1$. Define $E_i = V_i$ when $M_i^{(G)} = 1$ and $E_i = \perp$ otherwise, and write $\widetilde{V}_i = (M_i^{(G)}, E_i)$ for the observation with its flag. Then

$$\begin{aligned} \Pr(M_i^{(S)} = 0, M_i^{(G)} = 1) &= p(1-p), \\ I(S_i; \widetilde{V}_i \mid M_i^{(S)} = 0) &= I(S_i; M_i^{(G)}) + I(S_i; E_i \mid M_i^{(G)}) \\ &= (1-p) I(S_i; V_i). \end{aligned} \tag{29}$$

The first term vanishes by mask independence; the erased branch is constant and the visible branch reveals V_i. The channel carries positive information precisely when $I(S_i; V_i) > 0$. Network capacity alone cannot imply a positive lower bound: an encoder may discard information, and some local observations may be independent of the target.

Theorem A.2 (Erasure of the direct co-located channel)**.** *Let $M^{(S)} = M^{(G)} = M$, and fix $M = m$ with hidden set $\mathcal{M} = \{i : m_i = 0\}$. Suppose all inputs derived from hidden sequence or geometric evidence are erased, including incident pair descriptors and any frames constructed using hidden atoms. Then*

$$p(\widetilde{S}_{\mathcal{M}}, \widetilde{G}_{\mathcal{M}} \mid S, G, M = m) = \delta_{\texttt{[MASK]}^{|\mathcal{M}|}} \, \delta_{\perp^{|\mathcal{M}|}}, \qquad I(S_{\mathcal{M}}; \widetilde{S}_{\mathcal{M}}, \widetilde{G}_{\mathcal{M}} \mid M = m) = 0. \tag{30}$$

Let C_m collect all remaining visible inputs. If Z is computed only from (C_m, m) and independent encoder randomness, then

$$I(S_{\mathcal{M}}; Z \mid M = m) \le \min\{I(S_{\mathcal{M}}; C_m \mid M = m),\ I(C_m; Z \mid M = m)\}. \tag{31}$$

Proof. For the fixed mask, both hidden observation blocks are constants, giving the point mass and zero mutual information. By the encoder assumption, $p(z \mid S_{\mathcal{M}}, C_m, M = m) = p(z \mid C_m, M = m)$. Hence $S_{\mathcal{M}} \to C_m \to Z$ is a conditional Markov chain. Applying data processing in each direction yields both upper bounds in Eq. (31). □

Synchronization is a sufficient erasure rule, not a necessary one: any coupling with no visible co-located channel when the sequence is hidden has the same property. The theorem concerns the explicitly erased channel. Visible neighbors can retain local structural information, so it does not assert that all useful context must be distant, that global topology is necessarily learned, or that overfitting is eliminated. Our independent-mask control tests the empirical value of this particular channel removal (Table 6). At standard inverse-folding inference, complete target geometry remains legitimate conditioning information.

A.2.2 Variational MAP decoding and its approximation error

Flatten the latent to $z \in \mathbb{R}^{d_z}$, $d_z = Nd$, and let $X_\star = (G, \texttt{[MASK]}^N)$. With positive diagonal covariance Σ_ϕ, the Gaussian log density is

$$\log q_\phi(z \mid X_\star) = c(X_\star) - \tfrac{1}{2}(z - \mu_\phi)^\top \Sigma_\phi^{-1}(z - \mu_\phi). \tag{32}$$

The quadratic is uniquely minimized at $z = \mu_\phi$, proving that this is the variational mode. It is the MAP of the Gaussian encoder distribution, not necessarily the mode of a generative model's exact posterior.

Proposition A.5 (Mean-decoding approximation bound)**.** *Fix an output sequence S and let $f(z) = \log p_{\theta_s}(S \mid z)$. Suppose $f \in C^2(\mathbb{R}^{d_z})$ and $\|\nabla^2 f(z)\|_2 \le L_H$ on all line segments from μ to points in the support of $q = \mathcal{N}(\mu, \Sigma)$. Then*

$$|\mathbb{E}_q f(z) - f(\mu)| \le \frac{L_H}{2} \operatorname{tr} \Sigma. \tag{33}$$

If the Hessian is also L_3-Lipschitz on these segments, then

$$\mathbb{E}_q f(z) = f(\mu) + \tfrac{1}{2} \operatorname{tr}(\nabla^2 f(\mu)\Sigma) + \mathcal{R}, \qquad |\mathcal{R}| \le \frac{L_3}{6} \mathbb{E}_q \|z - \mu\|_2^3. \tag{34}$$

Proof. Set $\delta = z - \mu$, so $\mathbb{E}\delta = 0$ and $\mathbb{E}\delta\delta^\top = \Sigma$. Taylor's formula with integral remainder gives

$$f(\mu + \delta) = f(\mu) + \nabla f(\mu)^\top \delta + \int_0^1 (1-t)\delta^\top \nabla^2 f(\mu + t\delta)\delta \, dt. \tag{35}$$

The linear term has zero expectation. Bounding the quadratic form by $L_H\|\delta\|^2$ and using $\int_0^1(1-t)dt = 1/2$ proves Eq. (33). For Eq. (34), add and subtract $\nabla^2 f(\mu)$ inside the integral. The constant part has expectation $\frac{1}{2}\operatorname{tr}(\nabla^2 f(\mu)\Sigma)$. The remaining integrand has magnitude at most $L_3 t(1-t)\|\delta\|^3$; its integral is $L_3\|\delta\|^3/6$. □

The Gaussian has unbounded support; a bound only in a small neighborhood requires an additional tail estimate. These are conditional approximation results, not measured bounds for our trained network. They show that mean decoding preserves the expected first-order Taylor approximation, with a second-order discrepancy controlled by curvature and covariance. They concern the expected log score; Jensen's gap between $\mathbb{E}_q \log p(S \mid z)$ and $\log \mathbb{E}_q p(S \mid z)$ is a separate quantity.

What KL regularization controls. For a standard normal reference,

$$D_{\mathrm{KL}}(\mathcal{N}(\mu, \Sigma)\|\mathcal{N}(0, I)) = \tfrac{1}{2}\big(\|\mu\|^2 + \operatorname{tr}\Sigma - \log\det\Sigma - d_z\big). \tag{36}$$

Its covariance-only minimum is at $\Sigma = I$, not at zero. For a pointwise KL value K, the scalar inequality $u - \log u - 1 \geq u/2 - \log 2$ gives $\operatorname{tr}\Sigma \leq 4K + 2d_z \log 2$. This limits covariance growth at a fixed rate budget, but neither forces small variance nor controls decoder curvature. We do not assume that training makes $\sigma_\phi^2 \ll 1$ without measuring it.

We use the mode as a stable, reproducible conditioning value and retain stochastic reparameterized training. Conditional PLM sampling then controls sequence diversity at a stated temperature (Figure 4). Removing latent sampling eliminates its Monte Carlo variation, but the nonlinear plug-in score can be biased relative to the latent expectation. The reported foldable fraction measures the complete design pipeline; it does not establish an optimality theorem for mode decoding or isolate its effect from other components.

A.3 Geometric validity of the auxiliary outputs

We constrain the auxiliary frame head through the six-dimensional construction of Zhou et al. (2019):

$$\begin{aligned} r_1 &= a/\|a\|_2, \qquad v_2 = b - (r_1^\top b)r_1, \\ r_2 &= v_2/\|v_2\|_2, \qquad r_3 = r_1 \times r_2, \qquad \Gamma(a,b) = [r_1, r_2, r_3]. \end{aligned} \tag{37}$$

We state the exact guarantee on $\mathcal{U} = \{(a,b) : a \neq 0,\ v_2 \neq 0\}$.

Proposition A.6 (Rotation- and frame-valued reconstruction)**.** *For $(a_i, b_i) \in \mathcal{U}$ and $\widehat{x}_i \in \mathbb{R}^3$,*

$$\widehat{R}_i = \Gamma(a_i, b_i) \in SO(3), \qquad \widehat{T}_i = \begin{pmatrix} \widehat{R}_i & \widehat{x}_i \\ 0 & 1 \end{pmatrix} \in SE(3).$$

Frame-derived relative outputs $\widehat{A}_{ij} = \widehat{T}_i^{-1}\widehat{T}_j$ belong to $SE(3)$ and satisfy Eq. (25).

Proof. We have $\|r_1\| = \|r_2\| = 1$ and $r_1^\top r_2 = 0$. Their cross product is orthogonal to both, has unit norm, and gives a right-handed basis. Hence $\widehat{R}_i^\top \widehat{R}_i = I$ and $\det \widehat{R}_i = (r_1 \times r_2)^\top r_3 = 1$. This proves membership in $SO(3)$ and then $SE(3)$. Group closure gives relative-frame membership, while cancellation gives $\widehat{A}_{ij}\widehat{A}_{jk} = \widehat{T}_i^{-1}\widehat{T}_k = \widehat{A}_{ik}$. Identity and inverse relations follow similarly. □

Proposition A.7 (Smoothness and stability away from degeneracy)**.** *The map $\Gamma : \mathcal{U} \subset \mathbb{R}^6 \to SO(3)$ in Eq. (37) is real analytic. At any point where $\|a\| \geq \delta > 0$, $\|v_2\| \geq \delta$, and $\|b\| \leq B$, its differential satisfies*

$$\|d\Gamma\|_F \leq 2\left[\left(\delta^{-1} + 2B\delta^{-2}\right)\|da\| + \delta^{-1}\|db\|\right]. \tag{38}$$

Consequently it is locally Lipschitz away from zero and collinear inputs, with a uniform derivative bound on compact subsets of $\mathcal{U}$.

Proof. Normalization $v \mapsto v/\|v\|$ is real analytic for $v \neq 0$ and has derivative $(I - uu^\top)/\|v\|$, where $u = v/\|v\|$. Hence $\|dr_1\| \leq \|da\|/\delta$. Writing $v_2 = (I - r_1 r_1^\top)b$ gives $\|dv_2\| \leq \|db\| + 2B\|da\|/\delta$ and $\|dr_2\| \leq \|db\|/\delta + 2B\|da\|/\delta^2$. Finally $dr_3 = dr_1 \times r_2 + r_1 \times dr_2$, so $\|d\Gamma\|_F \leq \|dr_1\| + \|dr_2\| + \|dr_3\| \leq 2(\|dr_1\| + \|dr_2\|)$. All operations are real analytic on $\mathcal{U}$. Integrating the derivative gives a perturbation bound along paths that remain in a region with the stated lower bounds. □

The excluded set has Lebesgue measure zero: for each $a \neq 0$, collinear b form a one-dimensional subspace of $\mathbb{R}^3$, and $a = 0$ is itself null. An absolutely continuous six-dimensional output avoids exact degeneracy almost surely; a learned deterministic head need not satisfy that distributional assumption. The construction avoids choosing Euler-angle coordinates or a quaternion sign, but its derivatives can grow near the excluded set. Continuity does not give a global guarantee against large gradients.

This constraint acts at the output, not on hidden features. Near-zero or collinear vectors require explicit numerical handling; adding a positive constant only to a normalization denominator does not give exact orthonormality. We do not extend this guarantee to separately regressed $\widehat{D}_{ij}, \widehat{O}_{ij}$ supervision channels, which need not satisfy rotation or cycle constraints. The frame-derived construction identifies which outputs have exact geometric consistency. Valid $SE(3)$ frames alone also do not enforce bond lengths, bond angles, or steric feasibility.

The sequence head maps logits to the categorical probability simplex through softmax. FAPE instead compares local-frame displacements: a common rigid motion of all predicted frames cancels inside $\widehat{R}_i^\top(\widehat{x}_j - \widehat{x}_i)$, and the same holds separately for target frames. Thus the invariant decoder need not reproduce an arbitrary target world-coordinate gauge. We use this head during training and an external folding oracle for evaluation.

A.4 Why use all pairs when an anchor is sufficient?

Proposition A.8 (Linear-size sufficiency and pairwise reconstruction)**.** *For fixed k, the $N-1$ anchor-relative transforms $U_i = T_k^{-1}T_i$, $i \neq k$, determine the same quotient element as the full array through*

$$A_{ij} = U_i^{-1}U_j, \qquad U_k = I. \tag{39}$$

Relative transforms on any connected graph are also sufficient.

Proof. Equation (39) follows by cancellation, and row k recovers the U_i. On a connected graph, multiply edge transforms along a path from k to i, using inverses when traversing an edge backwards. The product telescopes to $T_k^{-1}T_i$, independent of the chosen path for exact transforms. In particular, chain transforms $C_i = T_i^{-1}T_{i+1}$ reconstruct $A_{ij} = C_i \cdots C_{j-1}$ for $i < j$. □

Absolute transforms T_i depend on the world frame; anchor-relative or connected-edge transforms remove that dependence. We choose all pairs for the attention interface, not for additional quotient information. They give an anchor-free tensor with a direct entry for each B_{ij}. Changing the anchor from k to m replaces every U_i by $U_m^{-1}U_i$, whereas the full relative array is unchanged. Residue relabeling permutes its rows and columns. Modifying only T_k changes only incident pairs in the raw array, although subsequent attention can propagate the effect globally.

At fixed feature widths, anchor or spanning-tree storage is $O(N)$, while storing all pairs is $O(N^2)$. Forming every dense bias from anchor coordinates still requires N^2 pair queries. Explicit pair features and materialized dense-attention scores therefore share a quadratic pair count; dense attention also takes $O(N^2)$ time at fixed width. Their constants and total costs differ: fused attention can avoid storing a full score matrix, while an explicit geometric bias remains quadratic. Sparsity and on-demand biases are implementation alternatives, not mathematically inferior representations.

B TRAINING ALGORITHM

Algorithm 1 Training GeoSeqAlign with Synchronized Corruption and Geometry-Aware Attention

Input: Paired dataset $\mathcal{D} = \{(G, S)\}$, frozen PLM f_{PLM}, trainable structure encoder $\mathcal{E}_{\mathrm{geo}}$, attention adapter, variational posterior q_ϕ, prefix adapter, and output heads.

For each minibatch $\{(G, S)\}_{b=1}^{B} \sim \mathcal{D}$:

1. Sample synchronized corruption seed ω and generate a shared Bernoulli residue mask M.
2. Compute jointly corrupted views $(\widetilde{G}, \widetilde{S}) \sim q(\widetilde{G}, \widetilde{S} \mid G, S)$.
3. Encode structure invariants: $(H_{\mathrm{geo}}, E_{\mathrm{geo}}) = \mathcal{E}_{\mathrm{geo}}(\widetilde{G})$.
4. Compute pairwise geometric descriptors D_{ij}, O_{ij} and map to scalar bias B_{geo}.
5. Embed sequence $H_{\mathrm{sem}} = \mathcal{E}_{\mathrm{sem}}(\widetilde{S})$.
6. Modulate PLM attention logits using the geometric bias:
   $A = \mathrm{Softmax}\left(QK^\top/\sqrt{d} + B_{\mathrm{geo}} + B_{\mathrm{mask}}\right)$.
7. Infer posterior parameters μ_ϕ, σ_ϕ from the aligned representations.
8. Sample latent representation $z \sim \mathcal{N}(\mu_\phi, \mathrm{diag}(\sigma_\phi^2))$ via the reparameterization trick.
9. Decode sequence logits and predict auxiliary geometry using f_{PLM} conditioned on z.
10. Compute $\mathcal{L}_{\mathrm{seq}}$ and $\mathcal{L}_{\mathrm{geo}}$.
11. Update trainable parameters by minimizing:
    $\mathcal{L} = \lambda_{\mathrm{seq}}\mathcal{L}_{\mathrm{seq}} + \lambda_{\mathrm{geo}}\mathcal{L}_{\mathrm{geo}} + \beta D_{\mathrm{KL}}(q_\phi(z \mid \widetilde{G}, \widetilde{S}) \,\|\, p(z))$.

C METRIC DEFINITIONS

Perplexity (PPL). For a native sequence $S = (s_1, \ldots, s_N)$ and target G, the reported decoder perplexity uses $z = \mu_\phi(G, \texttt{[MASK]}^N)$; native residues are scoring targets, not encoder inputs. This is the deterministic decoder score, rather than an exact marginal over stochastic latents:

$$\mathrm{PPL}(S) = \exp\left(-\frac{1}{N}\sum_{i=1}^{N}\log p_\theta(s_i \mid z)\right). \tag{40}$$

Amino-acid recovery (AAR). AAR denotes the exact match accuracy between the generated sequence $\widehat{S}$ and the native sequence S:

$$\mathrm{AAR} = \frac{100}{N}\sum_{i=1}^{N}\mathbb{I}[\widehat{s}_i = s_i]. \tag{41}$$

Self-consistency metrics (scTM and scRMSD). Given a designed sequence $\widehat{S}$, the ESMFold oracle predicts its 3D geometry $\widehat{G}_{\mathrm{oracle}}$. We compute the TM-score (scTM) and RMSD (scRMSD) using structural alignment between $\widehat{G}_{\mathrm{oracle}}$ and the original target backbone G.

Repetitiveness (Rep). Rep is a within-sequence statistic. Let $\mathcal{G}_n(S)$ be the multiset of its $N-n+1$ contiguous n-gram occurrences, and let $\mathrm{count}_S(g)$ count occurrences in S. Then the fraction of occurrences belonging to repeated n-gram types is:

$$\mathrm{Rep} = \frac{\sum_{g \in \mathcal{G}_n(S)}\mathbb{I}[\mathrm{count}_S(g) > 1]}{|\mathcal{G}_n(S)|}. \tag{42}$$

D LIMITATIONS AND BROADER IMPACT

Limitations. GeoSeqAlign fundamentally relies on a fixed-backbone assumption. Biological proteins are intrinsically dynamic, and our current formulation does not explicitly model conformational

ensembles, side-chain χ-angle flexibilities, ligand–protein interactions, or lipid membrane constraints. Furthermore, computing the fully dense pairwise geometric bias matrix B_{geo} scales quadratically with sequence length ($\mathcal{O}(N^2)$), which may pose memory constraints for ultra-large protein complexes. The functional evaluation is retrospective, and computational proxies do not establish prospective experimental activity. Matched ADFLIP/CoFlow comparisons, oracle-specific uncertainty estimates, and architecture-scaling controls remain incomplete. Threshold-based overlap filters cannot rule out remote homology or unknown PLM pretraining overlap.

The present conditional distribution is $p(S_b \mid G_b)$, not $p(S_b \mid G_b, G_t, S_t)$ for a binder and its target. A partner-conditioned extension would retain target residues and add chain identities and inter-chain edges. A faithful evaluation would use structure-disjoint complexes, binder-only and multichain baselines, interface recovery, ipTM, DockQ, clashes, diversity, target swaps, and an inter-chain-edge ablation. No partner-conditioned results are claimed here.

Broader impact. The ability to accurately design protein sequences conditioned on desired geometries has profound implications for accelerating therapeutic antibody discovery, optimizing industrial enzymes, and synthesizing novel biomaterials. Conversely, dual-use concerns exist: the same generative capabilities could theoretically be misappropriated to engineer pathogenic traits or toxic proteins. We urge the research community to enforce rigorous sequence screening procedures (e.g., against known pathogenic sequence databases), control access to high-risk structural targets, and adhere to emerging international biosecurity frameworks before executing physical DNA synthesis.

E Reproducibility Appendix

We describe the preprocessing, checkpoint labels, sampling, oracle settings, and hardware environment. For experiments repeated with seeds $\{0, 1, 2\}$, we report mean $\pm$ standard deviation in the corresponding tables. We report deterministic point evaluations across these extended ablation and sensitivity settings. For main benchmark comparisons, empirical standard deviations across three independent random seeds are reported in the respective tables.

E.1 Sequence-identity filtering with MMseqs2

To reduce sequence overlap, all train, validation, and test splits are filtered at a 30% sequence identity threshold. We employ MMseqs2 (v15.6f452) across the unified pool of CATH+ 4.3, TS50, and CASP15/TS45 chains.

Clustering and split verification. The target sequences are clustered using the following parameters:

```
mmseqs cluster \
  work/mmseqs/all_chains_db \
  work/mmseqs/all_chains_cluster30 \
  work/mmseqs/tmp \
  --min-seq-id 0.30 \
  -c 0.80 \
  --cov-mode 0 \
  --threads 64
```

We employ a custom validation script (`verify_split_leakage.py`) to check that no training chain shares a cluster identifier with any target in the test or out-of-distribution (OOD) sets.

E.2 Model checkpoints and source revisions

The baseline evaluation uses released weights where available. When official codebases lack a fixed-target evaluation interface, we introduce a minimal I/O wrapper without altering the model's internal computation graph.

Table 9: **Checkpoint manifest for evaluated baselines.** Revision strings are descriptive release/wrapper labels, not immutable commit hashes.

Model	Source	Relative Checkpoint Path	Code Revision
Struct. Trans.	Rerun	`ckpt/struct_trans/model.pt`	`local-unified-eval-v1`
GVP-GNN	Official	`ckpt/gvp_gnn/model.pt`	`gvp-pytorch-compatible`
ProteinMPNN	Official	`ckpt/proteinmpnn/vanilla.pt`	`proteinmpnn-2023-02`
PiFold	Official	`ckpt/pifold/model.pt`	`pifold-official-release`
ESM-IF1	Official	`ckpt/esm_if1/model.pt`	`fair-esm-2.0.0`
SaProt	Official	`ckpt/saprot/model.pt`	`saprot-official-release`
GRADE-IF	Official	`ckpt/grade_if/model.pt`	`grade-if-unified-eval`
SPDesign	Official	`ckpt/spdesign/model.pt`	`spdesign-unified-eval`
BC-Design	Official	`ckpt/bc_design/model.pt`	`bc-design-unified-eval`
MapDiff	Official	`ckpt/mapdiff/model.pt`	`mapdiff-unified-eval`
GeoSeqAlign	Ours	`ckpt/geoseqalign/model.pt`	`geoseqalign-v1`

E.3 Sampling and decoding settings

We use greedy decoding ($T = 0.0$) for single-sample AAR and sample eight sequences per backbone at $T = 0.10$ for Best@8. This fixed budget allows a controlled comparison of candidate-set recovery at a modest computational cost. For each backbone, we compute the maximum native-sequence recovery within its eight samples and then average these maxima over backbones. Because this selection uses the native sequence, we report Best@8 alongside greedy AAR and assess generative diversity separately with 50 samples per backbone.

Representative Decoding Command (Best@8).

```
python scripts/sample_inverse_folding.py \
  --model geoseqalign \
  --checkpoint ckpt/geoseqalign/model.pt \
  --split data/splits/cath43_test.txt \
  --num-samples 8 \
  --temperature 0.10 \
  --seed 0 \
  --out-fasta results/cath43/geoseqalign/seed0_best8.fasta
```

Robustness Evaluation Setup. Coordinate noise is injected *exclusively* into the input backbone parsed by the inverse folder ($\sigma \in \{0.0, 0.5, 1.0, 2.0\}$ Å). The pristine, uncorrupted backbone is always preserved for the final scTM and scRMSD oracle evaluation.

E.4 Folding oracle and structural scoring

Table 10: **ESMFold Oracle Specifications.**

Parameter	Configuration	Parameter	Configuration
Oracle Model	ESMFold (`esmfold_v1`)	Max Tokens/Batch	4096
Package	`fair-esm==2.0.0`	Chunk Size	128
Hardware	NVIDIA A100-80GB	Precision	FP16 (Inference)
PyTorch Ver.	`2.3.1` (CUDA 12.1)	Metric Accum.	FP32

Structural metrics (pLDDT, PAE) are extracted directly from the ESMFold output heads. Topological consistency (scTM, scRMSD) is computed via rigid alignment using a custom Python implementation interfacing with Biopython.

E.5 Rosetta $\Delta\Delta G$ and ProteinGym DMS scoring

Rosetta (`v2023.35`) serves as an auxiliary thermodynamic proxy. Structures undergo FastRelax before a Cartesian energy comparison with `ref2015_cart` (Alford et al., 2017):

$$\Delta\Delta G_{\mathrm{REU}} = \min_{r\in\mathcal{R}_{\mathrm{mut}}} E_{\mathrm{ref2015_cart}}(r) - \min_{r\in\mathcal{R}_{\mathrm{WT}}} E_{\mathrm{ref2015_cart}}(r)$$

For the 12 ProteinGym substitution assays (Notin et al., 2023), we rank variants by differences of the same deterministic, backbone-conditioned decoder log-scores; we provide neither native nor variant sequences to the encoder:

$$\Delta \log p(S') = \log p(S' \mid G_{\mathrm{WT}}) - \log p(S_{\mathrm{WT}} \mid G_{\mathrm{WT}})$$

E.6 Hardware environment

The reported environment is Ubuntu 22.04.4 LTS. Additional evaluations were distributed across 32 A100-80GB GPUs for parallel completion; single-device timing was measured separately. Deep learning frameworks rely on PyTorch 2.3.1 configured with CUDA 12.1 and cuDNN 8.9. All timing and latency measurements (reported in Table 7) are conducted on a single, isolated NVIDIA A100-80GB GPU. Latency is recorded strictly after 20 warm-up cycles and averaged over 100 synchronized forward passes to eliminate initialization artifacts.

F Additional controlled evaluations

We extend the main evaluation with analyses of diversity, structural overlap, folding-oracle sensitivity, generation cost, and KL regularization. We report aggregate point estimates and interpret the differences descriptively; we do not derive uncertainty intervals from summary values alone.

F.1 Cross-sample diversity and structural self-consistency

We evaluate $K = 50$ generated sequences for each of 100 pre-specified test backbones. For backbone b, exact uniqueness is $100|\{S_b^{(k)}\}_{k=1}^{K}|/K$. Mean pairwise identity averages the fraction of position-wise matches over all $\binom{K}{2}$ unordered sequence pairs within that backbone, then averages over backbones. A sample is called foldable in these tables when its oracle scTM against the target is at least 0.5; this is an operational in-silico threshold.

Table 11: **Cross-sample quality and diversity.** Fifty sequences per backbone, averaged over 100 backbones. Cluster counts in the last column use an 80% identity threshold among foldable samples.

Model	Unique (%)	Pairwise identity	Foldable (%)	Foldable clusters / 50
ProteinMPNN	99.8	0.35	88.2	21.4
MapDiff	99.5	0.32	82.1	23.1
Bridge-IF	97.6	0.41	91.5	16.2
GeoSeqAlign	99.2	0.38	93.4	18.2

Let C_τ denote the per-backbone cluster count at sequence-identity threshold τ, averaged over backbones. The multi-threshold analysis uses the same saved sequences without resampling. Entries in Table 12 give all-sample counts / counts after retaining foldable samples; the clustering threshold is applied within each backbone rather than across unrelated targets.

GeoSeqAlign has the largest reported foldable fraction in Table 11, but ProteinMPNN and MapDiff have more foldable clusters at 80% identity. At 50%, MapDiff has 3.4 versus our 3.3 foldable clusters. We therefore report a quality–diversity trade-off, not universal diversity superiority. These measurements describe variation among sequences conditioned on a shared backbone. They do not measure novelty relative to the training set, nor do they establish functional diversity. Exact reproduction of C_τ additionally requires the clustering algorithm, linkage rule, and saved sample set; the threshold alone does not specify a unique clustering.

Table 12: **Multi-threshold diversity.** All / foldable cluster counts. Exact uniqueness and C_{80} agree with the preceding analysis.

Model	Unique (%)	C_{50}	C_{60}	C_{70}	C_{80}	C_{90}
ProteinMPNN	99.8	2.5/2.1	5.2/4.4	10.9/9.2	23.1/21.4	36.2/31.0
MapDiff	99.5	3.8/3.4	8.2/7.3	15.6/13.9	25.5/23.1	41.5/36.8
KW-Design	98.9	2.9/2.5	6.4/5.5	12.8/11.0	22.0/19.2	36.8/31.8
Bridge-IF	97.6	3.1/2.6	6.9/5.9	13.6/11.5	18.0/16.2	36.0/31.0
GeoSeqAlign	99.2	3.5/3.3	7.8/7.3	13.5/12.6	19.8/18.2	36.0/33.0

F.2 Structure-filtered evaluation

In addition to sequence-identity filtering, Foldseek (van Kempen et al., 2024) was used to remove targets whose best training-structure hit had TM-score ≥ 0.5. The filter retained 321 of 500 pre-specified held-out targets (64.2%). This 500-target pool is a separate evaluation; it is not identified with the 45 CASP15/TS45 targets.

Table 13: **Results on the 321 structure-filtered targets.** AAR is in percent.

Model	AAR	scTM
ProteinMPNN	46.2	0.791
KW-Design	51.8	0.845
GeoSeqAlign	53.2	0.869

The observed margin over KW-Design is 1.4 AAR points and 0.024 scTM. The filter reduces detectable structural overlap with the searched training collection; it cannot certify absence of every remote structural/evolutionary relationship or overlap with external PLM pretraining data.

F.3 Folding-oracle sensitivity

We rescore the same generated sequences on the same targets with AlphaFold2 and Chai-1, without oracle-specific resampling. We observe aggregate differences of 0.018 scTM and 5.2 percentage points in success rate relative to Bridge-IF. We report these pooled summaries without per-oracle scores, uncertainty intervals, or an explicit success threshold, and therefore use them as an exploratory sensitivity check. We base the main structural comparisons on the explicitly defined ESMFold scTM evaluation; the pooled results do not establish experimental foldability.

F.4 Full generation cost

All generation components, including frozen PLMs, are included in resident parameter counts and FLOPs. One multiply–accumulate counts as two FLOPs. At batch size 1 and $N = 200$, the evaluated trajectories use 25 bridge steps for Bridge-IF, five refinement cycles for LM-Design, all evaluated KW-Design refinement operations, and one PLM decoding pass for GeoSeqAlign. The common folding oracle is excluded. Timing uses one isolated A100-80GB, 20 warm-ups, and 100 CUDA-synchronized runs.

Table 14: **Full-trajectory computation.** Approximately 700M is the total resident parameter count for GeoSeqAlign; 45M are trainable.

Model	Total	Trainable	GFLOPs/sequence	Latency (ms)
ProteinMPNN	1.9M	1.9M	8	28
PiFold	6.6M	6.6M	18	15
LM-Design	664M	11.9M	1,350	181
KW-Design (ESM2-650M)	678M	18.0M	2,200	309
Bridge-IF	698M	42.0M	6,500	242
GeoSeqAlign	700M	45.0M	380	225

The reported generation FLOPs for GeoSeqAlign are approximately 28% of LM-Design's and 6% of Bridge-IF's, but latency is not proportional to FLOPs. Pairwise geometry and adapter operations have different hardware utilization. At batch size 1, 1,000 sequences take approximately 225 s for GeoSeqAlign versus 28 s for ProteinMPNN, excluding folding. Thus parameter-efficient adaptation should not be interpreted as lightweight inference or a throughput advantage over graph models.

F.5 KL SENSITIVITY AND RATE–PERFORMANCE TRADE-OFF

The coefficient sweep uses $\beta \in \{0, 0.01, 0.05, 0.10, 0.20\}$. The reported empirical rate $\mathbb{E}_X D_{\mathrm{KL}}(q_\phi(z \mid X) \| p(z))$ decreases monotonically from 4.12 nats at $\beta = 0$ to 1.85 nats at $\beta = 0.20$ (55.1%). Performance is non-monotonic: the best reported robustness occurs at $\beta = 0.05$ (CASP AAR 42.4; scTM@2Å 0.48), compared with 40.2 and 0.42 at $\beta = 0$. At $\beta = 0.20$, CASP AAR is 39.1.

We report the evaluated sweep values without interpolating intermediate rates or metrics. We use the $\beta = 0.10$ annealing schedule for the main training description and treat the 0.05 result as a separate sensitivity setting. We do not infer checkpoint identity from agreement between rounded values. Direct reproduction and comparisons across latent dimensions require the rate normalization and the configuration of each run.

G ILLUSTRATIONS OF GEOMETRIC CONDITIONING

We use the following panels to connect the geometric definitions, model interfaces, and evaluation views. Figures 3 and 4 distinguish the quotient map from the neural implementation. Figures 5 and 6 serve as visual illustrations; we base quantitative comparisons on the experiment tables.

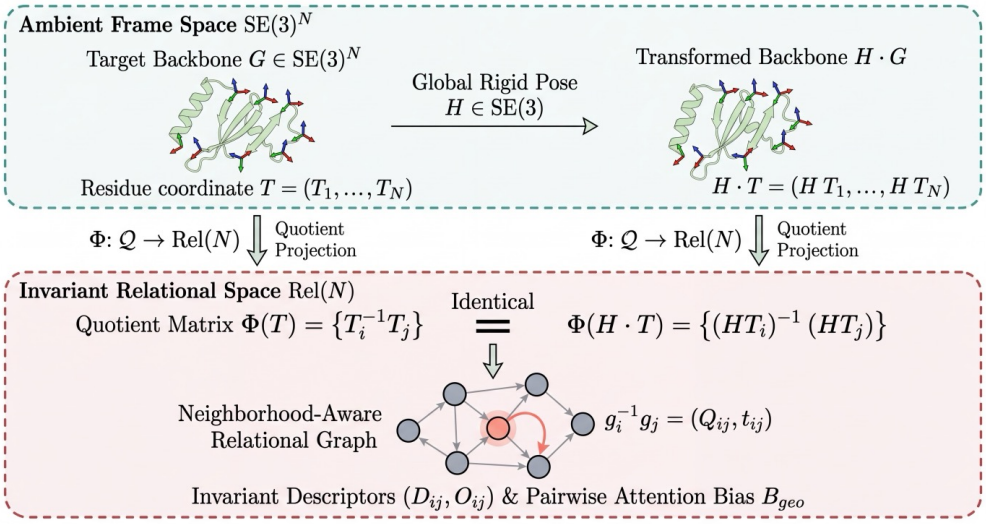


Figure 3: **Global-pose orbits and invariant relative frames.** We identify the two ambient configurations T and $H \cdot T$ through the same relative array. Formally, the quotient projection is $\pi : \mathcal{X} \to \mathcal{Q}$, the induced correspondence is $\bar{\Phi} : \mathcal{Q} \to \mathrm{Rel}(N)$, and the displayed ambient-to-relative arrows are their composition $\Phi = \bar{\Phi} \circ \pi$. The raw relative array is complete up to global pose; learned descriptors may compress it. Appendix A.1 proves the correspondence and its cycle constraints.

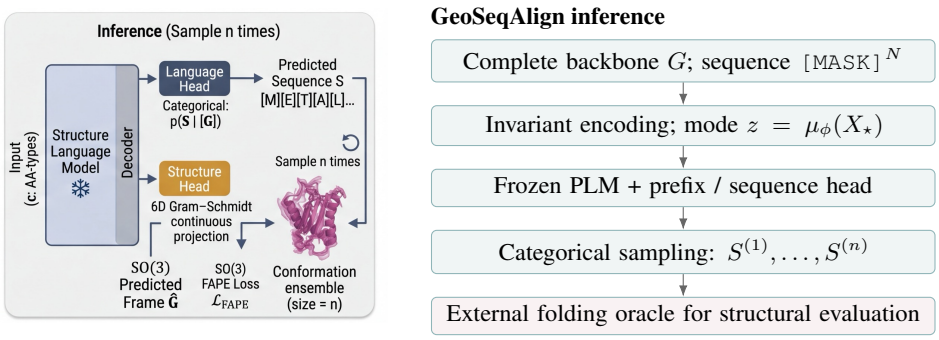


Figure 4: **Joint-output concept and implemented inference.** The left panel depicts a joint sequence/structure concept. The right panel gives our fixed-backbone inference: a frozen protein LM generates sequences, which an external oracle folds for scoring.

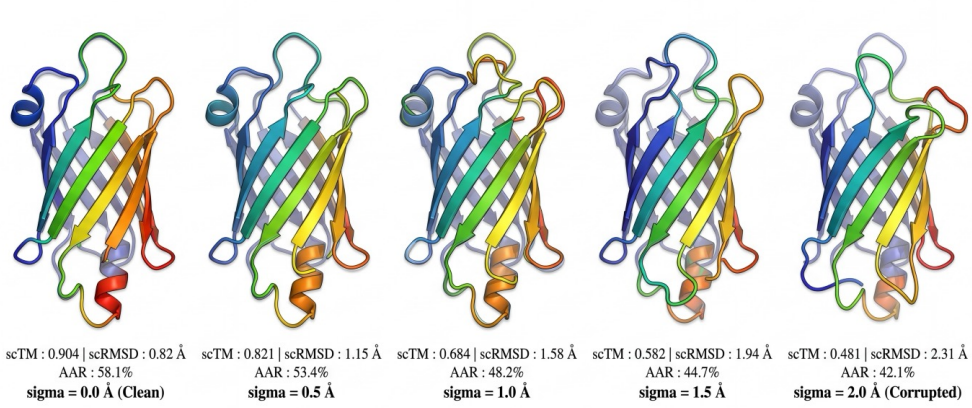
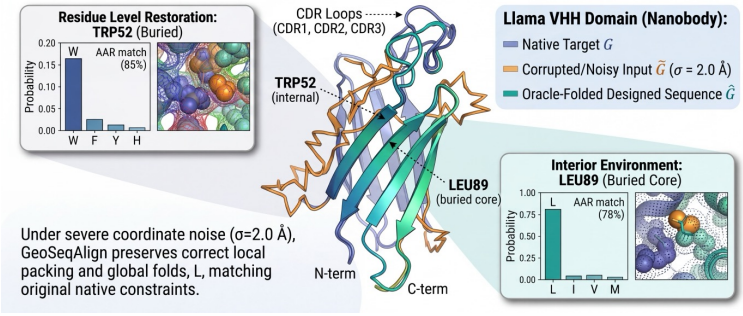


Figure 5: **Input-coordinate noise illustration.** Structural alignment across escalating coordinate noise on a representative target. Oracle-predicted folds ($\hat{G}$, rainbow) preserve overall tertiary topology against pristine backbones (G, gray) even under severe geometric perturbations.



Evaluation views

Native target: the fixed reference for structural scoring.

Perturbed input: the inverse folder's condition.

Folded design: an external oracle prediction.

Figure 6: **Local-packing illustration.** Local packaging preservation under 2.0 Å coordinate noise for the VHH domain (PDB: 1G9E). The model retains correct aromatic stacking for buried TRP52 (85% marginal probability) and hydrophobic cavity packaging for LEU89 (78% marginal probability) despite severe backbone distortion.