Jointly Embedding Protein Structures and Sequences through Residue Level Alignment

Foster Birnbaum[*]
Department of Biology
MIT
fosterb@mit.edu

Saachi Jain[*]
EECS
MIT
saachij@mit.edu

Aleksander Madry
EECS
MIT
madry@mit.edu

Amy E. Keating
Department of Biology
MIT
keating@mit.edu

Abstract

The relationships between protein sequences, structures, and their functions are determined by complex codes that scientists aim to decipher. While structures contain key information about the protein's biochemical functions, they are often experimentally difficult to obtain. In contrast, protein sequences are abundant but are a step removed from molecular function. In this paper, we propose Residue Level Alignment (RLA) — a self-supervised objective for aligning structure and sequence embedding spaces. By situating structure and sequence encoders within the same latent space, RLA allows the structure encoder to leverage large sequence databases and enriches the sequence encoder with spatial information. Moreover, our framework enables us to measure the similarity between a structure and sequence by comparing their RLA embeddings. We show how RLA similarity scores can be used for binder design by selecting true bound backbone structures from sets containing closely and distantly similar decoys.

1 Introduction

An important goal in biology is to determine the relationships between protein sequence, structure, and function. Protein structures underlie biochemical functions such as binding and catalysis, but they are often expensive and time-consuming to image (e.g., via x-ray crystallography, NMR, or cryo-electron microscopy). In contrast, the amino-acid sequences of proteins are far more accessible [1], but detailed information about function can be difficult to extract from sequence alone. There is thus considerable interest in matching the sequence of a protein to its structure using computational methods. Exciting progress has been made on problems that can be framed as sequence-structure alignment problems, such as structure prediction [2], sequence prediction [3, 4], and binder design [5]. To solve such problems using machine learning, practitioners need to build encoders that represent either the structure or the sequence in a meaningful way.

Several groups have developed large language model encoders of protein sequences [6, 7, 8]. These models learn evolutionary patterns from sequence data and generate powerful representations of the underlying protein. However, although in theory protein sequences and structures contain the same information, in practice some tasks benefit enormously from an embedding of an input structure. For example, binder design requires an encoding of the target structure to properly condition the structure of the binder to match the specific, high-resolution structure of the target [5].

In contrast, structure encoders are limited by data availability. Whereas the UniProt database contains 60 million protein sequences [1], the Protein Data Bank (PDB) — the most comprehensive public database of protein structures — contains only 200,000 structures (as of January 2023), many of which are redundant [9]. In theory this limitation could be addressed by augmenting training data

[*]Authors contributed equally to this work.

Machine Learning for Structural Biology Workshop, NeurIPS 2023.

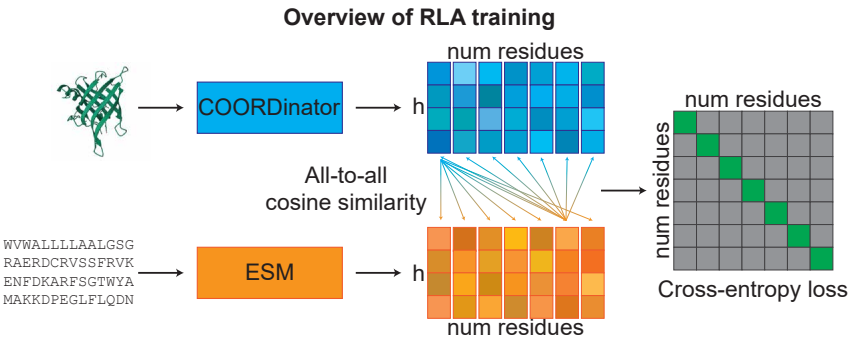


Figure 1: RLA training occurs by minimizing the cross-entropy loss of matching structure (blue) and sequence (orange) embeddings of dimension *h* for all *num residues* residues in a protein.

with millions of AlphaFold2 predicted structures: doing so can improve performance [10]. However, large-scale structure prediction and training on millions of structures is computationally intractable for most groups, and while AlphaFold2 predictions are generally good, they are not perfect and will introduce errors.

Sequence and structure encoders each have fundamental but complementary limitations, both in terms of data-availability and the richness of the modality. We thus ask the following question:

How can we leverage the availability of protein sequences and the spatial information in protein structures to jointly improve both sequence and structure encoders?

Contributions

We propose a new self-supervised objective, *Residue Level Alignment* (RLA) to align the structure and sequence embedding spaces. Specifically, starting with a pre-trained sequence encoder (a large language model such as ESM-2) and a randomly initialized structure encoder (a message-passing neural network, or MPNN), we encourage the latent representation of each residue in the structure embedding to match that residue's representation in the sequence embedding (and vice-versa). By aligning these two spaces, RLA enables the structure encoder to take advantage of large sequence databases while enriching the sequence encoder with high-resolution structural information. We demonstrate that RLA:

- **Injects spatial information into the sequence encoder.** RLA enhances the sequence encoder by indirectly providing structural supervision. RLA improves ESM's performance on unsupervised contact predictions and binding and stability energy predictions.
- **Identifies complementary sequences and structures.** Using RLA scores, we can rank structural decoys according to their similarity to a corresponding sequence at a comparable level of performance to AlphaFold2, at a fraction of the computational cost.
- **Facilitates binder design.** Using RLA, we demonstrate that we can screen for appropriate docking interactions better than AlphaFold2, the current state of the art, even when the sequence for a candidate binder is not specified [11, 12].

2 Methods

Our aim is to design two encoders — one that accepts structure and the other sequence — which each accepts only its own modality but whose latent space is informed by the other modality. Thus, our structure encoder takes advantage of large sequence databases without needing a specific protein sequence (which is not present in many design tasks), and our sequence encoder takes advantage of the spatial information present in structure data without requiring the actual protein structure (which is not known for most sequences).

The key idea is to leverage the fact that protein sequences and structures share the same underlying sub-unit: the residues, i.e., the linked amino acids. Each residue in the protein sequence has an

Method	Overall		Short		Medium		Long	
	Acc	TNR	Acc.	TNR	Acc.	TNR	Acc.	TNR
ESM-2	0.816	0.905	0.894	0.378	0.882	0.371	0.796	0.932
with RLA	**0.907**	**0.971**	**0.935**	**0.790**	**0.937**	**0.839**	**0.898**	**0.979**

Table 1: RLA improves unsupervised contact prediction maps derived when training a linear probe to predict residue contacts given language model attention maps. RLA improves on short-, medium-, and long-range contacts (see Appendix C for definitions of these ranges), and in particular, improves the True Negative Rate (TNR) by reducing the number of false positives.

associated position in the protein structure. Thus, if our individual encoders output an embedding for each residue, we enforce that a residue's structure embedding is more aligned with its corresponding sequence as compared to other residues in the same protein (Figure 1). Note that, unlike traditional cross-modal contrastive learning approaches (e.g., CLIP [13]), this residue level approach does not require a large batch size, making it much easier to scale.

Residue Level Alignment More formally, given a protein with T residues, let $\{U_T\}, \{V_T\} \in \mathbb{R}^d$ be the structure and sequence embeddings for each of the residues. We define the RLA alignment score for residues i, j as the cosine similarity $r_{RLA}(U_i, V_j) = \frac{U_i^T V_j}{\|U_i\| \|V_j\|}$ between the structure embedding for i and sequence embedding for j. We supervise a cross-entropy loss to maximize r_{RLA} for a structure/sequence pair for a single residue relative to the r_{RLA} between different residues.

$$\mathcal{L}_{RLA} = \frac{1}{T}\left(\sum_{i=1}^{T} \log \frac{r_{RLA}(U_i, V_i)}{\sum_{j=1}^{T} r_{RLA}(U_i, V_j)} + \sum_{j=1}^{T} \log \frac{r_{RLA}(U_j, V_j)}{\sum_{i=1}^{T} r_{RLA}(U_i, V_j)}\right)$$

Chain Shuffling Positional information can cause the model to "cheat," for example, by always giving the first residue the same embedding regardless of the inputted protein. To avoid this, we shuffle the chains between the sequence and structure encoder during training.

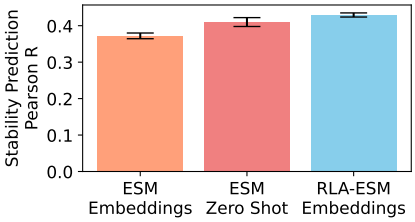
RLA Similarity Score We compute a similarity score to compare a candidate structure and sequence. For structure and sequence embedding $\tilde{U}$ and $\tilde{V}$, we define the similarity score by averaging the RLA scores over all residues $S_{RLA}(\tilde{U}, \tilde{V}) = \frac{1}{T}\sum_{i=1}^{T} r_{RLA}(\tilde{U}_i, \tilde{V}_i)$. For docking applications, we only average the RLA scores corresponding to the residues at the binding interface.

We leverage a pre-trained ESM-2 sequence encoder and a randomly initialized COORDinator structure encoder (see Appendix A.1 for details). We then perform RLA training on the structures within the Protein DataBank (PDB) (see Appendix A for training splits and hyperparameters).

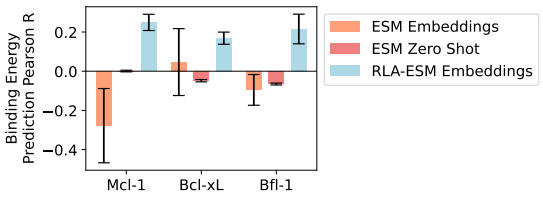
3 RLA injects spatial information into the sequence encoder

RLA aligns the sequence and structure embeddings together into a joint latent space. In Appendix B, we demonstrate that RLA successfully aligns native structure/sequence pairs together and allows the representations to gracefully diverge as noise is added to either modality. In this section, we assess how fine-tuning ESM with RLA can make the sequence model more structurally aware. We compare the performance of ESM and RLA-ESM embeddings on two structure-related tasks: predicting contacts between residues in the protein structure and predicting protein binding and stability energies.

Contact predictions using language model attention maps Rao et al. [14] found that the attention maps of ESM-2 contain structural contact information: i.e., whether two residues are in contact with each other in the folded protein structure. Specifically, they found that training a linear probe on top of the attention maps to predict residue contacts can compete with state-of-the-art contact prediction methods. However, ESM-2 often suffers from false positives. In Table 1, we show that using RLA-ESM significantly improves unsupervised contact prediction over short, medium, and long-range contacts. In particular, RLA-ESM reduces the number of false positives over all ranges. Further experimental details can be found in Appendix C.

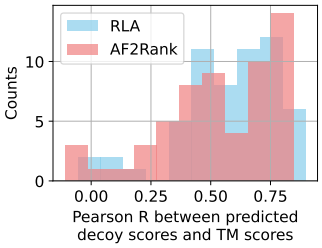


(a) Protein stability prediction.

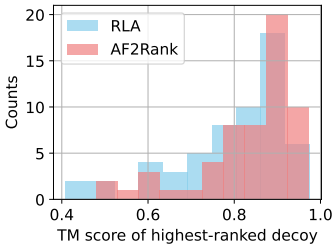


(b) Protein-peptide binding energy prediction.

Figure 2: RLA improves mutation effect prediction using language model embeddings. RLA-ESM embeddings are better able to predict **(a)** protein-peptide binding energies and **(b)** protein stability compared to baseline ESM embeddings. Data shown are mean ± SEM.



(a) Scoring methods correlation with TM scores.



(b) TM score of highest-ranked decoy.

Figure 3: RLA scores discriminate between real and decoy single-chain structures with similar performance to AlphaFold2. **(a)** RLA scores correlate with TM scores approximately as well as AlphaFold2 scores. **(b)** The highest-ranked decoy by RLA scores has approximately as high a TM score as the highest-ranked decoy by AlphaFold2.

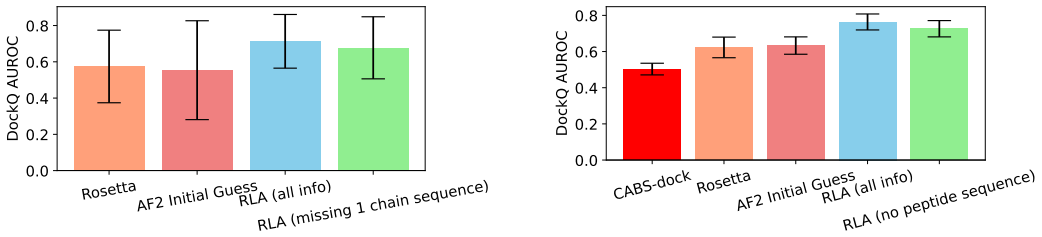
Protein binding energy and stability predictions using RLA scores Meier et al. [15] demonstrated that ESM embeddings can be used to predict the effect of mutations on various protein energy properties (see Appendix D). We find that RLA-ESM outperforms ESM embeddings on predicting protein stability [16] (Figure 2a). On the more difficult task of predicting the binding affinities for peptide mutants binding to 3 proteins [17, 16], RLA-ESM embeddings significantly outperform ESM embeddings (Figure 2b). We anticipate that including the learned RLA-COORDinator structure embeddings will further improve mutation effect prediction.

4 Using RLA to identify complementary sequences and structures

RLA positions both the sequence and the structure encoders within the same latent space. As a result, we can use RLA to identify complementary sequences and structures. In this section, we first evaluate RLA on the decoy ranking task of identifying native sequence-structure pairs in the presence of structural decoys. We then show how RLA can be used for binder design: i.e., identifying a new chain that stably binds to an existing chain.

Using RLA to rank structural decoys Roney and Ovchinnikov [11] showed that AlphaFold2 can be used to rank structural decoys. In Figure 3, we compare the efficacy of using AlphaFold2 and RLA scores to discriminate between native and decoy sequence-structure pairs. The decoys were sourced from structure predictions during the 2020 CASP competition and vary in quality according to the template modeling (TM) score between the prediction and the solved structure. RLA scores are approximately as good as AlphaFold2 scores in predicting the quality of a decoy while requiring significantly fewer compute resources (see Appendix E for details).

Using RLA for binder design We now turn to binder design, a longstanding problem where the task is to design the structure of a protein that binds a specified target. We apply RLA to a sub-task of binder design: discriminating good designs from bad ones. In the de novo design space, multiple sequence alignments (MSAs) are not available: while AlphaFold2 has made significant progress on



(a) Protein-protein complexes.

(b) Protein-peptide complexes.

Figure 4: RLA scores discriminate between bad (DockQ < 0.23) and good (DockQ >= 0.23) decoy complexes, even without peptide sequence information. RLA scores are better at classifying docked **(a)** protein-protein and **(b)** protein-peptide complexes. Classification performance is quantified by the area under the receiver operator curve (AUROC). Data shown are mean ± SEM.

this problem, it still struggles in the absence of MSAs [12]. We thus study whether RLA can be a more reliable method for binder discrimination when MSAs are not available.

We consider two datasets with decoy binder complexes: one with protein-protein interactions and the other with protein-peptide interactions. We evaluate how well RLA scores or AlphaFold2 confidence metrics discriminate between "bad" and "good" decoy complexes (according to DockQ score [18]). DockQ scores range from 0 to 1, with a score above 0.23 denoting an acceptable docked complex.

When designing a binder, the protein sequence and structure are both usually available, but often only the structure (and not the sequence) of the peptide is known. This is because the peptide sequence is often designed after the structure [19]. Thus, for design applications, an ideal decoy discriminator should be agnostic to the binder sequence. Accordingly, we assessed RLA's performance when the binder sequence is masked. (See Appendix F for details on benchmarks and additional results.)

We first consider a dataset of 10 protein-protein interaction structures sourced from the Critical Assessment of Predicted Interactions (CAPRI) [20]. Each structure has thousands of decoy structures of varying quality. RLA scores perform substantially better than AlphaFold2 or Rosetta scores in differentiating good and bad decoys, even when calculated without the sequence information for one of the chains (selected randomly) (Figure 4a).

Peptides are a binding partner in 15-40% of protein-protein interactions and have substantial therapeutic relevance [21, 22, 23, 24]. We identified 27 protein-peptide interaction structures and generated 10 decoy structures for each using CABS-dock, a docking method [25]. RLA scores are better able to predict which decoy complexes are acceptable binders compared to AlphaFold2 scores, Rosetta energies, or CABS-dock scores (Figure 4b), even when calculated without the peptide sequence.

5 Related Works

At least two other groups have applied contrastive learning in the protein space: [26, 27] developed PepPrCLIP, a contrastively-learned model that aligns the ESM-2 embeddings for the sequences of a protein and its peptide binding partner and uses the aligned score to design new binders; and S-PLM aligns the ESM-2 embeddings of a sequence with the ResNet50 embeddings of its contact map and shows the resulting sequence embeddings better predict structural features [28]. To our knowledge, we are the first group to apply contrastive learning at the residue level to align sequence and structure embeddings. (See Appendix G for more related works.)

6 Conclusion

Protein sequences and structures each have advantages: while structures contain richer spatial context, sequences are far more publicly accessible. By aligning these two latent spaces together, RLA enables us to indirectly supervise each encoder with the other modality. We demonstrate that RLA enriches both the sequence and structure encoders compared to their single-modality counterparts. Moreover, by positioning both encoders in the same space, RLA can identify complementary sequences and structures, making it especially suitable for design tasks.

References

[1] Uniprot: the universal protein knowledgebase in 2023. *Nucleic Acids Research*, 51(D1): D523–D531, 2023.

[2] 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.

[3] Alex J Li, Mindren Lu, Israel Desta, Vikram Sundar, Gevorg Grigoryan, and Amy E Keating. Neural network-derived Potts models for structure-based protein design using backbone atomic coordinates and tertiary motifs. *Protein Science*, 32(2):e4554, 2023.

[4] 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.

[5] 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.

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

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

[8] Ethan C Alley, Grigory Khimulya, Surojit Biswas, Mohammed AlQuraishi, and George M Church. Unified rational protein engineering with sequence-based deep representation learning. *Nature methods*, 16(12):1315–1322, 2019.

[9] Protein data bank: the single global archive for 3d macromolecular structure data. *Nucleic acids research*, 47(D1):D520–D528, 2019.

[10] 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. *ICML*, pages 8946–8970, 2022.

[11] James P Roney and Sergey Ovchinnikov. State-of-the-art estimation of protein model accuracy using AlphaFold. *Physical Review Letters*, 129(23):238101, 2022.

[12] Nathaniel R Bennett, Brian Coventry, Inna Goreshnik, Buwei Huang, Aza Allen, Dionne Vafeados, Ying Po Peng, Justas Dauparas, Minkyung Baek, Lance Stewart, et al. Improving de novo protein binder design with deep learning. *Nature Communications*, 14(1):2625, 2023.

[13] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PMLR, 2021.

[14] Roshan Rao, Joshua Meier, Tom Sercu, Sergey Ovchinnikov, and Alexander Rives. Transformer protein language models are unsupervised structure learners. *BioRxiv*, 2020.

[15] Joshua Meier, Roshan Rao, Robert Verkuil, Jason Liu, Tom Sercu, and Alex Rives. Language models enable zero-shot prediction of the effects of mutations on protein function. *Advances in Neural Information Processing Systems*, 34:29287–29303, 2021.

[16] Kotaro Tsuboyama, Justas Dauparas, Jonathan Chen, Elodie Laine, Yasser Mohseni Behbahani, Jonathan J Weinstein, Niall M Mangan, Sergey Ovchinnikov, and Gabriel J Rocklin. Mega-scale experimental analysis of protein folding stability in biology and design. *Nature*, 620(7973): 434–444, 2023.

[17] Vincent Frappier, Justin M Jenson, Jianfu Zhou, Gevorg Grigoryan, and Amy E Keating. Tertiary structural motif sequence statistics enable facile prediction and design of peptides that bind anti-apoptotic Bfl-1 and Mcl-1. *Structure*, 27(4):606–617, 2019.

[18] Sankar Basu and Björn Wallner. Dockq: a quality measure for protein-protein docking models. *PloS one*, 11(8):e0161879, 2016.

[19] Sebastian Swanson, Venkatesh Sivaraman, Gevorg Grigoryan, and Amy E Keating. Tertiary motifs as building blocks for the design of protein-binding peptides. *Protein Science*, 31(6): e4322, 2022.

[20] Marc F Lensink and Shoshana J Wodak. Score_set: a capri benchmark for scoring protein complexes. *Proteins: Structure, Function, and Bioinformatics*, 82(11):3163–3169, 2014.

[21] Victor Neduva, Rune Linding, Isabelle Su-Angrand, Alexander Stark, Federico de Masi, Toby J Gibson, Joe Lewis, Luis Serrano, and Robert B Russell. Systematic discovery of new recognition peptides mediating protein interaction networks. *PLoS biology*, 3(12):e405, 2005.

[22] Lei Wang, Nanxi Wang, Wenping Zhang, Xurui Cheng, Zhibin Yan, Gang Shao, Xi Wang, Rui Wang, and Caiyun Fu. Therapeutic peptides: Current applications and future directions. *Signal Transduction and Targeted Therapy*, 7(1):48, 2022.

[23] Ruei-Min Lu, Yu-Chyi Hwang, I-Ju Liu, Chi-Chiu Lee, Han-Zen Tsai, Hsin-Jung Li, and Han-Chung Wu. Development of therapeutic antibodies for the treatment of diseases. *Journal of biomedical science*, 27(1):1–30, 2020.

[24] Sameer Sachdeva, Hyun Joo, Jerry Tsai, Bhaskara Jasti, and Xiaoling Li. A rational approach for creating peptides mimicking antibody binding. *Scientific reports*, 9(1):997, 2019.

[25] Mateusz Kurcinski, Michal Jamroz, Maciej Blaszczyk, Andrzej Kolinski, and Sebastian Kmiecik. CABS-dock web server for the flexible docking of peptides to proteins without prior knowledge of the binding site. *Nucleic acids research*, 43(W1):W419–W424, 2015.

[26] Kalyan Palepu, Manvitha Ponnapati, Suhaas Bhat, Emma Tysinger, Teodora Stan, Garyk Brixi, Sabrina RT Koseki, and Pranam Chatterjee. Design of peptide-based protein degraders via contrastive deep learning. *BioRxiv*, 2022.

[27] Suhaas Bhat, Kalyan Palepu, Vivian Yudistyra, Lauren Hong, Venkata Srikar Kavirayuni, Tianlai Chen, Lin Zhao, Tian Wang, Sophia Vincoff, and Pranam Chatterjee. De novo generation and prioritization of target-binding peptide motifs from sequence alone. *BioRxiv*, 2023.

[28] Duolin Wang, Usman L Abbas, Qing Shao, Jin Chen, and Dong Xu. S-plm: Structure-aware protein language model via contrastive learning between sequence and structure. *BioRxiv*, 2023.

[29] Alexander Rives, Joshua Meier, Tom Sercu, Siddharth Goyal, Zeming Lin, Jason Liu, Demi Guo, Myle Ott, C Lawrence Zitnick, Jerry Ma, et al. 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.

[30] John Ingraham, Vikas Garg, Regina Barzilay, and Tommi Jaakkola. Generative models for graph-based protein design. *Advances in Neural Information Processing Systems*, 32, 2019.

[31] Helen M Berman, Tammy Battistuz, Talapady N Bhat, Wolfgang F Bluhm, Philip E Bourne, Kyle Burkhardt, Zukang Feng, Gary L Gilliland, Lisa Iype, Shri Jain, et al. The protein data bank. *Acta Crystallographica Section D: Biological Crystallography*, 58(6):899–907, 2002.

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

[33] Nicholas Bhattacharya, Neil Thomas, Roshan Rao, Justas Dauparas, Peter K Koo, David Baker, Yun S Song, and Sergey Ovchinnikov. Single layers of attention suffice to predict protein contacts. *BioRxiv*, 2020.

[34] Jun Cheng, Guido Novati, Joshua Pan, Clare Bycroft, Akvilė Žemgulytė, Taylor Applebaum, Alexander Pritzel, Lai Hong Wong, Michal Zielinski, Tobias Sargeant, et al. Accurate proteome-wide missense variant effect prediction with AlphaMissense. *Science*, 381(6664):eadg7492, 2023.

[35] P Benjamin Stranges and Brian Kuhlman. A comparison of successful and failed protein interface designs highlights the challenges of designing buried hydrogen bonds. *Protein Science*, 22(1):74–82, 2013.

[36] Sankar Basu and Björn Wallner. Finding correct protein–protein docking models using ProQ-Dock. *Bioinformatics*, 32(12):i262–i270, 2016.

[37] Zaixiang Zheng, Yifan Deng, Dongyu Xue, Yi Zhou, Fei Ye, and Quanquan Gu. Structure-informed language models are protein designers. *BioRxiv*, 2023.

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

[39] Minkyung Baek, Frank DiMaio, Ivan Anishchenko, Justas Dauparas, Sergey Ovchinnikov, Gyu Rie Lee, Jue Wang, Qian Cong, Lisa N Kinch, R Dustin Schaeffer, et al. Accurate prediction of protein structures and interactions using a three-track neural network. *Science*, 373(6557):871–876, 2021.

[40] Michael Heinzinger, Konstantin Weissenow, Joaquin Gomez Sanchez, Adrian Henkel, Martin Steinegger, and Burkhard Rost. ProstT5: Bilingual language model for protein sequence and structure. *BioRxiv*, 2023.

[41] 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*, 2023.

A RLA Training

A.1 Single Modality Encoders

For our sequence encoder, we fine-tune a pre-trained ESM-2 (150M) [6, 7]. This language model is trained on over 60 million protein sequences with a masked objective. ESM-2 residue embeddings reflect biochemical properties and evolutionary conservation [29]. The attention weights derived from ESM-2 sequence embeddings can be used as contact map predictions, and the embeddings themselves form the basis for the structure predictions of ESMFold [6, 7]. However, without structural information, ESM-2 contact maps are vulnerable to false positives [6, 7], and, without fine-tuning, can be insensitive to single amino acid substitutions [29].

For our structure encoder, we use COORDinator, an MPNN designed for sequence prediction [3]. COORDinator operates on a k-NN backbone structure graph with backbone residues as nodes and interactions between residues as edges. Following other MPNN-based sequence predictors, we use $k = 30$; node features are initialized as an encoding of the three local dihedral angles for each residue; and edge features are initialized as an encoding of the relative positions and orientations and all pairwise backbone atom distances for each pair of interacting residues [30, 4]. Node and edge features are updated by alternating edge-update and node-update message passing layers. For an edge, the update is computed based on the current edge feature and the current features of the nodes the edge connects; for a node, the update is computed based on the current node features of all k neighbors and the updated features of the edges that connect the node to its neighbors [3]. The COORDinator architecture can be used to learn a Potts model comprised of single- and pair-terms of interacting residues that describes the sequence-structure energy landscape for the input protein backbone. This Potts model can be used to power sequence and energy predictions [3]. We decided to use randomly initialized weights instead of pre-trained COORDinator weights to bias the resulting joint embedding space towards the ESM embeddings.

A.2 Dataset Details

We train on the Protein Data Bank (PDB) [31]. Following the procedure from ESM-2 we split the PDB examples based on a temporal cutoff: examples added to PDB before 2021-08-01 were put into the train split and those after into a test split. We also exclude examples from CAPRI and the 27 protein-peptide complexes used to generate CABS-dock decoys.

A.3 RLA Hyperparameters

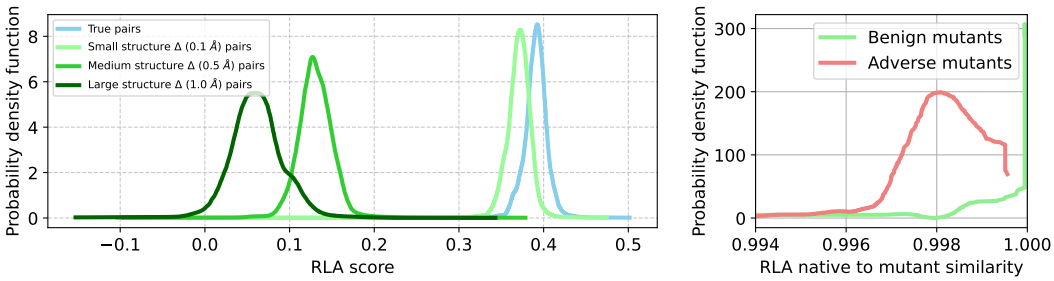
We train with the following hyperparameters:

Name	value
batch size	10
epochs	10
scheduler	cosine
optimizer	adam
learning rate	0.001
peak epoch	2
weight decay	0.001

B Exploring the joint RLA latent space

To explore the landscape of the joint latent space, we investigate how adding noise changes the alignment between sequence and structure pairs. As more noise is added to either the structure or the sequence, the representation of that modality should become progressively more misaligned. First, we add increasing amounts of Gaussian noise to the position of each atom in the input structure and observe that r_{RLA} degrades as the noise increases (Figure 5a).

Second, we make all possible "benign" and "adverse" single amino acid substitutions to all residues in the interface of 10 CAPRI protein-protein complexes [20]. The severity of each substitution is



(a) Adding Gaussian noise to structures.

(b) Making single amino acid substitutions in sequences.

Figure 5: RLA alignment diverges smoothly as noise is added. **(A)** RLA scores get progressively worse as structural noise is added. **(B)** RLA scores are changed more by adverse mutations than benign mutations.

classified according to its BLOSUM62 score, which reflects the mutation's likelihood based on the observed substitution rate in clusters of evolutionary related sequences. Substitutions with a positive BLOSUM62 score are labeled as benign while substitutions with a negative BLOSUM62 score are labeled as adverse [32]. The interface residues to mutate are defined as any residue for which a residue from the other chain is in its set of 30 nearest C-α neighbors Adverse substitutions decrease the RLA score substantially more than benign ones (Figure 5b). RLA thus creates a joint embedding of the sequence and structure of a protein that responds in a predictable manner to sequence or structure perturbations.

C Contact prediction

We follow the same protocol used for unsupervised contact prediction as described in Bhattacharya et al. [33] and Rao et al. [14]. Two residues are defined to be in contact in a structure if their C_α carbons are within 8 Åof each other. Contacts between residues i and j are separated into short ($6 < |i-j| < 12$), medium ($12 < |i-j| < 24$) and long contacts ($|i-j| > 24$) based on their position separation. The attention heads (with average product correction [14]) were extracted from the sequence encoder, and a linear probe was trained to classify the residue contacts. The training dataset consisted of 100 random proteins from the PDB training dataset, where negative contacts were downsampled so that the training dataset was balanced. Then, contact prediction on 100 random proteins from the test set was evaluated.

D ESM mutation effect predictions

Typically, ESM embeddings are used to predict the effect of mutations in a zero-shot fashion (ESM zero shot) using the language model head to compare the mutant and wild-type probabilities for a given residue [15]. Alternatively, the raw ESM embeddings of the mutant and the wild-type can be directly compared (e.g., using cosine similarity) to score the impact of the mutation. Because we did not fine-tune the language head to work with our RLA-ESM embeddings, we calculate mutation scores using the latter approach. For the PepPrCLIP comparison (Figure 7), we calculated the PepPrCLIP scores between the protein sequence and all mutant peptide sequences and computed the Pearson correlation between those scores and the observed binding energies.

We also compared the performance of RLA-ESM embeddings with that of AlphaMissense, an adaptation of AlphaFold trained to predict mutation pathogenicity [34]. Because AlphaMissense model weights are not publicly available, the validation set is limited to the mutations predicted by [34], which all correspond to single amino acid substitutions in human proteins. This ruled out using the protein-peptide binding data, because each mutant peptide has multiple substitutions, and all but 45 of the proteins in the stability dataset. AlphaMissense significantly outperforms RLA (Figure 6), which is not surprising considering that it was trained to predict pathogenicity, and pathogenic proteins are likely unstable. For an additional benchmark using the peptide binding energy data, we compared the performance of RLA-ESM embeddings with that of PepPrCLIP and found that RLA-ESM embeddings perform significantly better (Figure 7) [27].

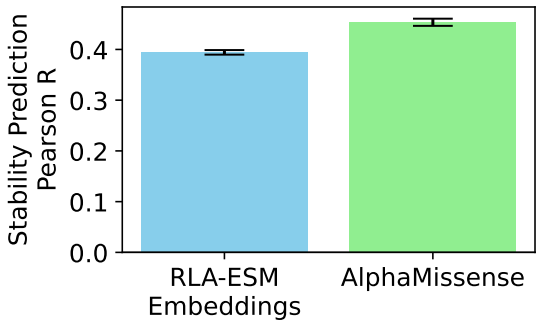


Figure 6: Prediction performance of RLA-ESM compared with AlphaMissense on a subset of the stability data.

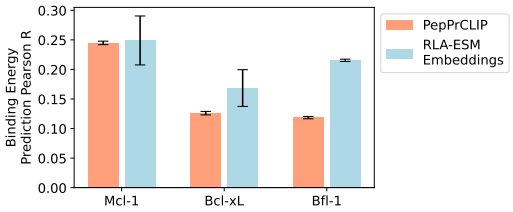


Figure 7: Protein-peptide binding energy predictions by PepPrCLIP and RLA-ESM.

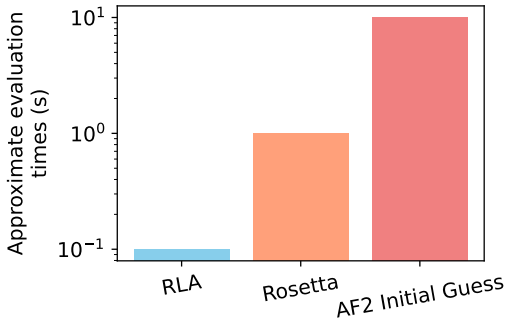


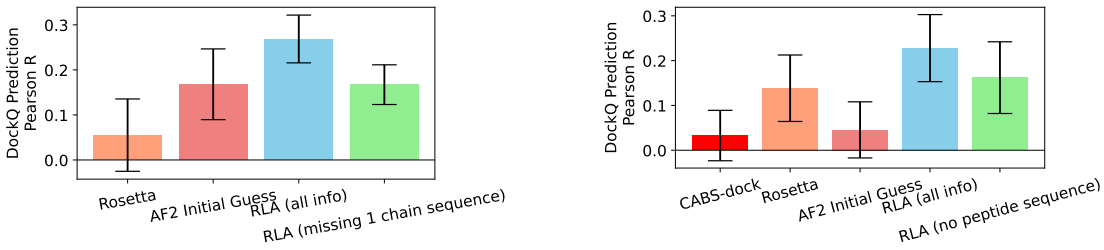
Figure 8: Approximate compute times for each method used in the design tests.

E Computational efficiency

Approximate compute times for each method used in the design tests are shown in Figure 8. RLA scores are approximately an order of magnitude faster to compute than Rosetta interface energies, which are themselves approximately an order of magnitude faster to compute than AF2 Initial Guess scores. The compute time tests were conducted using a single Nvidia Volta V100 GPU and Intel Xeon Gold 6248 20-CPU core processor.

F Design performance

We benchmark our performance on binder complex discrimination against Rosetta interface dG energies and AF2 Initial Guess scores [35, 12]. Rosetta energies were calculated using the InterfaceAnalyzer protocol with the following options: `out:file:score_only`, `add_regular_scores_to_scorefile`, and `tracer_data_print`. AF2 Initial Guess scores were calculated as described by Bennett et al., [12]. Both Rosetta and AF2 Initial Guess model the structure of side chains, which RLA does not. Side chain structure information is often important to binder prediction models [36]. This should provide Rosetta and AlphaFold2 with an advantage in this task.



(a) Protein-protein DockQ correlations.

(b) Protein-peptide DockQ correlations.

Figure 9: RLA scores discriminate between good and bad protein-protein docked complexes, even without sequence information for one of the chains. **(a)** RLA scores correlate better with DockQ scores than any other metric. **(b)** RLA scores are better at classifying docked complexes as incorrect (DockQ $<$ 0.23) or acceptable (DockQ $>=$ 0.23), quantified by the area under the receiver operator curve (AUROC). Data shown are mean $\pm$ SEM.

In addition to calculating AUROC values and quantifying the classification accuracies of each method, we calculate the Pearson correlation between the predicted scores and the DockQ scores. On this test, too, RLA scores outperform Rosetta energies and AF2 Initial Guess scores (Figure 9).

G Additional related works

Several other groups have developed models that combine sequence and structure information. Knowledge-Design and LM-Design combine learned structure encoders with ESM sequence encoders to improve protein sequence design [37, 38]. RFDiffusion, a state-of-the-art diffusion generative model that operates in structure space, is fine-tuned from RoseTTAFold, a structure predictor model that was trained with hundreds of thousands of sequences [5, 39]. ProstT5 uses a language model to embed both the protein sequence and a 1D string representation of the protein structure [40, 41]. The resulting embeddings can be used to predict protein properties and redesign sequences compatible with a fixed backbone structure [40].

H Code

Code for training RLA and calculating RLA scores is available here: https://github.com/MadryLab/rla.