Understanding Protein-DNA Interactions by Paying Attention to Protein and Genomics Foundation Models

Dhruva Abhijit Rajwade[1] **Erica Wang**[2] **Aryan Satpathy**[1] **Alexander Brace**[3, 4]

Hongyu Guo[5] **Arvind Ramanathan**[3, 4] **Shengchao Liu**[6] **Anima Anandkumar**[2] *

Abstract

Protein-nucleic acid (NA) interactions are key in controlling gene regulation. There lies a strong motivation in understanding these interactions, with a goal of engineering these interactions to solve biological problems. Current methods to quantify protein-nucleic acids are mainly experimental and require much time and money. To mitigate this, Deep learning methods have recently been applied to predict Protein-DNA contacts. Although promising, these methods are computationally expensive and face challenges in accuracy. To address these challenges, we propose Seq2Contact, a novel method to predict the protein-NA binding at a single nucleotide (DNA) and single amino acid (Protein) level. Seq2Contact is built on protein and DNA foundation models to obtain nucleotide and amino acid-specific embeddings and then introduces a cross-attention module to obtain the binding contact maps. We employ a sequence-similarity based clustering method to split the train-test data and empirically illustrate that Seq2Contact can achieve state-of-the-art performance, beating existing baselines by almost 20% (F1-Score) for Protein-DNA binding prediction. Our method is computationally more efficient, with up to 80% less memory cost and more than 90% less inference time. Code is available at https://github.com/DhruvaRajwade/Seq2Contact.

1 Introduction

Deep Learning has seen pivotal advances in the field of Structural Biology with the advent of AlphaFold2 (Jumper et al. (2021)) and RoseTTAFold (Baek et al. (2021)), marking never-seen-before progress in the task of tertiary protein structure prediction. More recently, RoseTTAFoldNA (RF2NA) (Baek et al. (2023)) and AlphaFold3(Abramson et al. (2024)) were released, which now support predicting 3D Structures for Protein-Nucleic acid complexes as well. However, these models require a large amount of computational resources for inference [2], while also facing overfitting and structure memorization issues (Chakravarty et al. (2024)).

Protein-nucleic acid (NA) interactions are crucial in many essential biological processes, including gene regulation, transcription, translation, and recombination. These interactions also hold significant potential for therapeutic applications (Bogdanove et al. (2018)). However, current methods for quantifying protein-NA interactions are predominantly experimental, requiring substantial time and resources. While models like RF2NA and AlphaFold3 can predict the 3D structures of protein-NA complexes, they often fall short in accurately modeling interactions at the resolution of individual amino acids and nucleotides, which is an important aspect for fully understanding these interactions.

*[1]Indian Institute of Technology Kharagpur, [2]Caltech, [3]University of Chicago, [4]Argonne National Laboratory [5]NRC Canada, [6]Independent

[2](RF2NA requires around 500GB of storage just for inference as they do MSAs on locally hosted databases, while AlphaFold3 is not open-source, also uses MSAs in its pipeline and has a job limit of 20 jobs per day)

Machine Learning for Structural Biology Workshop, NeurIPS 2024.

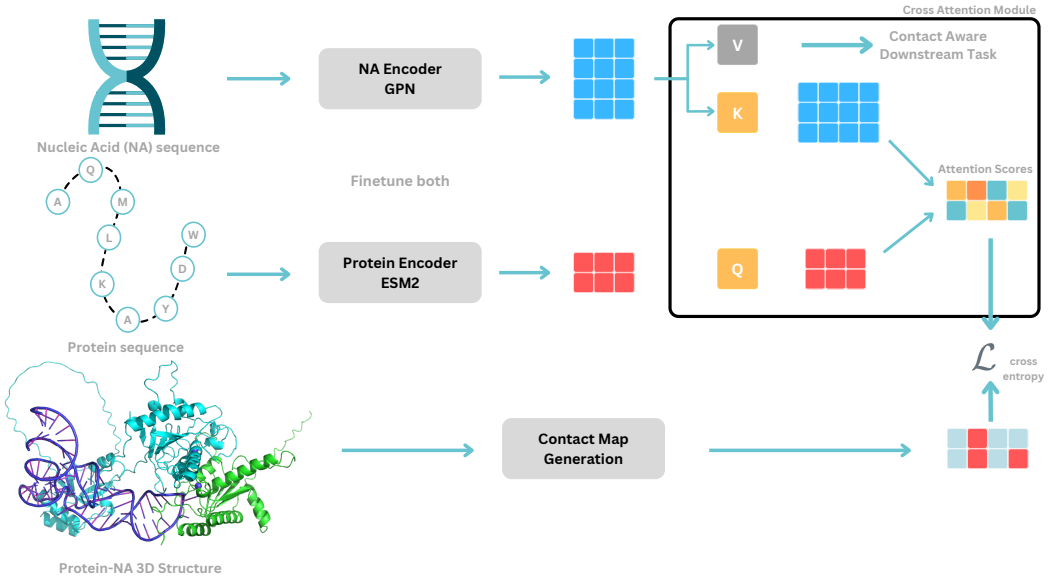


Figure 1: Overview of Seq2Contact: Protein and NA sequences are passed through FMs to get embeddings, which the Cross-Attention module uses to predict binding probabilities for each pair (amino-acid, nucleotide). 3D Structures are used to generate ground truth contact maps, which are compared with the predictions through a cross-entropy loss, and the whole ensemble (Cross-Attention Module + FMs) is backpropagated.

Our contributions: To address these challenges, we present Seq2Contact, a method for predicting protein-DNA binding at a single amino acid and single nucleotide level, exploiting the sequence data of the complexes while using the structural information purely as supervision. We leverage some recent advancements in the field of genomics and protein foundation models (FMs). We categorize these as FMs following Bommasani et al. (2021), who define FMs as models trained on a broad range of data, which can be applied to a wide range of downstream tasks. We use ESM2 (Lin et al. (2023)) and GPN (Benegas et al. (2023)) for this work, which are FMs for the Protein and DNA modalities, respectively. These FMs are trained on millions of sequences using a masked language modeling paradigm and have been demonstrated to capture information about the secondary structure and tertiary structure (Lin et al. (2023)). We exploit this fact and place embeddings produced from these models under a binding lens, which is a Cross-Attention module that predicts the binding probability of every pair of amino acids and nucleotides as a contact map. We partially finetune the FMs (Sec. 2.4) along with the training of our Cross-Attention module (Sec. 2.3). Seq2Contact requires just sequences during inference to obtain binding contact maps and takes only 1.2 GB of GPU memory and 50 seconds for inference (Sec. 3.3.1). Seq2Contact achieves state-of-the-art performance with a PR-AUC of 0.3372 and an F1 score of 0.3445 (Table 1). It surpasses existing methods not only in terms of inference time and memory efficiency but also in the quality of its predictions.

2 Method

2.1 Problem Formulation

For a given Protein sequence $\mathcal{P}$ of length L_p and Nucleic Acid sequence $\mathcal{N}$ of length L_n, our goal is to predict binding, same as a contact map $\mathcal{C} \in \mathbb{B}^{L_p \times L_n}$ such that

$$\mathcal{C}(i,j) = \begin{cases} 1 & \text{if the } i\text{th amino acid and } j\text{th nucleotide are in contact,} \\ 0 & \text{if the } i\text{th amino acid and } j\text{th nucleotide are not in contact.} \end{cases}$$

Given a dataset $\mathcal{D} = \{\mathrm{x}^1, \ldots, \mathrm{x}^{|\mathcal{D}|}\}$ containing triplets $\mathrm{x}_i = (\mathcal{P}^i, \mathcal{N}^i, \mathcal{C}^i)$, we first generate sequence-level embeddings for $\mathcal{P}$ and $\mathcal{N}$, which is then passed into our model $\mathcal{M}$ to predict the contact

probability $\hat{\mathcal{C}} = p(\mathcal{C} = 1|\mathcal{P}, \mathcal{N})$ containing values between [0, 1] (therefore representing binding in a continuous value).

2.2 Contact Map Generation from 3D Structures

We start by collecting all protein-NA complex structures available on the PDB (Berman et al. (2003)) and the NAKB (Lawson et al. (2023)) databases. Similar to Huang et al. (2024), given any Protein-NA complex structure, we represent all amino acids and all nucleotides as points in 3D space, based on the XYZ coordinates of the C_α and C_4' carbon atoms, respectively. We calculate Euclidean distances for all pairs of points (P_i, P_j) for all $i, j \in [L_p, L_n]$ and define a threshold distance [3] of 6Å to classify all pairs as binding or non-binding. This gives us a binary contact map $\mathcal{C} \in \mathbb{B}^{L_p \times L_n}$ with values 1 (binding) or 0 (non-binding). We use these contact maps as supervision for training our Cross-Attention module (Sec. 2.3).

2.3 Cross-Attention Module

We start by generating sequence-level embeddings for protein ($\mathbf{e}_p^i|_{i=1,...L_p}$) and NA ($\mathbf{e}_n^j|_{j=1,...L_n}$). We use Protein as the query and NA as the key in our cross-attention module and project both embeddings to a common embedding dimension d using the projectors $W_q \in \mathbb{R}^{d \times d_p}$ and $W_k \in \mathbb{K}^{d \times d_n}$ to get $\mathbf{q}$ and $\mathbf{k}$. The Attention map $\mathcal{A}$ (of shape (L_p, L_n)) is then obtained by multiplying query and key matrices.

$$\mathbf{q}^i = W_q \mathbf{e}_p^i \qquad \mathbf{k}^j = W_k \mathbf{e}_n^j$$

$$\mathcal{A}_{ij} = \frac{\mathbf{q}^i \cdot \mathbf{k}^j}{\sqrt{d}}$$

Instead of performing a **softmax** on $\mathcal{A}$ to get the attention scores, we apply **sigmoid** on $\mathcal{A}$ to get probabilities of contact $\hat{\mathcal{C}} = p(\mathcal{C} = 1|\mathcal{N}, \mathcal{P})$ (or binding strength) between [0, 1]. Since it is a Binary Classification problem, we optimize the weighted binary cross-entropy loss $\mathcal{L}_{bce}$ to train our model.

$$L_{ij} = -\frac{1}{w_c + 1}\left[w_c \mathcal{C}_{ij} \cdot \log\hat{\mathcal{C}}_{ij} + (1 - \mathcal{C}_{ij}) \cdot \log(1 - \hat{\mathcal{C}}_{ij})\right]$$

$$\mathcal{L}_{bce} = -mean(L)$$

where w_c is a hyperparameter that accounts for class imbalance.

In addition to contact map prediction, using cross-attention to model binding makes it possible to generate binding-aware joint embedding ($\mathbf{e} \in \mathbb{R}^{L_p \times d_o}$) that can be used for downstream tasks such as 3D complex structure decoding. Consider a value matrix $\mathbf{v} \in \mathbb{R}^{L_n \times d_o}$, obtained by projecting NA embeddings using a projection matrix $W_v \in \mathbb{R}^{d_o \times d_n}$.

$$\mathbf{v}^j = W_v \mathbf{e}_n^j \qquad \mathbf{e}^i = \text{softmax}(\mathcal{A}_i)\mathbf{v}$$

2.4 Finetuning Protein and NA foundation models

Figure 1 shows the end-to-end pipeline for Seq2Contact. We use ESM2 (Lin et al. (2023)) and GPN (Benegas et al. (2023))foundation models for our work to generate sequence-level representations of Protein and DNA, respectively, which are used as input to the Cross Attention module. However, these raw embeddings are unsuitable for our objective (Table 2) as they do not embed any information regarding the binding of protein and NA.

For downstream tasks using pre-trained protein language models, it has been shown (Valeriani et al. (2024)), Li et al. (2022)) that the last layer representations of pre-trained models might not be optimal. Recently, Schmirler et al. (2024) showed that finetuning Protein language models for residue-specific tasks (including secondary structure and disorder prediction) is generally beneficial. We finetune ESM2 and GPN by optimizing their last layers with a small learning rate (compared to the Cross Attention module) along with the Cross Attention module and notice a significant improvement in performance (Table 2). Finetuning encourages the foundation models to embed binding information by adding prior to the interaction information of individual nucleotides and amino acids.

[3] Our baselines (AlphaFold3 (Abramson et al. (2024)), RF2NA (Baek et al. (2023)) and FAFormer (Huang et al. (2024))) use threshold distances of 5,7 and 6 Å respectively

Table 1: Comparison of our method and baselines based on F1, PR-AUC, and MCC

Method	F1	PR-AUC	MCC
Seq2Contact (Ours)	**0.3445**	**0.3372**	**0.3638**
AlphaFold3	0.1615	0.1612	0.1627
RF2NA*	0.0850	0.1015	-
FAFormer*	0.1457	0.1279	-
Random	0.0058	0.0094	0.0012

* denotes that the values have not been reproduced and are taken from Huang et al. (2024)

Table 2: Metrics for Seq2Contact Under Different Gradient Configurations

Method	F1	PR-AUC	MCC	Loss
Both fine-tuned	**0.3445**	**0.3372**	**0.3638**	.1697
ESM2 frozen | GPN fine-tuned	0.1339	0.1021	0.1407	0.2076
ESM2 fine-tuned | GPN frozen	0.2326	0.2064	0.2549	0.1708
Both frozen	0.0979	0.0723	0.1022	0.2032

3 Experiments

We first introduce evaluation metrics. We use the PR-AUC metric, as it is particularly valuable in imbalanced datasets because it focuses on the model's performance in predicting the minority class, without being influenced by the abundance of the majority class. We also use F1-score and Matthew's correlation coefficient as metrics, and equations for all metrics are provided in appendix H

3.1 Baselines

To compare the efficacy of Seq2Contact, we use AlphaFold3 (Abramson et al. (2024)), RF2NA (Baek et al. (2023)), FAFormer (Huang et al. (2024)), and a random baseline (appendix E) as a comparison for the DNA-Protein contact prediction task. For AlphaFold3, we randomly choose 60 structures from the AlphaFold3 evaluation set, while for FAFormer and RF2NA we report metrics from Huang et al. (2024)[4]. For the random baseline, we sample contact maps randomly based on the unconditional Bernoulli distribution of binding.

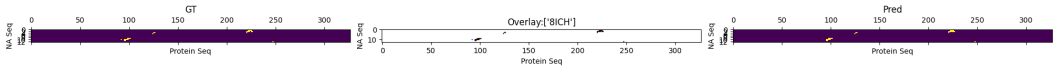
3.2 Experimental Details

For all experiments, we use a single RTX 4090 GPU which has 24 GB of memory, along with an AMD Ryzen 9 7900X 12-Core Processor and 64 GB of memory. We use the 8 Million parameter variant of ESM2 (which consists of 6 Transformer layers), and the 23 Million parameter variant of the GPN FM (which consists of 24 Transformer layers). For this work, we finetune the last layers of either of the models. For our dataset we process Protein-NA 3D structures from NAKB (Lawson et al. (2023)) and extract sequences, and generate contact maps as described in Sec. 2.2. We provide dataset statistics and processing details in appendix D.

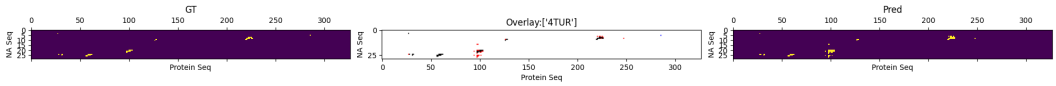
3.3 Results

We present results for the DNA-Protein Contact Prediction Task in Table 1. We see that Seq2Contact outperforms all the baselines, and achieves state-of-the-art metrics. To gain further insight into the influence of the Protein and DNA FMs in our task, we conduct further experiments, where we freeze either FM, freeze neither FM and freeze both FMs. We show the results for this analysis in Table 2, and show some good model predictions in Figure 2. We see that both the Protein and DNA FM play important roles in the binding prediction, and fine-tuning both FMs helps their representation spaces better align with the binding prediction task (Table 2).

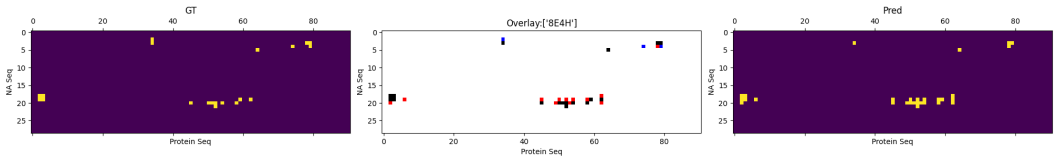
[4]Code and dataset details are unavailable for FAFormer, and they provide metrics for RF2NA as their baseline



(a) PDB: 8ICH



(b) PDB: 4TUR



(c) PDB: 8E4H

Figure 2: Contact maps (Ground-Truth (GT), Overlay of GT and Prediction (Pred), and Pred) shown for three complexes. Purple indicates no contact, and yellow indicates contact. In the overlays, black indicates a match, red indicates a Pred mismatch and blue indicates a GT mismatch

3.3.1 Memory and Inference Time Efficiency Comparison

We tracked inference time and GPU utilization for Seq2Contact and found out that our method takes less than 1.2 GB of cumulative GPU memory (for our entire evaluation dataset of 749 sequence pairs), and is able to infer a single contact map in less than 0.5 seconds (taking a total of 50 seconds for the 749 sequence pairs in our evaluation dataset). As a comparison, we tested inference on the RF2NA pipeline, after downloading the sequence databases required to do so (around 500 GB when zipped). We chose a single protein-DNA sequence pair of length 360 amino acids and 60 nucleotides (close to the mean sequence lengths in our evaluation dataset). For this single sequence pair, RF2NA took around 4 minutes for the MSA, and a further 1 minute and 8 GB of GPU memory for structure generation. This was expected since RF2NA is a structure prediction method with a lot more modules, layers, and parameters than our model. For AlphaFold3, although the code is not open-source, they report in their work that they use 16 A100 GPUs (which sums up to 1280 GB of total GPU memory) for inference, and for a single complex of token length 1024, it takes 22 seconds for structure prediction. For FAFormer, since the code is not public we cannot quantitatively compare memory usage and inference times, which is one of our goals for the future.

4 Discussion and Future Work

Seq2Contact achieves State-of-the-art performance on the DNA-Protein contact prediction task, as displayed in Table 1. Our method effectively leverages the highly informative single-element embeddings that ESM2 and GPN provide and also captures global correlations between each element and the rest of the complex through Cross-Attention. Finetuning the FMs along with the Cross-Attention module proves highly beneficial in overall performance as shown in Table 2, and also following the findings of Schmirler et al. (2024). Seq2Contact requires only sequence data for inference and is fast and computationally inexpensive as displayed in Sec. 3.3.1. There are many exciting future directions for this work. The value vector, which is the output of the Cross-Attention module (Figure 1), is a binding-aware joint embedding containing protein and NA information. One can use this joint-embedding space for many downstream tasks including structure prediction or sequence generation similar to ESMFold (Lin et al. (2023)). However, ESM2 (used by ESMFold) is not generative, and for de novo structure prediction applications, using ESM3 (Hayes et al. (2024)) would be much more useful. We are particularly interested in the conditional generation of DNA sequences given a protein sequence and a contact map, as this would allow us to design binding DNA sequences specific to a target protein, along with controlling which motifs in the protein the generated DNA sequence binds to. However, our method also faces some challenges and limitations which we describe in appendix G.

References

Abramson, J., Adler, J., Dunger, J., Evans, R., Green, T., Pritzel, A., Ronneberger, O., Willmore, L., Ballard, A. J., Bambrick, J., Bodenstein, S. W., Evans, D. A., Hung, C.-C., O'Neill, M., Reiman, D., Tunyasuvunakool, K., Wu, Z., Žemgulytė, A., Arvaniti, E., Beattie, C., Bertolli, O., Bridgland, A., Cherepanov, A., Congreve, M., Cowen-Rivers, A. I., Cowie, A., Figurnov, M., Fuchs, F. B., Gladman, H., Jain, R., Khan, Y. A., Low, C. M. R., Perlin, K., Potapenko, A., Savy, P., Singh, S., Stecula, A., Thillaisundaram, A., Tong, C., Yakneen, S., Zhong, E. D., Zielinski, M., Žídek, A., Bapst, V., Kohli, P., Jaderberg, M., Hassabis, D., and Jumper, J. M. Accurate structure prediction of biomolecular interactions with alphafold 3. *Nature*, 630(8016):493–500, May 2024. ISSN 1476-4687. doi: 10.1038/s41586-024-07487-w. URL http://dx.doi.org/10.1038/s41586-024-07487-w.

Alley, E. C., Khimulya, G., Biswas, S., AlQuraishi, M., and Church, G. M. Unified rational protein engineering with sequence-based deep representation learning. *Nature Methods*, 16 (12):1315–1322, October 2019. ISSN 1548-7105. doi: 10.1038/s41592-019-0598-1. URL http://dx.doi.org/10.1038/s41592-019-0598-1.

Baek, M., DiMaio, F., Anishchenko, I., Dauparas, J., Ovchinnikov, S., Lee, G. R., Wang, J., Cong, Q., Kinch, L. N., Schaeffer, R. D., Millán, C., Park, H., Adams, C., Glassman, C. R., DeGiovanni, A., Pereira, J. H., Rodrigues, A. V., van Dijk, A. A., Ebrecht, A. C., Opperman, D. J., Sagmeister, T., Buhlheller, C., Pavkov-Keller, T., Rathinaswamy, M. K., Dalwadi, U., Yip, C. K., Burke, J. E., Garcia, K. C., Grishin, N. V., Adams, P. D., Read, R. J., and Baker, D. Accurate prediction of protein structures and interactions using a three-track neural network. *Science*, 373(6557):871–876, 2021. doi: 10.1126/science.abj8754. URL https://www.science.org/doi/abs/10.1126/science.abj8754.

Baek, M., McHugh, R., Anishchenko, I., Jiang, H., Baker, D., and DiMaio, F. Accurate prediction of protein–nucleic acid complexes using rosettafoldna. *Nature Methods*, 21(1):117–121, November 2023. ISSN 1548-7105. doi: 10.1038/s41592-023-02086-5. URL http://dx.doi.org/10.1038/s41592-023-02086-5.

Benegas, G., Batra, S. S., and Song, Y. S. Dna language models are powerful predictors of genome-wide variant effects. *Proceedings of the National Academy of Sciences*, 120(44), October 2023. ISSN 1091-6490. doi: 10.1073/pnas.2311219120. URL http://dx.doi.org/10.1073/pnas.2311219120.

Berman, H., Henrick, K., and Nakamura, H. Announcing the worldwide protein data bank. *Nature Structural amp; Molecular Biology*, 10(12):980–980, December 2003. ISSN 1545-9985. doi: 10.1038/nsb1203-980. URL http://dx.doi.org/10.1038/nsb1203-980.

Bogdanove, A. J., Bohm, A., Miller, J. C., Morgan, R. D., and Stoddard, B. L. Engineering altered protein–dna recognition specificity. *Nucleic Acids Research*, 46(10):4845–4871, April 2018. ISSN 1362-4962. doi: 10.1093/nar/gky289. URL http://dx.doi.org/10.1093/nar/gky289.

Bommasani, R., Hudson, D. A., Adeli, E., Altman, R., Arora, S., von Arx, S., Bernstein, M. S., Bohg, J., Bosselut, A., Brunskill, E., Brynjolfsson, E., Buch, S., Card, D., Castellon, R., Chatterji, N., Chen, A., Creel, K., Davis, J. Q., Demszky, D., Donahue, C., Doumbouya, M., Durmus, E., Ermon, S., Etchemendy, J., Ethayarajh, K., Fei-Fei, L., Finn, C., Gale, T., Gillespie, L., Goel, K., Goodman, N., Grossman, S., Guha, N., Hashimoto, T., Henderson, P., Hewitt, J., Ho, D. E., Hong, J., Hsu, K., Huang, J., Icard, T., Jain, S., Jurafsky, D., Kalluri, P., Karamcheti, S., Keeling, G., Khani, F., Khattab, O., Koh, P. W., Krass, M., Krishna, R., Kuditipudi, R., Kumar, A., Ladhak, F., Lee, M., Lee, T., Leskovec, J., Levent, I., Li, X. L., Li, X., Ma, T., Malik, A., Manning, C. D., Mirchandani, S., Mitchell, E., Munyikwa, Z., Nair, S., Narayan, A., Narayanan, D., Newman, B., Nie, A., Niebles, J. C., Nilforoshan, H., Nyarko, J., Ogut, G., Orr, L., Papadimitriou, I., Park, J. S., Piech, C., Portelance, E., Potts, C., Raghunathan, A., Reich, R., Ren, H., Rong, F., Roohani, Y., Ruiz, C., Ryan, J., Ré, C., Sadigh, D., Sagawa, S., Santhanam, K., Shih, A., Srinivasan, K., Tamkin, A., Taori, R., Thomas, A. W., Tramèr, F., Wang, R. E., Wang, W., Wu, B., Wu, J., Wu, Y., Xie, S. M., Yasunaga, M., You, J., Zaharia, M., Zhang, M., Zhang, T., Zhang, X., Zhang, Y., Zheng, L., Zhou, K., and Liang, P. On the opportunities and risks of foundation models, 2021.

Brandes, N., Ofer, D., Peleg, Y., Rappoport, N., and Linial, M. Proteinbert: a universal deep-learning model of protein sequence and function. *Bioinformatics*, 38(8):2102–2110, February 2022. ISSN 1367-4811. doi: 10.1093/bioinformatics/btac020. URL http://dx.doi.org/10.1093/bioinformatics/btac020.

Chakravarty, D., Schafer, J. W., Chen, E. A., Thole, J. F., Ronish, L. A., Lee, M., and Porter, L. L. Alphafold predictions of fold-switched conformations are driven by structure memorization. *Nature Communications*, 15(1), August 2024. ISSN 2041-1723. doi: 10.1038/s41467-024-51801-z. URL http://dx.doi.org/10.1038/s41467-024-51801-z.

Chen, X., Castro, S. A., Liu, Q., Hu, W., and Zhang, S. Practical considerations on performing and analyzing clip-seq experiments to identify transcriptomic-wide rna-protein interactions. *Methods*, 155:49–57, February 2019. ISSN 1046-2023. doi: 10.1016/j.ymeth.2018.12.002. URL http://dx.doi.org/10.1016/j.ymeth.2018.12.002.

Dalla-Torre, H., Gonzalez, L., Mendoza-Revilla, J., Carranza, N. L., Grzywaczewski, A. H., Oteri, F., Dallago, C., Trop, E., de Almeida, B. P., Sirelkhatim, H., Richard, G., Skwark, M., Beguir, K., Lopez, M., and Pierrot, T. The nucleotide transformer: Building and evaluating robust foundation models for human genomics. January 2023. doi: 10.1101/2023.01.11.523679. URL http://dx.doi.org/10.1101/2023.01.11.523679.

Elnaggar, A., Heinzinger, M., Dallago, C., Rehawi, G., Yu, W., Jones, L., Gibbs, T., Feher, T., Angerer, C., Steinegger, M., Bhowmik, D., and Rost, B. Prottrans: Towards cracking the language of lifes code through self-supervised deep learning and high performance computing. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, pp. 1–1, 2021. doi: 10.1109/TPAMI.2021.3095381.

Ferruz, N., Schmidt, S., and Höcker, B. A deep unsupervised language model for protein design. *BioRxiv*, pp. 2022–03, 2022.

Hayes, T., Rao, R., Akin, H., Sofroniew, N. J., Oktay, D., Lin, Z., Verkuil, R., Tran, V. Q., Deaton, J., Wiggert, M., et al. Simulating 500 million years of evolution with a language model. *bioRxiv*, pp. 2024–07, 2024.

Huang, T., Song, Z., Ying, R., and Jin, W. Protein-nucleic acid complex modeling with frame averaging transformer, 2024.

Hwang, Y., Cornman, A. L., Kellogg, E. H., Ovchinnikov, S., and Girguis, P. R. Genomic language model predicts protein co-regulation and function. *Nature Communications*, 15(1), April 2024. ISSN 2041-1723. doi: 10.1038/s41467-024-46947-9. URL http://dx.doi.org/10.1038/s41467-024-46947-9.

Ji, Y., Zhou, Z., Liu, H., and Davuluri, R. V. Dnabert: pre-trained bidirectional encoder representations from transformers model for dna-language in genome. *Bioinformatics*, 37(15):2112–2120, February 2021. ISSN 1367-4811. doi: 10.1093/bioinformatics/btab083. URL http://dx.doi.org/10.1093/bioinformatics/btab083.

Jumper, J., Evans, R., Pritzel, A., Green, T., Figurnov, M., Ronneberger, O., Tunyasuvunakool, K., Bates, R., Žídek, A., Potapenko, A., Bridgland, A., Meyer, C., Kohl, S. A. A., Ballard, A. J., Cowie, A., Romera-Paredes, B., Nikolov, S., Jain, R., Adler, J., Back, T., Petersen, S., Reiman, D., Clancy, E., Zielinski, M., Steinegger, M., Pacholska, M., Berghammer, T., Bodenstein, S., Silver, D., Vinyals, O., Senior, A. W., Kavukcuoglu, K., Kohli, P., and Hassabis, D. Highly accurate protein structure prediction with alphafold. *Nature*, 596(7873):583–589, July 2021. ISSN 1476-4687. doi: 10.1038/s41586-021-03819-2. URL http://dx.doi.org/10.1038/s41586-021-03819-2.

Kingma, D. P. and Ba, J. Adam: A method for stochastic optimization, 2014.

Lawson, C. L., Berman, H. M., Chen, L., Vallat, B., and Zirbel, C. L. The nucleic acid knowledgebase: a new portal for 3d structural information about nucleic acids. *Nucleic Acids Research*, 52 (D1):D245–D254, November 2023. ISSN 1362-4962. doi: 10.1093/nar/gkad957. URL http://dx.doi.org/10.1093/nar/gkad957.

Li, F.-Z., Amini, A. P., Yang, K. K., and Lu, A. X. Pretrained protein language model transfer learning: is the final layer representation what we want. *Proc. Mach. Learn. for Struct. Biol. Work. NeurIPS*, 2022, 2022.

Lin, Z., Akin, H., Rao, R., Hie, B., Zhu, Z., Lu, W., Smetanin, N., Verkuil, R., Kabeli, O., Shmueli, Y., dos Santos Costa, A., Fazel-Zarandi, M., Sercu, T., Candido, S., and Rives, A. Evolutionary-scale prediction of atomic-level protein structure with a language model. *Science*, 379(6637):1123–1130, March 2023. ISSN 1095-9203. doi: 10.1126/science.ade2574. URL http://dx.doi.org/10.1126/science.ade2574.

Liu, S., Du, W., Li, Y., Li, Z., Zheng, Z., Duan, C., Ma, Z., Yaghi, O., Anandkumar, A., Borgs, C., Chayes, J., Guo, H., and Tang, J. Symmetry-informed geometric representation for molecules and proteins and and crystalline materials. *NeurIPS*, 2023a. URL https://openreview.net/forum?id=ygXSNrIU1p&referrer.

Liu, S., Li, Y., Li, Z., Gitter, A., Zhu, Y., Lu, J., Xu, Z., Nie, W., Ramanathan, A., Xiao, C., et al. A text-guided protein design framework. *arXiv preprint arXiv:2302.04611*, 2023b.

Liu, S., Wang, J., Yang, Y., Wang, C., Liu, L., Guo, H., and Xiao, C. Conversational drug editing using retrieval and domain feedback. In *The Twelfth International Conference on Learning Representations*, 2024.

Madani, A., Krause, B., Greene, E. R., Subramanian, S., Mohr, B. P., Holton, J. M., Olmos, J. L., Xiong, C., Sun, Z. Z., Socher, R., Fraser, J. S., and Naik, N. Large language models generate functional protein sequences across diverse families. *Nature Biotechnology*, 41(8):1099–1106, January 2023. ISSN 1546-1696. doi: 10.1038/s41587-022-01618-2. URL http://dx.doi.org/10.1038/s41587-022-01618-2.

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

Nguyen, E., Poli, M., Durrant, M. G., Thomas, A. W., Kang, B., Sullivan, J., Ng, M. Y., Lewis, A., Patel, A., Lou, A., et al. Sequence modeling and design from molecular to genome scale with evo. *BioRxiv*, pp. 2024–02, 2024.

Ollis, D. L. and White, S. W. Structural basis of protein-nucleic acid interactions. *Chemical Reviews*, 87(5):981–995, 1987.

Outeiral, C. and Deane, C. M. Codon language embeddings provide strong signals for use in protein engineering. *Nature Machine Intelligence*, 6(2):170–179, February 2024. ISSN 2522-5839. doi: 10.1038/s42256-024-00791-0. URL http://dx.doi.org/10.1038/s42256-024-00791-0.

Ramanathan, M., Porter, D. F., and Khavari, P. A. Methods to study rna–protein interactions. *Nature Methods*, 16(3):225–234, February 2019. ISSN 1548-7105. doi: 10.1038/s41592-019-0330-1. URL http://dx.doi.org/10.1038/s41592-019-0330-1.

Sagendorf, J. M., Mitra, R., Huang, J., Chen, X. S., and Rohs, R. Pnabind: Structure-based prediction of protein-nucleic acid binding using graph neural networks. *bioRxiv*, pp. 2024–02, 2024.

Sanabria, M., Hirsch, J., Joubert, P. M., and Poetsch, A. R. Dna language model grover learns sequence context in the human genome. *Nature Machine Intelligence*, 6(8):911–923, July 2024. ISSN 2522-5839. doi: 10.1038/s42256-024-00872-0. URL http://dx.doi.org/10.1038/s42256-024-00872-0.

Schmirler, R., Heinzinger, M., and Rost, B. Fine-tuning protein language models boosts predictions across diverse tasks. *Nature Communications*, 15(1), August 2024. ISSN 2041-1723. doi: 10.1038/s41467-024-51844-2. URL http://dx.doi.org/10.1038/s41467-024-51844-2.

Shadab, S., Khan, M. T. A., Neezi, N. A., Adilina, S., and Shatabda, S. Deepdbp: deep neural networks for identification of dna-binding proteins. *Informatics in Medicine Unlocked*, 19:100318, 2020.

Steinegger, M. and Söding, J. Mmseqs2 enables sensitive protein sequence searching for the analysis of massive data sets. *Nature Biotechnology*, 35(11):1026–1028, October 2017. ISSN 1546-1696. doi: 10.1038/nbt.3988. URL http://dx.doi.org/10.1038/nbt.3988.

Stormo, G. D. and Zhao, Y. Determining the specificity of protein–dna interactions. *Nature Reviews Genetics*, 11(11):751–760, September 2010. ISSN 1471-0064. doi: 10.1038/nrg2845. URL http://dx.doi.org/10.1038/nrg2845.

Szpotkowski, K., Wójcik, K., and Kurzyńska-Kokorniak, A. Structural studies of protein–nucleic acid complexes: A brief overview of the selected techniques. *Computational and Structural Biotechnology Journal*, 21:2858–2872, 2023. ISSN 2001-0370. doi: 10.1016/j.csbj.2023.04.028. URL http://dx.doi.org/10.1016/j.csbj.2023.04.028.

Tomaz da Silva, P., Karollus, A., Hingerl, J., Galindez, G., Wagner, N., Hernandez-Alias, X., Incarnato, D., and Gagneur, J. Nucleotide dependency analysis of dna language models reveals genomic functional elements. July 2024. doi: 10.1101/2024.07.27.605418. URL http://dx.doi.org/10.1101/2024.07.27.605418.

Valeriani, L., Doimo, D., Cuturello, F., Laio, A., Ansuini, A., and Cazzaniga, A. The geometry of hidden representations of large transformer models. *Advances in Neural Information Processing Systems*, 36, 2024.

Yuan, Q., Chen, S., Rao, J., Zheng, S., Zhao, H., and Yang, Y. Alphafold2-aware protein–dna binding site prediction using graph transformer. *Briefings in Bioinformatics*, 23(2):bbab564, 2022.

Zvyagin, M., Brace, A., Hippe, K., Deng, Y., Zhang, B., Bohorquez, C. O., Clyde, A., Kale, B., Perez-Rivera, D., Ma, H., et al. Genslms: Genome-scale language models reveal sars-cov-2 evolutionary dynamics. *The International Journal of High Performance Computing Applications*, 37(6):683–705, 2023.

A Acknowledgments

We thank Dr. Riddhiman Dhar for valuable feedback and compute resources. We would like to thank the Caltech SURF program for contributing to the funding of this project. This research was supported by the National Institutes of Health Award Number P01AI165077, Coalition for Epidemic Preparedness Innovations (CEPI), and the U.S. Department of Energy, Office of Science and Office of Advanced Scientific Computing Research under Contract No. DE-AC02-05CH11231 through the Biopreparedness Research Virtual Environment program: Elucidate Multiscale Ecosystem Complexities for Robust Epidemiological Modeling.

B Biological Preliminaries

Protein-Nucleic acid (NA) interactions play a key role in controlling gene regulation and, hence, life itself. Structural information on these interactions provides vital information on the nature and consequences of these interactions. Proteins are bio-polymeric compounds that are composed of an array of 20 possible amino acids. These amino acids, in a specific order, form the sequence of the protein, but proteins, in reality, exist as a 3D-shaped folded version of the sequence called the tertiary structure. NAs (DNA and RNA) can be annotated similarly but are composed of combinations of four possible nucleotides ([A, T, C, G] for DNA and [A, U, C, G] for RNA).

Protein-NA interactions are mainly of two types (Ollis & White (1987)): non-specific interactions where a positively charged amino acid is attracted to a negatively charged phosphate moiety on a nucleotide or specific interactions where specific patterns of nucleotides are recognized by sub-structural elements present in the protein (for e.g., Transcription Factors). These interactions are usually specific to a given DNA sequence, but proteins, on the other hand, display both specific and non-specific interactions (one protein may bind to more than one specific sequence of DNA).

Experimental Methods for mapping 3D Complex Structures: For a biological sample of a protein-NA complex, there are currently multiple experimental methods to obtain the tertiary structure (Szpotkowski et al. (2023)), including NMR-Spectroscopy, X-ray crystallography, and Cryo-EM microscopy. 3D structures of protein-NA complexes are deposited and are freely available on the Protein Data Bank (Berman et al. (2003)) database. However, isolating protein-NA complexes and solving structures experimentally is a difficult, expensive, and time-consuming task. Hence, there are very less Protein-NA structures available today, as compared to just Proteins, or just NAs. Similarly, a wide range of methods have been developed to isolate Protein and NA sequences that interact (Stormo & Zhao (2010), Ramanathan et al. (2019)), but these methods are again expensive, and often a bit noisy for highly accurate inference (Chen et al. (2019)).

Experimental Methods for quantifying Protein-NA interactions: Conventional methods of Protein-DNA binding investigation are performed in laboratory settings and include Electrophoretic Mobility Shift Assay, Chromatin Immuno-precipitation, SELEX (Systematic Evolution of Ligands by Exponential Enrichment), and more. As with experimentally solving protein-NA 3D structures, these methods are time-intensive and expensive, motivating a scalable computational approach.

C Related Work

C.1 Protein and Genomics Language Models

Protein Language Models: Protein language models have been developed and used for various downstream applications, including structure design, protein function prediction, and post-translational regulation identification, and can be trained by many sequence-modeling methods. Alley et al. (2019) use RNNs for unsupervised representation learning of protein sequences. ESM2 (Lin et al. (2023)) and ProteinBERT (Brandes et al. (2022)) are examples of transformer-based models trained on a masked language modeling (MLM) loss by filling in missing amino acids in protein sequences. The ESM2 model can learn dependencies among the amino acids and other biological information. With GPT-style architectures, Ferruz et al. (2022) and Madani et al. (2023) use language modeling for the generative design of protein sequences.

Rich textual data on protein functional description also exists, which can be incorporated to support more diverse tasks. ProteinDT Liu et al. (2023b) is built on ProtTrans Elnaggar et al. (2021), first

utilizing such a free-text format for protein design. ChatDrug Liu et al. (2024) utilizes the ChatGPT as the core agent for protein optimization. The most recent work is ESM3 Hayes et al. (2024). It incorporates all the modalities of data (sequence, structure, and textual description) for protein design with a wet lab verification.

Genomics Language Models: Genomics language models operate on DNA sequences and utilize MLM for a next token prediction task. However, there is a lot of freedom in choosing a token size as a K-mer (K nucleotides are assigned to a single token). GPN (Benegas et al. (2023)) is a model trained through an MLM task of predicting masked nucleotides (k=1) given a genomic context. GPN achieves SOTA performance on prediction of genome-wide variant effects and is a critical development in the prediction of genome-wide variant effects given a DNA sequence. Compared to GPN, which focuses on DNA sequences, Evo (Nguyen et al. (2024)) is a multi-modal model trained to generate DNA sequences, aiming to learn the relationships between and functions of DNA, RNA and proteins encoded in a genome. Ji et al. (2021), Dalla-Torre et al. (2023), Tomaz da Silva et al. (2024), Hwang et al. (2024), and Sanabria et al. (2024) are some recent works utilizing DNA language models for various downstream applications using genome annotation, understanding protein co-regulation and discovering genomic functional elements. Proteins are encoded by codons, which are specific sequences of 3 nucleotides that code for (produce) a specific amino acid. GenSLM (Zvyagin et al. (2023)) exploits this fact and uses a codon-level tokenization scheme (k=3) to quickly and accurately identify variants of SARS-CoV-2. Outeiral & Deane (2024) provides another example of a codon-based language model, which outperforms amino-acid-based models on downstream tasks for the protein modality.

C.2 Protein-NA Contact Prediction

To study Protein-NA interactions, a large focus has been on identifying DNA-binding proteins (DBP) given a protein sequence or structure. DeepDBP-ANN and DeepDBP-CNN (Shadab et al. (2020)) are two deep learning approaches using a traditional neural network and a convolutional neural network to identify DPBs given a protein sequence. Protein structures often offer a more comprehensive understanding of the protein's characteristics and may provide better results. Graphsite (Yuan et al. (2022)) implements a graph neural network combined with AlphaFold2 predicted structures to perform protein-DNA binding site prediction. PNABind (Sagendorf et al. (2024)) similarly incorporates a graph neural network that encodes spatial representations of physicochemical and geometric properties of the protein's surface to predict its binding function. DeepPBS (Mitra et al. (2024)) is a geometric deep learning model that captures physicochemical and geometric contexts of protein–DNA interactions to output a position weight matrix (PWM) that predicts binding specificity. DeepPBS is applied to both experimental and predicted structures (such as from AlphaFold2, etc.) and serves as a fundamentally new approach for protein-binding specificity. Alphafold3 (Abramson et al. (2024)) and RF2NA (Baek et al. (2023)) map Protein and NA sequences to 3D structures from which contacts can be inferred. FAFormer (Huang et al. (2024)) exploits the geometry of 3D complex structures by coupling Frame Averaging with a SE(3) equivariant transformer model to directly predict contact maps.

D Dataset Preparation

Table 3: Summary Dataset Statistics

Total Structures	Train set	Eval Set	ratio(binding/non-binding)
4021	3272	749	0.0035

This section explains our data preparation scheme for the contact map prediction task. We started with all available structures of Protein-DNA complexes until 12-01-2024, downloaded from the Nucleic Acid Knowledgebase (Lawson et al. (2023)). Similar to RF2NA, we do not include structures solved using NMR spectroscopy and also filter off structures with a resolution higher in magnitude than 4.0 Å. We also filter off DNA sequences that are less than 5 or more than 100 nucleotides in length and protein sequences that are more than 1000 amino acids in length, finally ending up with the final processed dataset of 4021 protein-DNA complexes. Figure 3 shows the distribution of lengths of Proteins and DNA in our data. We apply sequence-similarity-based clustering for our

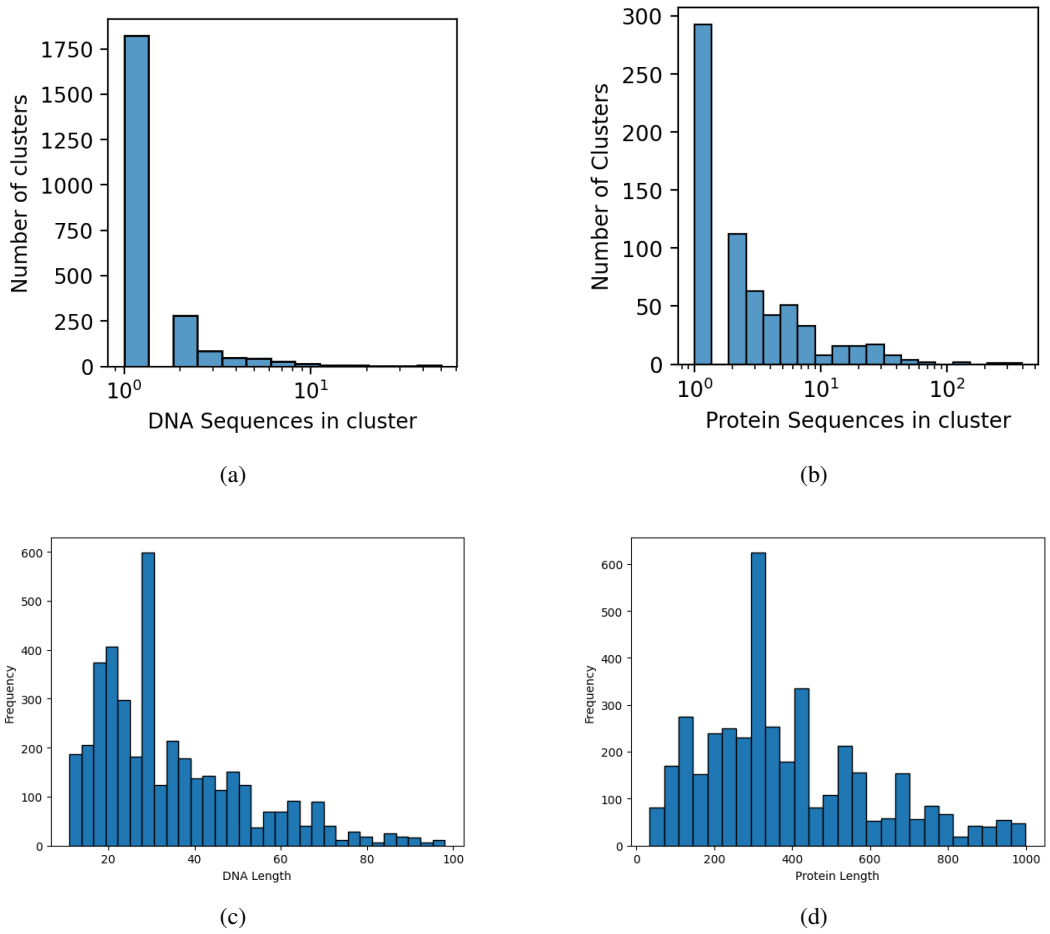


Figure 3: (a): DNA sequence similarity-based clustering, (b): Protein sequence similarity-based clustering. We see the DNA-based clustering to be more diverse than Protein-based clustering. (c): Histogram of DNA sequence lengths in our final dataset (Mean sequence length is 34.49 nucleotides), (d): Histogram of Protein sequence lengths in our final dataset (Mean sequence length is 389 amino acids)

DNA sequences using Steinegger & Söding (2017). We set a threshold of 30% similarity, assigning to the same cluster all possible sequence pairs in our dataset that share sequence similarity more than 30%. Our dataset statistics are summarized [5] in Table 3. We choose to apply clustering based on DNA sequences as DNA generally binds to proteins with high specificity. In contrast, proteins can display both specificity and non-specificity in their binding mode (Ollis & White (1987)). We show the frequency statistics of our data's Protein-sequence and DNA-sequence clustering in Figure 3 (where 30% similarity is set as the threshold for both). To create our train-test splits, we randomly sample clusters from our dataset till we get to 80% of the size of our dataset. We assign the remaining clusters as our validation set. This setup ensures that our model does not see any similar sequences (preventing data leakage). Also, it allows us to gauge model generalization because the evaluation set clusters are completely independent of the training set. For structures with multiple chains, we simply concatenate all chain sequences together, making our approach agnostic to number of chains in either the Protein or the corresponding DNA structure.

[5]Last column of the table indicates the sum of binding contacts over all structures divided by the sum of non-binding contacts

Table 4: Relevant hyperparameters

Hyperparameter	Value
d (Cross-Attention)	32
d_k (from ESM2)	320
d_q (from GPN)	512
Learning rate (Cross-Attention)	1×10^{-5}
Learning rate (ESM and GPN)	1×10^{-6}
Loss weight	20
Batch size	12

E Random Basline

Let us assume we have $\mathcal{D}$ complex structures, each with associated contact map information. For the i-th complex, the contact map is denoted as $C_i \in \mathcal{B}^{L_{p(i)} \times L_{n(i)}}$, where $L_{p(i)}$ and $L_{n(i)}$ are the lengths of the protein and nucleic acid sequences, respectively, and:

$$C_i \in \{0,1\}^{L_{p(i)} \times L_{n(i)}}, \forall i = 1, 2, \ldots, \mathcal{D}$$

Given an unknown sequence dataset $\mathcal{D}'$ with protein and nucleic acid sequence pairs $(P, \mathcal{N})$, where $L_{p(i)}$ and $L_{n(i)}$ $\forall i \in \mathcal{D}'$ are the lengths of the protein and nucleic acid sequences, respectively, and for which the contact map is unknown, our goal is to generate the contact maps $\mathcal{C}_i$ $\forall i \in \mathcal{D}'$ using the prior information from the $\mathcal{D}$ known contact maps without any parametrization or learning.

Since the classification problem of Protein-NA binding has a high class imbalance (sparse contacts), sampling contacts as 0 and 1 with equal probability would make a poor baseline. To this end, sample contacts from a Bernoulli distribution, considering class imbalance.

We define n as the total number of contacts across all $\mathcal{D}$ contact maps, expressed as follows:

$$n = \sum_{i=1}^{\mathcal{D}} \sum_{a=1}^{L_{p(i)}} \sum_{b=1}^{L_{n(i)}} \mathbb{1}\{\mathcal{C}_i(a,b) = 1\}$$

Where $\mathbb{1}\{\cdot\}$ is the indicator function that is 1 for contacts and 0 otherwise.

Next, we calculate the total number of elements across all contact maps:

$$N_{\text{total}} = \sum_{i=1}^{\mathcal{D}} L_{p(i)} \times L_{n(i)}$$

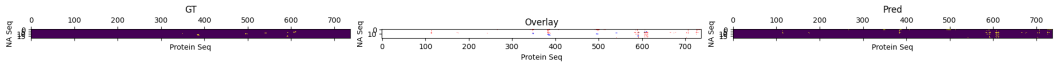
The probability of success, denoted as p, is then the ratio of the total number of contacts to the total number of elements:

$$p = \frac{n}{N_{\text{total}}}$$

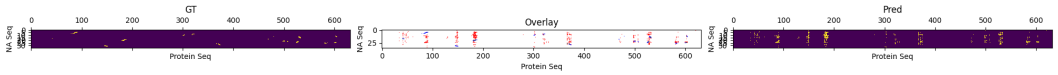
This probability p is used to define a Bernoulli distribution $\mathcal{B}(p)$ that models our prior as a Bernoulli process. To generate the target contact map $\mathcal{C}_i$ $\forall i \in \mathcal{D}'$, we sample each element $\mathcal{C}_i(a,b)$ from $\mathcal{B}$:

$$\mathcal{C}_i(a,b) \sim \mathcal{B}(p), \quad \text{for } a = 1, 2, \ldots, L_{p(i)} \text{ and } b = 1, 2, \ldots, L_{n(i)}$$

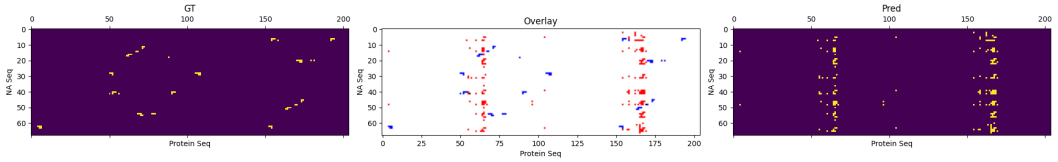
This ensures that the values in $\mathcal{C}_i$ are independent and identically distributed (i.i.d.) according to our prior, which is a fundamental property of the Bernoulli process.



(a) PDB: 8ICH



(b) PDB: 4TUR



(c) PDB: 8E4H

Figure 4: Poor Predictions: Contact maps (Ground-Truth (GT), Overlay of GT and Prediction (Pred), and Pred) shown for three complexes. Purple indicates no contact, and yellow indicates contact. In the overlays, black indicates a match, red indicates a Pred mismatch and blue indicates a GT mismatch

F Additional Experimental Details

Here, we provide our experimental details and related hyperparameters. We train Seq2Contact for 800 epochs and use the Adam optimizer (Kingma & Ba (2014)). We use a split of [80:20] to split our dataset into train and validation sets based on sequence-similarity clustering as described in Appendix D. We use weighted cross-entropy as our loss function and describe all relevant hyperparameters in Table 4

G Challenges and Limitations

In Figure 4 we show some results of Seq2Contact that fail to accurately predict contact maps. We observe that the model can learn binding with relatively high specificity along the protein axis. This observation matches closely with Biological evidence that protein binding is less specific and more general than DNA binding, and hence our model is able to learn binding from a protein sequence perspective quite well as compared to DNA.

Why did we use the GPN model?: There are many DNA language models or FMs available, with EVO (Nguyen et al. (2024)) and GenSLM (Zvyagin et al. (2023)) being some highly expressive models in our knowledge. However, GenSLM produces codon-level representations, meaning we cannot learn nucleotide-specific embeddings from GenSLM. EVO produces embeddings specific to individual nucleotides but has an embedding dimension of 4096 and consists of 7 Billion parameters, making it infeasible to fine-tune given the scarcity of available data. Hence we use GPN, which has an embedding dimension of 512 and is feasible to partially finetune in parallel with our protein language model.

We would like to note that we did not use any hyperparameter tuning at all, nor did we use the most expressive version of the ESM2 family (with 15 Billion parameters and 48 Transformer layers). Our reasoning is two-fold; firstly, in a low-data problem setting like ours, more parameters do not always lead to the best results and may cause overfitting. Second, just making the trainable part of our pipeline deeper increases the risk of our model memorizing information, which is not easy to detect and mitigate. (Chakravarty et al. (2024) recently showed that some AlphaFold2's better predictions are a result of memorization of training data).

A big challenge in learning contacts from 3D structures is that the present database of 3D structures does not simply capture all of the variances of binding interactions in nature, and we cannot accurately measure how good is the present database with respect to the heterogeneity of interactions. This means there are limitations on how generalizable any model trained on this data might be. For RNA-

Protein data, initial results indicate that the sparsity of contact maps is making learning meaningful contact-aware representations quite difficult, as RNAs are usually shorter than DNA, and we only have about 1000 structures of Protein-RNA complexes available post-filtering. One potential method to handle this is to inject more physics priors into the modeling, *i.e.*, the SE(3)-equivariant geometric models over proteins (Liu et al., 2023a), and we would like to leave this for our future exploration.

H Metrics

In this section, we describe the key evaluation metrics used to assess the performance of our binding prediction task, which is a highly class-imbalanced problem.

Precision and Recall Area Under the Curve (PR-AUC)

Precision (P) and Recall (R) are fundamental metrics used to evaluate the performance of a binary classifier:

$$\text{Precision} = \frac{TP}{TP + FP} \tag{1}$$

$$\text{Recall} = \frac{TP}{TP + FN} \tag{2}$$

where:

- TP denotes the number of true positives,
- FP denotes the number of false positives,
- FN denotes the number of false negatives.

The PR curve is plotted with Precision on the Y-axis and Recall on the X-axis. We define PR-AUC as the area under this curve.

F1 Score

The F1 score is the harmonic mean of Precision and Recall, providing a balance between these two metrics. It is defined as:

$$F1 = 2 \times \frac{\text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}} \tag{3}$$

Matthews Correlation Coefficient (MCC)

The Matthews Correlation Coefficient (MCC) is a measure of the quality of binary classifications, taking into account the four quadrants of the confusion matrix: true positives (TP), true negatives (TN), false positives (FP), and false negatives (FN). MCC is defined as:

$$\text{MCC} = \frac{(TP \times TN) - (FP \times FN)}{\sqrt{(TP + FP)(TP + FN)(TN + FP)(TN + FN)}} \tag{4}$$

where:

- TP denotes the number of true positives,
- TN denotes the number of true negatives,
- FP denotes the number of false positives,
- FN denotes the number of false negatives.

The MCC ranges from -1 to $+1$, where $+1$ indicates a perfect prediction, 0 indicates a random prediction, and -1 indicates a completely incorrect prediction.