A graph neural network approach to automated model building in cryo-EM maps

Kiarash Jamali, Dari Kimanius, & Sjors Scheres
MRC Laboratory of Molecular Biology
Cambridge, UK
{kjamali,dari,scheres}@mrc-lmb.cam.ac.uk

Abstract

Electron cryo-microscopy (cryo-EM) produces three-dimensional (3D) maps of the electrostatic potential of biological macromolecules, including proteins. At sufficient resolution, the cryo-EM maps, along with some knowledge about the imaged molecules, allow *de novo* atomic modelling. Typically, this is done through a laborious manual process. Recent advances in machine learning applications to protein structure prediction show potential for automating this process. Taking inspiration from these techniques, we have built *ModelAngelo* for automated model building of proteins in cryo-EM maps. ModelAngelo first uses a residual convolutional neural network (CNN) to initialize a graph representation with nodes assigned to individual amino acids of the proteins in the map and edges representing the protein chain. The graph is then refined with an equivariant graph neural network (GNN) that combines the 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. The final graph is post-processed with a hidden Markov model (HMM) search to map each protein chain to entries in a user provided sequence file. Application to 28 test cases shows that ModelAngelo outperforms 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, for favourable samples, cryo-EM structure determination of biological macromolecules is now possible to atomic resolution [1, 2]. For many other samples, such as large multi-component complexes and membrane proteins, often resolutions better than 3 Å are achieved [3]. In this method, transmission electron microscopy images are taken of many copies of the same molecules, which are frozen in a thin layer of vitreous ice. Each field of view contains hundreds of two-dimensional projections of the electrostatic potential of such molecules in unknown orientations. Radiation damage limits the amounts of electrons that can be used for imaging, which results in extremely low signal-to-noise ratios in the images. Consequently, averaging over many copies is necessary, and one often acquires thousands of field of views for a given sample. Dedicated softwares, like RELION [4] or cryoSPARC [5], implement iterative optimization algorithms to retrieve the unknown 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 then interpreted in terms of an atomic model of the corresponding macromolecular structures. Many samples contain only proteins; other samples also contain other biological molecules, like lipids or nucleotides. Proteins are linear chains of amino acids, or residues. There are twenty different amino acids. All amino acids have four atoms that make up the protein main chain. Often, but not always, the electron microscopist knows which

Machine Learning for Structural Biology Workshop, NeurIPS 2022.

proteins are present in the sample. The atomic model building task at hand is to identify the positions of all atoms for all proteins that are present in the reconstructed 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.

Atomic model building in cryo-EM maps is typically done manually using 3D visualisation software, [e.g. 6, 7], followed by refinement procedures that optimize the fit of the models in the map, [e.g. 8, 9, 10]. Building a reliable atomic model *de novo* in the reconstructed cryo-EM map is considered to be difficult for maps with resolutions worse than 4 Å resolution. 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 protein structure prediction [11, 12]. In these approaches, the sequence information of proteins and their evolutionary related homologues is used to predict the atomic structure of proteins 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 [13, 14]. These successes raise the question whether improvements to automated model building could be achieved by taking inspiration from novel machine learning approaches.

In this paper, we present a machine learning approach to automated model building in cryo-EM maps. Our approach combines modern GNN architectures and protein language models with novel techniques of incorporating the voxel-based information from cryo-EM maps alongside the amino-acid sequence and structural information for automated model building. In order to accomplish this, we implemented a pipeline that first seeds the map with the approximate locations of the individual residues using a CNN, and then uses a GNN to refine the full atom positions of each residue. This GNN has novel components that allow it to use information from the different modalities of text (the amino-acid sequence), 3D volumes (the cryo-EM grid), as well as graphs (the positions of the atoms) in one integrated neural network. Finally, a series of post-processing steps seek to improve the model and present it in a clean manner to users.

2 Prior Work

Automated approaches for atomic modelling in the related experimental technique of X-ray crystallography have existed for many years [for example, 15, 16, 17]. Although some of these approaches have also been applied to the more widely applicable technique of cryo-EM structure determination, their overall impact on the field has, thus far, been modest.

More recently, Deeptracer, the first deep learning approach for automated atomic modelling in cryo-EM maps, was reported to outperform previously existing approaches [18]. Deeptracer uses a U-Net [19], coupled with some heuristics, 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.

There have also been reports to use protein structure prediction programs, like AlphaFold2 [11], to morph their output predictions to fit the cryo-EM map [20, 21]. Such approaches are likely to propagate errors in the structure prediction. Thus, it seems sensible to design a neural network approach that integrates both the cryo-EM map and the sequence and protein structure primitives to produce a more reliable structure. This is the approach of ModelAngelo.

3 Methods

3.1 Residue segmentation

The first step in ModelAngelo is to identify where individual residues are placed. For proteins in our approach, this is modelled as the position of the C^{α} atom (see figure 1B) of each amino acid. This part of the pipeline is formulated as a straight-forward segmentation problem. That is, the cryo-EM

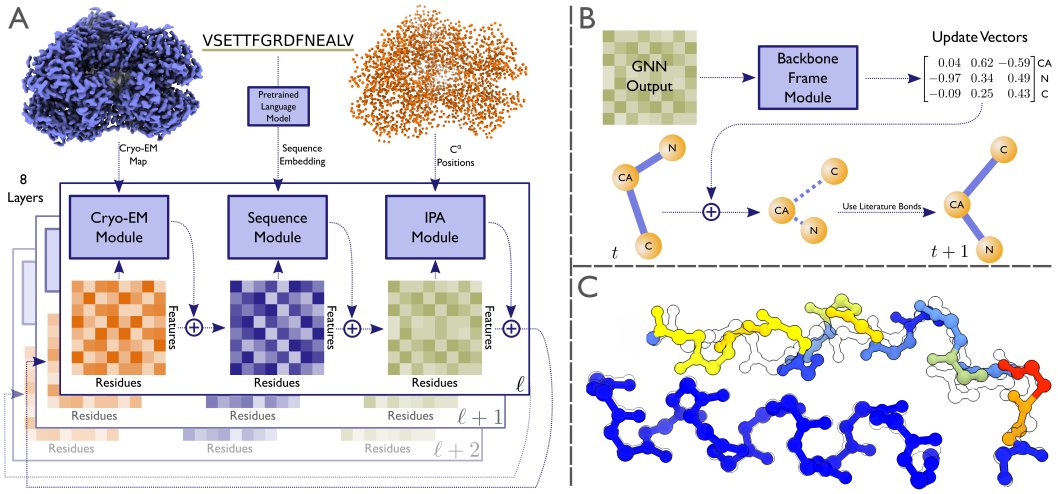

Figure 1: **(A)** shows the schematic of the graph neural network of ModelAngelo and how the 8 layers of the GNN iteratively refine the feature vectors per residue that are used to predict the atomic model. **(B)** illustrates how the Backbone Frame module updates the positions of the backbone atoms by first predicting a set of shift vectors, updating the positions, and then re-projecting the atoms back to literature bond lengths and angles. Finally, **(C)** contains two examples of high confidence (dark blue colour) and low confidence (yellow and red) predicted backbone regions. The confidence measure is a good predictor of fit to the deposited backbone model (shown in outline).

map $V \in \mathbb{R}^{N^3}$, where the N is the number of voxels, has an associated binary target T, 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 $f_\theta(V) \approx T$. We did this in a supervised manner using online deposited pairs of cryo-EM volumes and PDB models.

3.2 All-atom modelling

Starting from approximate positions for the C^α atoms, we need to get a complete atomic model. This consists of a few separate tasks. Since the input to the network is an unordered collection of C^α atoms, the amino-acid types, the chain directions, the side chain placements, and the correspondence to the given sequence need to be determined. In order to accomplish this, all residues are represented with a backbone affine frame and a series of torsion angles, similar to AlphaFold2 [11], with orientations initialized randomly and then optimized using the output of an $SE(3)$-equivariant graph neural network (GNN), as can be seen in figure 1.

Unlike AlphaFold2, which already has access to the residue-to-sequence relationship due to the nature of the protein structure prediction problem, our GNN cannot make any assumptions about the order of its input nodes. Therefore it also needs an amino-acid classification probability for each residue, along with learning the correspondence of the graph nodes to the sequence. The GNN consists of three main modules, all based on the attention algorithm [22], that are stacked 8 times in order to sequentially optimize the output structure.

The first of these modules 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 of neighbouring nodes, to update its representation. This is accomplished with a novel mix of a graph-based attention module and feature extraction using CNNs. A cube centered at each node is interpolated from the cryo-EM density and the orientation of this cube is defined by the backbone affine frame of each node. Similarly, rectangles of cryo-EM density are interpolated along the vectors that connect each node to its closest k neighbours. The query and value vectors are generated by the feature vector of each node, while the key vectors are generated from the cryo-EM rectangles between neighbours. Finally, the features generated from the centered cubes are concatenated with

the attention output and integrated to create the new features. This allows the network to see how strong the density between two nodes is before it makes decisions to mix features.

The second module is the Sequence Attention module, which lets each node search against the input sequence to find the relevant part of the sequence that matches its features. This is a conventional encoder-only transformer module, similar to that used in [23]. The sequence is first embedded using a pre-trained protein language model [24, 14].

Finally, there 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 is inspired by the similarly named module in AlphaFold2, however 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, 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 nodes, each node gets an attention score that is used to update its representation. Essentially, it 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 these modules are applied sequentially multiple times, the representations from each module allow 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 sequence 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 the application of the Backbone Frame module.

The Backbone Frame module (seen in figure 1B) takes as input the representation of each graph node that is the result of the sequential operation 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 [11]. After the backbone affine frame has been defined, we then use the known bond lengths and angles to get new positions for the C and N atoms.

Since this model does a series of tasks, there is no single loss function. Instead, the loss is split across different tasks and is 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 as follows: 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. 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. 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

$$\mathcal{L}_{C^\alpha} = \frac{1}{N} \sum_i \mathrm{RMSD}(\mathbf{x}_i, f_\theta(\mathbf{x}_i + \mathbf{e}_i)) \tag{1}$$

where $\mathbf{x}_i \in \mathbb{R}^3$ are the true C^α positions from the dataset, f_θ is the graph neural network, and $\mathbf{e}_i \sim \mathcal{N}(0, \frac{1}{\sqrt{3}})$. Note that $\mathbb{E}[\mathrm{RMSD}(\mathbf{e}_i, \mathbf{0})] \approx 0.9$ (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. 25].

The side-chain atoms are generated through prediction of their rotatable torsion angles with respect to the backbone frame. We noticed better results if the network predicted torsion angles for all 20 amino acids per residue. Then, based on the predicted amino-acid, we index into the torsion angle predictions and pick the set of angles that correspond to the residue. 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.

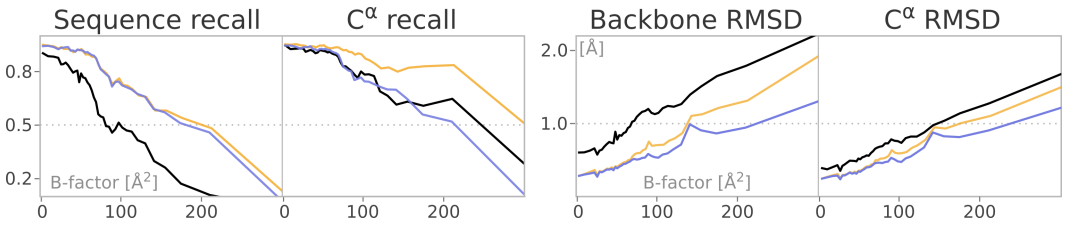

Figure 2: Average recall and RMSD as a function of B-factor labels from the deposited PDB entries. The results are coloured by method, with orange representing ModelAngelo's results before pruning and purple showing them after. The results for DeepTracer are plotted with a black line.

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 the threshold of 0.5.

There are two types of output from the model. In the unpruned output, chains are built more or less as-is from the output of the model, with minimal post-processing. This allows for a larger percentage of the model to be built (higher recall), however this also means that portions of the model that exist in lower resolution areas of the map are often wrong. This can be cumbersome for biologists to use as they have to manually prune and fix chains that are incorrect and closely analyze the model. The pruned output seeks to remedy this by removing portions of the model that do not have good matches to the sequence based on a hidden Markov model alignment. This leads to lower recall, but higher sequence matching percentages, correlating to the correctly modelled portion of the map.

4 Results

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. 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 [18].

We consider two metrics. Sequence recall is 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. Backbone RMSD is the root mean square deviation of all the backbone atoms in Ångstroms. We did not calculate all-atom RMSDs, as Deeptracer does not output side chain coordinates. Figure 2 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. Over the entire test dataset, there is little difference in sequence recall between the unpruned and the pruned model from ModelAngelo, whereas the backbone RMSD does improve after pruning. This implies that pruning removes incorrectly built parts of the model. We believe 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. 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. Given *de novo* model building takes on the order of weeks, we do not believe this to be a serious drawback.

References

[1] 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.

[2] 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.

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

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

[5] 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.

[6] 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.

[7] 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.

[8] 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.

[9] 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.

[10] 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.

[11] 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.

[12] 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.

[13] 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*, 2022.

[14] 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.

[15] 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.

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

[17] 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.

[18] 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.

[19] 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*, pages 234–241. Springer, 2015.

[20] 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.

[21] 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.

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

[23] 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.

[24] 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.

[25] 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.