ProtMAE: Masked Autoencoding of Protein Distance Maps for Structure-Aware Representation Learning

Stanford CS231N Final Report

Emilin Mathew
Department of Computer Science & Biology
Stanford University
emilinm@stanford.edu

Aya Aburous
Department of Computer Science
Stanford University
aaburous@stanford.edu

Jad Bitar
Department of Computer Science
Stanford University
jadbitar@stanford.edu

Abstract

We introduce ProtMAE, a domain-specific masked autoencoder for protein distance map reconstruction, designed to tackle the unique biological challenges of protein structure prediction. Our approach achieves high reconstruction quality with a Structural Similarity Index Measure (SSIM) of 0.847, significantly outperforming standard Vision Transformers (SSIM: 0.009), while maintaining computational efficiency with inference times of 0.05 ms per sample. By incorporating distance-aware positional embeddings, hybrid attention-convolution blocks, and protein-specific loss functions, we demonstrate that domain knowledge integration is crucial for effective self-supervised learning in such specialized applications. We validate our learned representations through two challenging downstream tasks: secondary structure classification achieving 78.1% accuracy, and protein contact prediction achieving 71.0% AUC and 0.443 AUPR despite severe class imbalance. Our results demonstrate that self-supervised pretraining on protein distance maps learns biologically meaningful features that transfer effectively to structure prediction tasks, opening new avenues for computational protein analysis.

1. Introduction

Protein structure prediction is essential for drug design and understanding disease mechanisms. While models like AlphaFold have predicted full protein structures with remarkable accuracy, little is known about how small fragments of amino acid fold. This local folding behavior is crucial for understanding novel sequences, mutations, and disordered regions, which are often misrepresented or poorly resolved by full protein models. These local structures have direct implications for drug discovery, as many binding sites involve flexible loops and disordered regions that challenge full protein modeling approaches.

We approach this problem through a computer vision lens, treating distance maps of protein fragments as images. Distance maps encode pairwise distances between amino acid residues in a protein's 3D structure, offering a compact yet information rich representation that captures both local secondary structure and aspects of global fold topology. Unlike natural images, these maps are symmetric matrices with structured diagonal patterns reflecting sequential proximity, and sparse off-diagonal elements capturing key long range interactions that stabilize protein conformation.

Our goal is to develop self-supervised models that can reconstruct complete distance maps from partially masked inputs, thereby learning the spatial patterns underlying protein folding. The input to our method is a 64×64 grayscale image representing a normalized pairwise distance map of a 40–64 residue protein fragment, with each pixel encoding the inter-residue distance (normalized to the range [0,1], corresponding to 0–20 Å). We designed a specialized Masked Autoencoder (MAE) architecture, ProtMAE, and compared it against three baseline models: a standard CNN, a masked CNN variant, and a Vision Transformer (ViT).

ProtMAE is designed specifically for protein distance maps, introducing several key innovations: (1) distance aware positional embeddings that encode diagonal and structural distance patterns critical for protein folding; (2) hybrid attention-convolution blocks that capture both local continuity and global dependencies; and (3) protein specific loss functions that enforce symmetry and encourage smooth, biologically plausible reconstructions.

To evaluate the effectiveness of the learned representations, we transfer them to two downstream tasks. For secondary structure classification, we predict one of three classes (alpha helix, beta sheet, or coil) for each fragment.

For contact prediction, we perform binary classification to determine whether residue pairs are within 8 Å. By framing protein folding as a computer vision problem with domain specific architectures and loss functions, we designed ProtMAE to learn meaningful structural representations and protein folding mechanisms.

2. Related Works

We organize existing methodologies into four categories, examining their relationship to ProtMAE and identifying gaps our approach addresses.

2.1. End-to-End 3D Structure Predictors

AlphaFold2 pioneered the Evoformer architecture and remains the gold standard for whole-protein modelling, reaching a median C_α RMSD of 0.96 Å on CASP14 targets [10]. Its reliance on deep evolutionary alignments and heavy supervision, however, leaves flexible or disordered segments poorly resolved an Achilles' heel that ProtMAE tackles with a fragment-centric view.

RoseTTAFold extends the idea with a three-track network that jointly reasons over sequence, distances and coordinates [2], while **ESMFold** shows that large protein language models can approach AlphaFold-level accuracy using sequence alone [13]. Although state-of-the-art at the protein level, these systems demand massive compute and exhibit uneven local accuracy, motivating a lighter self-supervised alternative.

2.2. Distance Map and Contact Prediction Methods

Supervised predictors that output inter-residue distances or contacts have converged near a 71 % AUC ceiling. **RaptorX-Contact** first applied very deep, dilated CNNs to this task [15]; **TripletRes** lifted precision to 71.6 % with triplet co-evolution features and attention [11]. **ProSPr** adopted 64×64 crops similar to ours and reached ∼70 % AUC using transformer-inspired attention [4]. All three hinge on multiple-sequence alignments (MSAs); by learning directly from geometry, ProtMAE circumvents the MSA bottleneck and is therefore better suited to orphan or de-novo sequences.

2.3. Self-Supervised Masked Representation Learning

The original **MAE** work showed that randomly masking 75 % of image patches yields powerful vision representations [9]. Off-the-shelf MAEs, however, collapse on protein distance maps (SSIM 0.009 in our tests) because they lack biological biases. Sequence-level models—**ProtTrans**, **ESM-2** and **ProteinBERT**—mask amino-acid tokens instead and excel at language-like tasks [8, 12, 5]. **GearNet** moved self-supervision into geometric space but requires full 3D coordinates [16]. None of these methods learns directly from 2D distance maps; ProtMAE fills that gap.

2.4. Hybrid Vision Transformer Architectures

We were inspired by two transformer papers: **Vision Transformer** introduced patch tokenization, which shaped our distance map inputs [7]. **Swin Transformer** proposed hierarchical attention, informing our progressive masking strategy [14].

CoAtNet provides theoretical foundation for ProtMAE's hybrid blocks by demonstrating that the combination of CNN spatial locality with Transformer global modeling achieves superior performance [6]. Guided by this, we created a hybrid architecture that combines local and global pattern recognition.

2.5. ProtMAE's Position and Key Innovations

ProtMAE bridges critical gaps dicussed above by (1) focusing on local fragment patterns crucial for disordered regions that global models miss; (2) using self-supervised pretraining without evolutionary data requirements; (3) explicitly capturing spatial geometry through distance maps.

Key innovations include distance-aware positional embeddings that encode diagonal patterns crucial for protein structure, hybrid attention-convolution blocks following CoAtNet principles, and protein specific loss functions that enforce biological constraints.

3. Methods

Our ProtMAE architecture consists of multiple specialized components to address the unique challenges of protein distance map analysis. We present the mathematical formulations and design rationale for each component below.

3.1. Protein Distance Map Embedding

Unlike standard patch embeddings that treat all image regions equally, protein distance maps exhibit specific patterns that require specialized processing. Given an input distance map $D \in \mathbb{R}^{1 \times 64 \times 64}$, we employ a three-stage convolutional embedding that progressively increases channels (64→128→256) using 3×3 kernels followed by a 4×4 strided convolution to generate patches. This design enables the model to build hierarchical features that capture both fine-grained distance variations and higher-level structural abstractions.

We apply batch normalization and GELU activations after each convolutional layer to stabilize training and maintain effective gradient flow. The final embedding is reshaped from spatial dimensions into a sequence format for transformer processing, resulting in 256 patches of dimension 256.

3.2. Distance-Aware Positional Encoding

Standard sinusoidal positional encodings fail to capture the structural significance of positions in distance maps. We design positional embeddings that explicitly encode three critical aspects: absolute position, diagonal distance (representing sequence separation) and the symmetric nature of the matrix. For each patch at position (i, j) in the 16×16 patch grid, we compute:

$$p_{i,j} = p_{\text{abs}} + p_{\text{diag}} + p_{\text{sym}}$$

The absolute position component uses standard sinusoidal encoding of the coordinates. The diagonal distance component represents how far each patch is from the main diagonal, which is important because the diagonal distance corresponds to the sequence separation between residues. The symmetry component ensures that patches along the diagonal receive special encoding, and that patches equidistant from the diagonal share similar encodings, enabling the model to leverage the symmetry in distance maps effectively. By embedding these protein specific spatial relationships directly into the positional encoding, we facilitate more efficient learning of structural patterns.

3.3. Hybrid Transformer Blocks

Pure attention mechanisms excel at capturing global dependencies but lack the inductive biases helpful for local structure modeling. We introduce hybrid blocks that alternate between standard Transformer layers and combined attention-convolution blocks.

The standard Transformer block is defined as:

$$z' = z + \text{MHSA}(\text{LN}(z))$$
$$z_{\text{out}} = z' + \text{MLP}(\text{LN}(z'))$$

where MHSA denotes multi-head self-attention, LN denotes layer normalization, and MLP is a two-layer feedforward network with GELU activation and an expansion ratio of 4.

The hybrid block additionally incorporates spatial convolutions. However, during masked pretraining, applying convolutions to incomplete sequences can introduce artifacts. Thus, the convolutional pathway is disabled during pretraining and only activated for downstream tasks where the full sequence is visible.

3.4. Multi-Scale Decoder Architecture

To reconstruct both fine grained distances and global structure, we employ a multi-scale decoder with separate prediction heads at different resolutions. Given the encoded representation from the transformer, we first project it into a lower dimensional space (256→128 dimensions) to reduce computational cost while maintaining representational capacity.

For masked patches, we introduce a learnable mask token that is inserted at the positions of missing patches using restore indices. After adding decoder specific positional embeddings and applying transformer blocks, we generate predictions at multiple scales.

The decoder produces two outputs: a finegrained prediction at full 64×64 resolution and a coarse prediction initially at 32×32 resolution that is upsampled using bilinear interpolation. The final output applies a sigmoid activation to ensure distance values remain in the valid [0, 1] range.

This multi-scale approach allows the model to capture both local distance variations and global structural patterns, leading to more accurate and structurally coherent reconstructions.

3.5. Protein Specific Loss Functions

Distance maps have unique structural properties not captured by standard reconstruction losses. We use a composite loss with three components:

1. Reconstruction Loss (MSE):

$$\mathcal{L}_{\text{recon}} = \frac{1}{|M|} \sum_{(i,j)\in M} (D_{ij} - P_{ij})^2$$

where M is the set of masked positions.

2. Smoothness Loss:

$$\mathcal{L}_{\text{smooth}} = \frac{1}{HW} \sum_{i,j} (|P_{i+1,j} - P_{i,j}| + |P_{i,j+1} - P_{i,j}|)$$

3. Symmetry Loss:

$$\mathcal{L}_{\text{sym}} = \frac{1}{HW} \sum_{i,j} (P_{ij} - P_{ji})^2$$

The total loss combines these with empirical weights:

$$\mathcal{L}_{\text{total}} = \mathcal{L}_{\text{recon}} + 0.1 \cdot \mathcal{L}_{\text{smooth}} + 0.1 \cdot \mathcal{L}_{\text{sym}}$$

We chose the weights to prioritize accurate reconstruction while ensuring the model learns smooth, symmetric outputs that are representative of real protein structures.

3.6. Training Strategy

We use progressive masking, starting at 50% and linearly increasing to 75% over the first 40% of training epochs. This schedule allows the model to learn easier tasks early on and gradually handle more challenging reconstructions. Masking follows a uniform random patch distribution.

We optimize using AdamW with weight decay of 10^{-5}, which improves generalization over standard Adam. The learning rate follows a cosine annealing schedule with linear warmup: it starts small, increases linearly to the maximum value during the warmup period, then gradually decreases following a cosine curve. This approach helps stabilize training in the early stages while allowing for finetuning in later epochs.

3.7. Downstream Task Adaptations

3.7.1 Secondary Structure Prediction

We adapt ProtMAE for secondary structure prediction by adding a classification head on top of the encoder. For this task, we utilize DSSP annotated protein structures from the Protein Data Bank (PDB). The downstream model takes a masked distance map $X \in \mathbb{R}^{64\times 64}$ as input and predicts secondary structure labels $Y \in \{H, E, C\}^L$, where H, E, and C denote α-helix, β-sheet, and coil structures respectively, and L is the sequence length.

The architecture consists of the pre-trained ProtMAE encoder followed by a global average pooling layer to reduce spatial dimensions, two fully connected layers with ReLU activations, and a final softmax classification layer for three-class prediction.

The loss function is a combination of cross-entropy loss for classification and an L2 regularization term:

$$\mathcal{L}_{\text{total}} = \mathcal{L}_{\text{CE}} + \lambda \mathcal{L}_{\text{reg}}$$

where

$$\mathcal{L}_{\text{CE}} = -\sum_i y_i \log(\hat{y}_i), \quad \mathcal{L}_{\text{reg}} = \|W\|^2$$

and λ is a hyperparameter controlling regularization strength.

3.7.2 Contact Map Prediction

To predict residue-residue contacts, we extract features from encoder layers 2, 4, 6, and 8 to capture a hierarchy of representations, from local atomic details to more global folding patterns. Each of these feature maps is passed through a separate linear layer, then concatenated to form a unified representation that preserves local and global structural information.

Our contact prediction head uses a dual branch design: one branch uses transposed convolutions to reconstruct the full 2D contact map, while the other predicts contact probabilities patch by patch. The outputs from both branches are combined using learned weights and trained with binary cross-entropy loss. This setup lets us make the most of our pretrained encoder, while adding only lightweight task-specific layers for efficient transfer to contact prediction.

3.8. Masked CNN Baseline Architecture

We also implement an optimized CNN tailored for protein distance map reconstruction. The architecture follows a U-Net design with an encoder-decoder structure and skip connections. The encoder consists of four blocks, each containing two convolutional layers (Conv $\rightarrow$ BatchNorm $\rightarrow$ ReLU), followed by max pooling. The channel progression is $[32, 64, 128, 256]$, culminating in a bottleneck with 512 channels.

A unique feature we added is channel attention in the bottleneck using global average pooling and learnable channel weights. This enhances the model's ability to capture long-range dependencies by reweighting feature channels based on global context.

The decoder mirrors the encoder, using transposed convolutions for upsampling and concatenated skip connections from the corresponding encoder layers at each level.

For the masked CNN variant, we apply structured masking during training by randomly removing square patches of size 4–12. This simulates realistic missing regions in protein structure data, encouraging the model to learn contextual reconstruction strategies.

The loss function combines mean squared error (MSE) for pixel-wise reconstruction, structural similarity index (SSIM) to preserve overall structural fidelity, and a symmetry regularization term that penalizes deviations from the natural symmetry of protein distance maps.

4. Dataset and Features

Our primary dataset consists of 822,122 protein fragment distance maps we extracted from structures in the Protein Data Bank [3]. Each map is a 64×64 matrix depicting pairwise Cβ distances between residues within a 40-residue window. A few representative examples are shown in (Figure 1. This resolution captures structural patterns across fragments up to 64 residues in length and distances up to 20 Å, balancing local detail with computational efficiency.

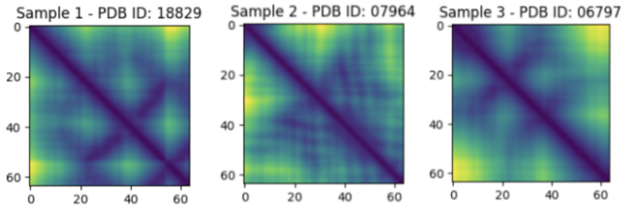

Figure 1. Examples of 64×64 distance maps extracted from PDB fragments.

To convert the raw PDB structures into normalized distance maps, we first extracted the 3D coordinates for each residue's representative atom. For missing residues and atoms with multiple conformations, we selected the highest occupancy positions. We then computed the Euclidean distance matrix, where each element $d_{ij} = \|\mathbf{r}_i - \mathbf{r}_j\|_2$ represents the distance between residues i and j. To ensure consistent input ranges suitable for neural network processing, we normalized the distances as follows: $D_{\text{norm}} = \min\left(\frac{D_{\text{raw}}}{20.0}, 1.0\right)$. We split our 822,122 samples into training, validation, and test sets with an 80-10-10 ratio: 657,697 training samples, 82,212 validation samples, and 82,213 test samples.

For our secondary structure prediction task, we used the CASP12 dataset splits from the ProteinNet benchmark of the PDB [1]. We applied strict filtering to the initial 50,914 records to ensure compatibility with ProtMAE. We retained only fragments with complete coordinate information, valid DSSP annotations, and sequence lengths ≤ 64 residues to match our encoder's input size. This resulted in a high quality subset of 3,302 protein fragments (6.5% retention). (Figure 2) shows an example of the original DSSP annotation for one of these fragments.

ID: `200L_1_A`
Sequence:
`MNIFEMLRIDEGLRLKIYKDTEGYYTIGIGHLLTKSPSLNAAK...`

DSSP:
`LLHHHHHHHHHLLEEEEEELTTSLEEEETTEEEESSSLHHHH...`

Figure 2. Example sequence and DSSP annotation (truncated) for a fragment.

Since our goal was to predict secondary structure, we mapped 8 states of DSSP labels into 3 classes we were ultimately interested in: alpha helices, beta sheets, and coils. For each protein fragment, we converted its DSSP string which consists of one character per residue into a sequence of numerical labels (0 for helix, 1 for sheet, 2 for coil). These label sequences were our training targets. Our model was trained to predict the secondary structure label for each individual amino acid residue within the 64-residue fragment.

For the contact map prediction task, we used a computationally manageable subset of 33,387 distance maps, split into 26,709 for training, 3,339 for validation, and 3,339 for testing. We derived contact map labels using the literature standard 8 Å threshold for residue-residue interactions in protein structures. To focus on long range interactions, we excluded residue pairs separated by fewer than 12 positions in the sequence since they do not typically contribute to tertiary structure. The resulting contact maps are sparse, with only 2–5% of residue pairs in contact.

5. Experiments, Results, and Discussion

5.1. Experimental Setup and Hyperparameters

All experiments were conducted on the full 822,122 protein fragment dataset unless otherwise specified. Hyperparameters were selected based on validation performance. A base learning rate of 1×10^{-4} was chosen for stable convergence, as lower values were too slow and higher ones led to instability in separate exploratory runs. A batch size of 256 was used for MAE training in order to balance gradient stability with GPU memory efficiency on T4 hardware (via GCP).

A progressive masking schedule (linearly increasing from 50% to 75%) was used after we observed that starting with high masking ratios hindered early learning. A warmup period of 2 epochs was added to decrease the likelihood of early training instability, especially given the model's custom encoder-decoder architecture and loss formulation. Weight decay of 0.05 provided effective regularization without overly constraining capacity.

5.2. Reconstruction Performance

Our primary metrics for evaluating reconstruction quality are Mean Squared Error (MSE) for pixel-level accuracy and Structural Similarity Index (SSIM) for perceptual quality. MSE measures the average squared difference between predicted and ground truth distance values:

$$\text{MSE} = \frac{1}{n}\sum_{i=1}^{n}(D_i - \hat{D}_i)^2$$

where D_i and $\hat{D}_i$ are the ground truth and predicted values. SSIM evaluates structural similarity by comparing luminance, contrast, and structure:

$$\text{SSIM}(x, y) = \frac{(2\mu_x\mu_y + c_1)(2\sigma_{xy} + c_2)}{(\mu_x^2 + \mu_y^2 + c_1)(\sigma_x^2 + \sigma_y^2 + c_2)}$$

where μ, σ^2, and σ_{xy} represent means, variances, and covariance respectively. We additionally report inference time to assess computational efficiency, a critical consideration for large scale protein analysis applications.

The results (Figure 3) reveal several critical insights. First, the standard Vision Transformer performs poorly on protein distance maps, achieving an SSIM of only 0.009, indicating failure to capture structural patterns. This highlights the importance of domain-specific design: using generic architectures developed for natural images to capture properties of protein data results in poor performance.

The Standard CNN achieves excellent reconstruction quality with the highest SSIM (0.983) and lowest MSE (0.000). This shows that convolutional architectures with proper design can effectively capture protein distance map patterns. However, this performance comes at a computational cost of 2.81 ms inference time.

The Masked CNN (75%) shows the impact of masking during training, with slightly reduced SSIM (0.960) compared to the standard version. That said, it maintains reasonable reconstruction quality while achieving 3× faster inference (0.930ms). This demonstrates that masking-based pretraining can learn meaningful representations even with significant information removal.

Our ProtMAE achieves competitive reconstruction quality (SSIM: 0.847) while being significantly more efficient, requiring only 0.05ms per sample (making it 56× faster than the standard CNN and over 1000× faster than the Vision Transformer). This efficiency stems from our optimized architecture with fewer parameters and streamlined processing, combined with the learned representations from masked pretraining.

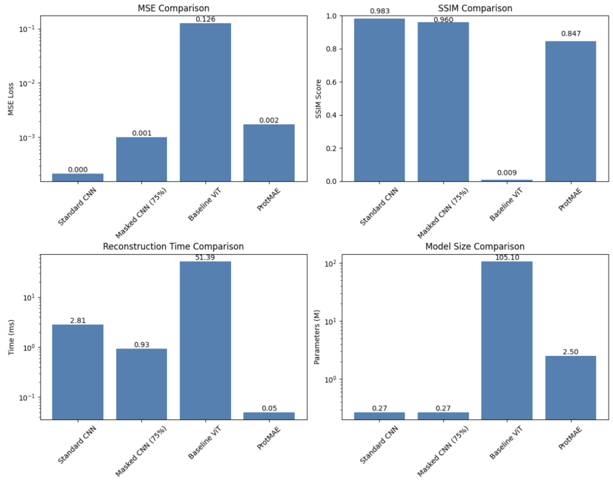

Figure 3. Reconstruction Performance Comparison

5.3. Training Dynamics and Convergence Analysis

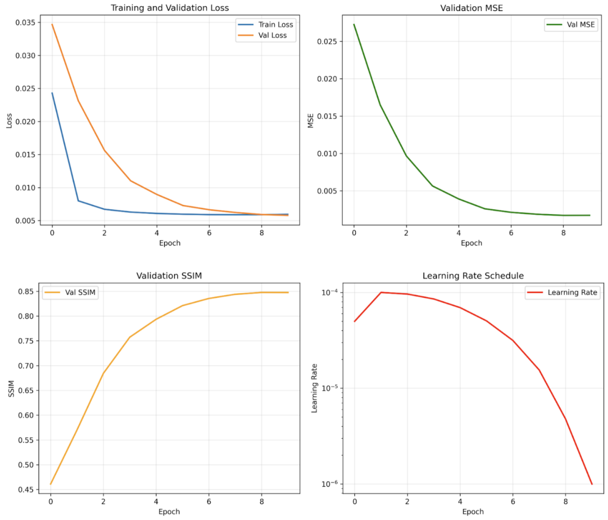

Figure 4. Training history showing (top left) training and validation loss convergence, (top right) validation MSE decrease, (bottom left) validation SSIM improvement, and (bottom right) learning rate schedule with warmup and cosine decay.

The training dynamics (Figure 4) show rapid initial convergence followed by steady improvement across all models. For ProtMAE, the validation loss closely mirrors training loss, indicating good generalization with no signs of overfitting. The SSIM improvement from 0.45 to 0.85 over 10 epochs demonstrates the model's ability to progressively capture finer structural details.

The learning rate schedule, with warmup followed by cosine decay, contributes to stable training and optimal final performance. The symmetry loss remains consistently near zero throughout training, indicating that our model naturally learns to enforce the fundamental symmetry property of distance maps without explicit constraint during reconstruction. This suggests successful internalization of protein-specific structural properties.

5.4. Qualitative Reconstruction Analysis

Qualitative analysis of reconstructions (Figure 5) reveals deeper insights into model behavior. The reconstructions show ProtMAE successfully recovers distance map structure from only 25% of visible patches. The model captures both the characteristic diagonal pattern (representing sequential proximity) and off-diagonal features (secondary structure elements like loops and turns appearing as bright yellow regions).

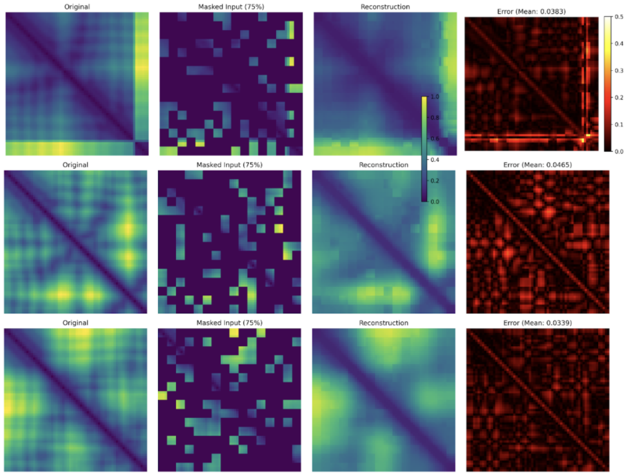

Figure 5. Qualitative reconstruction results from ProtMAE. Each row shows an example fragment: original distance map, masked input (75% masking), ProtMAE reconstruction, and corresponding error map. ProtMAE successfully recovers both diagonal and off-diagonal structural features from sparse input, with low reconstruction error.

The error maps indicate reconstruction errors remain relatively uniform across distance maps, with mean errors between 0.0383 and 0.0465. Errors follow a subtle pattern along the diagonal and at certain structural features, but importantly do not concentrate at patch boundaries—this suggests the model learns smooth transitions between masked and visible regions. The consistent red coloration in error maps shows the model slightly underestimates distances overall rather than exhibiting localized failure. The standard CNN, masked CNN, and ProtMAE models learned to enforce symmetry without explicit constraints during reconstruction, demonstrating internalization of this fundamental property. The CNN models excel at preserving fine-grained local patterns, while ProtMAE achieves better global consistency despite challenging masking ratios. This is clear in how ProtMAE maintains overall distance distribution and structural features even when reconstructing from sparse inputs, successfully recovering both local patterns (texture along the diagonal) and global structure (off-diagonal bright spots indicating long range contacts).

5.5. Secondary Structure Prediction

Through this downstream task, we evaluated ProtMAE's learned representations. We froze the pretrained encoder and trained a simple classification head to predict secondary structure classes (alpha helices, beta sheets, and coils). We observed 78.1% accuracy in predicting secondary structures which is impressive given that we used no additional information such as sequence or evolutionary data. This indicates our model was able to learn local folding patterns directly from geometric information alone.

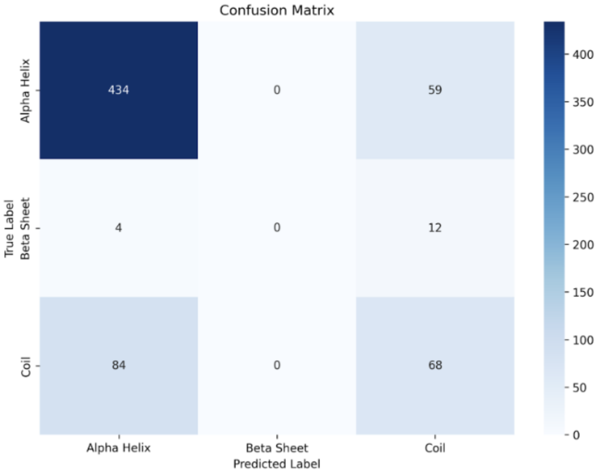

Figure 6. Confusion matrix for ProtMAE secondary structure prediction.

However, as we dug deeper into the specific performance metrics, we saw significant variability across classes: alpha helix prediction is remarkable (88% precision, 87% recall), coil prediction is moderate (52% precision, 53% recall), but beta sheet prediction fails completely (0% recall) (Figure 6). These results align with the structure of the training data: The coil class has a lot of structural heterogeneity since we labeled anything not alpha helix or beta sheet as a coil. Since they have to catch all non-regular classes, there is likely a tradeoff in precision. The beta sheet prediction failing is due to severe class imbalance with only 15 examples, which is insufficient for the model to learn reliable patterns (Figure 7). Nevertheless, the strong helix classification performance is a high signal that our pretrained features capture biologically relevant structure. With improved balancing strategies or more representative beta sheet data, we anticipate there would be substantial gains in overall and class specific performance.

5.6. Contact Prediction

Contact prediction is a challenging downstream task that evaluates whether our learned representations capture biologically meaningful structural information. Our contact prediction model achieves an AUC of 71% and an AUPR of 0.443, indicating that ProtMAE effectively encodes the structural relationships necessary to identify residue-residue contacts.

The AUPR of 0.443 is particularly noteworthy given the severe class imbalance inherent in contact prediction—typically only 2–5% of residue pairs are in contact. While the 71% AUC reflects solid discriminative performance, the modest AUPR underscores the inherent difficulty of the task, especially under our 64-residue window constraint, which limits the model's ability to detect long-range contacts.

To address this imbalance, we combine focal loss and Dice loss during training. We also incorporate transformer layers to capture spatial dependencies and apply contact-preserving augmentations to maintain structural coherence. Despite these optimizations, the results reveal fundamental challenges: the limited sequence window excludes many functionally relevant long-range contacts, and the 2D distance map representation may not fully capture the 3D spatial relationships that determine true physical contacts.

These limitations suggest that while ProtMAE learns meaningful structural features, achieving stronger performance on contact prediction may require access to longer sequence contexts or explicit 3D coordinate information.

6. Conclusion & Future Work

We designed ProtMAE, a masked autoencoder that reconstructs protein distance maps as a proxy for learning local folding patterns. We benchmarked four models: standard CNN, masked CNN, ImageNet-pretrained ViT, and ProtMAE and found that ProtMAE achieved competitive reconstruction quality.

Our core contribution is adapting masked autoencoding to this unique task of small fragment folding. ProtMAE uses a lightweight architecture with hybrid attention–convolution blocks to capture both local and long range folding patterns. We incorporate geometric priors through distance-aware positional embeddings and protein specific loss function. These inductive biases enable ProtMAE to generalize well across tasks, achieving 78% accuracy in secondary structure prediction and 71% AUC in contact map prediction. Compared to baselines, ProtMAE matches or exceeds performance (SSIM: 0.847) while being 56× faster than CNNs and over 1000× faster than the baseline ViT, demonstrating its strength as both an accurate and highly efficient representation learner.

In future work, we aim to explore two adaptations to ProtMAE. First, we are interested in integrating multimodal data, specifically sequence and functional information, in addition to the structural features this project focused on. Second, while our current approach is purely geometric, it would be interesting to incorporate evolutionary and biochemical constraints that influence protein folding. Both of these improvements could enrich the learned representations and improve generalization across diverse protein related tasks.

For the downstream tasks, we identified two challenges from limitations in fragment size and resolution. In the structure prediction task, we observed that the 64 residue input cap limited our ability to represent beta sheets, which often depend on long-range interactions. We likely require larger input contexts to capture more complex structural elements. Future work could explore expanding the input size or incorporating mechanisms to aggregate fragments in a computationally efficient way. Likewise, for the contact map task, we used an 8 Å threshold to define residue-residue contacts to predict tertiary interactions. Future work could explore adjusting this threshold to better capture a range of biologically meaningful spatial proximities. Overall, our findings suggest that distance map reconstruction is a powerful pretext task for learning rich, transferable protein representations. ProtMAE is both accurate and efficient, laying a strong foundation for future advances in structural biology and protein modeling.

7. Appendix

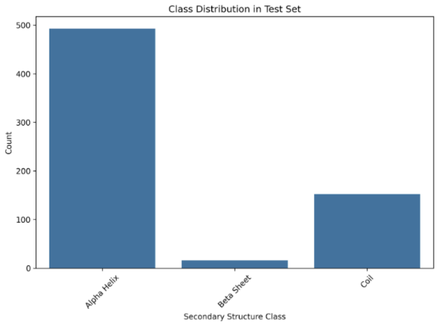

Figure 7. Class distribution in the secondary structure test set.

8. Contributions

All three authors contributed equally to this work. EM prepared the dataset and optimized the models to scale to the full set of 800,000 protein fragments. EM evaluated the baseline Vision Transformer (ViT), AA evaluated the baseline CNN. JB implemented the masked CNN. All three authors contributed to the development of the ProtMAE model. The downstream tasks were evenly divided: EM led the secondary structure prediction, while AA and JB jointly worked on the contact map prediction. This project was conducted independently and was not shared with any other course, did not involve external collaborators, and did not make use of public code.

Libraries Used

The following software libraries were used in this project:

- PyTorch (2.3.)
- NumPy (1.26.)
- Matplotlib (3.8.)
- scikit-image (0.23.)
- scikit-learn (1.4.)
- einops (Version 0.8.)
- tqdm (Version 4.66.)
- ProSPr [4]

References

[1] M. AlQuraishi. Proteinnet: a standardized data set for machine learning of protein structure. *BMC Bioinformatics*, 20(1):311, 2019.

[2] M. Baek, F. DiMaio, I. Anishchenko, J. Dauparas, S. Ovchinnikov, G. R. Lee, J. Wang, Q. Cong, L. N. Kinch, R. D. Schaeffer, et al. Accurate prediction of protein structures and interactions using a three-track neural network. *Science*, 373(6557):871–876, 2021.

[3] H. M. Berman, J. Westbrook, Z. Feng, G. Gilliland, T. N. Bhat, H. Weissig, I. N. Shindyalov, and P. E. Bourne. The protein data bank. *Nucleic Acids Research*, 28(1):235–242, 2000.

[4] W. M. Billings, B. Hedelius, T. Millecam, D. Wingate, and D. D. Corte. Prospr: democratized implementation of alphafold protein distance prediction network. *BioRxiv*, page 830273, 2019.

[5] N. Brandes, D. Ofer, Y. Peleg, N. Rappoport, and M. Linial. Proteinbert: a universal deep-learning model of protein sequence and function. *Bioinformatics*, 38(8):2102–2110, 2022.

[6] Z. Dai, H. Liu, Q. V. Le, and M. Tan. Coatnet: Marrying convolution and attention for all data sizes. *Advances in neural information processing systems*, 34:3965–3977, 2021.

[7] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*, 2020.

[8] A. Elnaggar, M. Heinzinger, C. Dallago, G. Rehawi, Y. Wang, L. Jones, T. Gibbs, T. Feher, C. Angerer, M. Steinegger, et al. Prottrans: Toward understanding the language of life through self-supervised learning. *IEEE transactions on pattern analysis and machine intelligence*, 44(10):7112–7127, 2021.

[9] K. He, X. Chen, S. Xie, Y. Li, P. Dollár, and R. Girshick. Masked autoencoders are scalable vision learners. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 16000–16009, 2022.

[10] J. Jumper, R. Evans, A. Pritzel, T. Green, M. Figurnov, O. Ronneberger, K. Tunyasuvunakool, R. Bates, A. Žídek, A. Potapenko, et al. Highly accurate protein structure prediction with alphafold. *Nature*, 596(7873):583–589, 2021.

[11] Y. Li, C. Zhang, E. W. Bell, W. Zheng, X. Zhou, D.-J. Yu, and Y. Zhang. Deducing high-accuracy protein contact-maps from a triplet of coevolutionary matrices through deep residual convolutional networks. *PLoS computational biology*, 17(3):e1008865, 2021.

[12] Z. Lin, H. Akin, R. Rao, B. Hie, Z. Zhu, W. Lu, A. dos Santos Costa, M. Fazel-Zarandi, T. Sercu, S. Candido, et al. Language models of protein sequences at the scale of evolution enable accurate structure prediction. *BioRxiv*, 2022:500902, 2022.

[13] Z. Lin, H. Akin, R. Rao, B. Hie, Z. Zhu, W. Lu, N. Smetanin, R. Verkuil, O. Kabeli, Y. Shmueli, et al. Evolutionary-scale prediction of atomic-level protein structure with a language model. *Science*, 379(6637):1123–1130, 2023.

[14] Z. Liu, Y. Lin, Y. Cao, H. Hu, Y. Wei, Z. Zhang, S. Lin, and B. Guo. Swin transformer: Hierarchical vision transformer using shifted windows. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 10012–10022, 2021.

[15] S. Wang, S. Sun, Z. Li, R. Zhang, and J. Xu. Accurate de novo prediction of protein contact map by ultra-deep learning model. *PLoS computational biology*, 13(1):e1005324, 2017.

[16] Z. Zhang, M. Xu, A. Jamasb, V. Chenthamarakshan, A. Lozano, P. Das, and J. Tang. Protein representation learning by geometric structure pretraining. *arXiv preprint arXiv:2203.06125*, 2022.