Semester Project: Protein Contact Prediction

Maria Yuffa
EPFL
maria.yuffa@epfl.ch

Luca Biggio
EPFL
luca.biggio@epfl.ch

Abstract

Protein contact maps offer valuable insights into the 3D structure and function of proteins. This project explores various methods for inferring amino acid couplings. Using synthetic data generated from a minimal model, we test traditional methods and compare the results with transformer approaches trained via Masked Language Modelling (MLM). Finally, we propose an encoder-decoder transformer that shows promising results in inferring structural contacts with less computational cost.

1 Introduction

Protein contact maps are graphical representations that illustrate which pairs of amino acids within a protein are in spatial proximity. Inferring structural contacts on Multiple Sequence Alignments (MSAs) has been a long-standing challenge. Traditionally, global inference methods such as the Potts model have been the most accurate in inferring the contact map. Nonetheless, such methods are not overall robust to biological noise, such as phylogeny Dietler et al. [2023] and their applicability is, therefore, limited on the real proteins.

Recent advancements in language modeling and transformers have opened new avenues in research aimed at protein contact inference. These machine learning methods have achieved significant breakthroughs in supervised contact inference, as evidenced by Jumper et al. [2021]. The core functionality of transformers, particularly their attention mechanisms, has been shown to mirror the computational aspects of the Potts model, offering potential enhancements in accuracy and efficiency Rende et al. [2024]. Recently, work by Caredda and Pagnani [2024] confirmed the equivalence of the transformer self-attention mechanism to PlmDCA. The advantage of transformers lies in leveraging fewer parameters when inferring contact matrices from real protein data. This suggests that transformers could be a computationally tractable solution to the problem, while also accommodating the specifics of proteins through architectural choices, such as feedforward layers.

The success of the Transformer architecture is partially attributed to the use of masked language modelling (MLM) for training. By masking tokens the model is able to infer the relations of each token to the others and extract the contact map. Specifically, MLM learns the conditional probability distribution of the surrounding spins in the sequence Rende et al. [2024].

In this project, we propose a sensible enhancement to the Vanilla transformer architecture for inferring the protein contacts from the attention map. To achieve that, we first generate a simple dataset of binary spins by sampling from Boltzmann distribution. This is done via the Metropolis-Hastings (MH) algorithm, utilizing the Cluster Ising flipping method. We revisit traditional approaches and implement a straightforward transformer model as described by Rende et al. [2024]. The performance of these methods is later compared to the proposed encoder-decoder transformer. We motivate the design of the proposed transformer by the hypothesis that information from the encoder will provide the decoder with an inductive prior, enabling faster learning of the data. Finally, this study addresses the short-comings of the proposed encoder-decoder transformer and outlines future directions for testing the model on more realistic data, involving 20 different amino acids and protein sequences of approximately 100-200 residues.

Preprint. Under review.

2 Background & Methodology

2.1 Data generation

The data in the project was simulated using Boltzmann distribution and presented as sequences of spins (-1 and +1). This was achieved via Metropolis-Hastings Algorithm with two types of mechanism discussed below.

Random spin flipping is based on energy criteria and stochasticity of the system. The former accounts for the difference of Hamiltonians between flipped sequence and the sequence before to ensure a more energetically favourable sequence is accepted. Such Hamiltonian is calculated according to the following equation:

$$H(\sigma) = -\sum_{i,j=1} J_{i,j}\sigma_i^L\sigma_j,$$

where L is the length of the sequence of spins. The stochasticity arises due to temperature: if the temperature is high, the spins are more likely to get flipped regardless of the whether the first criteria is satisfied. This is accounted by the following condition:

$$X < \exp^{-\frac{\triangle H}{T}},$$

where $X \sim Uniform(0, 1)$ and is sampled every iteration of the algorithm, or every time a new spin flip is considered.

Cluster Ising model is an extension of the traditional Ising model, used to study the collective behavior of systems with interacting spins more accurately by incorporating cluster-based interactions. Instead of flipping individual spins, the model considers flipping spins of the same orientation within the cluster. For protein structures, this involves representing the protein as a graph. A site in the graph is selected at random, and its neighboring sites (the cluster) are flipped if the following probability condition is satisfied:

$$X < 1 - \exp^{-\frac{2}{T}},$$

where $X \sim Uniform(0, 1)$ and is sampled randomly every time we consider flipping within the cluster.

2.1.1 Traditional approaches and Simple Transformer Data

In Figure 7, we present the average magnetization over 2048 sequences as a function of the number of flips for various temperatures. These plots validate the accuracy of our data sampling process. To compare across all the methods, the same generated data was used with the selected temperature equivalent to transition one of $T = 5.0$, according to Dietler et al. [2023]. We have set the probability of the connection between two spins equal to 0.3. The data therefore, was generated using the same J-coupling matrix, and consisted of 1000 sequences each containing 20 spins, flipped 100 times using Cluster Ising method.

2.1.2 Encoder-decoder Transformer Data

The creation of suitable data for the proposed encoder-decoder transformer presented several challenges and remains a work in progress. We generated five samples of 1000 sequences each, using different symmetric J matrices with a 0.3 probability of contact between sites. Additionally, we included sequences generated using the true contact map, representing the couplings the transformer aims to infer. The decoder input comprised of sequences, which were used for inference with other methods, enabling direct comparison between approaches.

2.2 Evaluation metric

The True-Positive (TP) fraction, indicating the proportion of correctly inferred protein contacts, is used as an evaluation metric for comparing traditional and transformer methods. The true positive fraction is described by the following rate $TPR = \frac{TP}{TP+FN}$, where TP is True positives and FN is False negatives. This is equivalent to calculating the ratio of contacts in the inferred map to the contacts in the true matrix. We utilize the TP fraction as a performance metric because it aligns

with the standard practice in natural sequence data analysis of predicting the top-scoring site pairs as contacts. Furthermore, it is important to evaluate how the quality of predictions varies with the number of predicted contacts.

2.3 Traditional approaches

The accuracies of traditional methods on both simulated and real protein data, with and without phylogenetic adversaries, have been extensively studied in Dietler et al. [2023]. The paper provides a detailed comparison of Mutual Information (MI), Covariance, Mean-field Direct Coupling Analysis (mfDCA), and Pseudo-likelihood Maximization Direct Coupling Analysis (PlmDCA). Below, we give a brief overview of each method.

Mutual Information (Weigt et al. [2009]) employs the single frequencies and two-body frequencies of occurrence of spins either by themselves or in pair. It delivers decent performance over temperature for simple cases that do not consider phylogeny which could be considered as "biological noise" for contact map inference.

Covariance (Jones et al. [1986]) considers paired and individual frequencies. To infer the coupling matrix J, the inverse of the covariance matrix with some regularisation is calculated. Overall, this approach achieves similar accuracy as mutual information. Both MI and Covariance methods perform badly for small temperatures without consideration of Average Product Correction and pose difficulties inferring the coupling matrix directly from the results.

mfDCA (Morcos et al. [2011]) uses a mean-field approximation to manage pairwise interactions among protein residues, rendering the task computationally manageable by assuming interactions between residue pairs are independent. It utilizes a pseudocount-corrected covariance matrix based on amino acid frequencies to compute direct coupling values, which reflect amino acid occurrences and correlations across protein families. Mean-field DCA creates an interaction matrix using inverse covariance or pseudoinverse methods to quantify the strength and direction of these couplings. High values in this matrix indicate direct correlations between residue changes, suggesting physical proximity within the protein's structure.

PlmDCA (Ekeberg et al. [2013]) employs pseudo-likelihood maximization for parameter estimation in scenarios where full likelihood calculations are too demanding. This technique approximates the full likelihood by assuming independence among variable subsets, focusing on maximizing conditional probabilities for each variable relative to its neighbors rather than the entire system. Particularly useful in protein structure prediction, it simplifies the problem by independently calculating the likelihood for each protein residue based on adjacent residues. This method offers significant computational efficiency and scalability, although it may sacrifice some accuracy due to its simplifying assumptions. Despite this, PlmDCA is notable for its state-of-the-art accuracy in contact inference within large datasets.

In this work, we study contact map predictions via Covariance, Mutual Information and mfDCA with intentions to draw comparison with coupling matrix obtained using plmDCA in the future.

2.4 Transformers

Transformers are a powerful type of neural network that have achieved state-of-the-art results in various tasks, including protein structure prediction. Unlike standard neural networks that function with a single input, transformers operate on sets of "tokens" such as words in a sentence or amino acids (spins) in the protein chain. Transformers are capable of capturing complex dependencies in sequential data making them useful for inferring structural contacts in proteins.

The work by Rende et al. [2024] demonstrates that the success of transformers in protein structure prediction is rooted in the self-attention mechanism. Self-attention (SA) analyzes amino acid sequences to predict interactions and proximities between residues by allowing each token in a sequence to weigh the importance of every other token, generating context-aware representations. SA effectively learns the Potts model during training with attention weights representing the contact map. Furthermore, the Factored Attention mechanism, which decouples positional encoding from the amino acid embedding, is shown to be equivalent to the Potts model (Rende et al. [2024]).

Relationships across sequences in a batch, can provide valuable insights for learning complex patterns from large-scale data. This principle underpins the success of MSA transformers Rao et al. [2021], which use both standard and row attention to leverage similarities and correlations across sequences. However, MSA transformers are computationally expensive, prompting the need for simpler, alternative approaches.

To train transformer architectures, Masked Language Modelling (MLM) is used. By randomly masking certain residues and predicting them based on their surrounding context, MLM enables the model to capture both direct and indirect interactions among residues. This method has proven to be especially effective as it trains models to grasp the contextual relationships between amino acids in a sequence. It helps identify which residues are likely to be in close proximity within the protein's three-dimensional structure, which is essential for accurately modeling protein folding and function.

With this in mind, we have designed an encoder-decoder transformer to infer correlations from simulated data blocks sampled using the Metropolis-Hastings algorithm via the Cluster Ising method (data generation and usage is described in 2.1 and 2.1.2, respectively), each with different contact maps. This inferred information serves as an inductive bias, assisting the decoder in predicting one or more masked tokens in the sequence. The inductive bias is described by the prior in the following Bayesian formula:

$$p(s_i|s_{\setminus i}, \text{data}) = \frac{p(s_{\setminus i}|s_i)p(\text{data}|s_i)p(s_i)}{p(s_{\setminus i}, \text{data})}$$

The encoder provides the inductive bias by optimally leveraging the sequence information for the decoder. The decoder then uses this bias to predict one or more masked spins in the sequence. The transformer's structure is presented in Diagram 1. The coupling matrix is further inferred through the self-attention weights of the decoder. This design aims to study the interaction between the encoder and decoder and its potential benefits for computationally demanding tasks.

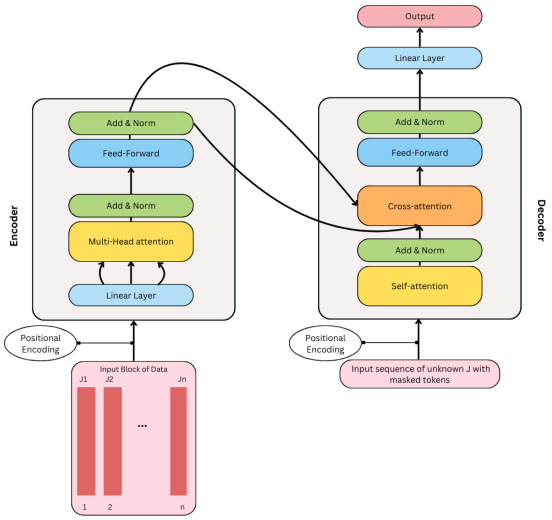

Figure 1: The proposed architecture of the encoder-decoder transformer. The pink blocks correspond to the input data. Each dark pink block within the light pink block for the encoder illustrates 1000 protein sequences of sequence length 20 with different J-coupling matrices with the same probability. The other pink module corresponds to the decoder data and contains the same chains used for traditional methods as well as simple transformer derived from the true contact map. Feed-forward network consists of 3 linear layers with ReLU non-linear activation function. The decoder contains standard vanilla attention mechanism.

To track the results we have used TensorBoard by TensorFlow Team [2015] and Weights & Biases by Weights & Biases Team [2020].

3 Experiments and Results

3.1 Covariance & Mutual Information

Initially we considered local inference methods, which give reliable results for simple 2-spin scenario. Specifically, we implemented covariance matrix of the data and calculated mutual information scores between each sites to obtain the MI matrix. By zeroing the diagonal and taking the 30% of sites with the highest value, since the set probability of the contact between two sites is equal to 0.3, we perform normalisation and obtain the coupling matrices observed in Figure 2 a) and b).

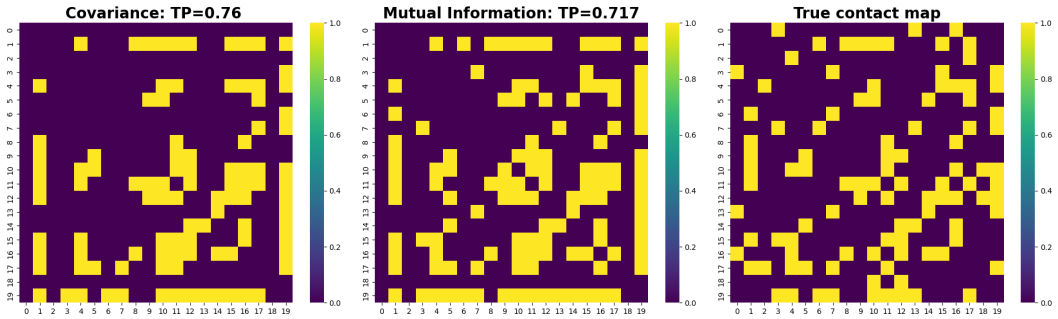

Figure 2: From left to right: a) Normalised covariance matrix for the chains generated by MH algorihtm with TP fraction of 0.76; b) The coupling matrix produced by calculating Mutual Information with TP fraction of 0.72; c) The true matrix.

3.2 MfDCA

The mean-field DCA performed better compared to local methods by more accurately distinguishing direct interactions from indirect correlations. For simple scenarios with sequences of size 20 and binary spins, mfDCA is expected to perform similar to PlmDCA, which is the state-of-the-art method for approximating the solution for Potts model.

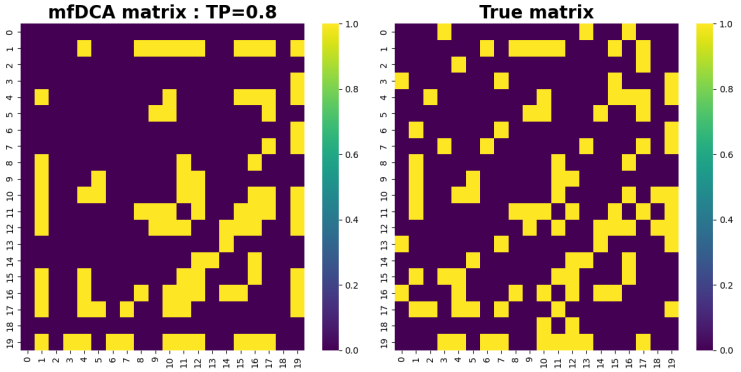

Figure 3: From left to right: a) mean-field DCA coupling matrix with regularization term $\lambda = 1e-5$ and calculated TP fraction of 0.8; b) True contact map.

3.3 Comparison

We have also compared the performance of the three methods across a temperature range. As the temperature increases, the accuracy of all inference techniques, after each reaching the top at different temperatures, drops. Overall, mfDCA outperforms other methods across temperature range.

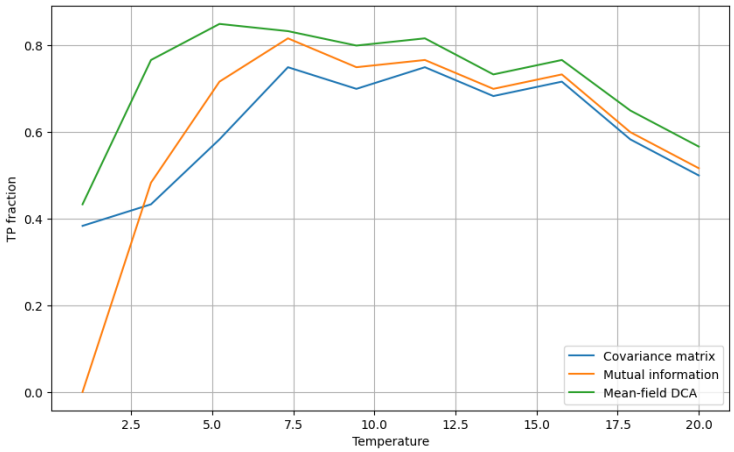

Figure 4: The performance of Inverse Covariance, Mutual Information and Mean-field DCA as the function of temperature.

3.4 Simple Transformer

For implementation of the simple transformer, we reproduced the model outlined in Rende et al. [2024]. We have tried both Vanilla and Factored attention mechanisms (by setting the coefficient of the positional encoding a to 0 for calculation of values). We demonstrate the results in figure 5 we demonstrate the results. An improved performance was observed, when considering Factored attention, confirming the results of the Rende et al. [2024] authors. This remained consistent with data generated using different temperatures, however, the difference in accuracies between the two methods was varied. Altering the architecture, we found significantly better results with a two-layer feedforward network following the attention mechanism. In other configurations, the network struggled to capture meaningful information Rende et al. [2024].

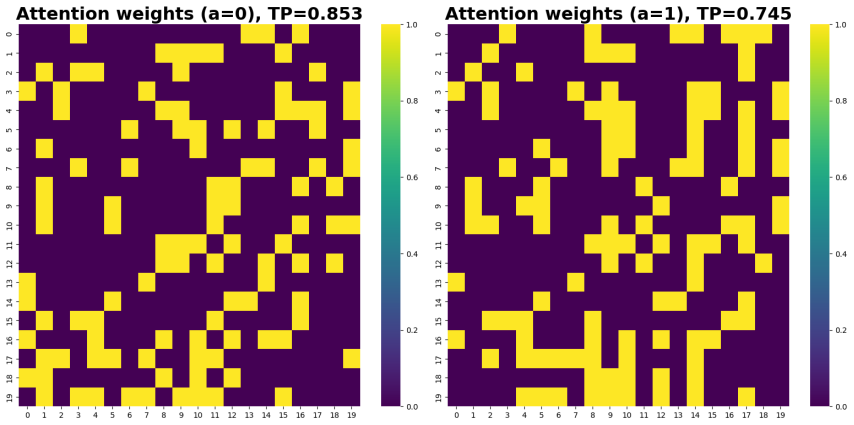

Figure 5: Comparison between the Factored and Vanilla attention for inferring protein contacts in fully trained setting (number of epoch equal to 300).

As expected, we have also observed improvement in performance when switching from the one-hot encoding of sequences to using Embedding module in PyTorch. This suggests that embedding sequences in higher dimensional space allows to better capture intricacies between the protein sequences. It is important to note that any discrepancies between our results and those of Rende (2024) may be due to using 1000 sequences instead of 3000. Additionally, variations in architecture, such as the number of attention layers, fully connected layers, or other hyperparameters, could have impacted the outcomes.

3.5 Proposed Transformer

To observe whether the encoder facilitates the learning of the decoder, we have trained the encoder-decoder transformer on varying number of epochs. We then compared the best performance as well as performance after few iterations of the simple transformer to such of the proposed encoder-decoder one. The proposed Transformer was able to overall slightly outperform the simple one while delivering significantly higher TP fraction score after only 50 epochs. Figure 6 demonstrates the comparison between contact maps derived from attention weights for simple transformer, the proposed encoder-decoder transformer and true contact map after training over 50 epochs. This aligns with our hypothesis that the encoder provides an inductive bias for contact map inference. Alternatively, encoder could be thought as a pre-trained model that is then fine-tuned by the decoder for a related task. Optimizing the encoder and refining the training data are promising directions for future research.

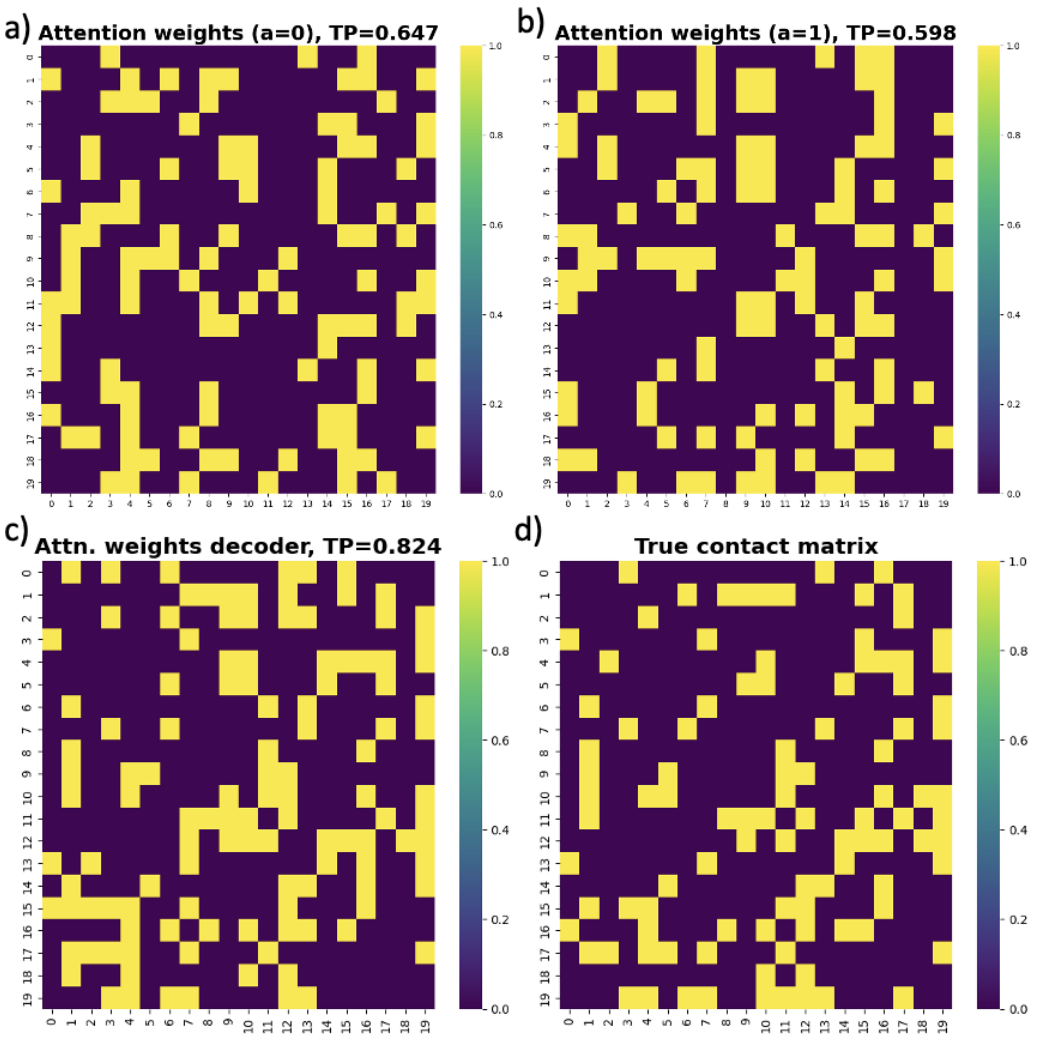

Figure 6: a) Factored attention transformer trained on 50 epochs with TP fraction equal to 0.65; b) Simple Transformer trained on 50 epochs with TP fraction of 0.60; c) Proposed encoder-decoder transformer after training on 50 epochs with TP fraction of 0.82; d) True contact map.

We have also experimented with the data closer mimicking realistic settings. Specifically, we increased the size of the protein sequence to 200 spins. This did not sustain good results, suggesting very high complexity of the task as well as the high randomness of the matrix (the contact map does not have patterns that are observed in real proteins, it is initialised randomly) (Figure 12). Increasing the size of the dataset will likely lead to better results but comes at higher computational cost.

Finally, it is worth mentioning that performance of the transformer methods varied significantly with the temperature used for generating the sequences.

Hyperparameters	**Vanilla Transformer**	**Enc-dec transformer**
Learning rate	1e-3	1e-3
Batch size	32	Enc: 160, Dec:32
Embedding dim	50	50
Hidden dim	50	50
Num. epochs	300	50

Table 1: Hyperparamters for the transformer models for data consisting of 1000 sequence of 20 spins. Note, that number of epochs in the table reflects the number of training steps required to achieve decent performance.

4 Discussion & Extensions

In our experiments, traditional methods using local inference performed poorly compared to mfDCA and Transformers, which had comparable performance. Simple transformer performed overall similar to the proposed Transformer, with the latter sometimes delivering better results (11). The proposed Transformer showed higher TP fraction score when trained for a small number of epochs compared to a simple Transformer, suggesting its potential use in cost-limiting settings.

In future, the following points should be addressed. To have a fairer comparison between all traditional methods, PlmDCA method should be implemented. Secondly, the simulated data should be extended to larger vocabulary of 20 different spins or amino acids. In this case the simulations would be close to the real case scenario. The required number of training steps as well as memory demand and robustness of each methods to both "biological" (e.g. phylogeny) and computational noise should be compared across different transformer architectures. Thirdly, it is crucial to study in more detail how temperature parameter affects the attention weights of the transformer, or in other words, the coupling matrix. Finally, an interesting direction to continue would be to use the proposed transformer architecture for efficient unsupervised protein homologue classification.

Real protein data exhibits distinct interaction patterns that are essential for accurate protein modeling. In contrast, synthetic datasets often employ randomly generated interaction matrices (J matrices), which lack the specificity and complexity of biological interactions. To effectively train a model, it is essential to employ sophisticated data generation techniques that can closely mimic the intricate patterns found in real protein data, thereby ensuring the model can generalize its learnings to new, unseen protein sequences.

A logical next step involves initially training the encoder and decoder together for several iterations, followed by focusing solely on optimizing the decoder. Ensuring the encoder is adequately trained is crucial, as an undertrained encoder can hinder effective learning of the contact map and introduce noise to the decoder. This approach could enhance the decoder's initial performance by providing an inductive bias from the encoder while progressively aligning the decoder's weights with the coupling matrix during training. Additionally, using factored attention instead of vanilla attention for the decoder could be considered. This adjustment is likely to improve results but should be tested on real protein data first.

Another modification to the proposed architectures could potentially improve the results. This alteration might involve generating multiple weighted outputs from the encoder, each corresponding to a typical or averaged sequence from the input and the assigned weight indicating the utility of such sequence. Further performing the dot product with the embedded decoder sequence and multiplying by the weight assigned to such sequence by the encoder can give more meaningful information for further optimisation. It is also worth considering another way of generating the data for the encoder to extract meaningful solutions from the decoder.

5 Conclusions

In this project we explored transformers for protein contact prediction task and attempted to design an encoder-decoder transformer for more efficient and accurate inference. We have noted certain interesting aspects, such as the necessity of the fully-connected layers for successful learning of the coupling matrix through attention mechanism and improvement of results with substitution of one-hot encoding with embedding matrix. We have also reproduced the result for the traditional methods with exception of PlmDCA and compared traditional methods to the novel ones involving MLM trained transformers. The proposed transformer led to higher TP fraction for smaller number of epochs, opposed to Vanilla transformer which introduces a step towards understanding how we could better leverage the capabilities of Transformers. Finally, we address the limitations of our work and proposed modifications to the transformer that minimises the computational noise of encoder output and helps transformer learn and extract more meaningful information.

References

F. Caredda and A. Pagnani. Direct coupling analysis and the attention mechanism. *bioRxiv*, pages 2024–02, 2024.

N. Dietler, U. Lupo, and A. F. Bitbol. Impact of phylogeny on structural contact inference from protein sequence data. *Journal of the Royal Society Interface*, 20(199):20220707, 2023. doi: 10.1098/rsif.2022.0707. URL `https://doi.org/10.1098/rsif.2022.0707`. Epub 2023 Feb 8. PMID: 36751926; PMCID: PMC9905998.

M. Ekeberg, C. Lövkvist, Y. Lan, M. Weigt, and E. Aurell. Improved contact prediction in proteins: Using pseudolikelihoods to infer potts models. *Physical Review E*, 87(1):012707, 2013. doi: 10.1103/PhysRevE.87.012707.

T. A. Jones, J. Hajdu, and M. Andersson. A new method for the display of protein structures on a two-dimensional map, with application to hemopexin. *EMBO Journal*, 5(11):2729–2734, 1986.

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

F. Morcos, A. Pagnani, B. Lunt, A. Bertolino, D. S. Marks, C. Sander, R. Zecchina, J. N. Onuchic, T. Hwa, and M. Weigt. Direct-coupling analysis of residue coevolution captures native contacts across many protein families. *Proceedings of the National Academy of Sciences*, 108(49):E1293–E1301, 2011. doi: 10.1073/pnas.1111471108.

R. Rao, N. Bhattacharya, N. Thomas, A. Derry, C. Tokheim, S. Ramachandran, A. Nambiar, A. Beyer, K. Choromanski, A. Klimovskaia, A. Luccioni, V. Saligrama, Y. Zhang, A. Green, A. Ryan, R. Chowdhury, C. Mayr, M. El-Kebir, O. Engkvist, K. Cho, Y. Bengio, R. Norel, N. Madhukar, E. Coffey, S. Bhuvanesh, L. Shen, P. Gallager, A. Wood, A. Nath, M. Wainberg, D. Kihara, and A. Rives. Msa transformer for protein structure prediction. *Nature*, 596(7873):583–589, 2021. doi: 10.1038/s41586-021-03819-2. URL `https://www.nature.com/articles/s41586-021-03819-2`.

R. Rende, F. Gerace, A. Laio, and S. Goldt. Mapping of attention mechanisms to a generalized potts model. *Physical Review Research*, 6(2), Apr. 2024. ISSN 2643-1564. doi: 10.1103/physrevresearch.6.023057. URL `http://dx.doi.org/10.1103/PhysRevResearch.6.023057`.

TensorFlow Team. Tensorboard. `https://www.tensorflow.org/tensorboard`, 2015. Accessed: 2024-07-10.

Weights & Biases Team. Weights & biases. `https://www.wandb.com/`, 2020. Accessed: 2024-07-10.

M. Weigt, R. A. White, H. Szurmant, J. A. Hoch, and T. Hwa. Identification of direct residue contacts in protein-protein interaction by message passing. *Proceedings of the National Academy of Sciences*, 106(1):67–72, 2009. doi: 10.1073/pnas.0805923106.

6 Appendix

Plots in Figure 7 showcase the magnetization of the sequences for different temperatures. This indicated the correct data generation process through Metropolis-Hastings algorithm.

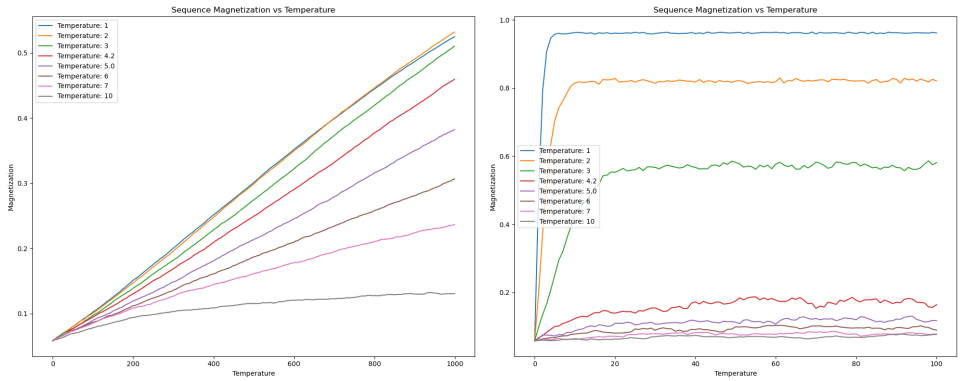

Figure 7: The average absolute magnetization per site versus the number of spin flips (left figure) and cluster flips (right figure).

Figure 8 illustrates the training of the proposed encoder-decoder transformer. Figure 9 demonstrates

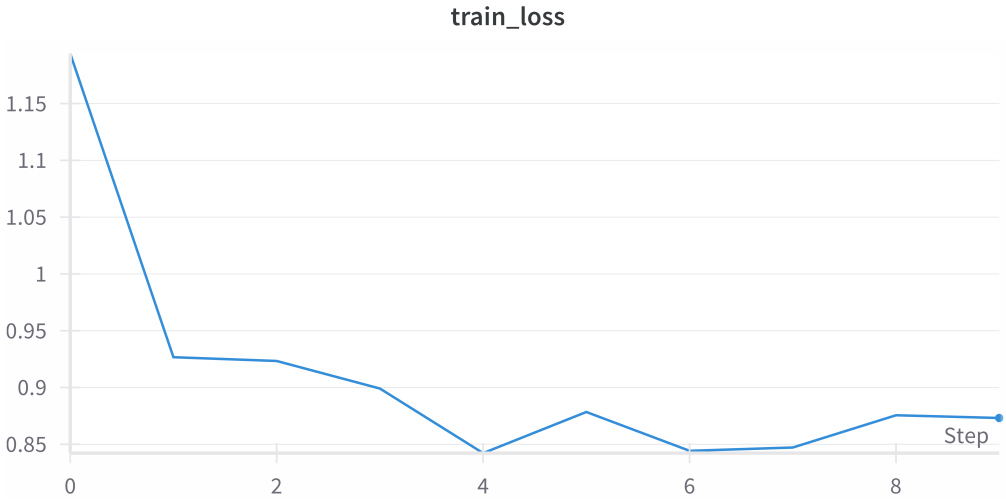

Figure 8: The loss function for encoder-decoder transformer over 10 iterations of training.

the mfDCA method applied to the data generated via real protein contact map and its comparison to the true couplings themselves

The similar result but for mutual information is demonstrated in Figure 10.

Performance of Proposed transformer on simple data with sequences of length 20 after 300 epochs is illustrated in Figure 11.

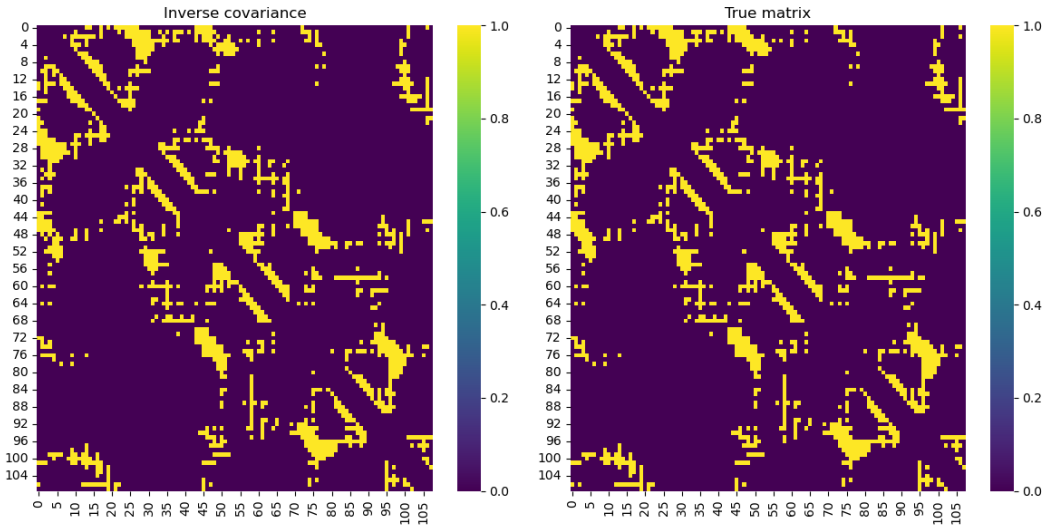

Figure 9: mfDCA matrix and the corresponding true matrix. TP fraction = 0.998

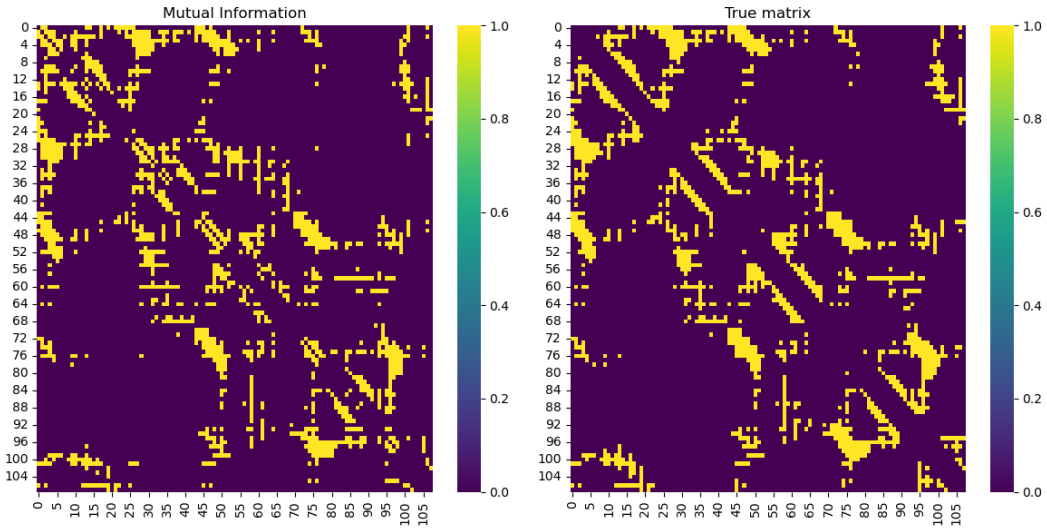

Figure 10: Mutual information coupling and the corresponding true real matrix. TP fraction = 0.966

Transformer results for sequences of length 200 with probability of contact in the matrix equal to 2%. 12.

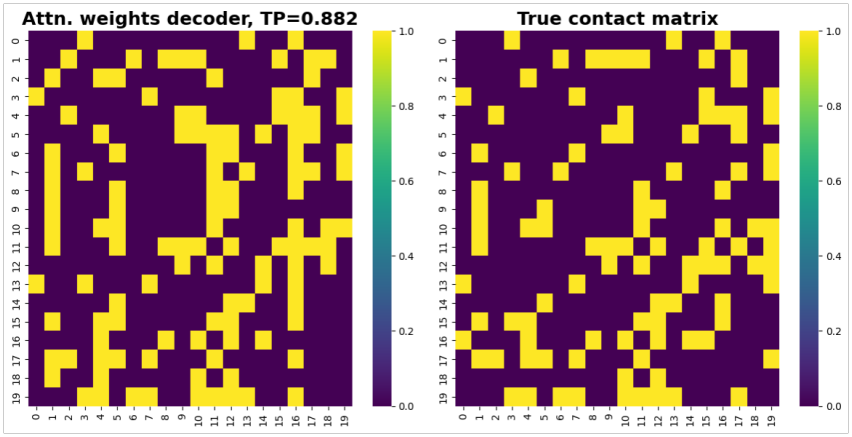

Figure 11: Performance of the proposed transformer trained on 300 epochs (equivalent to the fully trained regime of the simple transformer)

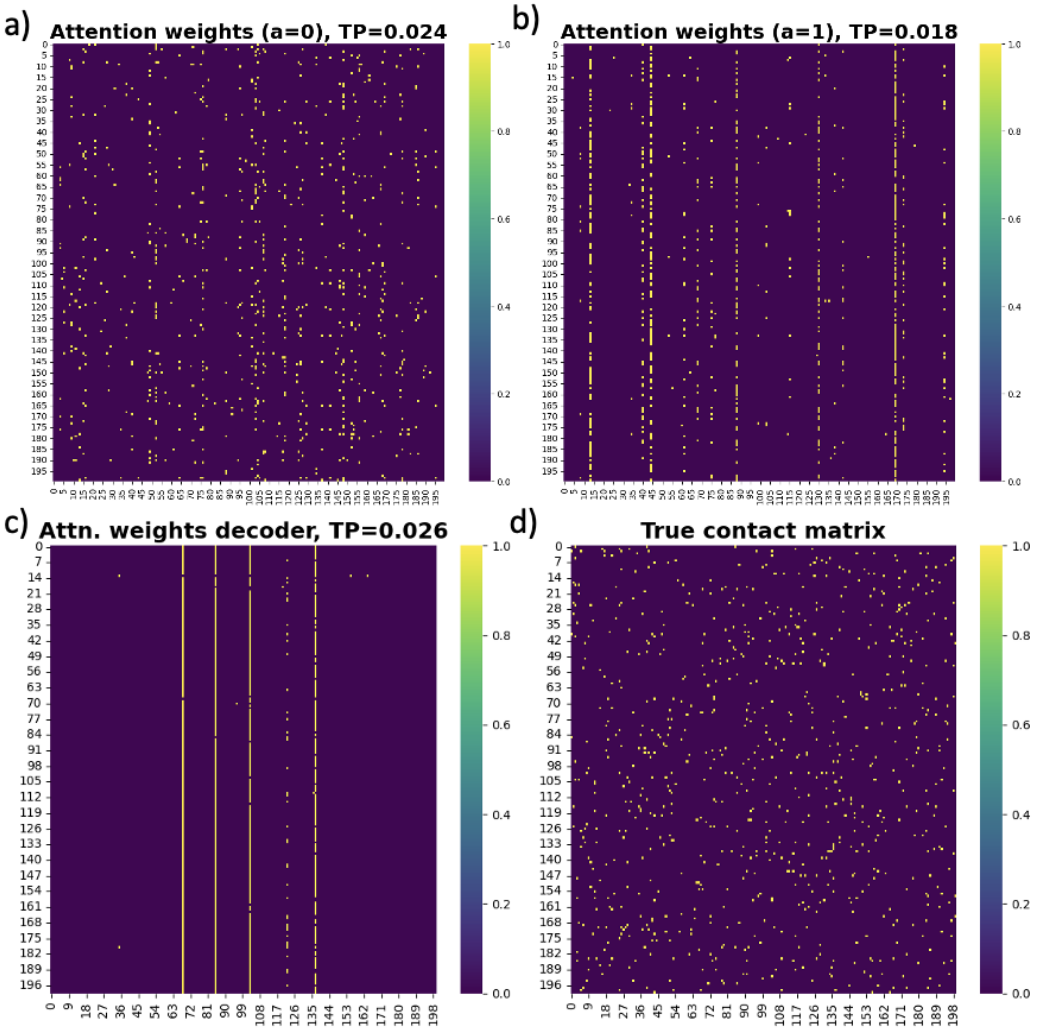

Figure 12: Performance of the simple transformer with factored and vanilla attention (parts a) and b), respectively) as well as proposed encoder-decoder transformer (part c)) on data consisting of 200 sequences of length 200 spins or amino acids. The simple transformer is trained on 300 epochs, whereas the proposed encoder-decoder one is trained on 20 epochs.