PEPPERMINT: STRUCTURE-CONDITIONED REPRESENTATION LEARNING FOR INTERACTING PROTEIN PAIRS

Anonymous authors
Paper under double-blind review

ABSTRACT

Protein–protein interactions depend on the sequences of both proteins and their spatial arrangement, motivating protein representations that incorporate molecular context. We introduce *Peppermint*, a structure-conditioned masked language model for interacting protein pairs. *Peppermint* maintains separate chain representations and alternates intra-chain self-attention with bidirectional cross-attention, preserving chain identity while allowing each residue to incorporate information from the other chain. The architecture assigns distinct roles to local conformation and interaction geometry: gated embeddings of discrete structural states and continuous backbone geometry enrich residue representations, while directional geometric biases guide attention between chains. Initialized from a pretrained sequence encoder, the model learns from paired sequences and protein complexes through masked amino-acid modeling and auxiliary structural prediction. Mixing examples with and without structure allows the same model to operate under either input condition. Supplying structure improves simultaneous recovery of masked interface residues by 9.3 percentage points relative to sequence-only input, with downstream benefits varying by task and metric. These results support incorporating complex geometry into paired protein representations while retaining sequence-only inference within the same encoder.

1 INTRODUCTION

Protein–protein interactions depend on the sequences of both proteins and their spatial arrangement. Protein language models such as ESM-2 and ESM-C learn structural and functional information from sequence pretraining (Lin et al., 2023; Candido et al., 2026), but independently encoding each chain produces representations that do not change with the interaction partner. Jointly encoding the two sequences allows each residue representation to depend on the other chain, while complex structures provide information about local conformation and inter-chain contacts. We aim to incorporate both sequence and structural information in a model that remains usable when complex structures are unavailable.

Sequence-based interaction models differ in how they combine information from the two chains. D-SCRIPT combines independently computed residue embeddings through a learned contact-map module (Sledzieski et al., 2021), whereas PLM-interact jointly encodes the two sequences using masked prediction and interaction classification (Liu et al., 2025). MINT and PPLM distinguish intra-chain and inter-chain attention during paired-sequence modeling (Ullanat et al., 2026; Liu et al., 2026). MSA Pairformer also exploits evolutionary information from multiple-sequence alignments (Akiyama et al., 2026).

Structural conditioning complements joint sequence encoding by supplying geometric information. SaProt and ProSST use discrete structural tokens (Su et al., 2024; Li et al., 2024), while LM-Design introduces structural adapters (Zheng et al., 2023). PST incorporates local graph features into the self-attention of a pretrained protein language model (Hartout et al., 2025). ProteinMPNN and ESM-IF1 condition sequence prediction on backbone geometry (Dauparas et al., 2022; Hsu et al., 2022), and MIF and MIF-ST use geometry to reconstruct masked sequences (Yang et al., 2023). PPIformer learns interface representations through masked amino-acid reconstruction on structural interaction graphs (Bushuiev et al., 2024). ESM3 and FlexRibbon learn jointly from sequence and structure (Hayes et al., 2025; Zhu et al., 2026). PPLM incorporates geometry from monomers and

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.

predicted complexes in downstream contact predictors (Liu et al., 2026). Independently positioned monomer structures describe local chain conformation but do not determine relative chain placement. We focus on conditioning the paired residue encoder itself on both chain-local conformation and the relative arrangement of the chains. Supplying or withholding these inputs within the same encoder lets us assess the value of explicit complex geometry for representations already informed by both sequences.

STRING v12.0 provides protein association networks across 12,535 genomes, with physical links supported by experiments, curated databases and text mining (Szklarczyk et al., 2023). These links support paired-sequence pretraining but do not specify residue-level interaction geometry. Experimental complexes from the Protein Data Bank provide complementary structural data (Berman et al., 2000). Cofolding methods such as AlphaFold-Multimer, AlphaFold 3 and ESMFold2 expand structural coverage, although prediction accuracy can be variable (Evans et al., 2021; Abramson et al., 2024; Candido et al., 2026). This motivates training on both paired sequences and complexes, allowing a single encoder to use geometry without requiring it for every example.

We introduce *Peppermint*, an ESM-C-initialized masked language model for interacting protein pairs. Building on MINT, the encoder maintains separate chain representations and alternates intra-chain self-attention with bidirectional cross-attention. Local structural embeddings describe each chain's conformation, while geometric attention biases encode the relative placement of the two chains. We train the model on STRING sequence pairs and experimental and predicted complexes, combining masked amino-acid modeling with auxiliary structural prediction. To expand the structural training data, we cofold approximately 1.5 million STRING pairs with Protenix-v1 (Team et al., 2026) and retain 175,821 complexes after confidence and clash filtering.

We evaluate simultaneous recovery of masked interface residues, sequence-only recovery on held-out STRING pairs, and downstream prediction from frozen representations, both on confidence-selected complex subsets and on full datasets using sequences alone. On 249 PINDER-S complexes, supplying structure increases interface recovery from 41.15% to 50.49%. The effects on downstream prediction vary by task and metric, while the same encoder remains competitive without structural inputs on the full datasets. Separate experiments examine how model size and structural inputs affect training dynamics.

2 PEPPERMINT

2.1 PAIRED SEQUENCE REPRESENTATIONS

Peppermint jointly encodes two interacting protein chains, using their sequences and optional complex geometry (Fig. 1c). For chain $c \in \{A, B\}$, let $x_c = (x_{c,i})_{i=1}^{L_c}$ denote its native amino-acid sequence, $\widetilde{x}_c$ its input sequence after masking, and $h_c^{(\ell)} \in \mathbb{R}^{L_c \times d}$ its residue states after ℓ blocks, with hidden width d.

Each block applies intra-chain self-attention, bidirectional cross-attention and a feed-forward update. Both cross-attention directions read the same post-self-attention states. Within each block, self-attention and feed-forward weights are shared between chains, and cross-attention weights between directions; transformer parameters are independent across blocks.

Self-attention uses chain-specific rotary positions. Cross-attention omits them because the chains have no shared sequence coordinate. MINT combines intra-chain and inter-chain logits in a joint softmax (Ullanat et al., 2026); *Peppermint* uses successive updates with separate normalizations, output projections and residual connections. We initialize the sequence backbone and amino-acid head from ESM-C 600M (Candido et al., 2026) and learn randomly initialized cross-attention during paired sequence training. The full model has approximately 769M parameters (Appendix A).

2.2 LOCAL AND INTER-CHAIN STRUCTURAL INPUTS

We compute 3Di states independently within each chain using a 20-state alphabet that describes spatial residue-neighbor geometry (Van Kempen et al., 2024). Let z_c denote the native states and $\widetilde{z}_c$ their input values after masking, X_c the residue-aligned N, Cα, C and O coordinates, and m_c the mask of residues with usable, unmasked geometry. Initial residue states combine sequence and

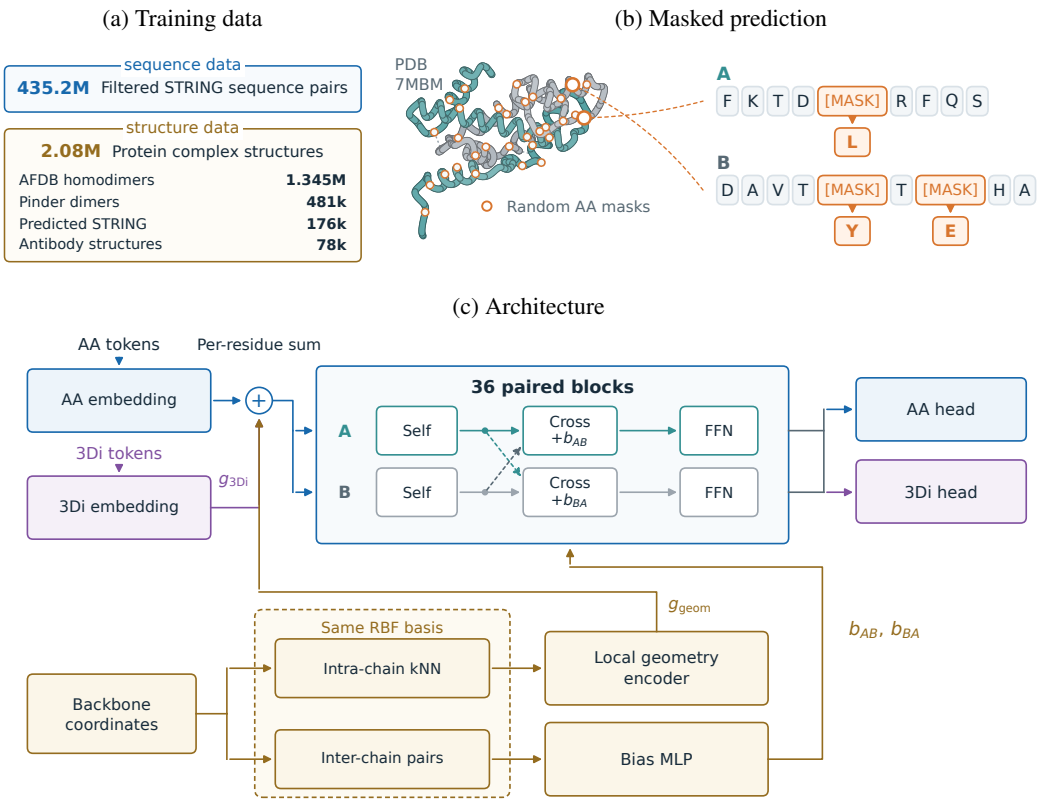


Figure 1: **Peppermint training data, masked prediction, and architecture. (a)** Training corpus sizes. **(b)** Masked language modeling illustrated on an example complex from the PINDER training split. **(c)** Paired-chain encoder combining intra-chain self-attention and bidirectional cross-attention, with optional structural embeddings and geometric attention biases.

structural representations

$$h_c^{(0)} = E_{\mathrm{AA}}(\widetilde{x}_c) + g_{\mathrm{3Di}} E_{\mathrm{3Di}}(\widetilde{z}_c) + g_{\mathrm{geom}} F_\theta(X_c, m_c), \tag{1}$$

where E_{AA} and E_{3Di} are embedding tables and F_θ is a continuous geometry encoder. Each term has width d. These modules and the learned, unconstrained scalar gates g_{3Di} and g_{geom} are shared between chains.

The continuous geometry encoder adapts ProteinMPNN message passing (Dauparas et al., 2022) within each chain, using backbone distances and residue offsets between spatial nearest neighbors. We train it from scratch, without ProteinMPNN's autoregressive decoder, and project its node states to width d. The 3Di embedding has an additional missing-state row fixed at zero.

Inter-chain geometry enters through an additive attention bias computed from ordered atom-pair distances. These distances require a common complex coordinate frame; independently positioned monomers do not specify relative chain placement. A shared multilayer perceptron maps radial basis function (RBF) distance features and a structure-presence indicator to one bias per head, using separate parameters from the local encoder. For one head in the $A \leftarrow B$ direction,

$$\mathrm{CrossAttn}_{A \leftarrow B} = \mathrm{softmax}\left(\frac{Q_A K_B^\top}{\sqrt{d_h}} + b_{AB} + M_B\right) V_B, \tag{2}$$

where Q_A contains chain A queries, K_B, V_B contain chain B keys and values, and d_h is the head width. The head-specific geometry bias is b_{AB}; M_B is zero for valid keys and $-\infty$ for padding. Softmax normalizes over key residues. The reverse bias b_{BA} uses separately constructed features because exchanging query and key also exchanges ordered atom-pair channels. Biases are reused across blocks; intra-chain self-attention receives no structural bias.

A rigid-body movement of one chain preserves within-chain distances and its 3Di states but changes cross-chain distances. A common rotation or translation of both chains leaves these geometric features unchanged, allowing them to describe relative arrangement without depending on the arbitrary orientation of the complex.

The gates and final bias projection are initialized to zero when structural fine-tuning begins, preserving the paired sequence model's initial amino-acid predictions. Without structure, residue-level structural additions are zero and the geometry bias is constant across valid keys, leaving attention weights unchanged. Appendix A specifies distance features, validity masks and structural masking.

2.3 Training objective

We reconstruct masked amino-acid identities and, during structural fine-tuning, masked 3Di states. Sequence and structure masks can coincide or differ according to the corruption modes in Section 3.

For modality $t \in \{\mathrm{AA}, \mathrm{3Di}\}$ and chain c, let $\mathcal{M}_c^t$ contain selected, valid target positions (n, i) across a local minibatch, where n indexes examples and i residues. Native targets are $y^{\mathrm{AA}}_{n,c,i} = x_{n,c,i}$ and $y^{\mathrm{3Di}}_{n,c,i} = z_{n,c,i}$. The paired input $\widetilde{\mathcal{I}}_n$ contains both corrupted sequences and the available, unmasked structural inputs. Prediction heads applied to final normalized residue states give softmax distributions $p^t_{\theta,c,i}$ with trainable parameters θ. Each component loss is

$$\mathcal{L}_{t,c} = -\frac{1}{|\mathcal{M}_c^t|} \sum_{(n,i)\in\mathcal{M}_c^t} \log p^t_{\theta,c,i}\left(y^t_{n,c,i} \mid \widetilde{\mathcal{I}}_n\right), \tag{3}$$

with $\mathcal{L}_{t,c} = 0$ when $\mathcal{M}_c^t$ is empty. The training objective is

$$\mathcal{L} = \sum_{c\in\{A,B\}} \left[\mathcal{L}_{\mathrm{AA},c} + \lambda_{\mathrm{3Di}}\, \mathcal{L}_{\mathrm{3Di},c}\right], \tag{4}$$

where $\lambda_{\mathrm{3Di}} = 0$ in sequence-only training and 0.2 in structural fine-tuning. Each component averages selected residues across the local minibatch for one chain, and the chain means are summed rather than pooled across chains. Gradient accumulation and distributed training average these local losses.

Structural targets must be valid before corruption and selected by the structural mask. We train all model parameters.

3 Training data and curriculum

The training corpus combines 435,185,984 STRING physical interaction records with combined score greater than 500 (Szklarczyk et al., 2023) and 2,080,090 structural examples: experimental Pinder dimers (Kovtun et al., 2024), predicted AFDB homodimers, predicted STRING complexes, and antibody complexes. Figure 1a and Appendix B give the source counts and curation details.

We use the high-confidence AFDB homodimers from the proteome-scale collection of Han et al. (2026), predicted with AlphaFold2-multimer (Evans et al., 2021). The antibody corpus combines experimental SAbDab structures (Dunbar et al., 2014) with predicted antibody–antigen complexes; predicted structures require ipTM > 0.7 and no flagged clash.

For cofolding, we select STRING physical links with combined scores greater than 950 and retain one representative link per unordered pair of MMseqs2 sequence clusters at 50% identity (Steinegger & Söding, 2017). We cofold the resulting 1,515,203 pairs with Protenix-v1 (Team et al., 2026). The designated best prediction per pair is retained if mean pLDDT ≥ 70, ipTM ≥ 0.6 and no clashes are flagged, yielding 175,821 complexes.

STRING uses a split based on MMseqs2 sequence clustering at 50% identity: training pairs contain two training-cluster proteins, whereas validation requires at least one validation-cluster protein. Pinder packaging excludes the official validation and test systems.

Starting from pretrained ESM-C 600M weights for the sequence backbone and amino-acid head, we first train on paired sequences with a combined crop budget of 512 tokens, increase the budget to 2,048 tokens, then retain this budget for structural fine-tuning (Table 1). The final stage interleaves STRING

Table 1: Training curriculum. Crops are combined token budgets for both chains. Batch sizes are nominal global sizes before length filtering. All stages use eight devices, with per-device batches 32/8/2 and accumulation factors 1/2/4, respectively. Updates and warmup count optimizer steps. Stage transitions retain model weights and reset the optimizer and learning-rate schedule.

Training stage	Crop	Batch	Updates	Warmup	Peak LR
Short-crop sequence	512	256	1,259,019	2,000	10^{-4}
Long-crop sequence	2,048	128	332,000	5,000	5×10^{-5}
Structural fine-tuning	2,048	64	400,000	5,000	5×10^{-5}

sequences, Pinder, AFDB homodimers, antibody examples and predicted STRING complexes with weights 0.60, 0.20, 0.12, 0.04 and 0.04, respectively. Sequence-only examples use random contiguous crops; structural examples use interface-centered crops when usable contacts are available.

All stages replace amino-acid identities by mask tokens with probability 15%, ensuring at least one target per nonempty chain without restricting masking to the interface. For AFDB homodimers, both chains share crop windows and amino-acid mask positions, preventing direct copying of a masked residue's identity from the other chain. During structural fine-tuning, 25% of structural examples mask sequence and structure at the same positions, 25% mask sequence while retaining available structure, and 50% draw the two masks independently with probability 15% each. Figure 1b illustrates amino-acid masking with backbone coordinates retained.

Training the full *Peppermint* model took 18 days on 8 NVIDIA B100 GPUs. We select the structural-stage checkpoint at step 360,000 from a 400,000 step run, using the combined amino-acid and 3Di validation objective in Eq. 4 on the Pinder validation stream. Validation samples crops and masks with coordinate noise disabled. Appendix B.2 specifies crop construction, optimization and validation frequency and budget.

4 Evaluation and results

We evaluate *Peppermint* on masked interface-residue recovery, held-out STRING sequence recovery and downstream prediction of interaction-related properties from frozen representations. Comparisons use ESM-2, ESM-C and MINT as sequence baselines, with ProteinMPNN as an additional structure-conditioned baseline for interface recovery. Where structures are available, we assess their value by supplying or withholding structural inputs; separate component ablations and model-size experiments examine their contribution to learning.

Single-chain ESM-2 650M and ESM-C 600M encode each chain independently, whereas MINT and *Peppermint* receive both chains. Pretrained baselines retain their token framing and embedding conventions for masked recovery. Scored positions are identical across models within each recovery benchmark.

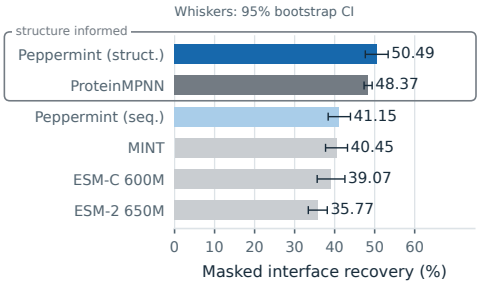
Recovery error bars show 95% bootstrap confidence intervals. We repeatedly sample whole PINDER-S complexes with replacement; for STRING, protein pairs linked through shared proteins are resampled together. We recompute pooled recovery using the same resampled examples for every model and take the middle 95% of the resulting scores or model differences. We reuse the original predictions and masked positions, so these intervals measure sensitivity to the evaluation sample rather than to retraining or changing the masks.

4.1 Masked interface-residue recovery

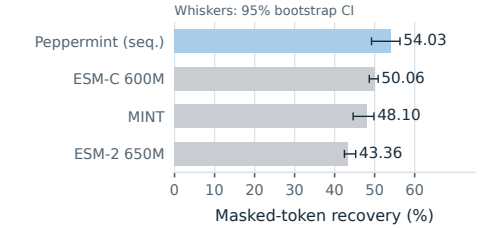
We evaluate 249 PINDER-S test complexes (Kovtun et al., 2024), defining an interface residue by a Cα within 10 Å of a Cα in the other chain. We simultaneously mask all eligible interface amino-acid identities on both chains, retain native non-interface sequence and supply backbone coordinates to structure-conditioned methods. Positions require finite N, Cα, C and O coordinates and a standard amino-acid target, giving 23,682 identical scored residues.

Masking both sides removes direct access to the scored interface identities in either chain. Retaining the remaining sequence and complex geometry tests residue prediction at a specified interaction

(a) Interface sequence recovery



(b) Held-out STRING sequence recovery



(c) Downstream representation tasks

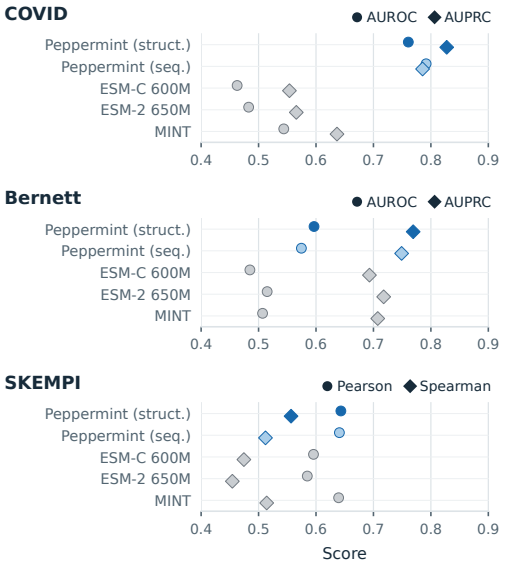


Figure 2: **Sequence recovery and downstream prediction with Peppermint. (a)** Interface-residue recovery on 249 PINDER-S complexes, with all eligible interface residues on both chains masked simultaneously and the remaining sequence provided unmasked. **(b)** Sequence-only recovery on 5,000 held-out STRING pairs, with residues randomly masked with probability 15% and recovery computed only at masked positions. **(c)** Downstream prediction on confidence-selected subsets of COVID, Bernett and SKEMPI. Peppermint uses the same encoder weights with and without structural inputs. Recovery error bars show 95% bootstrap confidence intervals.

surface; the STRING evaluation instead tests randomly masked sequence prediction without structural inputs.

Recovery is top-1 accuracy over the 20 standard amino acids, pooled across scored residues; negative log-likelihood and perplexity use probabilities renormalized over the same vocabulary. We compare *Peppermint* with and without structure, the sequence baselines and ProteinMPNN (Dauparas et al., 2022). ProteinMPNN inference settings are given in Appendix C.1.

With the same *Peppermint* encoder weights and crop windows, structural input increases interface recovery from 41.15% to 50.49%, a gain of 9.34 percentage points, and reduces perplexity from 6.7 to 5.7 (Fig. 2a). Recovery exceeds ProteinMPNN's point estimate of 48.37% by 2.12 percentage points, although the 95% interval for this difference includes zero. ProteinMPNN has lower perplexity (5.2).

Without structure, *Peppermint* exceeds ESM-C recovery by 2.08 percentage points. Differences in pretraining data, corruption objectives, model family and access to the other chain prevent attributing the cross-model differences to cross-attention alone.

4.2 Held-out STRING sequence recovery

We select 5,000 STRING pairs in deterministic stream order from the 80%-identity representatives on the validation side of the 50%-identity cluster split, retaining combined sequence lengths of at most 1,024 residues. Per-chain content-hash-derived masks replace residues with probability 15%, giving 436,943 identical scored positions. Recovery is pooled over masked residues; cross-entropy and perplexity use each model's native vocabulary, rather than the shared 20-amino-acid vocabulary used for interface recovery. This evaluation uses validation data; the corpus has no separate test split.

Without structure, *Peppermint* recovers 54.03% of masked residues, compared with 50.06% for ESM-C, 48.10% for MINT and 43.36% for ESM-2 (Fig. 2b). The gain over ESM-C is 3.97 percentage points, with perplexity 4.0 versus 5.1. *Peppermint* is trained on the STRING corpus family, so these

results do not isolate the effect of jointly encoding both chains. One linked group contains roughly half of the scored residues, limiting the reliability of the bootstrap intervals.

4.3 DOWNSTREAM PREDICTION

We evaluate frozen representations on COVID antibody cross-neutralization, Bernett interaction classification and SKEMPI mutation-induced binding-energy changes (Raybould et al., 2021; Bernett et al., 2024; Jankauskaitė et al., 2019). We predict complexes with ESMFold2-Fast (Candido et al., 2026) and retain those with interface confidence ipTM ≥ 0.5 and mean residue confidence pLDDT ≥ 70; both wild-type and mutant structures must pass for mutation tasks. Across all supplied partitions, the accepted fractions are 22.46%, 1.07% and 87.52% for COVID, Bernett and SKEMPI, respectively (Appendix C.2). MINT, ESM-2 and ESM-C provide sequence-only baselines.

We obtain one vector per chain by mean-pooling non-padded residue representations and concatenate the two vectors for prediction. Mutation tasks use the concatenated per-chain wild-type-minus-mutant differences. Task supervision fits only the prediction heads, leaving the residue encoder fixed.

The two *Peppermint* conditions use the same encoder checkpoint and accepted rows, withholding geometry for sequence-only inference. We fit heads separately for each representation: an MLP for COVID, logistic regression for Bernett and ridge regression with ten outer and five inner folds for SKEMPI. Above the 2,048-residue per-chain cap, the two input conditions may retain different sequence windows. Targets and fitting details are given in Appendix C.2.

On the confidence-selected COVID subset, both *Peppermint* conditions exceed the sequence baselines in AUROC and average precision (AUPRC). Structural input increases AUPRC from 0.7855 to 0.8276, while AUROC decreases from 0.7917 to 0.7607 (Fig. 2c). MINT has the highest baseline AUROC and AUPRC, 0.5438 and 0.6365.

On Bernett, structural input increases AUROC from 0.5746 to 0.5966 and AUPRC from 0.7490 to 0.7692. Both *Peppermint* conditions exceed the baseline point estimates, although this external comparison is descriptive: *Peppermint* retains 2,933 rows across all splits and the baselines retain 2,932. Confidence filtering also changes class prevalence, affecting AUPRC comparisons with the full dataset.

On SKEMPI, structural input gives the highest Pearson and Spearman correlation point estimates among the compared representations, increasing Pearson from 0.6407 to 0.6433 and Spearman from 0.5120 to 0.5566. MINT reaches 0.6394 and 0.5141, respectively. RMSE increases from 1.6051 to 1.6185 kcal mol^{-1} despite the correlation gains.

Without structural inputs, the same *Peppermint* checkpoint remains competitive on the full eligible COVID, Bernett and SKEMPI datasets (Appendix C.3, Table C.2). Eligibility does not depend on structure availability or confidence, and all models share the same evaluation examples and task-specific splits or folds. On COVID, *Peppermint* achieves the highest AUROC among the compared representations, 0.6781 versus 0.6617 for MINT, with similar AUPRC. It also gives the highest Bernett AUROC and AUPRC point estimates, although the margins over ESM-C are small. On SKEMPI, *Peppermint* exceeds both single-chain baselines in correlation and is close to MINT, which retains the highest correlations and lowest RMSE. These experiments use separately fitted heads on the full prescribed partitions; differences from the confidence-selected results therefore reflect changes in both training and evaluation populations. The results support using the shared encoder when reliable complex geometry is unavailable. On the additional TDC and PISTE T-cell receptor benchmarks, *Peppermint* has no consistent advantage.

4.4 STRUCTURAL INPUT ABLATION

In a separate 5,000-update study with a 183M-parameter paired-sequence model, 3Di tokens provide the largest reduction in held-out amino-acid loss among the considered configurations. Cross-chain distance biases give a smaller reduction alone and a small additional benefit with tokens; the configuration combining tokens, cross-chain bias and backbone coordinates achieves the lowest loss. These are smaller, short-budget variants, not a component-by-component ablation of the final model (Appendix D, Table D.1).

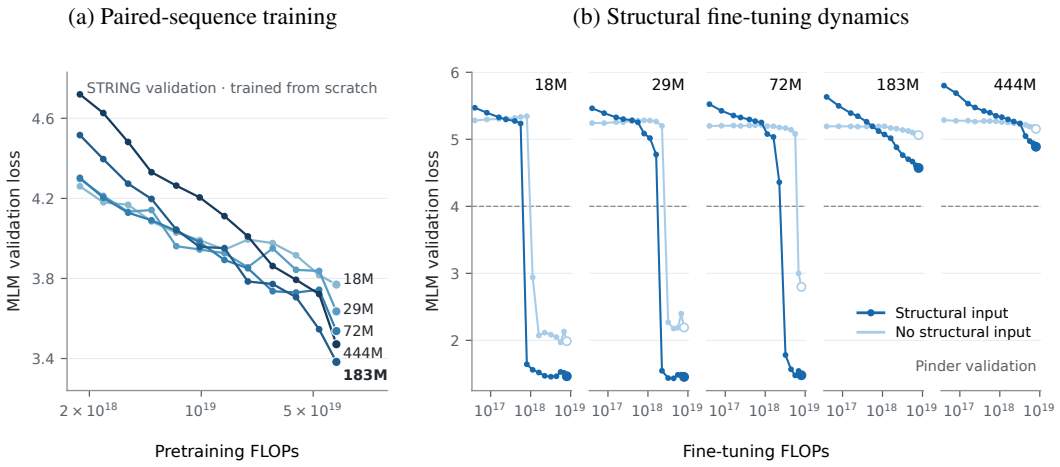


Figure 3: **Model size and structural fine-tuning dynamics. (a)** STRING MLM validation loss during sequence pretraining from scratch. **(b)** Pinder MLM validation loss during fine-tuning with and without structural inputs. Each pair starts from the same sequence checkpoint and receives equal updates. Points denote checkpoints. Horizontal dashed lines mark loss 4.0. MLM loss sums per-chain masked-amino-acid means and excludes auxiliary structural loss.

4.5 Model size and structural fine-tuning dynamics

In a separate study, we train five *Peppermint* models from scratch on STRING, with 18–444 million parameters, and examine sequence pretraining and structural fine-tuning. These runs are distinct from the ESM-C-initialized model used above. Held-out MLM loss generally decreases during sequence training, with model rankings changing as compute increases (Fig. 3a). At the common terminal budget of approximately 6.979×10^{19} nominal FLOPs, the 183M model achieves the lowest loss, 3.3832, compared with 3.4713 for the largest model.

We then fine-tune each initialization with or without structural inputs, matching update counts within model size and retaining auxiliary 3Di supervision in both conditions. Structural inputs advance the sharp decline in Pinder validation loss in the three smaller models (Fig. 3b). For the 72M model, loss first reaches 4.0 between updates 79,800 and 114,000 with structure, compared with 199,500–242,250 without it. The two smallest structure-conditioned models subsequently maintain losses near 1.5, compared with approximately 2.0–2.4 for their controls. The two larger models show no comparable transition within their evaluated budgets.

Structural inputs thus advance the loss decline in the three smaller models and lower the subsequent loss in the two smallest. Because both conditions retain auxiliary structural supervision, this comparison tests the effect of supplying geometry rather than the presence of an additional prediction objective. Appendix D gives the complete loss-crossing intervals, post-transition behavior, training protocols, nominal compute accounting and sequence-allocation analyses.

5 Discussion

Peppermint shows that explicit complex geometry remains useful after extensive sequence pretraining and paired-chain learning. Supplying structure improves prediction within the same encoder, supporting its use as conditioning information for representations already informed by both sequences. This connects paired-sequence learning (Ullanat et al., 2026; Liu et al., 2026) with structure-conditioned masked modeling (Bushuiev et al., 2024): a protein is represented in relation to the other chain and the available structural description of their interaction. Sequence-only and structure-conditioned inference are therefore complementary uses of a shared representation, rather than separate modeling problems. The separate model-size study suggests that geometry can also guide the acquisition of these representations, rather than only refine predictions at inference.

The downstream results give this combination a role beyond amino-acid reconstruction. The same frozen encoder supports prediction of distinct interaction-related properties, including interaction classification and mutation effects on binding. These tasks use features learned before supervision for the corresponding biological properties, supporting interaction-level pretraining as a common basis for studying whether proteins associate and how sequence variation affects binding. Structural conditioning adds predictive value across the main downstream tasks, although its benefit depends on the prediction target and evaluation criterion.

Optional structural input addresses the unequal coverage and resolution of interaction data. STRING provides broad evidence about protein pairs without specifying residue-level geometry; its physical links can also reflect membership in the same complex rather than direct contact (Szklarczyk et al., 2023). Experimental complexes supply observed arrangements, while cofolding translates interaction evidence into candidate molecular arrangements, allowing to scale up learned representations through predicted complex inputs. Predicted structures already support large-scale training, and expanding complex collections extend this opportunity to interaction representations (Hsu et al., 2022; Han et al., 2026). These geometries remain model-derived hypotheses. Accepting sequences with or without structure allows the encoder to use this growing resource without restricting its applicability to protein pairs with available complex structures.

The main downstream evidence concerns confidence-selected populations, and generalization to unfamiliar complex families remains to be established. A central next question is when geometry improves transfer across such families, and how uncertainty in the predictions affects that benefit. Evaluations that separate related complexes across data sources and span a range of structural confidence would be necessary to address these questions. This would test whether the structural information useful for familiar interactions also supports predictions when both the interacting proteins and their modes of association change.

Reproducibility statement

Section 3 specifies the training curriculum and checkpoint selection, and Appendix B.2 gives optimization, cropping and validation details; Section 4.2 specifies STRING recovery. Appendices A and B describe the architecture, masking, and data filters. Appendix C.2 specifies downstream fitting, and Appendix C.3 reports sequence-only results on the full datasets. Appendix C.1 defines the interface positions and scoring protocol; Appendix D gives the separate model-size experiments and their fitting procedures.

AI use statement

Generative AI tools assisted manuscript drafting and editing, preparation of figures, software code implementation and checks of reported methods and results against implementation and experimental records. The authors are responsible for the final text, claims, and artifacts.

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(8016):493–500, 2024.

Yo Akiyama, Zhidian Zhang, Olivia Tang, Rachel Seongeun Kim, Milot Mirdita, Martin Steinegger, and Sergey Ovchinnikov. Expanding the scope of protein language modeling to protein-protein interactions with msa pairformer. *Cell*, 2026.

Helen M Berman, John Westbrook, Zukang Feng, Gary Gilliland, Talapady N Bhat, Helge Weissig, Ilya N Shindyalov, and Philip E Bourne. The protein data bank. *Nucleic acids research*, 28(1): 235–242, 2000.

Judith Bernett, David B Blumenthal, and Markus List. Cracking the black box of deep sequence-based protein–protein interaction prediction. *Briefings in Bioinformatics*, 25(2):bbae076, 2024.

Anton Bushuiev, Roman Bushuiev, Petr Kouba, Anatolii Filkin, Marketa Gabrielova, Michal Gabriel, Jiri Sedlar, Tomas Pluskal, Jiri Damborsky, Stanislav Mazurenko, et al. Learning to design protein-protein interactions with enhanced generalization. In *International Conference on Learning Representations*, volume 2024, pp. 21010–21035, 2024.

Salvatore Candido, Thomas Hayes, Alexander Derry, Roshan Rao, Zeming Lin, Robert Verkuil, Bryan Z Wu, Jin Sub Lee, Elise S Bruguera, Jehan A Keval, et al. Language modeling materializes a world model of protein biology. *bioRxiv*, pp. 2026–06, 2026.

Xingyi Cheng, Bo Chen, Pan Li, Jing Gong, Jie Tang, and Le Song. Training compute-optimal protein language models. *Advances in Neural Information Processing Systems*, 37:69386–69418, 2024.

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

James Dunbar, Konrad Krawczyk, Jinwoo Leem, Terry Baker, Angelika Fuchs, Guy Georges, Jiye Shi, and Charlotte M Deane. Sabdab: the structural antibody database. *Nucleic acids research*, 42 (D1):D1140–D1146, 2014.

Richard Evans, Michael O'neill, Alexander Pritzel, Natasha Antropova, Andrew Senior, Tim Green, Augustin Žídek, Russ Bates, Sam Blackwell, Jason Yim, et al. Protein complex prediction with alphafold-multimer. *biorxiv*, pp. 2021–10, 2021.

Ziyan Feng, Jingyang Chen, Youlong Hai, Xuelian Pang, Kun Zheng, Chenglong Xie, Xiujuan Zhang, Shengqing Li, Chengjuan Zhang, Kangdong Liu, et al. Sliding-attention transformer neural architecture for predicting t cell receptor–antigen–human leucocyte antigen binding. *Nature Machine Intelligence*, 6(10):1216–1230, 2024.

Yewon Han, Maxim I Tsenkov, Niccolò AE Venanzi, Sooyoung Cha, Nilkanth Patel, Sreenath Nair, Razan Abbara, Damian Bertoni, Alejandro Chacón, Nick Dietrich, et al. Alphafold database expands to proteome-scale quaternary structures. *bioRxiv*, pp. 2026–03, 2026.

Philip Hartout, Dexiong Chen, Paolo Pellizzoni, Carlos Oliver, and Karsten Borgwardt. Endowing protein language models with structural knowledge. *Bioinformatics*, 41(11):btaf582, 2025.

Thomas Hayes, Roshan Rao, Halil Akin, Nicholas J Sofroniew, Deniz Oktay, Zeming Lin, Robert Verkuil, Vincent Q Tran, Jonathan Deaton, Marius Wiggert, et al. Simulating 500 million years of evolution with a language model. *Science*, 387(6736):850–858, 2025.

Jordan Hoffmann, Sebastian Borgeaud, Arthur Mensch, Elena Buchatskaya, Trevor Cai, Eliza Rutherford, Diego de Las Casas, Lisa Anne Hendricks, Johannes Welbl, Aidan Clark, et al. An empirical analysis of compute-optimal large language model training. *Advances in neural information processing systems*, 35:30016–30030, 2022.

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

Justina Jankauskaitė, Brian Jiménez-García, Justas Dapkūnas, Juan Fernández-Recio, and Iain H Moal. Skempi 2.0: an updated benchmark of changes in protein–protein binding energy, kinetics and thermodynamics upon mutation. *Bioinformatics*, 35(3):462–469, 2019.

Daniel Kovtun, Mehmet Akdel, Alexander Goncearenco, Guoqing Zhou, Graham Holt, David Baugher, Dejun Lin, Yusuf Adeshina, Thomas Castiglione, Xiaoyun Wang, et al. Pinder: the protein interaction dataset and evaluation resource. *bioRxiv*, pp. 2024–07, 2024.

Mingchen Li, Yang Tan, Xinzhu Ma, Bozitao Zhong, Huiqun Yu, Ziyi Zhou, Wanli Ouyang, Bingxin Zhou, Pan Tan, and Liang Hong. Prosst: Protein language modeling with quantized structure and disentangled attention. *Advances in Neural Information Processing Systems*, 37:35700–35726, 2024.

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

Dan Liu, Francesca Young, Kieran D Lamb, Adalberto Claudio Quiros, Alexandrina Pancheva, Crispin J Miller, Craig Macdonald, David L Robertson, and Ke Yuan. Plm-interact: extending protein language models to predict protein-protein interactions. *Nature Communications*, 16(1): 9012, 2025.

Jun Liu, Hungyu Chen, and Yang Zhang. A paired sequence language model for protein-protein interaction modeling. *Nature Communications*, 17(1):3733, 2026.

Mahdi Shafiei Neyestanak, Sarah M Burbach, Karenna Ng, Praneeth Gangavarapu, Jonathan Hurtado, Judie Magura, Nasreen Ismail, Daniel Muema, Thumbi Ndung'u, Andrew B Ward, et al. Data-optimal scaling of paired antibody language models. *bioRxiv*, 2025.

Matthew IJ Raybould, Aleksandr Kovaltsuk, Claire Marks, and Charlotte M Deane. Cov-abdab: the coronavirus antibody database. *Bioinformatics*, 37(5):734–735, 2021.

Samuel Sledzieski, Rohit Singh, Lenore Cowen, and Bonnie Berger. D-script translates genome to phenome with sequence-based, structure-aware, genome-scale predictions of protein-protein interactions. *Cell Systems*, 12(10):969–982, 2021.

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.

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*, volume 2024, pp. 6987–7009, 2024.

Damian Szklarczyk, Rebecca Kirsch, Mikaela Koutrouli, Katerina Nastou, Farrokh Mehryary, Radja Hachilif, Annika L Gable, Tao Fang, Nadezhda T Doncheva, Sampo Pyysalo, et al. The string database in 2023: protein–protein association networks and functional enrichment analyses for any sequenced genome of interest. *Nucleic acids research*, 51(D1):D638–D646, 2023.

Protenix Team, Yuxuan Zhang, Chengyue Gong, Hanyu Zhang, Wenzhi Ma, Zhenyu Liu, Xinshi Chen, Jiaqi Guan, Lan Wang, Yanping Yang, et al. Protenix-v1: Toward high-accuracy open-source biomolecular structure prediction. *bioRxiv*, pp. 2026–02, 2026.

Varun Ullanat, Bowen Jing, Samuel Sledzieski, and Bonnie Berger. Learning the language of protein-protein interactions. *Nature Communications*, 17(1):1199, 2026.

Michel Van Kempen, Stephanie S Kim, Charlotte Tumescheit, Milot Mirdita, Jeongjae Lee, Cameron LM Gilchrist, Johannes Söding, and Martin Steinegger. Fast and accurate protein structure search with foldseek. *Nature biotechnology*, 42(2):243–246, 2024.

Alejandro Velez-Arce, Michelle M Li, Wenhao Gao, Xiang Lin, Kexin Huang, Tianfan Fu, Bradley L Pentelute, Manolis Kellis, and Marinka Zitnik. Signals in the cells: multimodal and contextualized machine learning foundations for therapeutics. *bioRxiv*, 2024.

Kevin K Yang, Niccolò Zanichelli, and Hugh Yeh. Masked inverse folding with sequence transfer for protein representation learning. *Protein Engineering, Design and Selection*, 36:gzad015, 2023.

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

Jianwei Zhu, Yu Shi, Ran Bi, Peiran Jin, Chang Liu, Zhe Zhang, Haitao Huang, Zekun Guo, Pipi Hu, Fusong Ju, et al. Flexribbon: Joint sequence and structure pretraining for protein modeling. In *International Conference on Learning Representations*, volume 2026, pp. 65602–65628, 2026.

A Architecture and implementation details

A.1 Paired transformer

The encoder has 36 blocks, hidden width $d = 1{,}152$ and 18 heads of width $d_h = 64$. Each chain has a separate padding mask and uses residue tokens without CLS/EOS framing. We initialize token embeddings, self-attention, feed-forward layers, final normalization and the amino-acid head from ESM-C 600M.

Using the chain-state notation of Section 2, the updates for block $\ell = 0, \ldots, T-1$ are

$$u_A^\ell = h_A^\ell + s^{-1}\,\mathrm{SA}_\ell(h_A^\ell), \qquad u_B^\ell = h_B^\ell + s^{-1}\,\mathrm{SA}_\ell(h_B^\ell), \tag{5}$$

$$v_A^\ell = u_A^\ell + s^{-1}\,\mathrm{CA}_\ell(u_A^\ell, u_B^\ell; b_{AB}), \qquad v_B^\ell = u_B^\ell + s^{-1}\,\mathrm{CA}_\ell(u_B^\ell, u_A^\ell; b_{BA}), \tag{6}$$

$$h_A^{\ell+1} = v_A^\ell + s^{-1}\,\mathrm{FFN}_\ell(v_A^\ell), \qquad h_B^{\ell+1} = v_B^\ell + s^{-1}\,\mathrm{FFN}_\ell(v_B^\ell). \tag{7}$$

Here u and v are the states after self-attention (SA) and cross-attention (CA), FFN is the feed-forward operator, and b_{AB}, b_{BA} are the directional biases. The residual scale is $s = \sqrt{T/36}$ for a T-block ESM-C-shaped stack, giving $s = 1$ for the 36-block *Peppermint* model.

Each operator includes input layer normalization. Self-attention uses fused QKV projections. In both self-attention and cross-attention, queries and keys are layer-normalized across the model width before head splitting. Omitting batch indices, each cross-attention head has $Q_A \in \mathbb{R}^{L_A \times d_h}$ and $K_B, V_B \in \mathbb{R}^{L_B \times d_h}$, with M_B broadcast across query positions. Feed-forward updates use SwiGLU with expansion ratio $8/3$, rounded up to a multiple of 256. We exclude padded keys, zero padded output states, and apply a final layer normalization before the untied 64-token sequence head. Positional treatment and parameter sharing follow Section 2.

The auxiliary 3Di head is untied and predicts 20 structural classes. The missing input state is not a prediction class.

A.2 Local geometry and directional interface bias

The ProteinMPNN-derived local geometry encoder (Dauparas et al., 2022) has three layers of width 128 and a final projection to width 1,152. For each chain, it selects up to 32 nearest Cα neighbors per residue, including self. Neighbor search uses a dense distance matrix and remains quadratic in chain length; edge features and messages are stored only for the selected neighbors.

Each edge contains 16 radial basis features for each of the 25 ordered pairs of $\{\mathrm{N}, \mathrm{C}_\alpha, \mathrm{C}, \mathrm{O}, \mathrm{C}_\beta\}$ atoms, giving 400 distance features. The local encoder recomputes virtual C$_\beta$ from N, C$_\alpha$ and C. For atom-pair distance r, the radial basis features are

$$\phi_k(r) = \exp\left[-\left(\frac{r - \mu_k}{1.25\,\text{Å}}\right)^2\right], \qquad \mu_k = 2\,\text{Å} + \frac{20k}{15}\,\text{Å}, \quad k = 0, \ldots, 15. \tag{8}$$

Here μ_k is the center of radial basis channel k. We concatenate these channels with a learned 16-dimensional encoding of the clipped residue offset $i - j \in [-32, 32]$, where i and j index the query and neighbor residues. We project the resulting 416-dimensional edge vector to width 128 and normalize it. Node states start at zero. Each layer updates nodes using an MLP over the query node, edge and neighbor node, a masked neighbor sum divided by 30, and a feed-forward residual; it then updates the edge states. We share one encoder between chains and evaluate it once before the transformer stack.

The interface bias appends a structure-present indicator to the ordered distance expansion for each cross-chain residue pair. This indicator records finite C$_\alpha$ coordinates at both non-padded endpoints; structural masking also sets it to zero. Unlike the local encoder, the cross-chain feature builder uses supplied C$_\beta$ coordinates when available and otherwise reconstructs virtual C$_\beta$. A shared $401 \to 64 \to 18$ MLP with GELU activation produces per-head logits. Reversing direction exchanges both residue axes and ordered atom-pair channels. The projected biases have shapes $[B_{\mathrm{mb}}, H, L_A, L_B]$ and $[B_{\mathrm{mb}}, H, L_B, L_A]$, where B_{mb} is the local minibatch size, $H = 18$ and L_A, L_B are the padded chain lengths. Pair features are constructed on the model device before the transformer stack, and the biases are reused across blocks.

The structural input modules use geometry or geometric states without amino-acid identities; sequence information enters cross-attention through queries, keys and values.

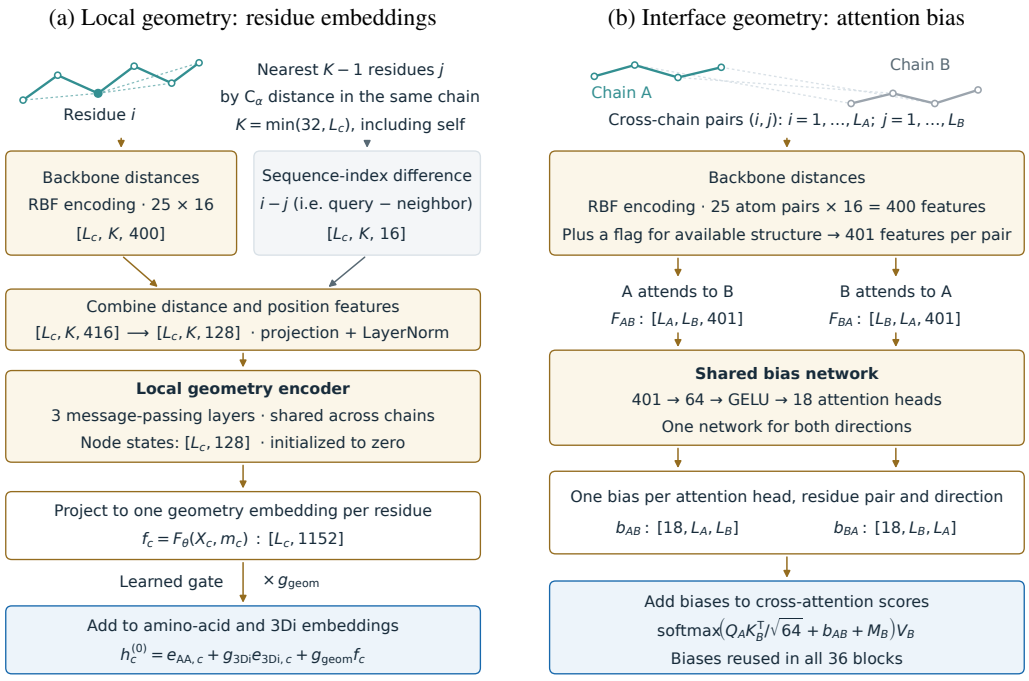


Figure A.1: Two routes for backbone geometry in *Peppermint*. **(a)** A shared local encoder produces per-residue geometry embeddings, gated into the amino-acid and 3Di inputs. **(b)** A shared network converts directional interface features into per-head attention biases, reused across all 36 blocks. Both routes use radial basis function (RBF) distance features but separate learned parameters. Shapes omit the batch axis; L_c is the padded length of chain c and K its neighbor count. Here e denotes input embeddings and M_B masks padded keys; the attention equation applies separately to each head.

A.3 Masking and missing structure

The zero-initialized gates and interface-bias output projection described in Section 2 preserve the sequence model's amino-acid predictions when structural modules are introduced. The auxiliary structural loss changes the training objective.

An encoder node is usable when it is not padded, has finite Cα coordinates and is not structure-masked. Both message endpoints must be usable, and invalid output rows are zero. Other non-finite atom coordinates are zero-filled before local distance calculations; the cross-chain feature builder zeros the RBF channels of unavailable atom pairs and records structural availability explicitly.

Structural masking zeros the cross-chain features and removes the pair from the output-bias mask, so its final bias is zero even after training. Missing coordinates alone need not produce a zero MLP output. For fully structure-free inputs, all cross-chain features are zero, so the bias is constant across valid keys and leaves attention weights unchanged (Section 2).

We use the corruption probabilities and mode mixture in Section 3. Structural masking replaces 3Di inputs by the missing state, removes the corresponding geometric messages, and masks cross-bias features wherever either endpoint is hidden. The auxiliary head predicts only selected positions with a valid native 3Di target, as defined in Eq. 3. AFDB homodimer chains share sequence masks and crop windows, preventing an unmasked copy in the other chain from revealing a target amino-acid identity.

We add Gaussian coordinate noise with standard deviation 0.2 Å after cropping during training and use unperturbed coordinates for validation. We crop and reuse precomputed 3Di states without

recomputing them after noise. Because these states are computed before structural masking, retained states may still depend on a structurally masked residue's coordinates.

Table A.1: Parameter allocation in *Peppermint*. Shared directional parameters are counted once.

Component	Parameters
ESM-C backbone, embeddings and amino-acid head	575,036,992
Bidirectional cross-attention	191,268,864
Auxiliary 3Di head	1,353,620
Local geometry encoder	1,094,448
Interface-bias MLP	26,898
Input 3Di embedding	24,192
Two scalar input gates	2
Total	768,805,016

B Data curation

B.1 Training sources and sampling

For sequence-only STRING data, selection and cluster splitting follow Section 3. The short-crop curriculum uses pairs of total length at most 1,024; the long-crop curriculum extends the limit to 2,048. Structural fine-tuning uses a combined crop budget of 2,048 tokens and a total-length filter of 4,096. Table B.1 reports source weights and length distributions.

To curate a dataset of predicted structures, we select physical interaction links from STRING v12.0 with combined scores greater than 950 (Szklarczyk et al., 2023). Individual protein sequences are clustered with MMseqs2 at 50% sequence identity, retaining one representative link per unordered pair of sequence clusters. We generate complex predictions for 1,515,203 selected pairs using Protenix-v1 (Team et al., 2026). For each pair, we retain the upstream-designated best prediction if pLDDT ≥ 70, ipTM ≥ 0.6 and no clashes are flagged in the prediction metadata. These filters retain 175,821 complexes (11.6%) for structural fine-tuning. No additional sequence clustering is applied during structure packaging.

We select experimentally determined two-chain complexes from Pinder (Kovtun et al., 2024), requiring at least four backbone atom types and experimental resolution at most 3.5 Å (including entries with a reported resolution of zero). We further require predicted biological probability at least 0.5, at least five intermolecular contacts, no more than five missing interface residues under the 4 Å definition, resolved fraction at least 0.5 in each chain, one component per chain and total length at most 4,096. We exclude the official validation and test systems and use the observed, coordinate-aligned sequences. These system exclusions do not impose a separate chain-homology split.

We use the high-confidence AFDB homodimers released by Han et al. (2026), predicted with AlphaFold2-multimer (Evans et al., 2021). This source is already filtered for interface-confidence score ipSAE ≥ 0.6 in both chain directions, mean pLDDT ≥ 70 and at most ten backbone clashes. Packaging removes incomplete records and duplicate entry identifiers and excludes structures that cannot be loaded or parsed, retaining 1,344,906 complexes for training.

The antibody corpus combines experimental SAbDab structures (Dunbar et al., 2014) with antibody–antigen complexes predicted using Protenix-v1 (Team et al., 2026). For each predicted complex, we select the highest-ranked prediction and retain it if ipTM > 0.7 and no clashes are flagged; no additional pLDDT threshold is applied. Experimental records require resolution below 3.5 Å and a single antigen chain.

For each retained antibody complex, we concatenate the heavy- and light-chain sequences, in that order and without a linker, into one model input and use the antigen as the other. Backbone coordinates are concatenated in the same order and retain the complex coordinate frame. Nanobodies retain their single antibody chain. Complexes containing both heavy and light chains also yield a separate example pairing these two chains, without the antigen.

Table B.1: Packaged training sources and structural fine-tuning sampling weights. Lengths are numbers of residues across both chains.

Source	Pairs	Weight	Median length	95th percentile
STRING sequences	435,185,984	0.60	662	2,041
Pinder	481,503	0.20	487	1,066
AFDB homodimers	1,344,906	0.12	736	1,610
Antibody complexes	77,860	0.04	268	1,071
Predicted STRING complexes	175,821	0.04	509	790

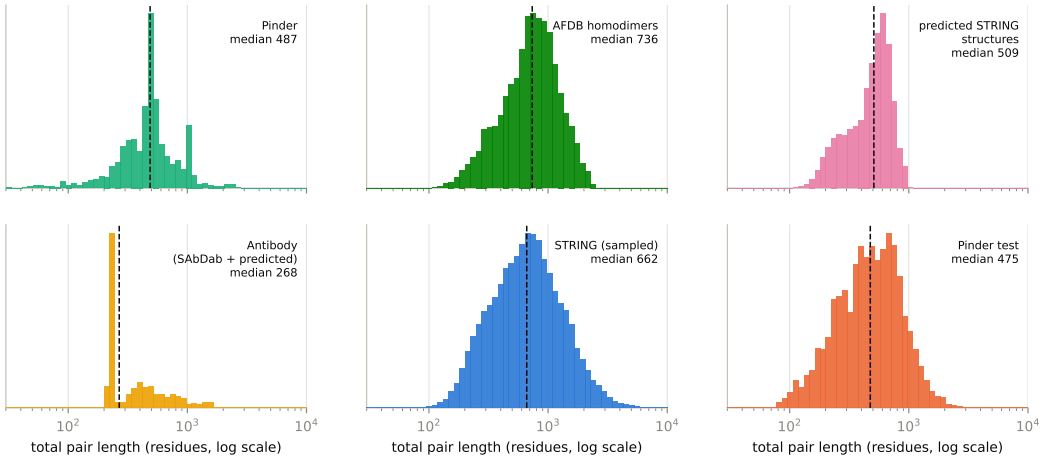


Figure B.1: Combined chain lengths in the packaged corpora of Table B.1. Training applies additional length filters and crops.

B.2 Cropping, optimization and checkpoint selection

Structural examples contain residue-aligned N, C_α, C and O coordinates and 3Di states. The combined crop budget is divided approximately equally between chains, with unused capacity transferred from a shorter chain. Structural crops contain the interface when it fits; otherwise they maximize retained interface residues, breaking ties randomly. Contacts are defined by inter-chain C_α distances below 10 Å; inputs without usable contacts use random contiguous crops. Coordinate noise and handling of precomputed 3Di states follow Appendix A.3.

We use AdamW with $(\beta_1, \beta_2) = (0.9, 0.95)$, $\epsilon = 10^{-8}$ and weight decay 0.01, excluding biases and normalization layers. We clip gradient norms at 1.0 and use bf16 computation. Each stage uses linear warmup, a constant peak learning rate, and cosine decay over the final 16.67% of updates to 10% of the peak. Table 1 gives the stage budgets, learning rates and batch configuration.

For structural-stage checkpoint selection, validation runs every 10,000 updates on at most 50 microbatches per device, averaging their local losses across batches and devices. Coordinate noise is disabled, but crops and masks remain stochastic. Selection uses the combined amino-acid and 3Di objective in Eq. 4 on the Pinder validation stream; update 360,000 is selected from the run completed at update 400,000.

C Evaluation protocols and additional results

C.1 Masked interface recovery

Using the interface and eligibility definitions of Section 4.1, we crop chains to at most 1,022 residues with deterministic interface-centered windows, recompute interfaces on the retained chains and slice the full-chain precomputed 3Di states with the crop. Two complexes reach the crop cap. The retained

structures contain 23,689 geometric interface positions, of which seven lack complete backbone coordinates; we score 23,682 positions, with 12,510 on chain A and 11,172 on chain B.

We use the simultaneous masking and shared 20-amino-acid scoring vocabulary of Section 4.1. All models and input conditions use the same systems, precomputed crop windows and scored positions. Recovery and negative log likelihood pool scored residues across complexes; perplexity is the exponential of this pooled mean negative log likelihood, not of the summed chain losses in Eq. 4.

The ProteinMPNN reference uses its full-backbone model trained with 0.2 Å coordinate noise and 48 neighbors, with no coordinate noise at evaluation (Dauparas et al., 2022). The sequence baselines and their access to the two chains are specified in Section 4.

Sequence-only interface recovery is 39.07% for ESM-C, 35.77% for ESM-2 and 40.45% for MINT.

C.2 Protocols for confidence-selected prediction

We freeze each representation model, mean-pool non-padded residue states separately for each chain and concatenate the pooled vectors. Mutation features concatenate the per-chain wild-type-minus-mutant differences. SKEMPI targets are mutant-minus-wild-type binding free-energy changes, $\Delta\Delta G = \Delta G_{\mathrm{mut}} - \Delta G_{\mathrm{WT}}$, in kcal mol^{-1}, computed at 298.15 K; positive values indicate weaker binding. Complex-level confidence filters follow Section 4.3. Individual low-confidence residues are not removed. Both input conditions use the accepted complexes, with separately fitted heads on the same splits.

We cap each chain at 2,048 residues during downstream inference. Longer structured chains use a deterministic interface-centered crop; longer structure-free chains use a contiguous crop selected using a seed derived from the sequence content. The conditions retain identical rows but can retain different residue windows for overlength chains.

For COVID cross-neutralization, we concatenate heavy and light antibody sequences into one chain input and use spike as the other. We fit a 512-unit MLP to standardized pair embeddings with a fixed random seed, at most 200 iterations and early stopping using 10% of the training data. The accepted subset contains 848 training and 144 test examples.

For Bernett, we fit logistic regression to standardized features with inverse regularization strength $C = 1$, using the predefined sequence-identity splits. For both classification tasks, we compute AUPRC as average precision and use a probability threshold of 0.5 for accuracy and F1.

For SKEMPI, we use ten shuffled outer folds with a fixed seed. Within each outer training fold, five-fold cross-validation selects ridge regularization from ten logarithmically spaced values between 10^{-4} and 10^{4} using inner-fold R^2, fitting standardization within each fold. We report the mean metric across outer folds.

Table C.1: Whole-complex structural coverage for the *Peppermint* evaluation. Counts aggregate all supplied splits and therefore are not test-set sizes. Both input conditions use the accepted subset; withholding structure changes the model input after subset selection.

Benchmark	All rows	Accepted rows	Coverage
COVID	4,417	992	22.46%
Bernett	274,319	2,933	1.07%
SKEMPI	6,706	5,869	87.52%

External-baseline Bernett results retain 2,932 rows across splits, compared with 2,933 for *Peppermint*; its cross-model comparison is descriptive. Confidence filtering changes class balance, so accepted-subset and full-data AUPRC are not directly comparable.

On the confidence-selected Bernett data, the external sequence baselines range from 0.4849 to 0.5150 AUROC and from 0.6931 to 0.7178 AUPRC. On SKEMPI, the structure-conditioned *Peppermint* Pearson and Spearman correlations have outer-fold standard deviations of 0.0385 and 0.0478, respectively; these describe fold variability rather than confidence intervals.

C.3 SEQUENCE-ONLY PREDICTION ON FULL DATASETS

We evaluate all eligible examples in the prescribed COVID, Bernett and SKEMPI partitions, and the TDC and PISTE T-cell receptor binding benchmarks (Velez-Arce et al., 2024; Feng et al., 2024), without selection by structure availability or confidence. All four representation models receive sequences only; *Peppermint* uses the same validation-selected checkpoint as in the main text, with structural inputs withheld. Each model's downstream head is fitted on its full prescribed training partition and evaluated on the corresponding held-out examples, independently of the confidence-selected fits. Differences between the two evaluations therefore also reflect changes in training and evaluation populations.

COVID and Bernett retain the head families and preprocessing in Appendix C.2. COVID uses 3,365 training and 1,052 test examples; Bernett uses 163,093 training, 59,195 validation and 52,031 test examples. SKEMPI uses all 6,706 examples with ten outer and five inner folds. For each task, the four models share the same evaluation examples, labels and split or fold assignments, verified from the saved predictions. Table C.2 reports test metrics for COVID and Bernett and mean outer-fold metrics for SKEMPI.

Table C.2: Full-data downstream prediction with sequence-only inputs. No examples are selected by structural confidence. COVID and Bernett report test metrics; SKEMPI reports means over ten outer folds. RMSE is in kcal mol^{-1}; lower is better. The *Peppermint* column uses the same weights as the main-text evaluation with structure withheld.

Task and metric	MINT	ESM-2 650M	ESM-C 600M	*Peppermint* (seq.)
COVID AUROC	0.6617	0.5285	0.5616	0.6781
COVID AUPRC	0.5321	0.4006	0.3992	0.5312
Bernett AUROC	0.5395	0.5529	0.5613	0.5632
Bernett AUPRC	0.5270	0.5395	0.5418	0.5453
SKEMPI Pearson	0.6419	0.5776	0.5855	0.6400
SKEMPI Spearman	0.5043	0.4455	0.4489	0.5012
SKEMPI RMSE	1.5804	1.6906	1.6710	1.5864

On the full Bernett dataset, *Peppermint* leads ESM-C by 0.0019 AUROC and 0.0034 AUPRC. The comparisons in Table C.2 are point estimates and do not establish statistical significance; Section 4.3 discusses the results across tasks.

TCR benchmarks and structural coverage. TDC provides CDR3β and epitope sequences; PISTE additionally provides an MHC pseudo-sequence. These inputs omit the full receptor scaffold and, for PISTE, the complete MHC molecule, limiting reliable full-complex modeling. We therefore evaluate both tasks without predicted structures.

Both tasks use frozen representations and a 256-unit ReLU MLP, trained with Adam at learning rate 10^{-3}, batch size 32 and binary cross-entropy with positive-class weight 2. TDC encodes CDR3β and the epitope as the two sequence inputs, with dropout 0.1 and ten training epochs. We report means over five provided random-split repetitions, each with 70,236–73,541 test examples. For PISTE, we append the MHC pseudo-sequence to CDR3β with an `X` separator and use the epitope as the second input, with dropout 0.2 and 14 training epochs.

PISTE uses three negative-sampling schemes: Random (`random`) shuffles sequence triples; Reference TCR (`reftcr`) pairs epitopes with TCRs from healthy-donor reference data; and Unified peptide (`unipep`) samples epitopes according to their frequencies in the positive data (Feng et al., 2024). We fit a separate head for each scheme. Following MINT's choice of evaluation cohort (Ullanat et al., 2026), we report the external cancer cohort (`dbpepneo`, PISTE test set II), not the separate pMTnet-derived test set I. The three schemes retain 2,806, 2,533 and 2,955 evaluation examples, respectively (Table C.3).

TDC AUROCs are similar across models (0.5454–0.5481), with differences smaller than the variation across repetitions. On PISTE's cancer cohort, ESM-C has the highest AUROC under Random and

Table C.3: Sequence-only TCR binding prediction without structure filtering. TDC reports means over five random-split repetitions; PISTE uses the external cancer cohort (`dbpepneo`) under three negative-sampling schemes. AUPRC is average precision.

Benchmark / sampling	MINT		ESM-2 650M		ESM-C 600M		*Peppermint* (seq.)	
	AUROC	AUPRC	AUROC	AUPRC	AUROC	AUPRC	AUROC	AUPRC
TDC	0.5481	0.3813	0.5477	0.3770	0.5454	0.3783	0.5479	0.3813
PISTE: Random	0.6628	0.1725	0.6193	0.1512	0.6697	0.1972	0.6603	0.1591
PISTE: Reference TCR	0.6729	0.2113	0.6388	0.1813	0.6344	0.2147	0.6396	0.1794
PISTE: Unified peptide	0.6769	0.1617	0.6618	0.1677	0.7136	0.1698	0.7061	0.1606

Unified peptide sampling and the highest AUPRC under all three schemes; MINT leads AUROC under Reference TCR sampling. *Peppermint* does not lead either metric under any scheme. Sequence-only evaluation therefore extends coverage beyond structurally resolved examples, but does not yield a consistent advantage on these TCR benchmarks.

D Supporting ablations and model-size analyses

This section gives the structural component ablation supporting Section 4.4 and the training protocols, compute accounting and allocation analyses supporting Section 4.5. The model-size studies use the *Peppermint* architecture with sequence pretraining from scratch; the component ablation varies its structural inputs. Model sizes are rounded to the nearest million in the prose and Fig. 3; appendix tables, appendix figures and numerical analyses retain higher precision.

Studies of compute allocation in language and protein modeling (Hoffmann et al., 2022; Cheng et al., 2024) and scaling with fixed amounts of paired antibody data (Neyestanak et al., 2025) motivate the model-size analyses below.

D.1 Structural component ablation

We compare ten configurations, as well as a control without structural input in Table D.1. Each configuration shares an ESM-C-shaped paired sequence initialization with width 704, 22 blocks and 11 heads, containing 183 million sequence-model parameters which has been pretrained on STRING. We fine-tune each for 5,000 updates with effective batch 32, peak learning rate 5×10^{-5}, 500 warmup updates and a warmup–stable–decay schedule. The mixture is 70% Pinder and 10% each AFDB homodimers, antibody complexes and predicted STRING complexes. We retain the auxiliary 3Di head with weight 0.2, use 1,024-token interface-centered crops and evaluate loss on the 1,922-pair Pinder validation stream with the training structural-mask mixture.

Combining cross RBF bias, 3Di tokens and backbone coordinates gives the lowest held-out loss, 5.1157, a reduction of 0.0892 nats from the control and 0.0079 nats from bias plus tokens (Table D.1). Among the ten configurations, input 3Di tokens yield the largest reductions; adding cross RBF bias to tokens lowers loss by 0.0016 nats. Interface-restricted amino-acid accuracy is 18.23% for the control, 19.61% for tokens, 19.80% for tokens plus cross bias and 18.49% for cross bias alone.

D.2 Sequence model-size study

The five models in Fig. 3a use 1,024-token crops, 15% pure-mask corruption, nominal global batch 256 and AdamW with $(\beta_1, \beta_2) = (0.9, 0.95)$. Learning rates are transferred from the 72-million-parameter reference using width and depth. For U optimizer updates, we define

$$D_{\text{nominal}} = U \times 256 \times 1{,}024, \qquad C_{\text{proxy}} = 6 N_{\text{compute}} D_{\text{nominal}}, \tag{9}$$

where D_{nominal} counts configured token capacity and N_{compute} is the rounded sequence-model parameter count used in the analysis. Padding and post-batching length filtering reduce retained exposure. The FLOP axes in Fig. 3 use this parameter-based estimate rather than measured hardware FLOPs; fine-tuning token accounting is specified below. The allocation plots abbreviate these quantities as D and C.

Table D.1: Structural component ablation. Values are the summed per-chain masked amino-acid losses after 5,000 fine-tuning updates; the auxiliary structural loss is excluded. The ten bias/token configurations lack the local geometry encoder and learned input gates; the first row supplies backbone coordinates alongside cross RBF bias and 3Di tokens. No replicated uncertainty estimate is available.

Structural input	Held-out loss	Difference from control
Cross RBF bias + 3Di tokens + backbone coordinates	5.1157	−0.0892
Cross RBF bias + 3Di tokens	5.1236	−0.0813
3Di tokens only	5.1252	−0.0797
Cross RBF bias	5.1908	−0.0141
Cross + intra-chain RBF bias	5.1915	−0.0134
Intra-chain RBF bias	5.2033	−0.0016
Intra-chain binned-distance bias	5.2038	−0.0011
Cα cross bias	5.2041	−0.0008
Cross binned Cα-distance bias	5.2045	−0.0004
Cross binned backbone-distance bias	5.2045	−0.0004
No structural input	5.2049	0

All five runs use the full STRING validation stream on the held-out side of the 50%-identity cluster split, with combined sequence length at most 1,024 residues. This differs from the final STRING-recovery evaluation on 80%-identity representatives (Section 4.2). Validation averages local minibatch losses across batches and devices, with freshly sampled masks. Per-chain loss means follow Eq. 3; their weighting can differ across model sizes with different retained microbatch sizes.

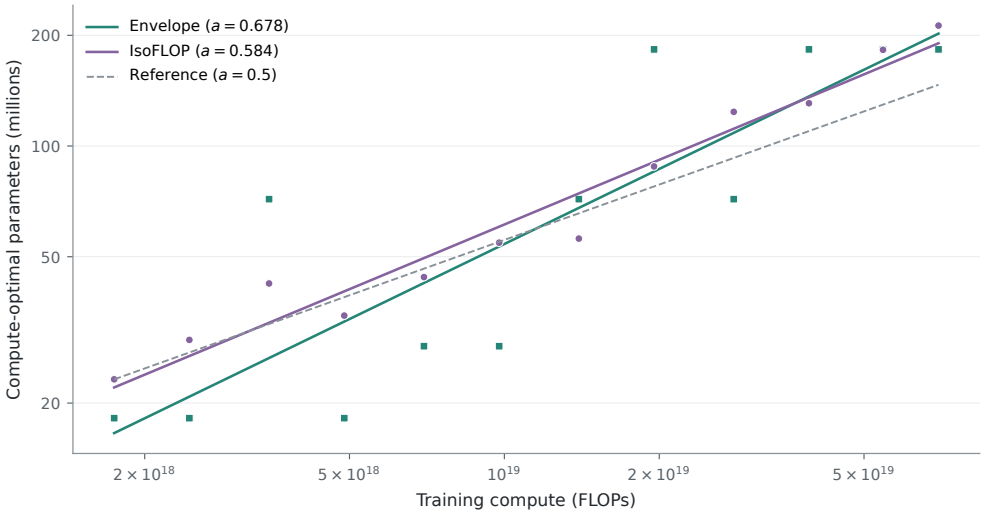


Figure D.1: Loss-minimizing sequence-model size estimated from the empirical loss envelope and IsoFLOP minima. Compute uses C_{proxy} from Eq. 9; the slope-0.5 line is a reference. Training trajectories are shown in Fig. 3a.

Table D.2: Completed sequence model-size study at $C_{\text{proxy}} \simeq 6.979 \times 10^{19}$. D_{nominal} is the configured token budget, not counted nonpadding residues. Total parameters N_{total} count each shared weight once.

Parameters	D_{nominal}	Final loss	$D_{\text{nominal}}/N_{\text{total}}$
18.20 M	6.391×10^{11}	3.7695	35,118.0
28.56 M	4.073×10^{11}	3.6354	14,261.9
71.71 M	1.622×10^{11}	3.5372	2,261.7
183.13 M	6.351×10^{10}	3.3832	346.8
443.70 M	2.621×10^{10}	3.4713	59.1

Following Hoffmann et al. (2022), we estimate loss-minimizing model size from the lower envelope of training curves, selecting the lowest loss across model sizes at each budget, and from quadratic fits in $\log_{10} N$ at 12 fixed compute budgets (IsoFLOP analysis; Figs. D.1 and D.2a). We also fit a loss surface

$$L(N, D_{\text{nominal}}) = E + AN^{-\alpha} + BD_{\text{nominal}}^{-\beta}. \tag{10}$$

Here L is the held-out loss, $N = N_{\text{compute}}$, E a fitted loss floor, and A, B, α, β fitted coefficients and exponents. The surface uses 55 stable-phase evaluations with Huber loss in log space, excluding fully decayed endpoints; the envelope and IsoFLOP analyses include those endpoints. All three methods reuse the same five trajectories.

The fitted allocation is $N_{\text{opt}} \propto C_{\text{proxy}}^{0.592}$ and $D_{\text{nominal,opt}} \propto C_{\text{proxy}}^{0.408}$; the parameter exponent has a 10th–90th percentile checkpoint-resampling interval of 0.358–0.660. The envelope and IsoFLOP estimates are 0.678 and 0.584. These intervals describe sensitivity to resampling correlated checkpoints, not variation across independent training runs.

Table D.3: Sequence loss-surface parameters and 10th–90th percentile checkpoint-resampling intervals.

Parameter	Estimate	10th percentile	90th percentile
E	2.5576	2.1639	3.0271
A	28.777	15.313	1,216.8
α	0.1986	0.1488	0.4561
B	587.31	250.21	1,416.3
β	0.2887	0.2436	0.3347
$\beta/(\alpha+\beta)$	0.5924	0.3577	0.6604

Under nominal pair-draw accounting, only the smallest model enters a second pass through the packaged training corpus: the 18M and 29M models reach approximately 1.43 and 0.91 passes, respectively. These counts precede length filtering and are distinct from retained residue exposure.

D.3 Structure conditioning across model sizes

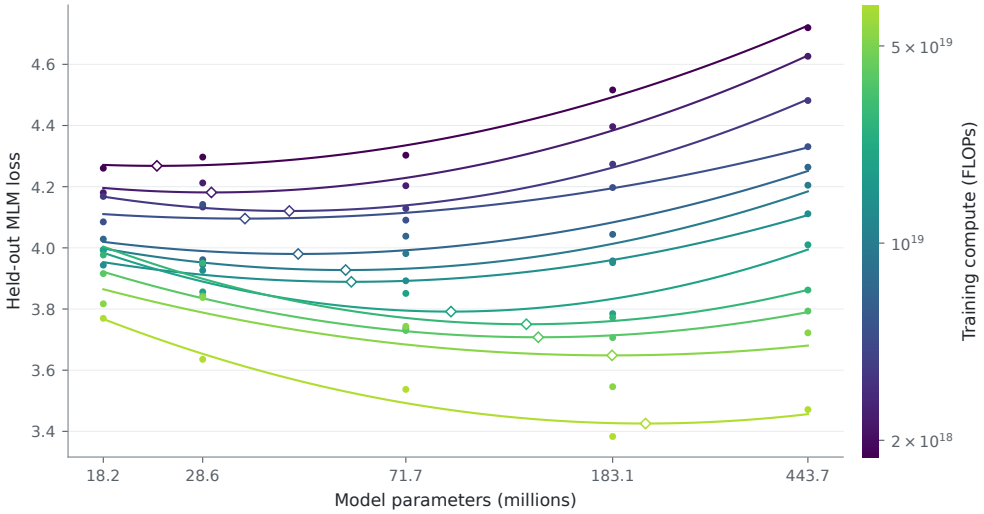
For the fine-tuning trajectories in Fig. 3b, each structural-input/control pair starts from the same sequence checkpoint and retains auxiliary 3Di supervision. Both conditions use the 70/10/10/10 structural mixture specified above and 1,024-token interface-centered crops. Validation samples the training structural-mask modes rather than always exposing the full structure. Peak learning rates are $(7.420, 6.770, 5.860, 5.000, 4.280) \times 10^{-5}$ in increasing model-size order, with warmup occupying 10% of each update budget.

Here $D_{\text{nominal}} = 65{,}536\,U$ for U updates. Equation 9 uses the same rounded sequence-model parameter count for both input conditions, without additional structural-module FLOP accounting. Terminal budgets are approximately 8.0×10^{18} nominal FLOPs across sizes, giving the update counts in Table D.4.

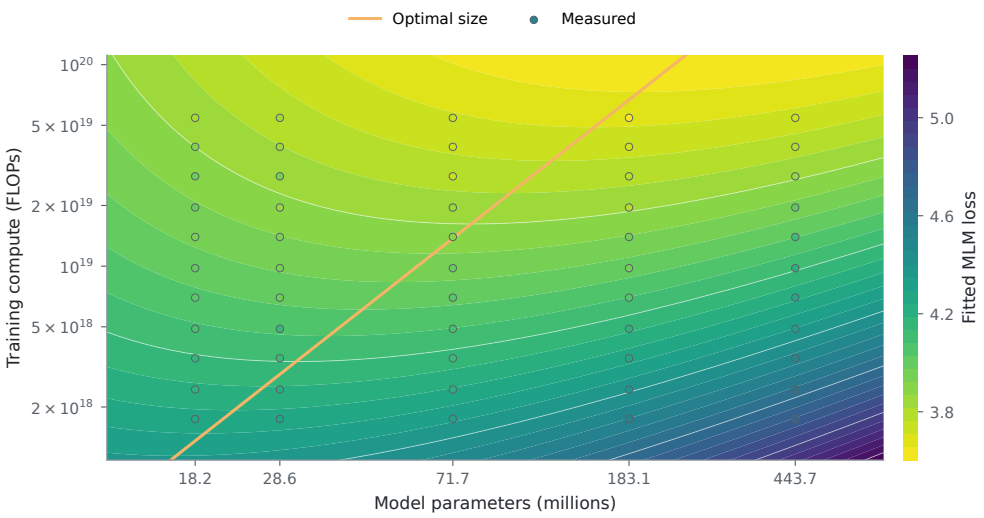
Table D.4: Fine-tuning budgets and first attainment of held-out MLM loss 4.0. Intervals give adjacent recorded updates bracketing the first crossing, not uncertainty intervals. Both input conditions receive the same final update budget; "not reached" denotes no crossing within that budget.

		First loss-4.0 crossing (updates)	
Parameters (M)	Final update	Structural input	No structural input
18.20	1,122,000	78,540–112,200	112,200–157,080
28.56	714,000	142,800–199,920	199,920–285,600
71.71	285,000	79,800–114,000	199,500–242,250
183.13	111,000	Not reached	Not reached
443.70	45,900	Not reached	Not reached

Table D.4 brackets each crossing by the last recorded loss above 4.0 and the first at or below it; all later recorded losses remain below the threshold. The threshold lies within the sharp decline in the three smaller models and measures attainment of a common loss, without estimating an exact



(a) IsoFLOP profiles.



(b) Parametric loss surface.

Figure D.2: Sequence model-size diagnostics under C_{proxy} from Eq. 9. Quadratic fits locate loss minima at matched proxy budgets. The surface fit uses the stable-phase evaluations in Eq. 10; the orange curve is its analytic loss-minimizing size.

inflection point or final plateau. Brackets are not uncertainty intervals and use no interpolation or extrapolation.

After the loss decline, the two smallest models with structural inputs maintain losses near 1.5, compared with approximately 2.0–2.4 for their controls without structural inputs. The 72M control continues to improve after its decline, leaving its plateau unresolved.