A GRAPH NEURAL NETWORK APPROACH TO AUTOMATED MODEL BUILDING IN CRYO-EM MAPS

Anonymous authors
Paper under double-blind review

ABSTRACT

Electron cryo-microscopy (cryo-EM) produces three-dimensional (3D) maps of the electrostatic potential of biological macromolecules, including proteins. Along with knowledge about the imaged molecules, cryo-EM maps allow *de novo* atomic modelling, which is typically done through a laborious manual process. Taking inspiration from recent advances in machine learning applications to protein structure prediction, we propose a graph neural network (GNN) approach for automated model building of proteins in cryo-EM maps. The GNN acts on a graph with nodes assigned to individual amino acids and edges representing the protein chain. Combining information from the voxel-based cryo-EM data, the amino acid sequence data and prior knowledge about protein geometries, the GNN refines the geometry of the protein chain and classifies the amino acids for each of its nodes. Application to 28 test cases shows that our approach outperforms the state-of-the-art and approximates manual building for cryo-EM maps with resolutions better than 3.5 Å.

1 INTRODUCTION

Following rapid developments in microscopy hardware and image processing software, cryo-EM structure determination of biological macromolecules is now possible to atomic resolution for favourable samples (Nakane et al., 2020; Yip et al., 2020). For many other samples, such as large multi-component complexes and membrane proteins, resolutions around 3 Å are typical (Cheng, 2018). Transmission electron microscopy images are taken of many copies of the same molecules, which are frozen in a thin layer of vitreous ice. Dedicated softwares, like RELION (Scheres, 2012) or cryoSPARC (Punjani et al., 2017), implement iterative optimization algorithms to retrieve the orientation of each molecule, and perform 3D reconstruction to obtain a voxel-based map of the underlying molecular structure.

Provided the cryo-EM map is of sufficient resolution, it is interpreted in terms of an atomic model of the corresponding molecules. Many samples contain only proteins; other samples also contain other biological molecules, like lipids or nucleic acids. Proteins are linear chains of amino acids, or residues. There are twenty different amino acids that make up proteins. All of these amino acids have four (non-hydrogen) atoms that make up the protein main chain. The different amino acids have different numbers, types and geometrical arrangements of their side-chain atoms. The smallest amino acid, glycine, has no side chain atoms; the largest amino acid, tryptophan, has ten side chain atoms. Typical proteins range in size from tens to more than a thousand residues. Typically, the electron microscopist knows which protein sequences are present in the sample. The task at hand is to build the atomic model, which identifies the positions of all atoms for all proteins that are present in the cryo-EM map. For each residue, there are two rotational degrees of freedom in the conformation of its main chain. Distinct orientations of the side chains provide additional conformational possibilities, the number of which depends on the type of amino acid (figure 1).

Atomic model building in cryo-EM maps is typically done manually using 3D visualisation software, (e.g. Emsley et al., 2010; Pettersen et al., 2021), followed by refinement procedures that optimize the fit of the models in the map, (e.g. Murshudov et al., 2011; Croll, 2018; Liebschner et al., 2019). Often, in areas of weak density in the map, one cannot discern the amino acid identity of residues from the map alone and sequence information has to be used to make an accurate assessment. Manually building a reliable atomic model *de novo* in the reconstructed cryo-EM map is considered to

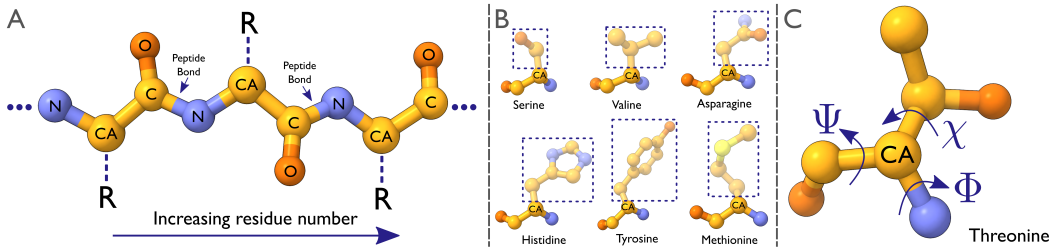

Figure 1: **(A)** shows a backbone where the chain is ordered from left to right. Arrows mark the peptide bonds between the C atom of one residue and the N atom of the next. **(B)** shows six amino-acids, pointing out the differences in the side chains, marked by the outline. **(C)** shows the Φ and Ψ angles of the backbone and the additional rotatable bond of the side chain for threonine.

be difficult for maps with resolutions worse than 4 Å. Although the task is more straightforward for maps with resolutions better than 3 Å, it still typically requires large amounts of time and a high level of expertise.

Machine learning has recently achieved a major step forward in structure prediction for individual proteins (Jumper et al., 2021; Baek et al., 2021). In these approaches, the sequence information of proteins and their evolutionary related homologues is used to predict their atomic structure without the use of experimental data. In addition, protein language models, which are trained in an unsupervised fashion on the amino acid sequences of many proteins, have also provided useful results in protein structure prediction (Lin et al., 2022; Wu et al., 2022a). Although these techniques are not yet capable of predicting structures of the larger complexes that are typically studied by cryo-EM, their success for individual proteins inspired us to explore similar approaches for automated model building in cryo-EM maps.

In this paper, we present a single integrated GNN that combines the voxel-based information from the cryo-EM map with information from the protein sequence through a protein language model, and information from the topology of the graph through invariant point attention (IPA) (Jumper et al., 2021). For 28 test cases, we demonstrate that our approach approximates the accuracy of manual model building for maps with resolutions better than 3.5 Å.

2 Prior Work

Automated model building. Automated approaches for atomic modelling in the related experimental technique of X-ray crystallography have existed for many years (for example, Perrakis et al., 1999; Cowtan, 2006; Terwilliger et al., 2008). Some of these approaches have also been applied to cryo-EM maps. For example, the PHENIX package builds models that are on average 47% complete for cryo-EM maps with resolutions worse than 3 Å (Terwilliger et al., 2018). For similar maps, MAINMAST, an approach that was designed to build C^{α} main-chain traces in cryo-EM maps, often produces models with root means square deviations (RMSDs) in the range of tens of Å (Terashi & Kihara, 2018). Relatively incomplete models, with large residuals, have limited the impact of these techniques on automated model building in the cryo-EM field thus far.

More recently, Deeptracer (Pfab et al., 2021), the first deep learning approach for automated atomic modelling in cryo-EM maps, was reported to outperform these earlier approaches. Deeptracer uses a U-Net (Ronneberger et al., 2015) to construct an atomic model *de novo* in the cryo-EM map. In contrast to our work, Deeptracer does not integrate the sequence information with the U-Net, and it does not use a graph representation of the protein chain during model refinement. Instead, Deeptracer treats the entire problem as a segmentation and classification problem. Thereby, it also does not have support for refining already built models or performing multiple recycling steps. Although Deeptracer predicts amino acid types for each residue, it only builds atoms for the main chains.

There have also been reports to dock and morph the output from protein structure prediction programs, like AlphaFold2 (Jumper et al., 2021), to fit cryo-EM maps (He et al., 2022; Terwilliger et al., 2022). Such approaches suffer when proteins change their conformation in complex with others, and are likely to propagate errors in the structure prediction. Thus, it seems sensible to design a neu-

ral network approach that integrates both the cryo-EM map and the sequence and protein structure primitives to produce a more reliable structure. This is essence of our approach.

GNNs for proteins. A number of different approaches to modelling proteins with GNNs have been proposed recently. This includes modelling the protein with torsion angles (Jing et al., 2022), SE(3) equivariant graph neural networks (Ganea et al., 2022), and SE(3) invariant GNNs (Dauparas et al., 2022). In our approach, the ordering and connectivity of residues are unknown and have to be inferred, hence representations that require this to be known *a priori*, such as the torsion angle representation, are inappropriate. Furthermore, the relative orientation of the model and the cryo-EM map is important in model building, which makes SE(3) invariant representations ill-suited. Thus, the most natural graph representation is an SE(3) equivariant one. In this project, we choose the backbone frame representation that was first introduced in AlphaFold2, but is now also used in other protein prediction networks (e.g. Wu et al., 2022a; Lin et al., 2022).

3 METHODS

3.1 GRAPH INITIALIZATION

We start by identifying the positions of the C^{α} atoms [1] of all residues in the map, which will form the nodes of our graph. This part of the pipeline is formulated as a straight-forward segmentation problem, similar to prior work discussed in section 2. That is, the cryo-EM map $V \in \mathbb{R}^N$, where N is the number of voxels, has an associated binary target $T^* \in \{0,1\}^N$, where 1 represents the existence of a C^{α} atom in the voxel and 0 the lack of it. Since the minimum distance of two C^{α} atoms is 3.8 Å, resampling the cryo-EM voxel maps with a pixel size of 1.5 Å ensures that there is no voxel that contains more than one such atom. The goal then becomes to train a neural network $f_\theta(V) \approx T^*$.

3.1.1 NETWORK ARCHITECTURE

We implemented $f_\theta(V)$ as a residual network (He et al., 2016) inspired by the Feature Pyramid Network (Lin et al., 2017a). First, we changed all convolutions to 3D convolutions and changed batch normalization layers (Ioffe & Szegedy, 2015) to instance normalization layers (Ulyanov et al., 2016), since local boxes of cryo-EM density might be from different sections of the map with large differences in scale, so the statistics should not be averaged across instances of a batch. No noticeable effects on performance was observed between ReLU and LeakyReLU (Xu et al., 2015), so ReLU was selected due to its improved computational efficiency. We also shifted the network parameters from the low resolution part of the model to the high resolution part, and we changed the order of operations, so that global information about structure is more directly integrated into the model at an early stage (for detailed pseudo-code that contains the exact number of parameters and order of operations, see algorithm 1 in A.3.1). In this manner, large scale structural features, which are recognized at lower resolutions, become easier to integrate with the classification task.

3.1.2 TRAINING DATASET

The dataset starts with 6351 cryo-EM maps with resolution higher than 4 Å, downloaded from the EMDB (Lawson et al., 2016) on 01/04/2022, and the corresponding PDB files that were downloaded from RCSB (Burley et al., 2021). A portion of these were manually checked for orientation issues, large unmodelled regions, and existence of large modelled regions that do not correspond to the cryo-EM map. This led to $\sim$700 manually curated pairs for the first round of training. Then, using the first trained model, we were able to automatically detect issues with the rest of the map-structure pairs and prune them to $\sim$3200 structures. The pruning was based on a cutoff of 70% precision and 70% recall of the model output C^{α} positions compared to the ground-truth PDB coordinates. This produced the dataset that was used for training the model.

[1] Atom names in this document follow the PDB (Protein Data Bank) naming convention. See the PDB atomic coordinate and bibliographic entry format description

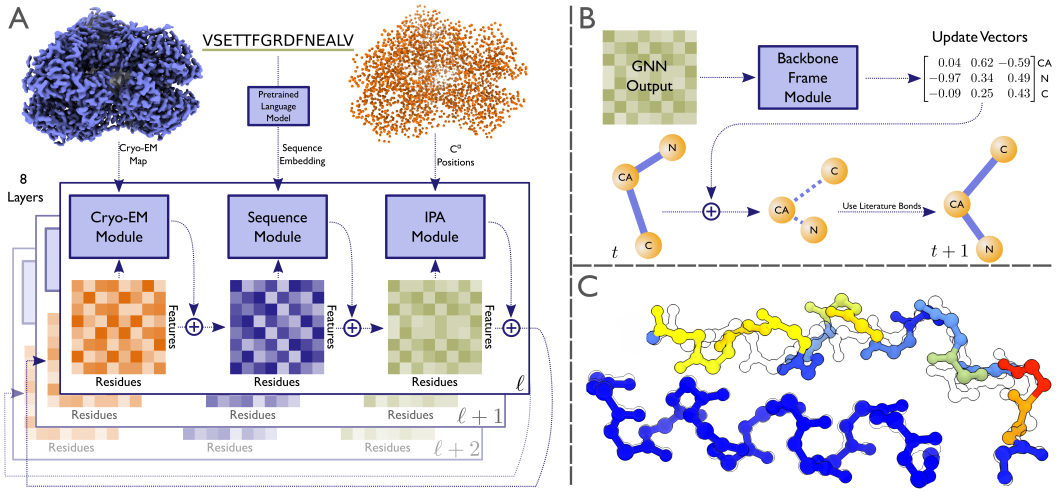

Figure 2: **(A)** shows the schematic of the GNN and how the 8 layers iteratively refine the feature vectors. **(B)** illustrates how the Backbone Frame module updates the positions of the backbone atoms. Finally, **(C)** contains two examples of high confidence (dark blue colour) and low confidence (yellow and red) predicted backbone regions for PDB entry 7Z1M. The confidence measure is a good predictor of fit to the deposited backbone model (shown in outline).

3.1.3 Training

We define the true number of the C^{α} atoms, or residues, in a map as M^*. To ameliorate the issue with class imbalance between the number of C^{α} atoms and the number of voxels ($M^* \ll N$), we chose focal loss (Lin et al., 2017b) over binary cross entropy loss. Additionally, we up-weighted the loss of voxels with C^{α} atoms and used a combination of auxiliary loss functions to get an acceptable trade-off between precision and recall. We up-weight the C^{α} containing voxels in the focal loss with the ratio between the true negatives and true positives, using the following formula:

$$w_x = \frac{N - M^*}{M^*}\chi(x) + (1 - \chi(x)) \tag{1}$$

where w_x is the weight for voxel x and χ is the characteristic function of whether the voxel x contains a C^{α} atom. This formula ensures that half of the loss of a map comes from empty voxels and half from the C^{α} atom containing voxels. We found that also using Tversky loss (Hashemi et al., 2018) in the later parts of training allowed higher recall at the expense of precision. Lastly, to improve generalization to variations in experimental and data acquisition settings, we applied data augmentation schemes to the cryo-EM maps. We added random colored noise to every map, and performed random sharpening/dampening by sampling a B-factor (see Rosenthal & Henderson, 2003) from a uniform distribution between -30 Å^2 and 30 Å^2. Lastly, we applied random rotations that are integer multiples of 90° to both the cryo-EM map and the targets in order to have the model learn different orientations, while avoiding interpolation effects.

3.2 Graph Refinement

Next, we define a GNN g_ϕ that is trained to refine the position of all C^{α} atoms in the graph, to map each of them to an individual residue in the user-provided input sequence, and to provide coordinates for all the atoms in the residues:

$$g_\phi\left(X^{(n-1)}, F^{(n-1)}, V, S\right) = \left(X^{(n)}, F^{(n)}, G^{(n)}, P^{(n)}, O^{(n)}\right) \tag{2}$$

In the above, $X^{(n)} \in \mathbb{R}^{M\times 3}$ are the C^{α} positions at iteration n, $F \in \mathbb{R}^{M\times 3\times 4}$ are the affine frames defined by the C^{α}, C, and N atoms of the backbone, $G \in \mathbb{R}^{M\times 4\times 2}$ are the torsion angles of the side chains, $P \in \mathbb{R}^{M\times 20}$ is a probability vector for all twenty amino-acids for each residue, and $S \in \mathbb{R}^{M\times 1280}$ are protein sequence embeddings of all residues in the input sequence, and $O \in \mathbb{R}^{M}$ is a per-residue confidence prediction. The backbone affine frames (F) and the side-chain torsion

angles (G) are similar to what is used in AlphaFold2 (Jumper et al., 2021), see figure 2. Together (X, F, G, P) provide enough information to calculate all-atom coordinates for the proteins in the cryo-EM map.

The overall training objective is to acquire $g_\phi\left(X^{(n)}, F^{(n)}, V, S\right) \approx (X^*, F^*, G^*, P^*)$, where X^* is the set of C^α positions in the training data, F^* and G^* are calculated from the atom coordinates for every residue in the training data, and $P^* \in \{0,1\}^{M\times 20}$ is a one-hot encoding of the amino acid classes in the training data. At iteration $n = 0$, the graph is initialized with M nodes with random $F^{(0)}$, and with $X^{(0)}$ set to the coordinates of those voxels predicted to contain C^α atoms by the graph initialization step in section 3.1.

3.2.1 Network Architecture

The GNN consists of eight consecutive layers, each of which contains three main modules that are based on the attention algorithm (Bahdanau et al., 2014). Each module extracts information relevant to its modality, and updates the feature encoding of the graph nodes with a residual connection, (see figure 2A and algorithm 9 in A.4.2).

The first module is the *Cryo-EM Attention module*, which allows the GNN to look at the density around each C^α atom with convolutional neural networks (CNN), as well as the density linking it to its neighbouring nodes, to update its representation. This is accomplished with a mix of a graph-based attention module and feature extraction using CNNs. The edge features that determine the attention keys are calculated based on CNN embeddings of cuboids (elongated cubes in the direction of the neighbour) extracted from the cryo-EM map spanning each edge that connect a node with each of its neighbours. These regions of the cryo-EM map capture the connectivity between residues, e.g. peptide bond or side-chain:side-chain interactions, and they inform the attention scores. The query and value vectors are generated by the feature vector of each node. Additionally, a cube centered at each node is extracted from the cryo-EM density with an orientation defined by the backbone affine frame of each node and passed through a convolutional feature extraction network. Finally, the extracted features from the centered cubes are concatenated with the attention output and projected to create the new feature representation. For precise details about the algorithm, see algorithm 10 in A.4.2.

The second module is the *Sequence Attention module*, where each node searches against the input sequence embedding to find the relevant entries that best fit its features. This is a conventional encoder-only transformer module, similar to that used in Devlin et al. (2018). The user-provided sequence is embedded using a pre-trained protein language model (ESM-1b) (Rives et al., 2021; Lin et al., 2022), which only use the primary sequence, instead of multiple sequence alignments (MSAs). Generally, MSAs improve results of protein prediction algorithms by giving them access to co-evolutionary information (Jumper et al., 2021), and they need less learnable parameters than algorithms that rely on a single sequence only (Rao et al., 2021; Lin et al., 2022; Wu et al., 2022b). However, because the cryo-EM map already provides sufficient information about the global fold of the proteins, we chose not to use MSAs, as their calculation would have made using our approach more difficult. For additional details about this module, see algorithm 11 in A.4.2.

The third module is the *Spatial Invariant Point Attention* (IPA) module, which allows the network to update its representation based on the geometry of the nodes in the graph. This module is inspired by the module with the same name in AlphaFold2 (Jumper et al., 2021), although simplifications have been made to better fit the problem at hand. Each node predicts query points based on its current representation in its own local affine frame; these points are transformed into the global affine frame (by application of F to the predicted point); the distance is calculated between each node's query points and its neighbours; and based on the sum of the distances of the query points to the neighbouring nodes, each of these nodes get an attention score that is used to update the central node's representation (see algorithm 12 in A.4.2). Essentially, this module queries parts of the graph where it expects specific nodes to be and then uses the distance of the neighbouring nodes to its query point to collect information from the other nodes.

Since the three main modules are applied sequentially in eight layers, the representations from each module allow the other modules to gradually extract more information from their inputs. For example, using the cryo-EM density, the network is able to find a better orientation for its backbone as well as a more accurate set of probabilities for its amino acid identity, which lets it search the se-

quence more accurately with the sequence attention module. This process of improvement continues while the positions of the atoms also get optimized using these representations through application of the Backbone Frame module.

The Backbone Frame module (seen in figure 3.2B) takes as input the representation of each graph node that is the result of the sequential operation of the three main modules described above, and outputs three vectors that describe the change in position of the C^{α}, C, and N atom positions with respect to the network's current backbone affine frame. The shift in position is applied to the backbone atoms and the new backbone affine frame is calculated using Gram-Schmidt, similar to Algorithm 21 in AlphaFold2 (Jumper et al., 2021). Because the three shifts may distort the geometry of the peptide plane, the new coordinates for the backbone are calculated by aligning a peptide plane with ideal geometry (from literature bonds) with the shifted positions (see algorithm 16 in A.3.2).

3.2.2 TRAINING

Multiple loss functions define the tasks of the different modules and the resulting losses are optimized jointly with gradient descent. Most losses are calculated at each intermediate layer of the GNN, so that it is able to learn the correct structure as early in the layers as possible. The most important losses are: C^{α} root mean squared deviation (RMSD) loss; backbone RMSD loss; amino-acid classification loss; local confidence score loss; torsion angles loss; and full atom loss. A full definition of all losses is given in Appendix A.1.

The main training loop consists of taking a PDB structure, extracting just the C^{α} atoms, distorting them with noise, initializing the backbone frames for each node randomly, and then having the network predict the original PDB structure. Cryo-EM maps are augmented with the addition of noise and sharpening/dampening, similar to the training of the graph initialization network as described in section 3.1.3. Because the initial C^{α} positions are noisy, with an RMSD to the deposited model of 0.9 Å on average, one important loss function is

$$\mathcal{L}_{C^{\alpha}} = \frac{1}{N} \sum_{i} \text{RMSD}(\mathbf{x}_i, g_{\phi}(\mathbf{x}_i + \mathbf{e}_i)) \tag{3}$$

where $\mathbf{x}_i \in \mathbb{R}^3$ are the true C^{α} positions from the dataset, g_{ϕ} is the graph neural network, and $\mathbf{e}_i \sim \mathcal{N}(0, \frac{1}{\sqrt{3}})$. Note that $\mathbb{E}[\text{RMSD}(\mathbf{e}_i, \mathbf{0})] \approx 0.9$ (this comes from the average norm of a Gaussian distributed vector). Denoising node positions has been shown to be a powerful training paradigm in other use cases as well (e.g. Godwin et al., 2021). In addition to noise added to the C^{α} positions, sometimes the graph initialization step misses some residues, or adds extra ones. In order to have the GNN be able to deal with these scenarios, during training 10% of the residues are randomly removed and replaced with randomly generated peptides that are between 2-5 residues long. The network is then also trained to be able to predict whether or not a node actually exists in the model, or if it is extra. All classification losses use focal loss (Lin et al., 2017b). This includes the amino-acid classification loss, the sequence match loss, and the edge prediction loss.

An important feature of this network is that it also gives a measure of its confidence in its output per residue. This is trained by having the network predict its backbone loss per residue. This output (referred to in 3.2 as O) is then normalized and saved in the B-factor section of the mmCIF file it outputs. This output is useful in pruning regions of the model where the network is not confident in the postprocessing step (see section 3.3). Generally, we observe that well-ordered, high-resolution parts of the cryo-EM map have higher confidence values than regions with disordered and lower resolution (see figure 2C).

The side chain atoms are generated through prediction of their rotatable torsion angles with respect to the backbone frame (for an example, see figure 1C). We noticed better results if the network predicted torsion angles for all 20 possible amino acids assignments for each residue. Then, we index into the torsion angle predictions for each residue and pick the set of angles that correspond to the predicted amino acid. To train this part of the model, for each layer, the mean squared loss of the torsion angles of the target amino-acid against the true torsion angles is calculated, and at the last layer the all-atom RMSD to the target structure is also calculated.

3.2.3 RECYCLING

The output of one round of the GNN denoises the positions of the C^{α} nodes and gives better orientations for the backbone frames, and we observed that a subsequent round of the GNN, starting from the output of the previous round, improved the results further. We therefore train the GNN with recycling. For every training step, we randomly pick an integer $r \in \{1, 2, 3\}$ and run the GNN $r - 1$ times with gradients turned off, and then use the output to run the GNN one more time with gradients. This allows the GNN to learn to keep the input approximately unchanged when the positions are correct. We do not recycle the GNN features so they are recalculated with the corrected positions and orientations. The same recycling scheme, but with $r = 3$, is also used during inference.

3.3 POSTPROCESSING

The GNN processes the C^{α} atoms into a set of unordered residues. Next, we connect the residues into chains that define the full atomic model. In the strictest sense, not even the direction of the chains is defined by the GNN. However, using the fact that $\|C_{t-1} - N_t\|_2 < 1.4$ Å (known as peptide bonds, see figure 1A), we can combine the atomic coordinates predicted by the network as well as the edge prediction probabilities as a heuristic to connect residues. More concretely, the residues are tied so that the sum of peptide bond lengths across all nodes is minimized, ignoring links where the edge prediction is below a threshold of 0.5.

After the chains are connected, we use the amino acid prediction probabilities to construct an HMM profile. We then perform a sequence search using HMMER (Mistry et al., 2013) against the given set of sequences of the model. The uncertainty of the predicted amino-acid probabilities P can vary due to several factors, e.g. map resolution or characteristics of the sequence itself, which can limit an accurate mapping of the sequence onto the structure by the GNN. Due to this uncertainty, doing a probabilistic search against the sequence after postprocessing gives superior performance over just assigning the highest probability amino acid. After alignment with the sequence search, residues that correspond to a "match" state (as defined in Krogh et al., 1994) are mutated to the amino acids that exist in the sequence. Based on the sequence search, we also connect separate chains that should be connected depending on both the matched sequence gap and the proximity of the chains.

Lastly, chains shorter than 4 residues are pruned and the resulting coordinates are used as the input to the GNN network again. This process continues for 3 recycling iterations and the end result undergoes a final "relaxation" step that uses physical restraint-based losses to optimize the positions of the model atoms using an L-BFGS optimizer (Liu & Nocedal, 1989). This step mainly alleviates unnatural side chain distance violations, and does not noticeably affect the distance metrics in section 4, which are all based on main chain atoms.

4 RESULTS

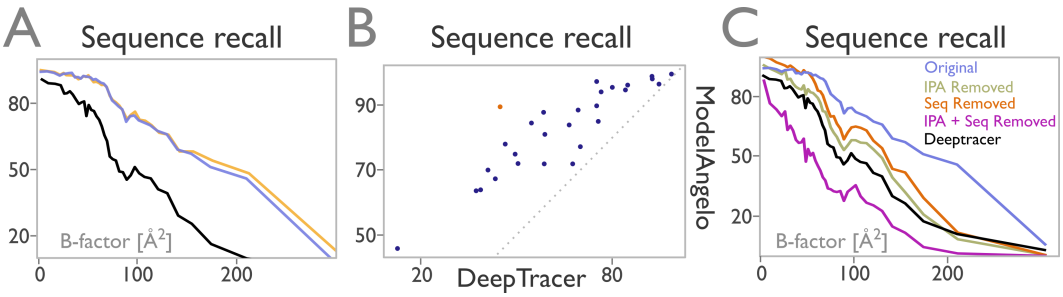

Figure 3: **(A)** shows sequence recall for all residues in the test dataset as a function of B-factor labels for final model outputs (after postprocessing), for Deeptracer (black) and ModelAngelo (before pruning in orange; after pruning in purple). **(B)** shows the same results, but averaged for each PDB entry, with ModelAngelo's pruned prediction (y-axis) versus Deeptracer (x-axis). The dotted line marks the identity line. The orange marker represents PDB entry 8DTM, which is shown in figure 4. **(C)** shows the result of the ablation experiment, similar in format to (A).

Here we report results for a test dataset of 28 map-model pairs that were deposited to the PDB and EMDB after the cutoff date for training. Results on a smaller test set without protein chains with more than 30 % sqeuence similarity with any model in the training set are described in A.5. We implemented our approach in an open-source software package called ModelAngelo. Generally, the atomic models built by ModelAngelo are close to the deposited PDB structures and they degrade with the resolution of the cryo-EM map. The overall resolution of the 28 test maps ranges from 2.1 to 3.8 Å. However, flexibility in parts of the protein structures also leads to local variations in resolution across the maps. The latter are reflected in the refined B-factors of the individual residues in the deposited PDB coordinate files, where higher B-factors indicate lower local resolution. We compare our results against the current state-of-the-art method for automated model building, Deeptracer (Pfab et al., 2021). A comparison with results obtained with the Phenix software (Terwilliger et al., 2018), which performs worse than DeepTracer, is available in table 4 of A.3.

Our main metric is sequence recall: the percentage of residues for which the C^{α} atom is within 3 Å of the deposited model, and the amino acid prediction is correct. Figure 3 compares the performance of ModelAngelo and Deeptracer for each of the structures in the test dataset, and as a function of B-factor averaged over all residues. A more comprehensive list of metrics, for each of the 28 map-model pairs, is shown in Appendix A.2. Over the entire test dataset, there is little difference in sequence recall between the unpruned and the pruned model from ModelAngelo, which implies that pruning removes incorrectly built parts of the model.

Ablation studies (figure 4C) show that the improved results of ModelAngelo versus Deeptracer are because ModelAngelo is able to combine different modalities of information to build the model, rather than just the cryo-EM map. Removing the sequence module or the IPA module results in worse results that are still better than Deeptracer. However, removing both of these modules, such that the GNN only relies on the Cryo-EM module, results in predictions that are worse than Deeptracer.

Figure 4 illustrates the quality of the built models from ModelAngelo's and Deeptracer for one example from the test dataset (PDB entry 8DTM); more examples are given in A.6. Because of its increased complexity, ModelAngelo is considerably slower than Deeptracer. Still, execution times for the test dataset are in the range of several minutes to one hour and a half, depending on the size of the structure (see A.4.3). Given that manual *de novo* model building takes on the order of weeks, we do not believe this to be a serious drawback.

5 Discussion

Our results illustrate that combining the voxel-based information from the cryo-EM map with sequence information and graph topology is useful for automating the intensive task of atomic modelling in cryo-EM maps. Below, we consider limitations of the current implementation of ModelAngelo, and outline lines of future research to overcome them.

Sensitivity to resolution. Even though the network has multiple different modalities of input, relatively low resolutions of the cryo-EM map will affect the results. The graph initialization by the CNN, and the amino acid classification that provides the information for mapping the sequence onto the main chain, are obvious examples that benefit from higher resolution maps. Poor amino acid classifications may also lead to errors in the sequence assignment in the postprocessing step, which may then feed into the subsequent HMM sequence alignment and lead to incorrect chain assignments. This is more likely to happen for complexes with many similar sequences. In practice, we observe that ModelAngelo's performance starts degrading at resolutions worse than 3.5 Å, see Appendix A.3. A similar trend is also typical for manual model building. It may be possible to combine the embedding of information from relatively low-resolution cryo-EM maps (e.g 10 Å) with methods for protein structure prediction, such as AlphaFold2 (Jumper et al., 2021), which would extend beyond what is possible for manual building. Despite the observation that ModelAngelo already uses some ideas from AlphaFold2, such an approach for building in low-resolution maps, which would blur the boundaries between experimental structure determination and prediction, would require major changes to the approaches outlined in this paper.

Nucleic acids. Many large complexes that are solved with cryo-EM comprise nucleic acids as well as proteins, e.g. ribonucleic acids (RNA) in spliceosomes and ribosomes, or deoxyribonucleic acids

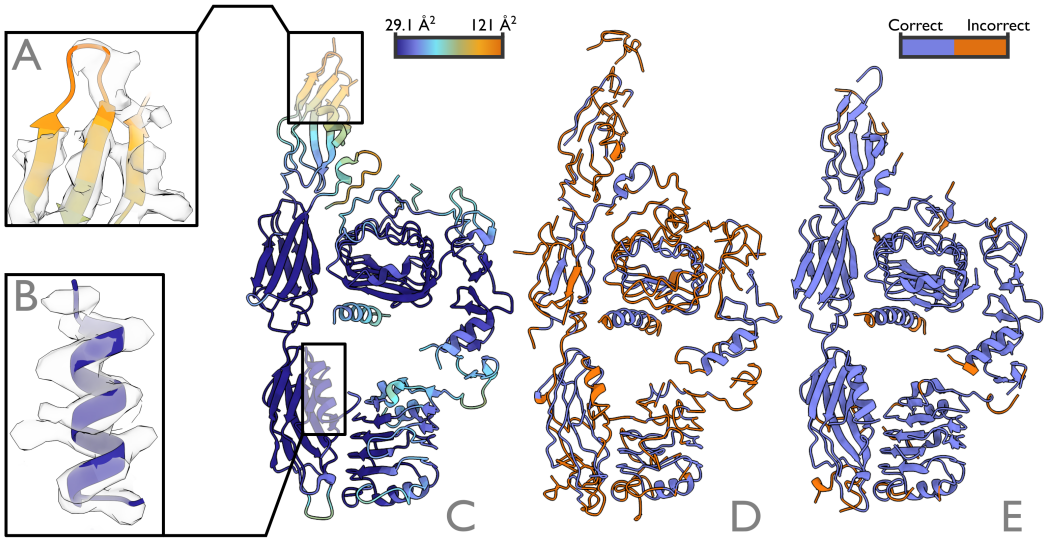

Figure 4: Comparison of the deposited model for PDB entry 8DTM **(C)**, Deeptracer's prediction **(D)** and ModelAngelo's pruned model **(E)**. The deposited model is coloured according to the refined B-factor. Meanwhile, the predictions are coloured orange where their amino acid prediction is different from the deposited structure, and purple where it is the same. **(A)** shows an iso-surface of the cryo-EM map and the deposited model for a high B-factor region, with correspondingly poor density. **(B)** shows the same for a low B-factor region, where side chains are well resolved in the density.

(DNA) in gene replication or transcription machinery. The backbone of DNA or RNA strands is made from alternating phosphate and sugar groups. The phosphorus atoms has high contrast in the cryo-EM map, which makes the segmentation problem of nucleic acids easier than that of protein residues. However, the main difficulty lies in identifying the correct sequence for the nucleobases that make up the equivalent of side chains for RNA or DNA strands. There are four typical bases for both RNA and DNA: two purines and two pyrimidines. At resolutions around 3.5 Å, one can distinguish the purines from the pyrimidines, but not the two purines or pyrimidines from each other. This makes sequencing DNA or RNA strands in 3.5 Å cryo-EM maps difficult. Yet, already the ability to automatically model nucleotide backbones, together with a classifier to distinguish purines from pyrimidines, would alleviate the task of manual model building. Therefore, we plan to add support for building RNA or DNA to ModelAngelo in the near future.

Unknown sequences. Because cryo-EM can be performed on samples that are extracted from native cells or tissues, it is not always obvious which proteins are present in a cryo-EM map, (for an example, see Schweighauser et al., 2022). Our current implementation depends on a user-provided sequence file that defines all proteins present in the map. Recently, semi-automated tools to identify proteins in cryo-EM maps have been reported. For example, findMySequence (Chojnowski et al., 2022) can search protein seqeuence databases, using amino acid classifications after a backbone model has been built in the cryo-EM map. We plan to extend our approach with a sequence-free model to fully automate this process, by combining automated model building with searches through large sequence databases like Uniclust (Mirdita et al., 2017) using our HMM search procedure combined with tools such as HHblits (Remmert et al., 2012).

REFERENCES

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

Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. Neural machine translation by jointly learning to align and translate. *arXiv preprint arXiv:1409.0473*, 2014.

Stephen K Burley, Charmi Bhikadiya, Chunxiao Bi, Sebastian Bittrich, Li Chen, Gregg V Crichlow, Cole H Christie, Kenneth Dalenberg, Luigi Di Costanzo, Jose M Duarte, et al. Rcsb protein data bank: powerful new tools for exploring 3d structures of biological macromolecules for basic and applied research and education in fundamental biology, biomedicine, biotechnology, bioengineering and energy sciences. *Nucleic acids research*, 49(D1):D437–D451, 2021.

Yifan Cheng. Single-particle cryo-em—how did it get here and where will it go. *Science*, 361(6405): 876–880, 2018.

Grzegorz Chojnowski, Adam J Simpkin, Diego A Leonardo, Wolfram Seifert-Davila, Dan E Vivas-Ruiz, Ronan M Keegan, and Daniel J Rigden. findmysequence: a neural-network-based approach for identification of unknown proteins in x-ray crystallography and cryo-em. *IUCrJ*, 9(1), 2022.

Kevin Cowtan. The buccaneer software for automated model building. 1. tracing protein chains. *Acta crystallographica section D: biological crystallography*, 62(9):1002–1011, 2006.

Tristan Ian Croll. Isolde: a physically realistic environment for model building into low-resolution electron-density maps. *Acta Crystallographica Section D: Structural Biology*, 74(6):519–530, 2018.

Justas Dauparas, Ivan Anishchenko, Nathaniel Bennett, Hua Bai, Robert J Ragotte, Lukas F Milles, Basile IM Wicky, Alexis Courbet, Rob J de Haas, Neville Bethel, et al. Robust deep learning based protein sequence design using proteinmpnn. *bioRxiv*, 2022.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*, 2018.

Paul Emsley, Bernhard Lohkamp, William G. Scott, and Kevin Cowtan. Features and development of coot. *Acta Crystallographica Section D - Biological Crystallography*, 66:486–501, 2010.

Octavian-Eugen Ganea, Xinyuan Huang, Charlotte Bunne, Yatao Bian, Regina Barzilay, Tommi S. Jaakkola, and Andreas Krause. Independent SE(3)-equivariant models for end-to-end rigid protein docking. In *International Conference on Learning Representations*, 2022. URL https://openreview.net/forum?id=GQjaI9mLet.

Jonathan Godwin, Michael Schaarschmidt, Alexander Gaunt, Alvaro Sanchez-Gonzalez, Yulia Rubanova, Petar Velickovic, James Kirkpatrick, and Peter W. Battaglia. Very deep graph neural networks via noise regularisation. *CoRR*, abs/2106.07971, 2021. URL https://arxiv.org/abs/2106.07971.

Seyed Raein Hashemi, Seyed Sadegh Mohseni Salehi, Deniz Erdogmus, Sanjay P. Prabhu, Simon K. Warfield, and Ali Gholipour. Tversky as a loss function for highly unbalanced image segmentation using 3d fully convolutional deep networks. *CoRR*, abs/1803.11078, 2018. URL http://arxiv.org/abs/1803.11078.

Jiahua He, Peicong Lin, Ji Chen, Hong Cao, and Sheng-You Huang. Model building of protein complexes from intermediate-resolution cryo-em maps with deep learning-guided automatic assembly. *Nature Communications*, 13(1):1–16, 2022.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 770–778, 2016.

Sergey Ioffe and Christian Szegedy. Batch normalization: Accelerating deep network training by reducing internal covariate shift. In *International conference on machine learning*, pp. 448–456. PMLR, 2015.

Bowen Jing, Gabriele Corso, Jeffrey Chang, Regina Barzilay, and Tommi Jaakkola. Torsional diffusion for molecular conformer generation. *arXiv preprint arXiv:2206.01729*, 2022.

John Jumper, Richard Evans, Alexander Pritzel, Tim Green, Michael Figurnov, Olaf Ronneberger, Kathryn Tunyasuvunakool, Russ Bates, Augustin Žídek, Anna Potapenko, et al. Highly accurate protein structure prediction with alphafold. *Nature*, 596(7873):583–589, 2021.

Anders Krogh, Michael Brown, I Saira Mian, Kimmen Sjölander, and David Haussler. Hidden markov models in computational biology: Applications to protein modeling. *Journal of molecular biology*, 235(5):1501–1531, 1994.

Catherine L Lawson, Ardan Patwardhan, Matthew L Baker, Corey Hryc, Eduardo Sanz Garcia, Brian P Hudson, Ingvar Lagerstedt, Steven J Ludtke, Grigore Pintilie, Raul Sala, et al. Emdatabank unified data resource for 3dem. *Nucleic acids research*, 44(D1):D396–D403, 2016.

Dorothee Liebschner, Pavel V Afonine, Matthew L Baker, Gábor Bunkóczi, Vincent B Chen, Tristan I Croll, Bradley Hintze, L-W Hung, Swati Jain, Airlie J McCoy, et al. Macromolecular structure determination using x-rays, neutrons and electrons: recent developments in phenix. *Acta Crystallographica Section D: Structural Biology*, 75(10):861–877, 2019.

Tsung-Yi Lin, Piotr Dollár, Ross Girshick, Kaiming He, Bharath Hariharan, and Serge Belongie. Feature pyramid networks for object detection. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 2117–2125, 2017a.

Tsung-Yi Lin, Priya Goyal, Ross B. Girshick, Kaiming He, and Piotr Dollár. Focal loss for dense object detection. *CoRR*, abs/1708.02002, 2017b. URL http://arxiv.org/abs/1708.02002.

Zeming Lin, Halil Akin, Roshan Rao, Brian Hie, Zhongkai Zhu, Wenting Lu, Allan dos Santos Costa, Maryam Fazel-Zarandi, Tom Sercu, Sal Candido, and Alexander Rives. Language models of protein sequences at the scale of evolution enable accurate structure prediction. *bioRxiv*, 2022. doi: 10.1101/2022.07.20.500902. URL https://www.biorxiv.org/content/early/2022/07/21/2022.07.20.500902.

Dong C Liu and Jorge Nocedal. On the limited memory bfgs method for large scale optimization. *Mathematical programming*, 45(1):503–528, 1989.

Valerio Mariani, Marco Biasini, Alessandro Barbato, and Torsten Schwede. lddt: a local superposition-free score for comparing protein structures and models using distance difference tests. *Bioinformatics*, 29(21):2722–2728, 2013.

Milot Mirdita, Lars Von Den Driesch, Clovis Galiez, Maria J Martin, Johannes Söding, and Martin Steinegger. Uniclust databases of clustered and deeply annotated protein sequences and alignments. *Nucleic acids research*, 45(D1):D170–D176, 2017.

Jaina Mistry, Robert D Finn, Sean R Eddy, Alex Bateman, and Marco Punta. Challenges in homology search: Hmmer3 and convergent evolution of coiled-coil regions. *Nucleic acids research*, 41 (12):e121–e121, 2013.

Garib N Murshudov, Pavol Skubák, Andrey A Lebedev, Navraj S Pannu, Roberto A Steiner, Robert A Nicholls, Martyn D Winn, Fei Long, and Alexei A Vagin. Refmac5 for the refinement of macromolecular crystal structures. *Acta Crystallographica Section D: Biological Crystallography*, 67(4):355–367, 2011.

Takanori Nakane, Abhay Kotecha, Andrija Sente, Greg McMullan, Simonas Masiulis, Patricia MGE Brown, Ioana T Grigoras, Lina Malinauskaite, Tomas Malinauskas, Jonas Miehling, et al. Single-particle cryo-em at atomic resolution. *Nature*, 587(7832):152–156, 2020.

Anastassis Perrakis, Richard Morris, and Victor S Lamzin. Automated protein model building combined with iterative structure refinement. *Nature structural biology*, 6(5):458–463, 1999.

Eric F Pettersen, Thomas D Goddard, Conrad C Huang, Elaine C Meng, Gregory S Couch, Tristan I Croll, John H Morris, and Thomas E Ferrin. Ucsf chimerax: Structure visualization for researchers, educators, and developers. *Protein Science*, 30(1):70–82, 2021.

Jonas Pfab, Nhut Minh Phan, and Dong Si. Deeptracer for fast de novo cryo-em protein structure modeling and special studies on cov-related complexes. *Proceedings of the National Academy of Sciences of the United States of America*, 118, 1 2021. ISSN 10916490. doi: 10.1073/PNAS.2017525118/SUPPL_FILE/PNAS.2017525118.SAPP.PDF.

Ali Punjani, John L Rubinstein, David J Fleet, and Marcus A Brubaker. cryosparc: algorithms for rapid unsupervised cryo-em structure determination. *Nature methods*, 14(3):290–296, 2017.

Roshan M Rao, Jason Liu, Robert Verkuil, Joshua Meier, John Canny, Pieter Abbeel, Tom Sercu, and Alexander Rives. Msa transformer. In *International Conference on Machine Learning*, pp. 8844–8856. PMLR, 2021.

Michael Remmert, Andreas Biegert, Andreas Hauser, and Johannes Söding. Hhblits: lightning-fast iterative protein sequence searching by hmm-hmm alignment. *Nature methods*, 9(2):173–175, 2012.

Alexander Rives, Joshua Meier, Tom Sercu, Siddharth Goyal, Zeming Lin, Jason Liu, Demi Guo, Myle Ott, C Lawrence Zitnick, Jerry Ma, et al. Biological structure and function emerge from scaling unsupervised learning to 250 million protein sequences. *Proceedings of the National Academy of Sciences*, 118(15):e2016239118, 2021.

Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. In *International Conference on Medical image computing and computer-assisted intervention*, pp. 234–241. Springer, 2015.

Peter B Rosenthal and Richard Henderson. Optimal determination of particle orientation, absolute hand, and contrast loss in single-particle electron cryomicroscopy. *Journal of molecular biology*, 333(4):721–745, 2003.

Sjors HW Scheres. Relion: implementation of a bayesian approach to cryo-em structure determination. *Journal of structural biology*, 180(3):519–530, 2012.

Manuel Schweighauser, Diana Arseni, Mehtap Bacioglu, Melissa Huang, Sofia Lövestam, Yang Shi, Yang Yang, Wenjuan Zhang, Abhay Kotecha, Holly J Garringer, et al. Age-dependent formation of tmem106b amyloid filaments in human brains. *Nature*, 605(7909):310–314, 2022.

Genki Terashi and Daisuke Kihara. De novo main-chain modeling for em maps using mainmast. *Nature communications*, 9(1):1–11, 2018.

Thomas C Terwilliger, Ralf W Grosse-Kunstleve, Pavel V Afonine, Nigel W Moriarty, Peter H Zwart, L-W Hung, Randy J Read, and Paul D Adams. Iterative model building, structure refinement and density modification with the phenix autobuild wizard. *Acta Crystallographica Section D: Biological Crystallography*, 64(1):61–69, 2008.

Thomas C Terwilliger, Paul D Adams, Pavel V Afonine, and Oleg V Sobolev. A fully automatic method yielding initial models from high-resolution cryo-electron microscopy maps. *Nature methods*, 15(11):905–908, 2018.

Thomas C. Terwilliger, Billy K. Poon, Pavel V. Afonine, Christopher J. Schlicksup, Tristan I. Croll, Claudia Millán, Jane. S. Richardson, Randy J. Read, and Paul D. Adams. Improved alphafold modeling with implicit experimental information. *bioRxiv*, 2022. doi: 10.1101/2022.01.07.475350. URL `https://www.biorxiv.org/content/early/2022/01/30/2022.01.07.475350`.

Dmitry Ulyanov, Andrea Vedaldi, and Victor Lempitsky. Instance normalization: The missing ingredient for fast stylization. *arXiv preprint arXiv:1607.08022*, 2016.

Ruidong Wu, Fan Ding, Rui Wang, Rui Shen, Xiwen Zhang, Shitong Luo, Chenpeng Su, Zuofan Wu, Qi Xie, Bonnie Berger, Jianzhu Ma, and Jian Peng. High-resolution de novo structure prediction from primary sequence. *bioRxiv*, 2022a. doi: 10.1101/2022.07.21.500999. URL `https://www.biorxiv.org/content/early/2022/07/22/2022.07.21.500999`.

Ruidong Wu, Fan Ding, Rui Wang, Rui Shen, Xiwen Zhang, Shitong Luo, Chenpeng Su, Zuofan Wu, Qi Xie, Bonnie Berger, et al. High-resolution de novo structure prediction from primary sequence. *bioRxiv*, 2022b.

Bing Xu, Naiyan Wang, Tianqi Chen, and Mu Li. Empirical evaluation of rectified activations in convolutional network. *arXiv preprint arXiv:1505.00853*, 2015.

Ka Man Yip, Niels Fischer, Elham Paknia, Ashwin Chari, and Holger Stark. Atomic-resolution protein structure determination by cryo-em. *Nature*, 587(7832):157–161, 2020.