# AtoMAE: Learning Protein Structure Representations from Atomic Voxel Grids via Masked Autoencoders

**Namuk Park** [1] **Pedro O. Pinheiro** [1] **Nathan C. Frey** [1] **Sidney Lisanza** [1] **Andrew M. Watkins** [1] **Arian R. Jamasb** [1] **Matthieu Kirchmeyer** [1] **Richard Bonneau** [1] **Saeed Saremi** [1] **Vladimir Gligorijević** [1]

## Abstract

We propose AtoMAE (Atomistic Transformer with Masked Autoencoder) for deciphering three-dimensional protein structures using limited biological prior knowledge. Rather than relying on amino acid identifiers or backbone markers, the model uses voxelized protein structures with atom types as its sole input. These atomic voxels allow for the use of a Vision Transformer architecture pre-trained via Masked Autoencoder framework. Through its self-supervised reconstruction approach, AtoMAE preserves spatial context while achieving superior performance and scalability without strong inductive biases or complicated modules. In structural classification, AtoMAE outperforms both protein language modeling and graph neural networks by effectively capturing both short- and long-range relationships. Furthermore, AtoMAE can predict residue identities from backbone structures alone, achieving accuracy comparable to inverse folding models while preserving architectural simplicity. These results encourage a design shift towards models that autonomously learn multi-level biological understanding, from structure to residue, instead of relying on architectures with deeply encoded domain knowledge.

## 1. Introduction

Are the extensive biological priors we embed in protein models necessary? Recent neural network approaches typically hard-wire biological knowledge into architectures by assuming it is essential for performance. Graph representations for protein structures serve as a canonical example. Nodes typically encode spatial coordinates of alpha carbon atoms ($C_\alpha$) alongside amino acid identities (Ingraham et al., 2019; Zhang et al., 2023) and are sometimes augmented with physicochemical properties (Xia et al., 2021), while edges represent covalent bonds or spatial proximity (Gligorijević et al., 2021; Zhang et al., 2023). This paradigm inherently incorporates multiple biological premises, including the prioritization of backbone atoms and residue interactions. Though apparently natural, this approach steers architectures toward constrained components, e.g., message-passing and graph convolutional frameworks, that may limit representational expressivity. Appendix A further discusses how prior works have approached protein structure encoding.

Lessons from modern deep learning challenge this approach. Extensive prior domain knowledge in neural networks can sometimes prove not only unnecessary but detrimental to performance and scaling (Tay et al., 2023; Dehghani et al., 2023). While appropriate inductive biases ideally benefit both small and large models and data regimes (Park & Kim, 2022), even seemingly appropriate biases may become counterproductive because neural networks can learn to internalize these principles naturally (Gruver et al., 2023).

This work demonstrates that protein structural understanding can emerge from models with minimal embedded priors. In particular, we address three questions:

- **What information should inform our data representation?** We use only *atomic types*, deliberately omitting conventional biological priors such as amino acid type, backbone identifiers such as $C_\alpha$, $C_\beta$, sidechain labels, distance metrics, surface curvature, and dihedral angles (Zhang et al., 2023; Dauparas et al., 2022; Hayes et al., 2025; Yuan et al., 2023). Neural networks learn structural characteristics during training.
- **Which architectural frameworks should be used?** We employ *Vision Transformers* (ViTs) (Dosovitskiy et al., 2021) for their weak inductive bias, pre-trained using *Masked Autoencoders* (MAEs) (He et al., 2022) that offer simplicity, scalability, and exceptional performance while preserving spatial context.
- **How should we represent three-dimensional protein structures?** Our models use *voxelized featurization* of

[1]Prescient Design, Genentech. Correspondence to: Vladimir Gligorijević <gligorijevic.vladimir@gene.com>.

*Proceedings of the Workshop on Generative AI for Biology at the 42nd International Conference on Machine Learning*, Vancouver, Canada. PMLR 267, 2025. 

protein structures, in line with prior work (Ragoza et al., 2017; Derevyanko et al., 2018; Bhadra-Lobo et al., 2023). This voxel featurization seamlessly integrates with ViTs, captures rich structural information, and sidesteps atom-count assumptions (Pinheiro et al., 2023; 2024).

Rather than explicitly introducing biological knowledge through multiple features and architectural constraints, we minimize prior biases while allowing models to discover relevant biological principles as detailed in Section 2.

The proposed AtoMAE (Atomistic Transformer with Masked Autoencoder) model extracts hierarchical information from protein structures, revealing insights at both macromolecular and residue-specific levels. To provide a multifaceted assessment of protein understanding, we investigate two complementary tasks: protein structure classification and residue-type identification.

**Structure classification (Section 3).** Despite their remarkable sequence diversity, proteins can adopt only a limited repertoire of folds. Classifying such structural homology has been a central focus of structural biology. AtoMAE surpasses conventional protein language modeling approaches in structural classification by directly incorporating three-dimensional spatial information, thereby enabling the capture of short- and long-range interactions crucial for accurate structural assessment. More importantly, unlike Graph Neural Networks which face challenges with parameter scaling (Kipf & Welling, 2017; Li et al., 2018; Oono & Suzuki, 2020; Liu et al., 2024) and global information aggregation (Alon & Yahav, 2021; Wu et al., 2021; Dwivedi et al., 2022), AtoMAE achieves superior performance while maintaining computational efficiency across varying model sizes.

**Residue Identification (Section 4).** Inverse folding (IF), i.e., the challenge of predicting amino acid sequences from desired backbone structures, represents one of the cornerstone tasks in protein design. AtoMAE also demonstrates versatility in residue identification tasks, performing effectively not only with complete side-chain information but also with backbone-only input. In particular, AtoMAE fine-tuned for IF approaches the performance of specialized IF models while maintaining architectural simplicity, eliminating the need for complex components typically required by competing approaches. This suggests that our streamlined approach captures key relationships between protein structure and residue identity without relying on explicitly encoded biological priors.

## 2. Method

A wide range of protein tasks—from structure classification to protein design—primarily depend on structures. Although protein language models can encode certain structural elements (Rao et al., 2020; Rives et al., 2021; Lin et al., 2023a), we follow the hypothesis that incorporation of 3D structural data as input features will significantly outperform sequence-only approaches. This intuition has driven the increasing adoption of neural networks and pre-trained models for protein structures across the field (Gligorijević et al., 2021; Jing et al., 2021; Zhang et al., 2023; Dauparas et al., 2022; Watson et al., 2023). Currently, graph embedding dominates structural featurization, representing proteins as atomic point clouds and interaction networks. However, our observations reveal that despite encoding spatial information, graph neural network approaches fail to capture the complete structural complexity of proteins, particularly regarding computational and memory efficiency. Furthermore, these methods suffer from limitations common to graph-based paradigms such as scaling challenges (Kipf & Welling, 2017; Li et al., 2018; Oono & Suzuki, 2020; Liu et al., 2024) and global information aggregation difficulties (Alon & Yahav, 2021; Wu et al., 2021; Dwivedi et al., 2022).

We present a novel approach to protein structure modeling that captures the richness of three-dimensional information via voxelization. Remarkably, our approach introduces no explicit biological inductive bias, unlike conventional protein modeling paradigms (Gligorijević et al., 2021; Zhang et al., 2023; Lin et al., 2023b; Hayes et al., 2025; Dauparas et al., 2022). Despite this, our experimental results demonstrate that biological knowledge can be derived from the pre-training process. The proposed architecture employs vanilla Transformers with weak machine learning inductive bias, a design choice offering scalability similar to advances witnessed in language and vision domains (Kaplan et al., 2020; Zhai et al., 2022). This work demonstrates that superior performance in protein-related tasks can be achieved without relying on strong priors from either biological or machine learning domains (see Figure 1).

***Featurization*: Voxelized atom-level protein structures.** We propose an approach using voxel-based representations to featurize solely atoms from protein structures, exploiting the advantages of a 3D grid. We primarily focus on three common protein atom types: carbon, nitrogen, and oxygen, as including sulfur sometimes degrades the performance due to its rareness. We use `PyUUL` (Orlando et al., 2022) voxelizer at a resolution of 0.5Å to balance computational efficiency with performance. We truncate protein structures to 48Å×48Å×48Å volumes, yielding an input dimensionality of `[3,96,96,96]`. One problem with voxels is that they contain many sparse regions. To address this issue, we remove sparse tokens to boost predictive performance and minimize both memory and computational overhead. See the selective prompting paragraph below for more details.

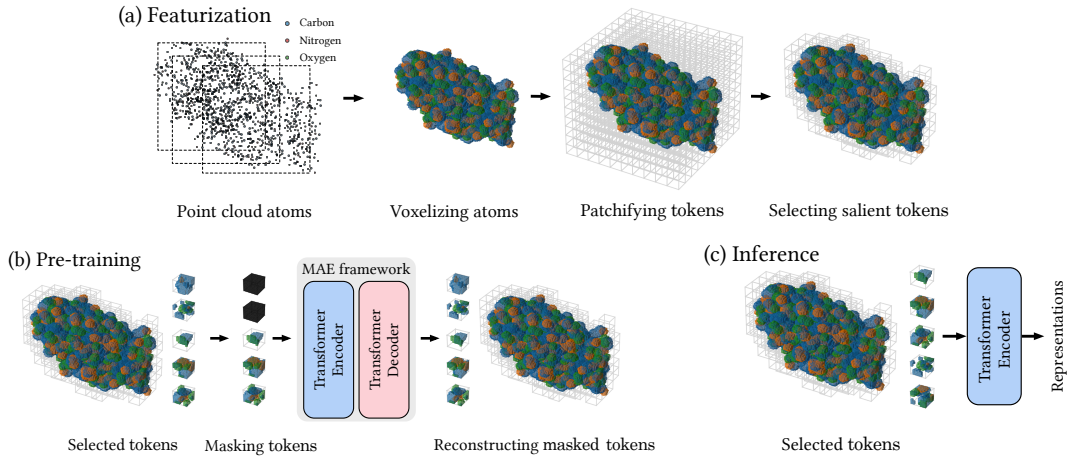



Figure 1: **AtoMAE consists of a Vision Transformer that takes patchified voxelized 3D protein structures as input.** (a) After applying data augmentation to atom point clouds, including random cropping for protein truncation, the method voxelizes the atoms and patchifies the large voxel space into small cube *tokens* for the ViT. A large portion of empty tokens is excluded, and occupied tokens are primarily used. (b) MAE is applied to pre-train the ViT. After masking out 90% of tokens, the ViT reconstructs the input protein structures to learn representations. (c) Only the encoder is used for downstream tasks.

***Architecture*: ViTs for 3D voxel spaces.** We employ Vision Transformers (ViTs) (Dosovitskiy et al., 2021) instead of convolutional neural networks (CNNs) (He et al., 2016; Liu et al., 2022) as they rely on weaker inductive biases. This choice is justified by the difficulty of predefining the best interaction range for each biological task; consequently, the extensive receptive field of ViTs (Dosovitskiy et al., 2021; Raghu et al.; Park & Kim, 2022) offers a promising path toward a robust protein structure foundation model. In our approach, we divide protein-containing voxels into $12^3$ subcubes, each with dimensions of $8^3$ ($4^3$Å$^3$), balancing predictive performance and computational efficiency. To accommodate 3D voxel data, we extend 2D cosine positional embeddings (Dosovitskiy et al., 2021) into three-dimensional space (c.f. Pang et al. (2022); Hess et al. (2022)).

***Pre-training*: Masked Autoencoders for 3D.** For pre-training, we adapt masked autoencoders (MAEs) (He et al., 2022) with very few modifications. Recent findings indicate that masked token modeling significantly surpasses contrastive learning methods in scalable and dense prediction scenarios (Park et al., 2023), suggesting that MAEs can effectively model biological information from local protein geometry. During pre-training, we mask $p = 90\%$ of all tokens, which is higher than typical vision settings ($p = 75\%$). Because the voxel featurization contains substantial redundant regions, this high masking ratio remains effective (Mirza et al., 2023; Feichtenhofer et al., 2022; Tong et al., 2022). The decoder takes both the representation tokens produced by the encoder and mask tokens associated with masked-out input tokens, aiming to reconstruct these masked tokens into their original protein structures. Departing from the vanilla MAE approach, the proposed method omits mask tokens for the removed input tokens to improve the performance.

***Selective prompting*: attention masking and sparse token removal improve the performance.** A key improvement over vanilla MAE is selective prompting. By recognizing that not all tokens are equally informative, selective prompting—which consists of two techniques: *self-attention masking* and *token removal*—enables models to focus on the most salient tokens (c.f. Rao et al. (2021); Goyal et al. (2020)). This technique can be applied to both training and inference without extensive modifications, and can be integrated into existing pipelines or already trained models as an off-the-shelf method. First, we incorporate a mask, $\mathbf{M}$, into the self-attention mechanism (Vaswani et al., 2017; Dosovitskiy et al., 2021) of ViT as follows:

$$\text{Attention}(Q, K, V) = \text{Softmax}\left(\frac{QK^\top}{\sqrt{d_k}} + \mathbf{M}\right)V \quad (1)$$

where $Q$, $K$, and $V$ are query, key and value derived from inputs, $\sqrt{d_k}$ is the dimensionality of the key vectors. Here, each element of the self-attention mask $\mathbf{M}$ is set to $-\infty$ (for tokens to be disregarded) or $0$ (for tokens to be preserved) under a specified token selection algorithm (c.f. Vaswani et al. (2017); Devlin et al. (2019)). This binary mask excludes marked tokens from the attention map by nullifying their attention scores. The selection algorithm we employ is rule-based random selection: This approach selects $q_+\%$ of salient tokens and $q_-\%$ of non-salient empty tokens, which

implies acknowledging the usefulness of a few non-salient tokens. In particular, we average each token's values and apply a hyperparameter cutoff $\varepsilon = 0$ to identify salient *occupied* tokens. We employ $q_+ = 60\%$ and $q_- = 40\%$ for zero-shot, and $q_+ = 100\%$ and $q_- = 20\%$ for finetuning.

Although the attention mask ignores non-salient tokens in self-attention, the overall computational and memory complexity remains unchanged. Because one of the main computational burdens in ViTs is the calculation of self-attention maps, we propose removing non-salient tokens first using a hyperparameter ratio $\rho$. While a small $\rho$ preserves the most informative regions but yields minimal savings in training time and memory, a larger $\rho$ can inadvertently remove some salient tokens while enabling larger models and larger batch sizes. This allows us to exclude empty regions of voxels from computation, enabling memory and computationally efficient handling of large proteins.

***Loss*: Classification objective improves the performance.** We also modify the reconstruction loss. Our target atom type-wise densities $x_{i,j}$ are bounded between 0 and 1. Interpreting these as target occupancy probabilities allows the use of a binary cross-entropy (BCE) loss, instead of a regression-based $\ell_2$ loss common in MAE for continuous values. Pre-training is thus reframed as predicting voxel occupancy probability per atom type, with the model's decoder outputting corresponding probability predictions via a final sigmoid activation. This approach is akin to performing per-voxel, per-atom-type logistic regression. To address the challenge of atom type class imbalance, we integrate importance weights $w_j$ for each atom type $j$. Moreover, small batch sizes and low learning rates can also mitigate class imbalance and improve training stability (Shwartz-Ziv et al., 2023). In sum, the reconstruction loss $\mathcal{L}_{\text{rec}}$ is defined as:

$$\mathcal{L}_{\text{rec}} = \frac{1}{|N|} \sum_{i \in N} \frac{1}{C} \sum_{j=1}^{C} w_j \cdot \ell_{\text{BCE}}(x_{i,j}, \hat{x}_{i,j}) \qquad (2)$$

where $w_j$ is the class weight, $\ell_{\text{BCE}}$ is the BCE between the ground truth occupancy probability $x_{i,j}$ (a value in $[0, 1]$ for atom type $j$ in voxel $i$) and the decoder's predicted probability $\hat{x}_{i,j}$. $C$ is the number of atom types, and $|N|$ is the total number of voxels.

***Dataset*: Protein structure curation.** In our data curation process, we retrieve 218k protein structures from the Protein Data Bank (PDB) (Berman et al., 2000) (cutoff date: September 2024). To ensure a purely protein dataset, we remove all entries containing non-protein components. We then further limit our selection to structures solved by X-ray crystallography or electron microscopy.

One practical concern is that experimentally resolved protein structures often have missing atoms (Gall et al., 2007). We addressed this by using `PDBFixer` (Eastman et al., 2017) (Eastman et al., 2017) to repair minor gaps, such as one missing amino acids or missing atoms of residues. However, for regions with extensive gaps, we did not attempt to reconstruct the missing segments, leaving them unfilled in our final dataset. We lastly remove duplicate protein sequences using MMSeq2 (Steinegger & Söding, 2017).

***Data augmentation*.** We employ four data augmentation strategies—rotation, random cropping, coordinate noise, and protein substructure sampling—during the pre-training stage. These methods not only mitigate the lack of training data issue in comparison to amino acid sequences but also increase robustness against each form of transformation.

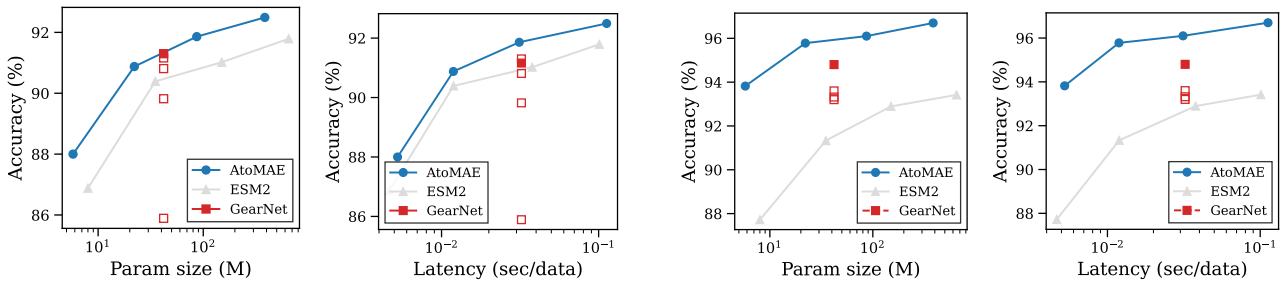
ViTs for proteins rely heavily on the rotation of protein structures as a data augmentation strategy in contrast to many popular protein architectures that employ SE(3)-equivariant modules, e.g., Fuchs et al. (2020); Satorras et al.; Ganea et al. (2022). Such augmentation, alongside with large-scale protein structure datasets, enhances both predictive performance and rotational invariance (Krizhevsky et al., 2012; Cubuk et al., 2020; Park & Kim, 2022).

Proteins are often so large that they surpass the voxel capacity dictated by the input dimensions. To address this limitation, we randomly crop regions of the protein structures so they fit the input volume. This technique with rotation lets the model remain equivariant with respect to translation transformation, free from reliance on any particular coordinate frame.

Experimentally determined protein structures intrinsically contains prediction errors (Hooft et al., 1996; Read et al., 2011). We strengthen neural networks' atomic coordinate robustness by adding random noise as spatial jitter to each atom (c.f., Dauparas et al. (2022)). Controlled by the hyperparameter $\sigma$, the noise is drawn from a three-dimensional Gaussian distribution, i.e., $\boldsymbol{\xi} \sim \mathcal{N}(\mathbf{0}, \sigma^2 \mathbf{I})$ where $\sigma = 0.2$Å.

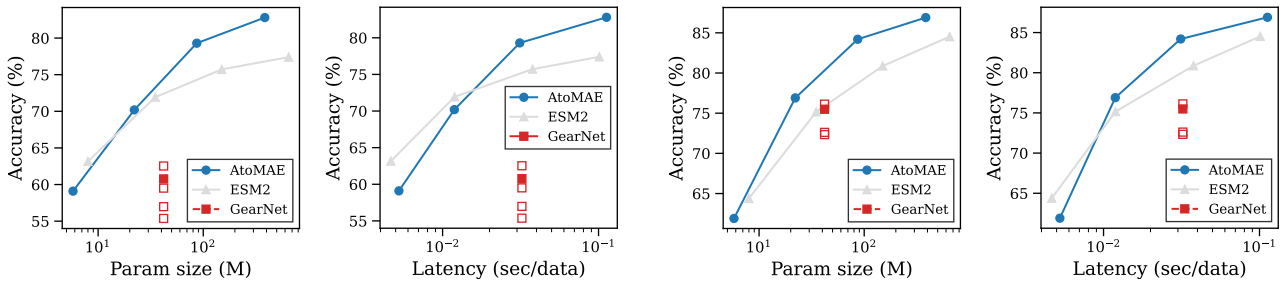
While the dataset contains entire protein structures, inputs may appear in diverse formats, creating a domain mismatch. To address this challenge, we introduce protein substructure sampling, a novel technique that strategically fragments whole proteins into smaller functional units like chains or domains for use as inputs. In practice, rather than explicitly decomposing proteins into domains, we can generate inputs by integrating samples with the CATH training dataset (Orengo et al., 1997) for simplicity.

***Optimizer*.** We use AdamW (Loshchilov & Hutter, 2019) for training, by following the vanilla MAE setting (He et al., 2022). See Appendix B for detailed configuration and optimizer information for downstream tasks.




(a) Zero-shot Class classification

(b) Fine-tuning Class classification




(c) Zero-shot Architecture classification

(d) Fine-tuning Architecture classification

Figure 2: **AtoMAE outperforms baselines in protein structure classification problems on the CATH dataset.** The filled square and open red squares represent GearNet with distance prediction and other pre-training methods, respectively. (a) In zero-shot (kNN) Class-level classification (mainly secondary structure prediction), AtoMAE surpasses ESM2, achieving better accuracy-to-parameter and accuracy-to-latency. GearNet shows comparable parameter-accuracy trends but with higher latency. (b) These patterns persist in finetuning: structure-aware models (AtoMAE, GearNet) outperform ESM2 in accuracy-parameter ratio, though GearNet requires significantly more computational resources. (c) For Architecture-level classification (secondary structure global arrangement), AtoMAE exceeds ESM2's performance while GearNet underperforms, suggesting GearNet's limitation in capturing global information. (d) Models leveraging structural information (AtoMAE, GearNet) demonstrate significant performance improvements during finetuning compared to ESM2.

## 3. Structure Classification

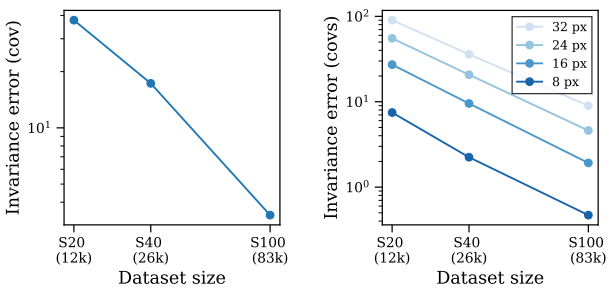
This section shows that the proposed method distills structural cues across scales from local patterns to fold topology. In particular, it surpasses both protein language modeling, ESM2 (Lin et al., 2023b), and a graph neural network approach, GearNet (Zhang et al., 2023), on the CATH classification benchmark (Orengo et al., 1997; Sillitoe et al., 2019), which probes local secondary structure makeup alongside global spatial arrangement of the elements.

Leveraging the Transformer backbone, the proposed models scale similarly to those in vision and language fields. In particular, compared to a GNN, our streamlined architecture demonstrates superior computational and memory efficiency relative to parameter count. While Transformers lack built-in SE(3) invariance (Dosovitskiy et al., 2021; Rojas-Gomez et al., 2024), extensive data augmentation allows it to learn such symmetry, an advantage that compounds with more data.

**A preliminary task: structural moieties identification.** The proposed neural networks demonstrate the capability to identify critical protein structural elements, including backbone conformations and $C_\alpha$ positions, without receiving explicit structural annotations during pre-training. This indicates the models' understanding of structural components, which guides their inference of structurally informed output representations. See Appendix C.1 for more details.

**CATH classification as a structural benchmark.** The CATH dataset organizes protein structural domains into a hierarchical classification system with several levels (i.e., Class, Architecture, Topology, and Homology) of increasing specificity. The evaluation framework in this section leverages this multi-tier taxonomy, with particular emphasis on the top two crucial classification levels to assess our model's ability to discern both local structures and global arrangements.

We evaluated the proposed method against established baselines using two distinct approaches: zero-shot (k-nearest




(a) Rotation (b) Translation

Figure 3: **Pre-trained AtoMAE learn SE(3) symmetry.** (a) The invariance error of AtoMAE's representation with respect to rotation. Lower invariance error indicates higher SO(3) invariance. As the training dataset size increases, the model better learns SO(3) invariance. (b) AtoMAE's invariance error with respect to translation. Consistent with the previous result, increasing training data improves translation invariance. This improvement is consistent across various translation magnitudes.

neighborhood) and fine-tuning. For the k-NN approach, we used $k = 20$. The fine-tuning protocol involves training for 50 epochs on the S20 dataset (12k training data points) which is the smallest similarity-based subset of CATH.

**CATH: Class-level results.** The highest level taxonomy of the CATH dataset depends on overall secondary-structure composition, which reflects the proteins' local geometry. The classes contain mainly-$\alpha$, mainly-$\beta$, a mixture of $\alpha/\beta$, and proteins with few secondary structures. We include only the three major classes in the evaluation.

Figure 2a illustrates top-1 zero-shot accuracies of the proposed method and the baselines. From tiny (6M) to large (385M) parameter models, the proposed AtoMAE consistently outperforms (sequence-only) ESM2 counterparts. With base-sized (86M) parameters, AtoMAE achieves performance comparable to the 650M-parameter ESM2. A (graph-based) GearNet pre-trained on distance prediction (the filled square) follows AtoMAE's performance curve, delivering impressive results relative to its parameter count.

However, GearNet requires significant computational and memory resources due to its mechanisms such as message passing (Zhang et al., 2023), offering only a limited latency advantage over ESM2 despite incorporating structural information. Additionally, unlike Transformers, GearNet exhibits performance degradation with increased layer depth in other tasks such as Enzyme Commission number prediction. This scalability limitation has been observed across protein fields (Gligorijević et al., 2021; Hsu et al., 2022) and other GNN applications (Kipf & Welling, 2017; Li et al., 2018; Oono & Suzuki, 2020; Liu et al., 2024). We were unable to report GearNet's performance on deeper architectures for CATH due to checkpoint unavailability.

Figure 2b shows the top-1 accuracy of the fine-tuned models. In this task, AtoMAE achieves significant improvements when fine-tuned by effectively encoding structural information from the training dataset; even the smallest (6M) variant of AtoMAE outperforms large (650M) ESM2. GearNet maintains its parameter efficiency advantage over ESM2, though at the cost of increased computational latency.

These results demonstrate that AtoMAE's pre-training process instills a robust understanding of structural information, particularly local characteristics. AtoMAE achieves this despite its constrained training dataset, enabling transfer to downstream tasks that depend on protein structural properties, surpassing other baseline approaches even under data-limited conditions.

**CATH: Architecture-level results.** The second level of CATH classification, exemplified by structures such as 2-layer sandwich and $\alpha$-$\beta$ barrel configurations, primarily depends on the 3-D arrangement of secondary structures. To evaluate the global properties of pre-trained representations, we compare methods at this level using the 40 subclasses derived from the three categories used in the previous task.

Figure 2c shows the zero-shot top-1 accuracy on architecture-level classification. AtoMAE outperforms ESM2 in large parameter regimes (86M, 385M), aligning with MAE's capacity to encode global information (He et al., 2022; Park et al., 2023). While it falls behind ESM2 in small parameter regimes (6M), it demonstrates superior scalability. GearNet's pre-training methods perform below ESM2's benchmark; we assume that it is due to the explicit encoding of edges introducing locality bias that constrains the model's ability to represent long-range dependencies (c.f. Alon & Yahav (2021); Wu et al. (2021); Dwivedi et al. (2022)).

Figure 2d shows the top-1 fine-tuning accuracy. Consistent with previous results, AtoMAE demonstrates significant performance improvements through fine-tuning and outperforms ESM2 and GearNet. GearNet also exhibits substantial gains when transitioning from zero-shot to fine-tuned evaluation, matching ESM2's accuracy within the parameter regime.

These results demonstrate that AtoMAE approach successfully captures both local and global architectural information. This ability contrasts with GNN frameworks' inherent limitations in global feature acquisition (Alon & Yahav, 2021; Wu et al., 2021; Dwivedi et al., 2022), while Vision Transformers excel in modeling long-range dependencies (Dosovitskiy et al., 2021; Raghu et al.; Park & Kim, 2022).

**Symmetry can be a learnable property.** By incorporating rotation and random crop data augmentation in both pre-training and fine-tuning stages, we expect AtoMAE to learn SE(3) symmetry. To quantify this property, we measure the invariance error of representations, $\mathrm{Cov}\big(f(\boldsymbol{x}),\, f(\mathcal{T}(\boldsymbol{x}))\big)$, which is the covariance between the representation and the transformed representation with respect to a transformation $\mathcal{T}$, after applying rotation and translation to the input 3D voxelized protein structures.

Figure 3 (a) and (b) show an inverse relationship between training dataset size and invariance error of rotation and translation, respectively. This confirms that AtoMAE acquires SE(3) symmetry with larger and more diverse datasets. However, ViT's architecture, lacking explicit SE(3) equivariance modules, cannot achieve *perfect* equivariance regardless of data abundance (Gruver et al., 2023). To address this limitation, we implement rotation test-time augmentation ensemble (c.f. Krizhevsky et al. (2012); Ayhan & Berens (2018)), which is particularly useful when processing larger proteins that require truncation via random crop. Practical examples of this ensemble approach are provided in Section 4 and Appendix C.2.

## 4. Residue Identification

This section demonstrates that the proposed method can understand both residue information and structural features. Notably, despite receiving only atomic type information as input, the pre-training process effectively instills meaningful biological knowledge into the neural networks. This inherent understanding is further refined through fine-tuning with residue labels, resulting in high predictive accuracy.

**A preliminary task: residue identification from sidechains.** Our analysis confirms that neural networks extract residue information from datasets encompassing both backbone and sidechain elements, even during self-supervised pre-training without explicit residue labels. By utilizing a frozen pre-trained neural network with a minimal four-layer Transformer for token-level decoding, we achieve an accuracy of $92\%$. Fine-tuning substantially enhances this performance, culminating in high discrimination between residue types at an accuracy of $99\%$. See Appendix C.2 for a practical use case of this task.

This framework can be extended to the inverse folding (IF) problem, where the task requires predicting viable residue types using only backbone structures. Initial experiments with frozen backbone yielded promising results with an accuracy of 23%. The label-informed fine-tuning achieves comparable accuracy to state-of-the-art IF methods such as ProteinMPNN (Dauparas et al., 2022) and ESM-IF (Hsu et al., 2022) without using complex and task-specific decoders, e.g., U-Net (Ronneberger et al., 2015) or SigFormer (Xie et al., 2021). Below, we provide a detailed explanation of our approach and present the results for this problem.

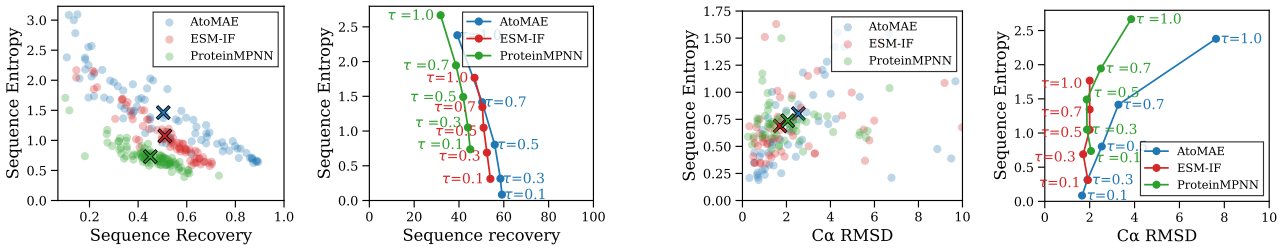
**Semantic segmentation approach.** Our approach diverges from conventional inverse folding paradigms. We treat proteins as voxelized structures and consider only backbone carbon and nitrogen atoms, excluding additional biological features such as $C_\alpha$ labels from the input, thereby maximizing input flexibility. Instead of producing sequence-level probabilities, our architecture first generates spatially-resolved predictions analogous to semantic segmentation, maintaining dimensional context between input and output spaces. We then aggregate voxel-wise probabilities into residue-specific predictions through averaging, enabling direct comparison with conventional sequence-based inverse folding methods.

To verify whether our method correctly identifies residues, we adopt multinomial sampling with temperature $\tau$ to obtain diverse sequences from the estimated probabilities. Even with this simple approach, a substantial number of sequences fold correctly. For example, in terms of folding success rate, i.e., the proportion of structures with RMSD below 2.0Å among those with pLDDT exceeding 80%, AtoMAE demonstrates competitive performance at 69.7% (72.3% for ESM-IF and 61.5% for ProteinMPNN). For training and testing, we use the same dataset used for pre-training.

To accommodate larger protein structures that surpass our model's input capacity, we implement test-time augmentation via random cropping and rotation. By averaging sequence probabilities across 8 ensemble iterations, we achieve comprehensive coverage of proteins spanning about 500 amino acids. The large ensemble size also results in enhanced recovery rates. See Appendix C.2 for details.

**Metrics for fidelity and diversity.** We use metrics to evaluate the generated sequences from fidelity and diversity perspectives. For fidelity, we use two metrics: (1) sequence recovery rate, which is a standard sequence-level accuracy metric as well as the training objective, and (2) mean $C_\alpha$ RMSD between the input reference backbone structure and the backbones of predicted structures via ESMFold (Lin et al., 2023a). This $C_\alpha$ RMSD demonstrates whether the generated sequences are biologically plausible and preserve structure. For diversity, we use mean residue-wise Shannon entropy, $H = -\frac{1}{L}\sum_{i=1}^{L}\sum_{a\in\mathcal{A}} p_{i,a}\,\log p_{i,a}$, where $L$ is the sequence length, $\mathcal{A}$ is the set of 20 standard amino acids, and $p_{i,a}$ is the empirical probability of observing amino acid $a$ at position $i$ across the sampled sequences.

**Inverse folding results.** Figure 4a shows the sequence recovery rate and entropy for the sequences generated by the proposed method and baselines. While not specifically designed for inverse folding, AtoMAE demonstrates com-




(a) Sequence perspective (sequence recovery rate)

(b) Structure perspective ($C_\alpha$ RMSD)

Figure 4: **AtoMAE can predict residue types from protein backbone structures.** (a) AtoMAE shows a sequence recovery rate similar to baseline inverse folding methods that infer sequences from the backbone. Left figure shows the distribution; each point represents a protein, and $\times$ represents the mean value. AtoMAE exhibits similar patterns in the recovery rate versus sequence entropy relationship across sampling temperatures ($\tau$) compared to other models. (b) The sequences generated by AtoMAE are refoldable. The $C_\alpha$ RMSD values are comparable at low temperatures.

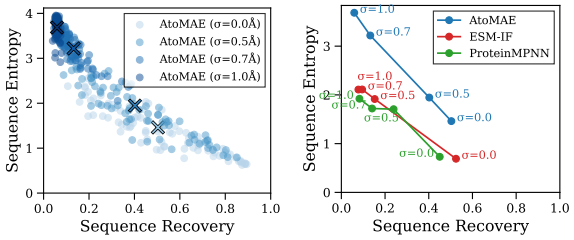



Figure 5: **AtoMAE is robust against coordinate noise.** Adding coordinate noise to the backbone results in decreased fidelity while increasing diversity (where $\sigma$ represents the Gaussian noise magnitude). In the left figure, each point represents a protein. AtoMAE demonstrates greater robustness compared to baselines in the low noise regime.

parable performance to specialized algorithms in backbone-to-sequence prediction. Without an autoregressive decoder, AtoMAE generates more diverse sequences, with the trade-off between recovery rate and diversity controllable by temperature adjustment. The trend with respect to varying temperature aligns with other inverse folding methods, suggesting AtoMAE's simple structure captures similar characteristics to specialized models.

Beyond sequence-level metrics, we also evaluated the structural quality of the generated sequences. Figure 4b shows the three methods demonstrated comparable $C_\alpha$ RMSD values at low sampling temperatures ($0.1 \leq \tau \leq 0.5$). This result suggests that AtoMAE effectively encodes biological priors at a level consistent with other specialized inverse folding methods. Despite its success, AtoMAE exhibits high $C_\alpha$ RMSD values at high sampling temperatures, which we attribute to the absence of a specialized sampling decoder.

**Robustness to random noise.** Input backbones may contain coordinate noise. On one hand, this can increase the diversity of generated sequences, while on the other hand, excessive sensitivity to noise can degrade performance and generalization abilities (Naseer et al., 2021). By introducing Gaussian random noise to the backbone atoms during *inference*, we evaluate the robustness of the methods. Figure 5 shows performance changes with respect to noise magnitude. As expected, stronger random noise tends to decrease fidelity while increasing sequence diversity. AtoMAE demonstrates higher robustness against weak noise ($\sigma = 0.5$Å) perturbations compared to ProteinMPNN and ESM-IF.

## 5. Discussion

The proposed atomic voxel-based neural network approach for protein structure modeling demonstrates significant promise across multiple perspectives. It surpasses both protein language models and graph neural networks in capturing structural information while learning essential biological knowledge. Performance metrics and properties scale with increasing parameter count. The transformer-only design delivers efficiency without sacrificing performance, while leaving room for future optimization through advanced techniques, such as FlashAttention (Dao et al., 2022). Moreover, the voxelized atom featurization method provides advantages over traditional representations by encoding rich 3D structure information. This approach captures atomic detail, encompassing both protein backbones and sidechains, which is essential for accurate molecular representation. Such detailed modeling potentially benefits applications involving molecular interfaces. The limitations and future research directions are discussed in Appendix D. Consequently, these findings underscore the importance of key design paradigms: models with weak inductive bias can autonomously construct biological domain knowledge via representation learning, thereby enabling flexible input, scalability, and superior performance.

## Acknowledgements

We are grateful to the Prescient Design team for their support and discussion throughout this project. Special thanks to Kyunghyun Cho for his invaluable guidance and feedback. We also thank Sungmin Cha for the insightful discussions that contributed to this work.

## Impact Statement

This paper presents work whose goal is to advance the field of Machine Learning and AI for biology. There are many potential societal consequences of our work, none which we feel must be specifically highlighted here.

## References

Alon, U. and Yahav, E. On the bottleneck of graph neural networks and its practical implications. In *International Conference on Learning Representations*, 2021.

Ayhan, M. S. and Berens, P. Test-time data augmentation for estimation of heteroscedastic aleatoric uncertainty in deep neural networks. In *Medical Imaging with Deep Learning*, 2018.

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

Bhadra-Lobo, S., Subedy, A., Khare, S. D., and Lamoureux, G. Se3lig: Se (3)-equivariant cnns for the reconstruction of cofactors and ligands in protein structures. In *Advances in Neural Information Processing Systems*, 2023.

Cubuk, E. D., Zoph, B., Shlens, J., and Le, Q. V. Randaugment: Practical automated data augmentation with a reduced search space. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition workshops*, pp. 702–703, 2020.

Dao, T., Fu, D., Ermon, S., Rudra, A., and Ré, C. Flashattention: Fast and memory-efficient exact attention with io-awareness. In *Advances in Neural Information Processing Systems*, 2022.

Dauparas, J., Anishchenko, I., Bennett, N., Bai, H., Ragotte, R. J., Milles, L. F., Wicky, B. I., Courbet, A., de Haas, R. J., Bethel, N., et al. Robust deep learning–based protein sequence design using proteinmpnn. *Science*, 378 (6615):49–56, 2022.

Dehghani, M., Djolonga, J., Mustafa, B., Padlewski, P., Heek, J., Gilmer, J., Steiner, A. P., Caron, M., Geirhos, R., Alabdulmohsin, I., et al. Scaling vision transformers to 22 billion parameters. In *International Conference on Machine Learning*. PMLR, 2023.

Derevyanko, G., Grudinin, S., Bengio, Y., and Lamoureux, G. Deep convolutional networks for quality assessment of protein folds. *Bioinformatics*, 34(23):4046–4053, 2018.

Devlin, J., Chang, M.-W., Lee, K., and Toutanova, K. Bert: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 conference of the North American chapter of the association for computational linguistics: human language technologies, volume 1 (long and short papers)*, pp. 4171–4186, 2019.

Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., Uszkoreit, J., and Houlsby, N. An image is worth 16x16 words: Transformers for image recognition at scale. In *International Conference on Learning Representations*, 2021.

Dwivedi, V. P., Rampášek, L., Galkin, M., Parviz, A., Wolf, G., Luu, A. T., and Beaini, D. Long range graph benchmark. In *Advances in Neural Information Processing Systems*, 2022.

Eastman, P., Swails, J., Chodera, J. D., McGibbon, R. T., Zhao, Y., Beauchamp, K. A., Wang, L.-P., Simmonett, A. C., Harrigan, M. P., Stern, C. D., et al. Openmm 7: Rapid development of high performance algorithms for molecular dynamics. *PLoS computational biology*, 13(7): e1005659, 2017.

Feichtenhofer, C., Li, Y., He, K., et al. Masked autoencoders as spatiotemporal learners. In *Advances in Neural Information Processing Systems*, 2022.

Fuchs, F., Worrall, D., Fischer, V., and Welling, M. Se (3)-transformers: 3d roto-translation equivariant attention networks. In *Advances in Neural Information Processing Systems*, 2020.

Gall, T. L., Romero, P. R., Cortese, M. S., Uversky, V. N., and Dunker, A. K. Intrinsic disorder in the protein data bank. *Journal of Biomolecular structure and dynamics*, 24(4):325–341, 2007.

Ganea, O.-E., Huang, X., Bunne, C., Bian, Y., Barzilay, R., Jaakkola, T. S., and Krause, A. Independent SE(3)-equivariant models for end-to-end rigid protein docking. In *International Conference on Learning Representations*. PMLR, 2022.

Gligorijević, V., Renfrew, P. D., Kosciolek, T., Leman, J. K., Berenberg, D., Vatanen, T., Chandler, C., Taylor, B. C., Fisk, I. M., Vlamakis, H., et al. Structure-based protein function prediction using graph convolutional networks. *Nature communications*, 12(1):3168, 2021.

Goyal, S., Choudhury, A. R., Raje, S., Chakaravarthy, V., Sabharwal, Y., and Verma, A. Power-bert: Accelerating bert inference via progressive word-vector elimination. In *International Conference on Machine Learning*. PMLR, 2020.

Gruver, N., Finzi, M. A., Goldblum, M., and Wilson, A. G. The lie derivative for measuring learned equivariance. In *International Conference on Learning Representations*, 2023.

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. *Science*, pp. eads0018, 2025.

He, K., Zhang, X., Ren, S., and Sun, J. Deep residual learning for image recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2016.

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

Hess, G., Jaxing, J., Svensson, E., Hagerman, D., Petersson, C., and Svensson, L. Masked autoencoders for self-supervised learning on automotive point clouds. *arXiv preprint arXiv:2207.00531*, 3(4):5, 2022.

Hooft, R. W., Vriend, G., Sander, C., and Abola, E. E. Errors in protein structures. *Nature*, 381(6580):272–272, 1996.

Hsu, C., Verkuil, R., Liu, J., Lin, Z., Hie, B., Sercu, T., Lerer, A., and Rives, A. Learning inverse folding from millions of predicted structures. In *International Conference on Machine Learning*. PMLR, 2022.

Ingraham, J., Garg, V., Barzilay, R., and Jaakkola, T. Generative models for graph-based protein design. In *Advances in Neural Information Processing Systems*, 2019.

Jamasb, A. R., Morehead, A., Joshi, C. K., Zhang, Z., Didi, K., Mathis, S., Harris, C., Tang, J., Cheng, J., and Liò, Pietro, T. B. Evaluating representation learning on the protein structure universe. In *International Conference on Learning Representations*, 2024.

Jing, B., Eismann, S., Suriana, P., Townshend, R. J. L., and Dror, R. Learning from protein structure with geometric vector perceptrons. In *International Conference on Learning Representations*, 2021.

Kaplan, J., McCandlish, S., Henighan, T., Brown, T. B., Chess, B., Child, R., Gray, S., Radford, A., Wu, J., and Amodei, D. Scaling laws for neural language models. *arXiv preprint arXiv:2001.08361*, 2020.

Kipf, T. N. and Welling, M. Semi-supervised classification with graph convolutional networks. In *International Conference on Learning Representations*, 2017.

Krizhevsky, A., Sutskever, I., and Hinton, G. E. Imagenet classification with deep convolutional neural networks. In *Advances in Neural Information Processing Systems*, volume 25, 2012.

Li, Q., Han, Z., and Wu, X.-M. Deeper insights into graph convolutional networks for semi-supervised learning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, 2018.

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

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

Liu, J., Mao, H., Chen, Z., Zhao, T., Shah, N., and Tang, J. Towards neural scaling laws on graphs. In *The Third Learning on Graphs Conference*, 2024.

Liu, Z., Mao, H., Wu, C.-Y., Feichtenhofer, C., Darrell, T., and Xie, S. A convnet for the 2020s. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022.

Loshchilov, I. and Hutter, F. Decoupled weight decay regularization. In *International Conference on Learning Representations*, 2019.

Mirza, M. J., Shin, I., Lin, W., Schriebl, A., Sun, K., Choe, J., Kozinski, M., Possegger, H., Kweon, I. S., Yoon, K.-J., et al. Mate: Masked autoencoders are online 3d test-time learners. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 16709–16718, 2023.

Naseer, M. M., Ranasinghe, K., Khan, S. H., Hayat, M., Shahbaz Khan, F., and Yang, M.-H. Intriguing properties of vision transformers. In *Advances in Neural Information Processing Systems*, 2021.

O'Boyle, N. M., Banck, M., James, C. A., Morley, C., Vandermeersch, T., and Hutchison, G. R. Open babel: An open chemical toolbox. *Journal of cheminformatics*, 3: 1–14, 2011.

Oono, K. and Suzuki, T. Graph neural networks exponentially lose expressive power for node classification. In *International Conference on Learning Representations*, 2020.

Orengo, C. A., Michie, A. D., Jones, S., Jones, D. T., Swindells, M. B., and Thornton, J. M. Cath–a hierarchic classification of protein domain structures. *Structure*, 5(8):1093–1109, 1997.

Orlando, G., Raimondi, D., Duran-Romaña, R., Moreau, Y., Schymkowitz, J., and Rousseau, F. Pyuul provides an interface between biological structures and deep learning algorithms. *Nature communications*, 13(1):961, 2022.

Pang, Y., Wang, W., Tay, F. E., Liu, W., Tian, Y., and Yuan, L. Masked autoencoders for point cloud self-supervised learning. In *European conference on computer vision*, pp. 604–621. Springer, 2022.

Park, N. and Kim, S. How do vision transformers work? In *International Conference on Learning Representations*, 2022.

Park, N., Kim, W., Heo, B., Kim, T., and Yun, S. What do self-supervised vision transformers learn? In *International Conference on Learning Representations*, 2023.

Pinheiro, P. O., Rackers, J., Kleinhenz, J., Maser, M., Mahmood, O., Watkins, A., Ra, S., Sresht, V., and Saremi, S. 3d molecule generation by denoising voxel grids. In *Advances in Neural Information Processing Systems*, 2023.

Pinheiro, P. O., Jamasb, A., Mahmood, O., Sresht, V., and Saremi, S. Structure-based drug design by denoising voxel grids. In *International Conference on Machine Learning*. PMLR, 2024.

Raghu, M., Unterthiner, T., Kornblith, S., Zhang, C., and Dosovitskiy, A. Do vision transformers see like convolutional neural networks? In *Advances in Neural Information Processing Systems*.

Ragoza, M., Hochuli, J., Idrobo, E., Sunseri, J., and Koes, D. R. Protein–ligand scoring with convolutional neural networks. *Journal of chemical information and modeling*, 57(4):942–957, 2017.

Rao, R., Meier, J., Sercu, T., Ovchinnikov, S., and Rives, A. Transformer protein language models are unsupervised structure learners. *Biorxiv*, 2020.

Rao, Y., Zhao, W., Liu, B., Lu, J., Zhou, J., and Hsieh, C.-J. Dynamicvit: Efficient vision transformers with dynamic token sparsification. In *Advances in Neural Information Processing Systems*, 2021.

Read, R. J., Adams, P. D., Arendall, W. B., Brunger, A. T., Emsley, P., Joosten, R. P., Kleywegt, G. J., Krissinel, E. B., Lütteke, T., Otwinowski, Z., et al. A new generation of crystallographic validation tools for the protein data bank. *Structure*, 19(10):1395–1412, 2011.

Rives, A., Meier, J., Sercu, T., Goyal, S., Lin, Z., Liu, J., Guo, D., Ott, M., Zitnick, C. L., Ma, J., et al. Biological structure and function emerge from scaling unsupervised learning to 250 million protein sequences. *Proceedings of the National Academy of Sciences*, 118(15):e2016239118, 2021.

Rojas-Gomez, R. A., Lim, T.-Y., Do, M. N., and Yeh, R. A. Making vision transformers truly shift-equivariant. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 5568–5577, 2024.

Ronneberger, O., Fischer, P., and Brox, T. U-net: Convolutional networks for biomedical image segmentation. In *Medical image computing and computer-assisted intervention–MICCAI 2015: 18th international conference, Munich, Germany, October 5-9, 2015, proceedings, part III 18*, pp. 234–241. Springer, 2015.

Satorras, V. G., Hoogeboom, E., and Welling, M. E (n) equivariant graph neural networks. In *International Conference on Machine Learning*. PMLR.

Shwartz-Ziv, R., Goldblum, M., Li, Y., Bruss, C. B., and Wilson, A. G. Simplifying neural network training under class imbalance. In *Advances in Neural Information Processing Systems*, 2023.

Sillitoe, I., Dawson, N., Lewis, T. E., Das, S., Lees, J. G., Ashford, P., Tolulope, A., Scholes, H. M., Senatorov, I., Bujan, A., et al. Cath: expanding the horizons of structure-based functional annotations for genome sequences. *Nucleic acids research*, 47(D1):D280–D284, 2019.

Steinegger, M. and Söding, J. Mmseqs2 enables sensitive protein sequence searching for the analysis of massive data sets. *Nature biotechnology*, 35(11):1026–1028, 2017.

Tay, Y., Dehghani, M., Abnar, S., Chung, H. W., Fedus, W., Rao, J., Narang, S., Tran, V. Q., Yogatama, D., and Metzler, D. Scaling laws vs model architectures: How does inductive bias influence scaling? In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 2023.

Tong, Z., Song, Y., Wang, J., and Wang, L. Videomae: Masked autoencoders are data-efficient learners for self-supervised video pre-training. In *Advances in Neural Information Processing Systems*, 2022.

Varadi, M., Anyango, S., Deshpande, M., Nair, S., Natassia, C., Yordanova, G., Yuan, D., Stroe, O., Wood, G., Laydon, A., et al. Alphafold protein structure database: massively expanding the structural coverage of protein-sequence space with high-accuracy models. *Nucleic acids research*, 50(D1):D439–D444, 2022.

Varadi, M., Bertoni, D., Magana, P., Paramval, U., Pidruchna, I., Radhakrishnan, M., Tsenkov, M., Nair, S., Mirdita, M., Yeo, J., et al. Alphafold protein structure database in 2024: providing structure coverage for over 214 million protein sequences. *Nucleic acids research*, 52(D1):D368–D375, 2024.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł., and Polosukhin, I. Attention is all you need. In *Advances in Neural Information Processing Systems*, 2017.

Watson, J. L., Juergens, D., Bennett, N. R., Trippe, B. L., Yim, J., Eisenach, H. E., Ahern, W., Borst, A. J., Ragotte, R. J., Milles, L. F., et al. De novo design of protein structure and function with rfdiffusion. *Nature*, 620(7976): 1089–1100, 2023.

Wu, Z., Jain, P., Wright, M., Mirhoseini, A., Gonzalez, J. E., and Stoica, I. Representing long-range context for graph neural networks with global attention. In *Advances in Neural Information Processing Systems*, 2021.

Xia, Y., Xia, C.-Q., Pan, X., and Shen, H.-B. Graphbind: protein structural context embedded rules learned by hierarchical graph neural networks for recognizing nucleic-acid-binding residues. *Nucleic acids research*, 49(9): e51–e51, 2021.

Xie, E., Wang, W., Yu, Z., Anandkumar, A., Alvarez, J. M., and Luo, P. Segformer: Simple and efficient design for semantic segmentation with transformers. In *Advances in Neural Information Processing Systems*, 2021.

Yuan, M., Shen, A., Fu, K., Guan, J., Ma, Y., Qiao, Q., and Wang, M. Proteinmae: masked autoencoder for protein surface self-supervised learning. *Bioinformatics*, 39(12): btad724, 2023.

Zhai, X., Kolesnikov, A., Houlsby, N., and Beyer, L. Scaling vision transformers. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022.

Zhang, Z., Xu, M., Jamasb, A., Chenthamarakshan, V., Lozano, A., Das, P., and Tang, J. Protein representation learning by geometric structure pretraining. In *International Conference on Learning Representations*, 2023.

# A. Related Work

Prior work to obtain protein representations primarily focused on finding appropriate inductive biases. Graph and equivariant networks, while encoding rich priors, ultimately struggle to scale; conversely, language models scale effectively by leveraging amino acid sequences but overlook 3-D structural details, particularly local geometry. Although voxel approaches show promise, they were frequently overlooked due to their resource requirements; voxel CNNs can capture atoms but have lacked strong pre-training methods. AtoMAE unifies these approaches by marrying atom type voxel inputs with a weak-bias Transformer and a masked-reconstruction objective for pre-training, thus enabling data-driven discovery of protein geometry while maintaining scalability.

## A.1. Graph-based protein structure modeling

Graph neural networks (GNNs) have emerged as the dominant paradigm for structure-aware protein modeling (Ingraham et al., 2019; Jing et al., 2021; Gligorijević et al., 2021; Xia et al., 2021; Zhang et al., 2023). Researchers have widely demonstrated that residues can naturally serve as graph nodes—enriched with coordinates, amino-acid identity, and physicochemical descriptors—while covalent bonds or distance-based contacts define edges (Ingraham et al., 2019; Gligorijević et al., 2021). Early successes, including the design-graph model of Ingraham et al. (2019), GVP-GNN's rotation-aware vector features (Jing et al., 2021), and structure-GNNs for function and binding-site prediction (Gligorijević et al., 2021; Xia et al., 2021), demonstrated that message passing can capture local geometry and chemical context with modest computational cost. Follow-on systems such as GearNet introduce contrastive learning or reconstruction-based pre-training methods to capture detailed information from protein structures and accelerate fine-tuning, cementing graph-based methods as a practical, biologically intuitive choice for many protein tasks (Zhang et al., 2023). Jamasb et al. (2024) benchmark GNN efforts in a unified scenario, providing comprehensive evaluation frameworks for the field.

Despite these merits, standard message-passing graphs struggle to propagate long-range information. As depth increases, oversmoothing and bottleneck effects erode expressive power, causing accuracy to plateau or even decline while memory and latency grow sharply (Li et al., 2018; Oono & Suzuki, 2020; Alon & Yahav, 2021). Sophisticated variants like GearNet mitigate but do not eliminate these issues (Zhang et al., 2023). These scaling limits, coupled with the need to hand-design edge lists and interaction radii, have motivated newer methods to capture both local and global geometric properties of proteins (Dwivedi et al., 2022; Wu et al., 2021; Liu et al., 2024).

## A.2. SE(3)-equivariant neural networks

Equivariant models incorporate 3-D symmetry directly into their layers so outputs move in step with rotated or translated inputs. The SE(3)-Transformer adds this property to self-attention by using steerable tensor features, yielding frame-invariant performance on point-cloud and molecular tasks (Fuchs et al., 2020). The E(n)-GNN maintains the same rotational and translational equivariance without heavy tensor mathematics, achieving strong results on particle-dynamics and molecular benchmarks with lower computational cost (Satorras et al.). Extensions to rigid-body docking and ligand reconstruction further underscore the benefits of explicit symmetry (Ganea et al., 2022; Bhadra-Lobo et al., 2023). These works show that enforcing geometric symmetry can reduce sample complexity and boost robustness, though their specialized operations still add overhead compared with isotropic Transformer backbones that approximate the same invariance through random-rotation augmentation (Jing et al., 2021; Zhang et al., 2023).

## A.3. Protein language modeling

Transformer language models trained on raw amino-acid sequences treat proteins as sentences, allowing them to internalize co-evolutionary patterns and "grammatical" rules of folding without any 3-D supervision (Rao et al., 2020; Rives et al., 2021). Sequence-only systems such as ESM-1b (Rives et al., 2021) and ESM-2 (Lin et al., 2023b) scaled to billions of parameters and hundreds of millions of sequences—learn residue contacts well enough to solve annotation tasks in a zero-shot setting (Lin et al., 2023b) and, when augmented with lightweight structure heads (e.g., ESMFold), can predict atomic coordinates directly from a single sequence (Lin et al., 2023b). Yet because these models see only 1-D inputs, they still lack fine-grained geometric context and often lag behind structure-aware architectures on fold-classification or design benchmarks, motivating hybrid or explicitly 3-D approaches (Zhang et al., 2023; Dauparas et al., 2022).

### A.4. Voxel-based 3-D representations

These methods "rasterize" atomic coordinates into a regular Cartesian grid: each voxel stores occupancy or density for a handful of atom channels, allowing generic vision backbones to process a protein structure as a 3-D image (Ragoza et al., 2017; Derevyanko et al., 2018). Early 3-D CNN systems for binding-site scoring and fold-quality assessment pioneered this volumetric view of pockets and whole proteins (Ragoza et al., 2017; Derevyanko et al., 2018). Compared with graph models, the grid approach preserves full atomic detail and offers a wide receptive field that naturally captures long-range context, and recent diffusion-style generators demonstrate that the same voxel format scales to molecular design tasks (Pinheiro et al., 2023; 2024). The main drawbacks remain memory overhead and the absence of built-in SE(3) equivariance, but token-selection and data-augmentation strategies outlined in our work partially offset these limitations.

## B. Hyperparameter

Table 1 enumerates the key hyper-parameters to pre-train neural networks. Training is executed on 8 NVIDIA A100 GPUs.

Table 1: Hyper-parameters for pre-training.

| | |
|---|---|
| *Training* | |
| Batch size | 32 |
| Total epochs | 800 |
| Warm-up epochs | 10 |
| *Optimizer & LR schedule* | |
| Optimizer | AdamW |
| Learning rate | $1.0 \times 10^{-2}$ |
| Weight decay | $7.0 \times 10^{-2}$ |
| $\beta$-values | $(0.9,\ 0.95)$ |
| *Model* | |
| Input voxel grid | $96^3$ |
| Voxel kernel / stride | $8/8$ |
| Mask ratio | 90% |
| Loss function | BCE + channel weights |
| *Data & augmentation* | |
| Voxel channels (#) | 3 (C, N, O) |
| Resolution (Å) | 0.5 |
| Coord. noise (std. Å) | 0.2 |

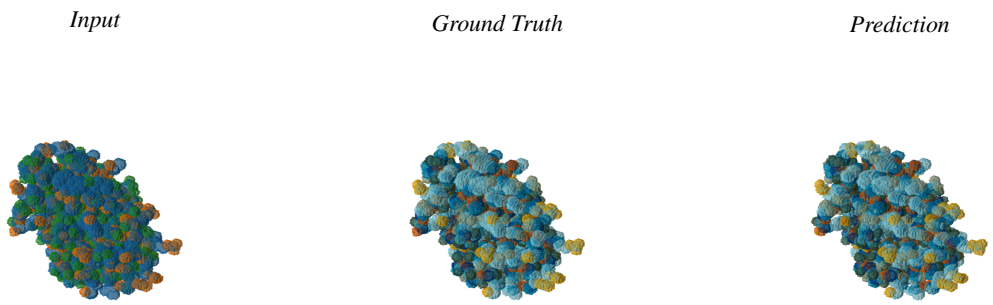
## C. Structural Pattern Identification

AtoMAE demonstrates the ability to learn patterns from protein structures during pre-training, including atom types and residue identification. To evaluate whether the pre-trained model understands protein structural features, we first examine if the model can infer coarse-grained atom-type categories from atomic inputs alone. Second, assessing its capacity to extract biological meaning by inferring amino acid identities from atomic arrangements.

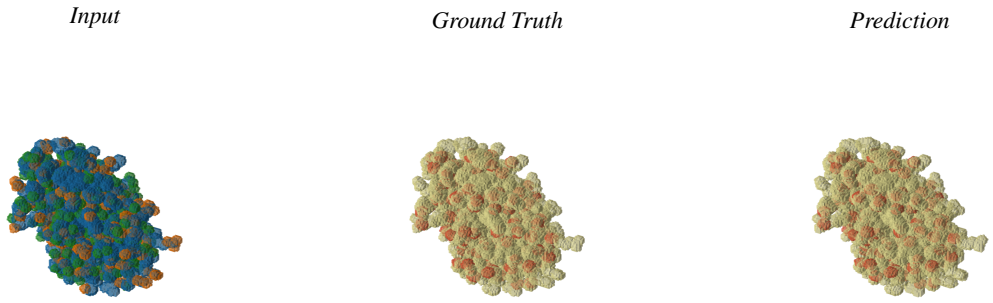
### C.1. coarse-grained atom-type category identification

We use eight coarse-grained atom-type categories: alpha carbon, beta carbon, backbone carbon, sidechain carbon, backbone nitrogen, sidechain nitrogen, backbone oxygen, and sidechain oxygen. These categories capture essential structural and chemical distinctions in protein architecture.

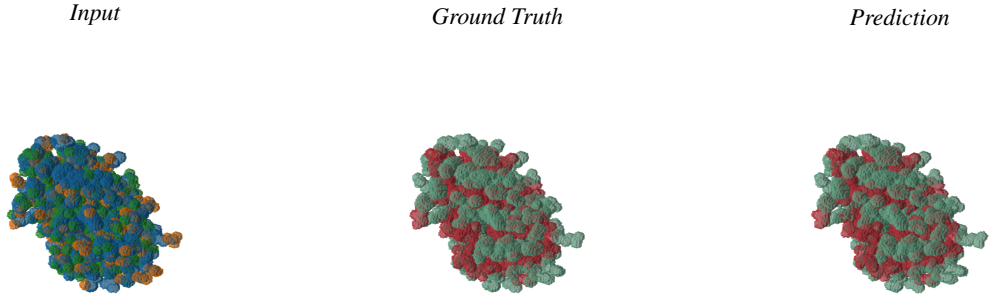
The experiment in this section follows the semantic segmentation approach described in Section 4, however, we freeze the pre-trained encoder and train only a streightforward shallow 4-layer Transformer decoder instead of U-Net or other semantic segmentation specialized decoders. This design choice ensures that structural understanding emerges from the pre-trained




(a) Eight coarse-grained atom-type categories




(b) $C_\alpha$ (red) and others (yellow)




(c) Backbone (red) and sidechain (green)

Figure 6: **Pre-trained AtoMAE can capture structural patterns in protein structures in a zero-shot manner.** The first column displays the input voxelized protein structure (with blue, yellow and green represent carbon, nitrogen, and oxygen, respectively), the second column presents the ground truth, and the final column shows the prediction. (a) We use a pre-trained AtoMAE with a 4-layer Transformer decoder to predict eight coarse-grained atom type categories. Here, the pre-trained backbone is frozen and only a thin decoder is trained to assess how much structural information the backbone captured during pre-training. The figure shows that the ground truth closely resembles the prediction, which suggests that the backbone possesses such capabilities.

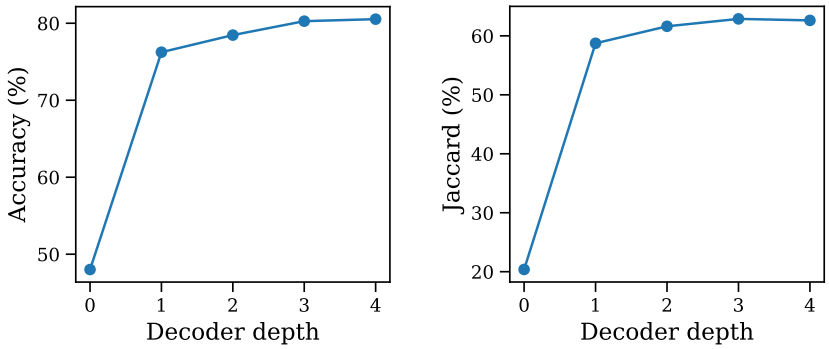



Figure 7: **Shallow Transformer decoders are enough for structure identification.** To assess the impact of architectural depth on structural pattern identification across the 8 atomic categories, we very the Transformer decoder depth from 0 (linear probing) through 4 layers, evaluating performance in terms of accuracy and Jaccard similarity metrics. The results reveal that a single Transformer layer yields substantial performance gains, while 4 decoder layers reach the performance plateau.

representations rather than being learned during fine-tuning. The decoder in the framework primarily serves to reconstruct fine-grained, high-resolution representations from coarse patch-level embeddings.

**Quantitative results: Shallow Transformer decoders are enough for structure identification.** We systematically vary decoder depth from 0 (linear layer) to 4 layers while maintaining consistent encoder architecture. Performance is evaluated on the atom-type classification task described above, providing a direct measure of how decoder capacity affects structural pattern recovery. Figure 7 shows that adding even a single decoder layer results in dramatic accuracy improvements compared to direct classification from encoder outputs. This suggests that the patch-level representations contain rich structural information that requires minimal processing to extract atom-level predictions. Performance reaches near-optimal levels at the accuracy of 79% with just 2 decoder layers, and further increases beyond 4 layers show marginal gains. This saturation behavior indicates that shallow Transformer decoders are sufficient for structure identification tasks, consistent with the relatively local nature of atom-type prediction.

**Qualitative results: AtoMAE understands protein structurs.** The qualitative results demonstrate consistent and interpretable patterns in the model's predictions. The pre-trained AtoMAE model shows clear separation between the eight atom-type classes in Figure 6a, with visually coherent spatial clustering that aligns with chemical expectations. For improved interpretability, we also provide separate visualizations focusing on alpha carbon atoms in Figure 6b and backbone/sidechain distinctions in Figure 6c. These targeted visualizations more clearly reveal the structural trends captured by the model, demonstrating that AtoMAE develops a hierarchical understanding of protein architecture from fine-grained atomic details to broader structural motifs.

### C.2. Residue identification for rescuing AI-generated proteins

Beyond distinguishing coarse structural elements such as $C_\alpha$ and backbone atoms, a fine-tuned AtoMAE can full residue identities directly from voxelised atomic protein structures. Because the architecture consumes only raw atom types and 3-D positions, it remains agnostic to missing or misplaced atoms that frequently plague *de novo* structures designed by generative AI pipelines. This section demonstrates how such a model can rescue noisy all-atom designs that defeat rule-based parsers (e.g. `OpenBabel` (O'Boyle et al., 2011)) and sequence-only language models.

**Experimental set-up.** Starting from high-resolution experimentally obtained proteins, we inject 0.4Å Gaussian coordinate noise and random atom deletions or duplications with 4% probability. These corruptions alter about 60% of all residues, mimicking the idiosyncratic artefacts observed in *de novo* all atom AI-generated protein structures.

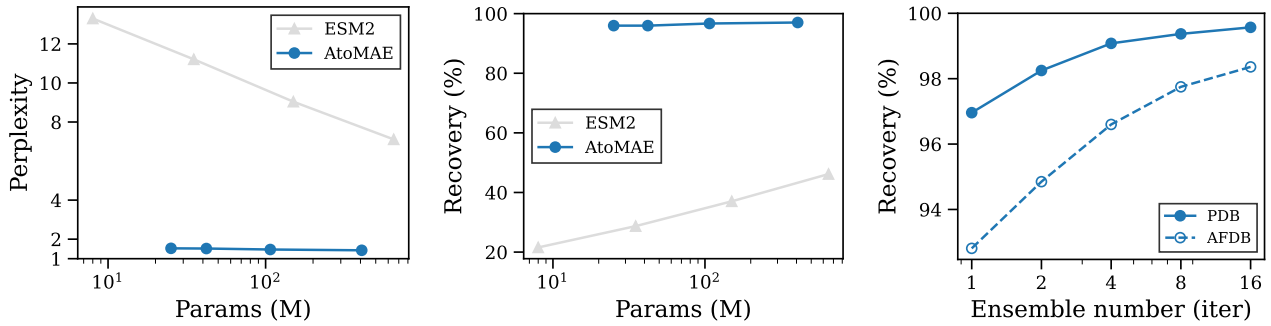



Figure 8: **AtoMAE demonstrates the capability to rescue all-atom AI-generated proteins suffering from atomic loss, duplication, and coordinate noise.** (a, b) To evaluate the capacity for converting corrupted AI-generated proteins back to their original sequences, we design a controlled scenario involving deliberate insertion of atomic loss, duplication, and coordinate noise into protein structures, followed by sequence recovery. Despite these substantial perturbations, AtoMAE achieves highly accurate sequence recovery by effectively leveraging structural information. In contrast, ESM2 exhibits considerably lower recovery rates even under significantly more lenient conditions, reflecting its reliance solely on sequence contextual information. (c) AtoMAE employs test-time augmentation strategies, incorporating random cropping and rotation, to effectively process very large proteins. The inference performance exhibits dramatic improvement with increasing ensemble size.

**Structure-aware models outperform sequence-only baselines.** Despite the severe perturbations, the fine-tuned AtoMAE achieves 99 % residue accuracy as shown in Figure 8, vastly exceeding the ∼40 % accuracy obtained by an ESM2 language model that restores 10% masked tokens using only sequence context. The result underscores the advantage of exploiting explicit geometric information when rescuing AI-designed proteins. These results position AtoMAE as a practical drop-in module for sanitising AI-generated proteins, enabling downstream pipelines (e.g. sequence design or functional screening) to operate without costly manual curation.

**Test-time augmentation (TTA) boosts the performance.** Large proteins sometimes exceed the receptive field of a single voxel unit ($48^3$Å$^3$). We therefore apply random crops and rotations at inference, and aggregate results across views. As shown in Appendix C.1, we observe a monotonic improvement up to an ensemble size of 16. Balancing compute and accuracy, we adopt an ensemble of 8 crops in subsequent experiments.

## D. Limitations and Future Work

While the proposed atomic voxel-based approach demonstrates significant potential, several limitations present opportunities for future development.

**Limitations.** A fundamental limitation lies in our approach's structural dependency. Unlike sequence-based protein language models, our method requires three-dimensional structural information as input. This constraint is challenging given the substantial data requirements for large parameter models and the relative scarcity of experimentally determined protein structures compared to available sequences. While we limited biological priors in our current work to establish a baseline, integrating domain knowledge and developing novel training techniques offers a promising direction for addressing structural data scarcity.

**Future work.** We aim to verify whether these favorable scaling trends persist in substantially higher parameter regimes. Additionally, we will assess the approach's generalizability to computationally predicted structural datasets such as AlphaFoldDB (Varadi et al., 2022; 2024), which could dramatically expand available training data points. Finally, we will evaluate our voxel featurization method across diverse downstream tasks to comprehensively benchmark it against established sequence-based and graph-based approaches.