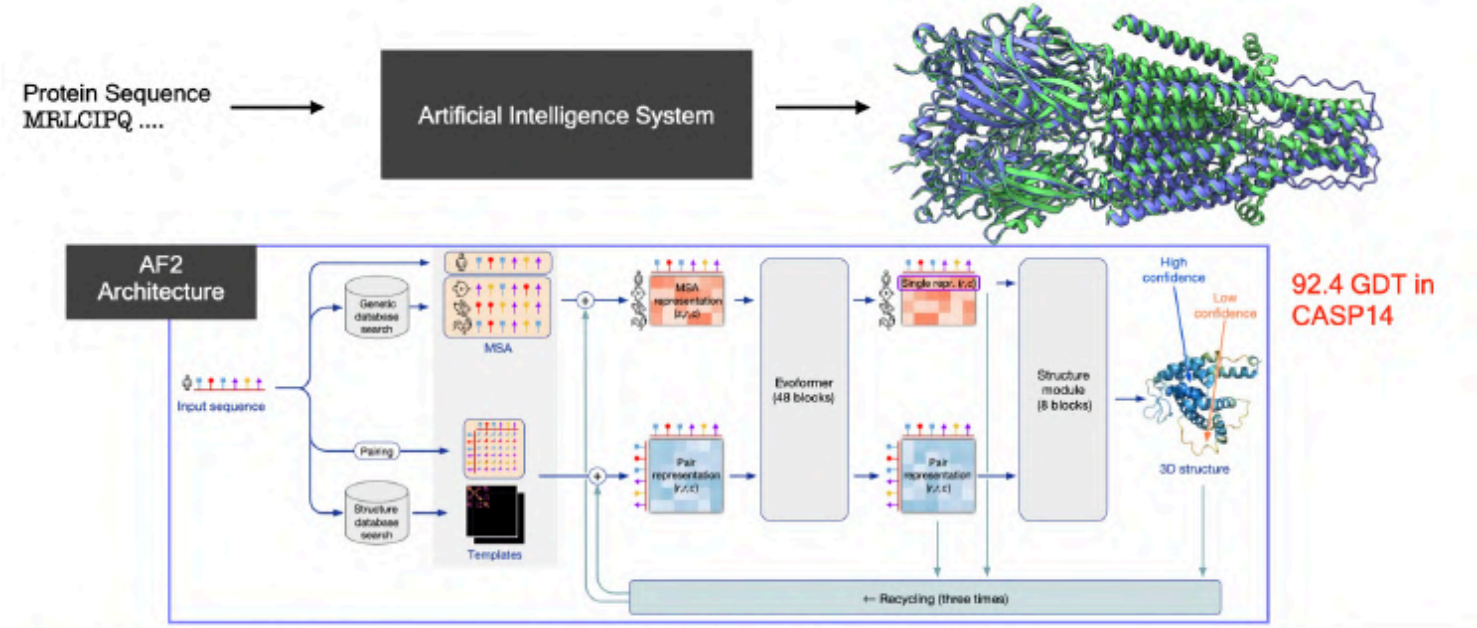
Modeling Atomic Structures from Cryo-EM Maps

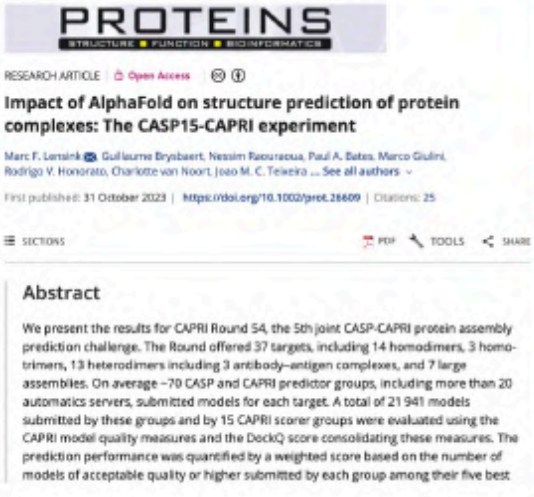
Nabin Giri
PhD Defense Presentation

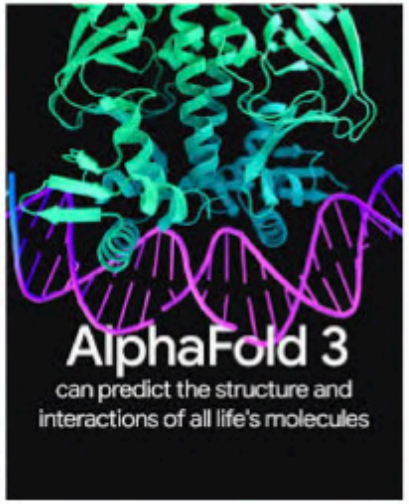
May, 2025

Electrical Engineering
& Computer Science
University of Missouri

Machine Learning in Structural Biology (MLSB)

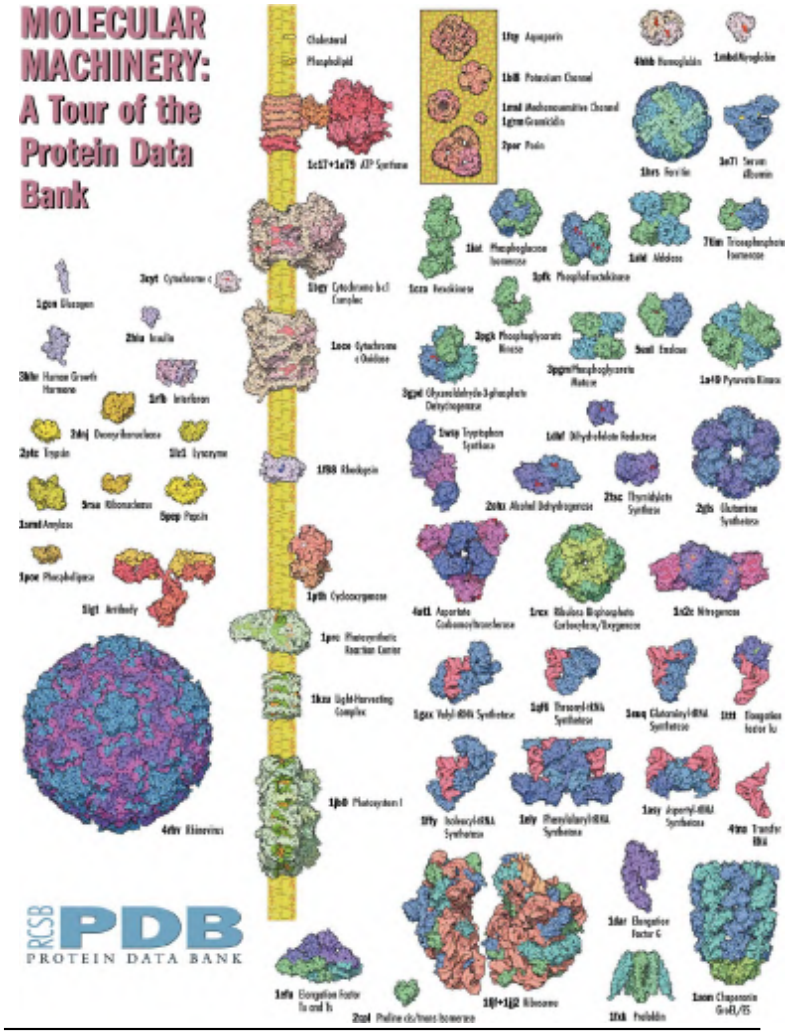






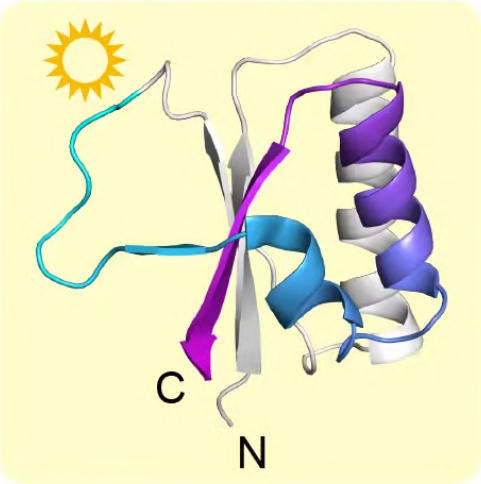
Jumper, John, et al. "Highly accurate protein structure prediction with AlphaFold." Nature 596.7873 (2021): 583-589.

Protein Structures and Their Conformations



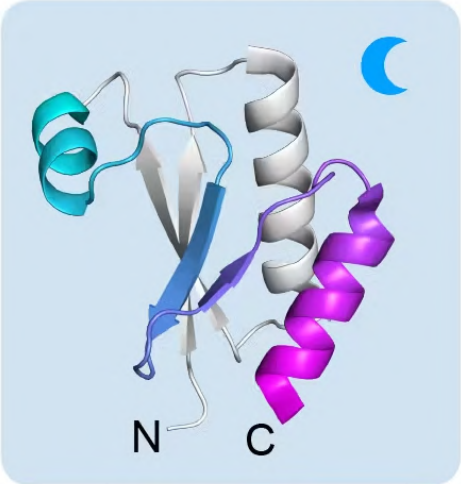
KaiB: a bacterial circadian rhythm protein

Ground State



in vitro: 90%

Fold-switched state



10%

But! Default AlphaFold only predicts this state

Berman, Helen M. "The protein data bank: a historical perspective." Acta Crystallographica Section A: Foundations of Crystallography 64.1 (2008)

Wayment-Steele, Hannah K., et al. "Predicting multiple conformations via sequence clustering and AlphaFold2." Nature 625.7996 (2024): 832-839.

Studying Molecular Motions

- **Nuclear Magnetic Resonance (NMR)**: Limited to small molecular sizes due to signal overlap and sensitivity issues.
- **X-ray Crystallography**: Requires crystallization, which may not reflect native dynamics.
- **Molecular Dynamics (MD) Simulations**: Dependent on computational power and the accuracy of force fields such as CHARMM and AMBER.
- **Modify AlphaFold Code** for generating different conformations of a protein. e.g.: clustering MSAs by sequence similarity enables AlphaFold2 to sample alternative states of KaiB, metamorphic clock protein with high confidence.
- **Cryo-Electron Microscopy (cryo-EM)**: Enables high-resolution imaging of large complexes.

Cryo-EM Workflow

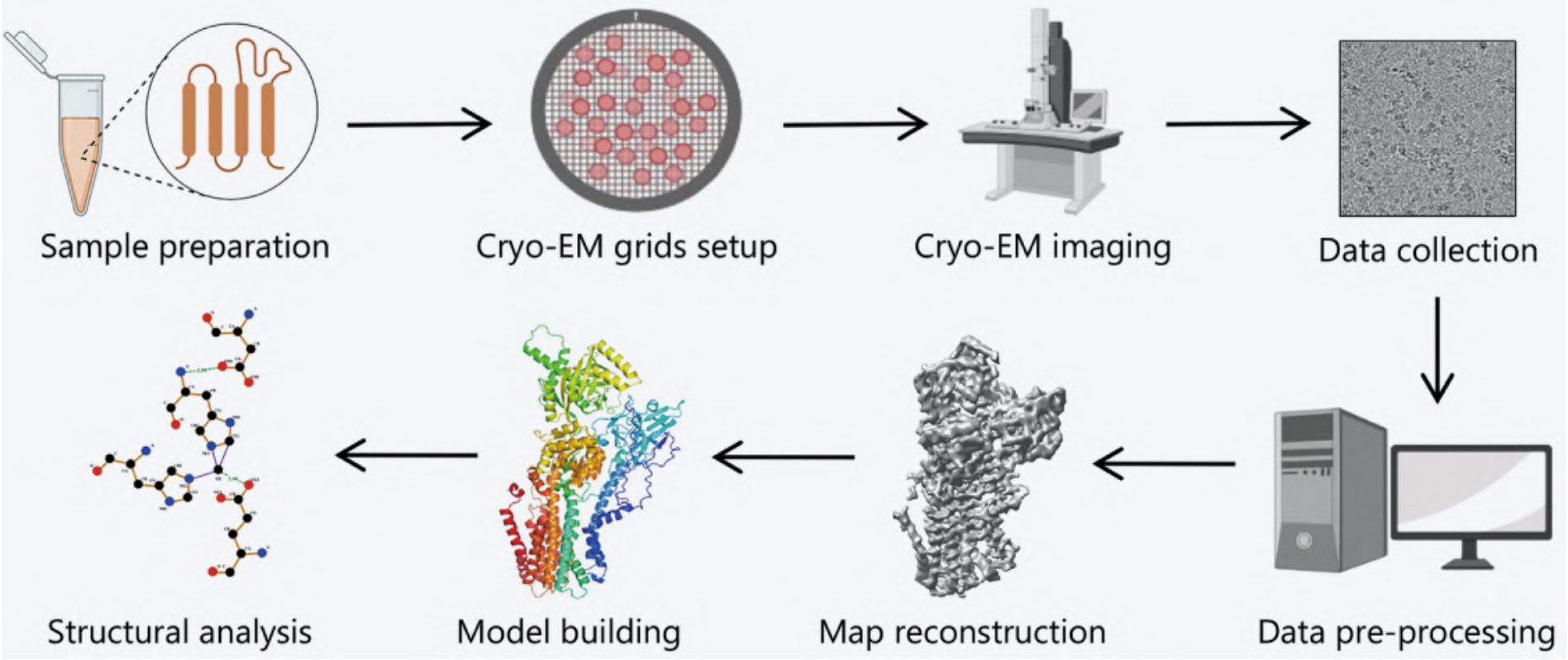
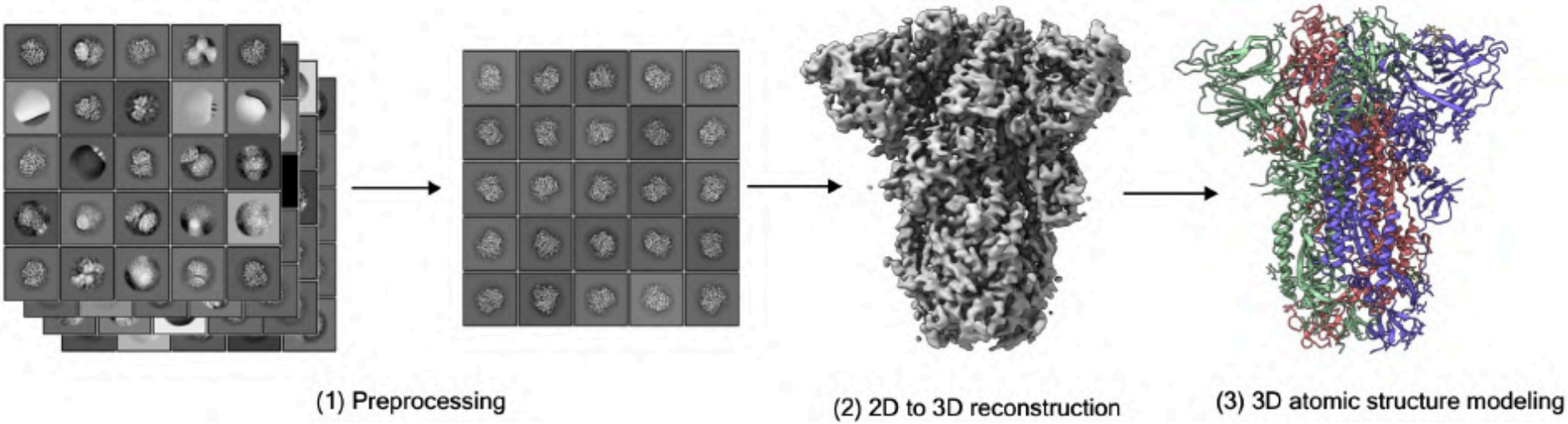


Image Source: https://blog.delmic.com/cryo-electron-microscopy-for-drug-discovery

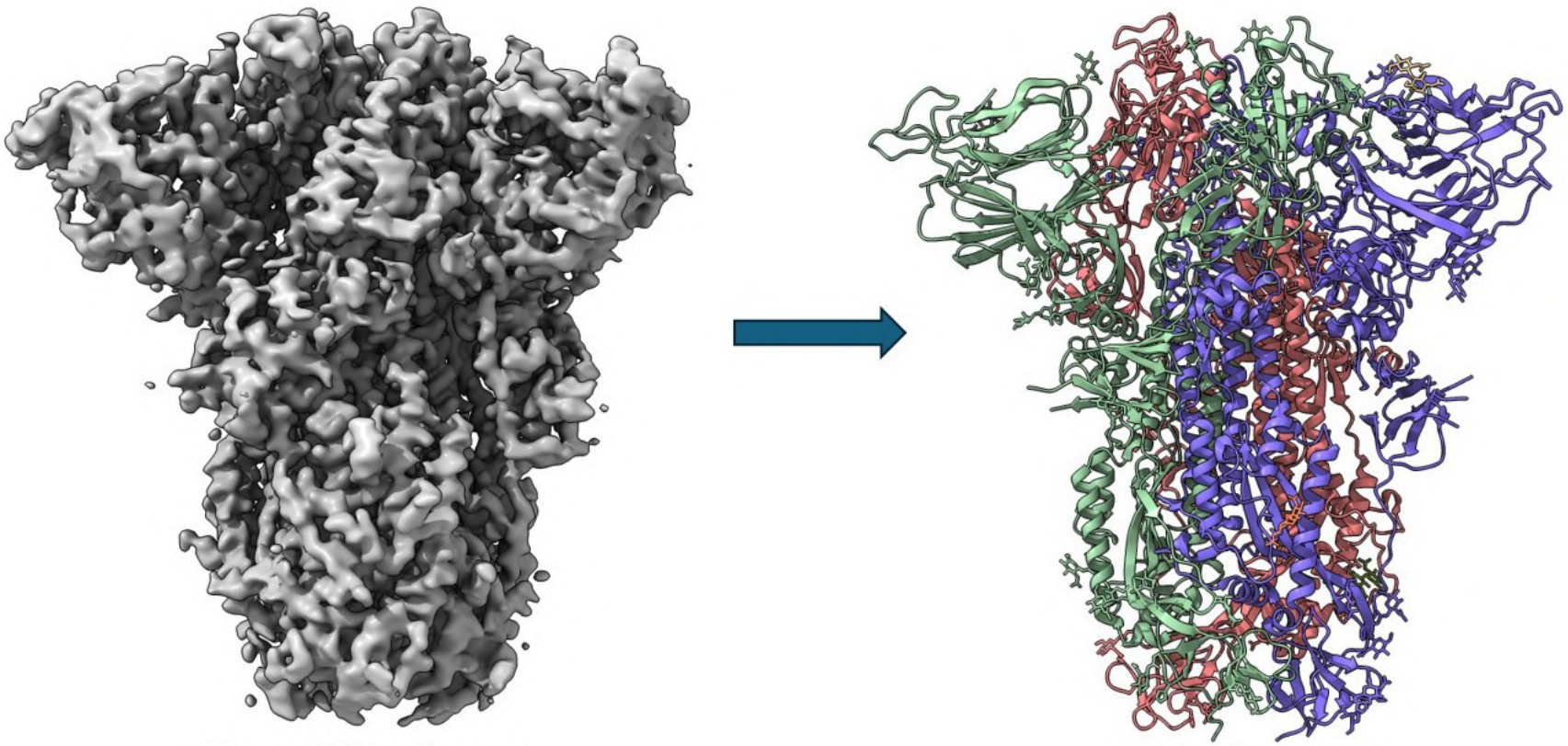
Cryo-EM Image Processing Workflow



3D Atomic Structure Modeling Task

Challenges:

- Low signal-to-noise ratio and local low resolution areas.
- Factorial growth in the number of possible Cα atom connections.
- Increased complexity with multi-chain protein structures.



3D cryo-EM density map
EMD ID: 38778

3D Atomic structure
PDB Code: 8XYM

Why *De Novo* Model Building in Cryo-EM?

- **Structures of Novel Proteins**
  - No homologous structures available in Protein Data Bank (PDB).
  - *De novo* modeling allows atomic-level structure determination.
- **Correcting Errors in Homology Models**
  - Homology models can introduce inaccuracies and biases.
  - *De novo* modeling reduces template bias.
- **Revealing Protein Dynamics**
  - Allows for the exploration of multiple conformational states of proteins.
  - Understanding dynamic processes like protein folding or interactions with ligands.
- **Structural Insights for Drug Design**
  - Many important drug targets are solved by cryo-EM, e.g. membrane proteins and ribosomes.
  - Accurate insights into protein mechanisms essential for rational drug design.

Cumulative Growth of Cryo-EM maps (as of April 2025)

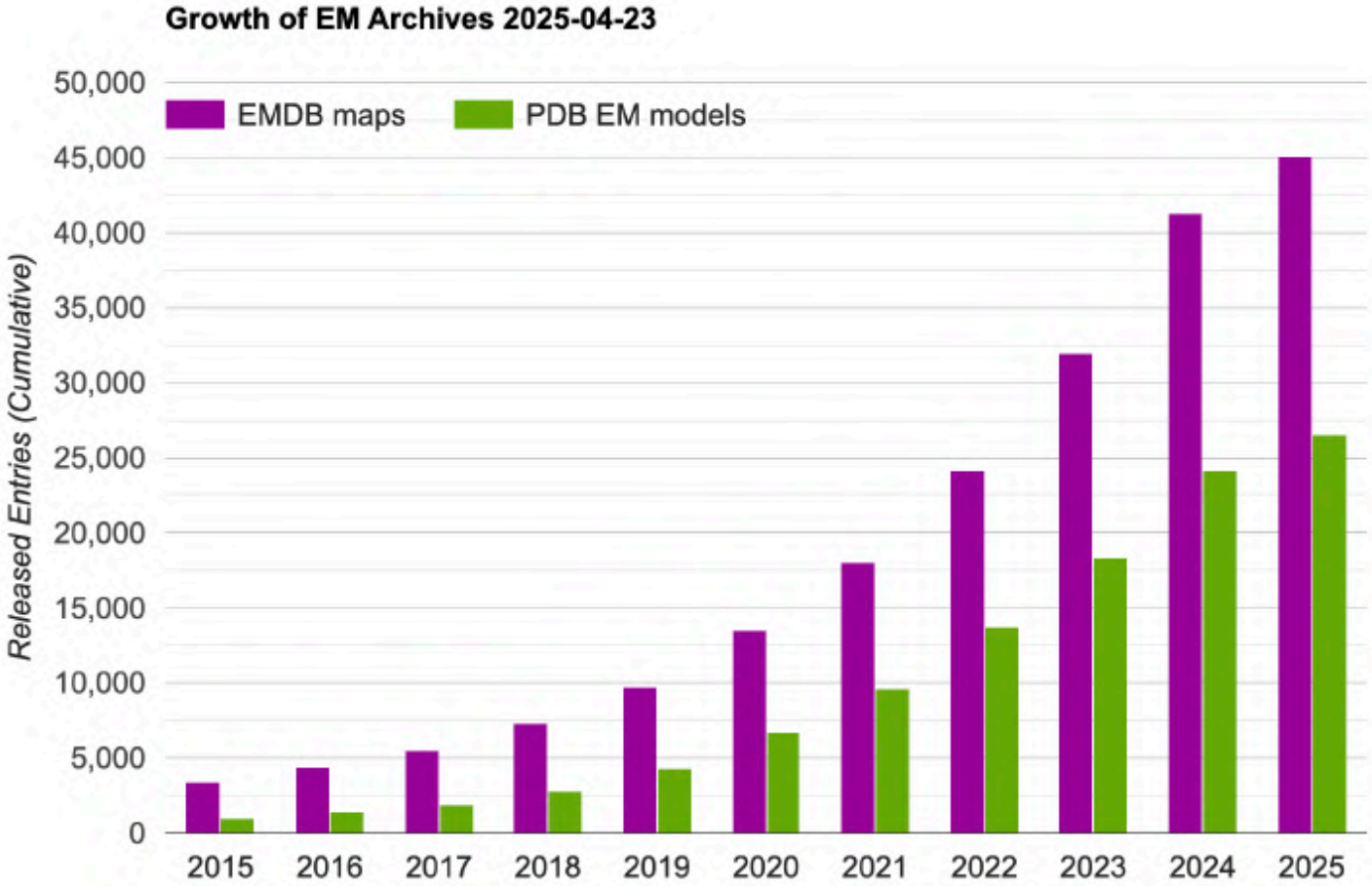


Figure: As of 2025, a total of 45,089 cryo-EM density maps have been deposited, with only 26,519 associated PDB structures which represents approximately **58.8%** available atomic structures. Data from `www.emdataresource.org`

Task Overview

Objective

Predict the 3D atomic structure of a protein from its 3D cryo-EM density map and associated amino acid sequence.

Given:

- A 3D cryo-EM density map
- An amino acid sequence

Predict:

- The 3D coordinates (x, y, z) of each Cα atom
- Form the Cα backbone structure of the protein

Existing Automated Methods for Atomic Modeling

De novo Setting:

- No reference structure.
- No manual interventions.
- Uses only **cryo-EM maps** and **amino acid sequences**.

Automated & De Novo Modeling Approaches:

- **Phenix** (phenix.map_to_model): Automated atomic structure modeling.
- **MAINMAST**: De novo modeling from cryo-EM maps using Minimum Spanning Tree and Tabu Search algorithm.
- **DeepTracer**: AI-based de novo modeling for cryo-EM maps using heuristic Traveling Salesman Problem (TSP) algorithm.
- **Cryo2Struct**: AI-based de novo modeling with Hidden Markov Model (HMM) and Viterbi Algorithm.
- **EModelX**: AI-based de novo modeling from cryo-EM integrating AlphaFold.

First: Preparing Cryo-EM Dataset for AI-based modeling

Why?

- **Instrumental & Experimental Variability**: Differences in microscope models, electron doses, detectors, and imaging conditions.
- **Heterogeneous Density Distributions**:
  - Variations in `density values`. e.g: $[-2.32, 3.91]$, $[-0.553, 0.762]$.
  - Variation in `voxel sizes`. e.g: 0.7 Å, 1.6 Å.
- **Noise and Artifacts**: Inconsistent local resolution and noise intensity affecting structure interpretation.

How?

- **Normalization Challenges**: How to standardizing data for AI-based modeling?
- **Labels**: How to generate labels for AI-based training?

Cryo2StructData

Cryo2StructData: A Large Labeled Cryo-EM Density Map Dataset for AI-based Modeling of Protein Structures

Nabin Giri, Liguo Wang and Jianlin Cheng. Scientific Data (2024).

https://doi.org/10.1038/s41597-024-03299-9

Cryo2StructData: **Comparison**

Table: Summary of the currently used number of experimental cryo-EM density maps in the datasets for various AI-based methods. **Cryo2StructData** contains cryo-EM density maps with resolution **up to 4 Å**. Data collected on 27 Match 2023.

	EMNUSS	Haruspex	DT	MA	**Cryo2StructData**
Approach	U-Net++	U-Net	U-Net	CNN+GNN	**Cryo2Struct**
Dataset Size	163	415	1,800	3,892	**7,600**
Train Size	120	293	1,440	3,715	**7,392**
Test Size	43	122	360	177	**208**
Released	×	×	×	×	✓

DT = DeepTracer & MA = ModelAngelo

*Pfab, Jonas, Nhut Minh Phan, and Dong Si. "DeepTracer for fast de novo cryo-EM protein structure modeling. ." PNAS 118.2 (2021)

*Jamali, Kiarash, et al. "Automated model building and protein identification in cryo-EM maps." Nature 628.8007 (2024): 450-457.

*Mostosi, Philipp, et al. "Haruspex: a neural network .. in cryo-electron microscopy maps." Angewandte Chemie International Edition 59.35 (2020)

*He, Jiahua, and Sheng-You Huang. "EMNUSS: a deep learning framework .. in cryo-EM maps." Briefings in bioinformatics 22.6 (2021)

Trends in Cryo-EM Map Resolution (as of April 2025)

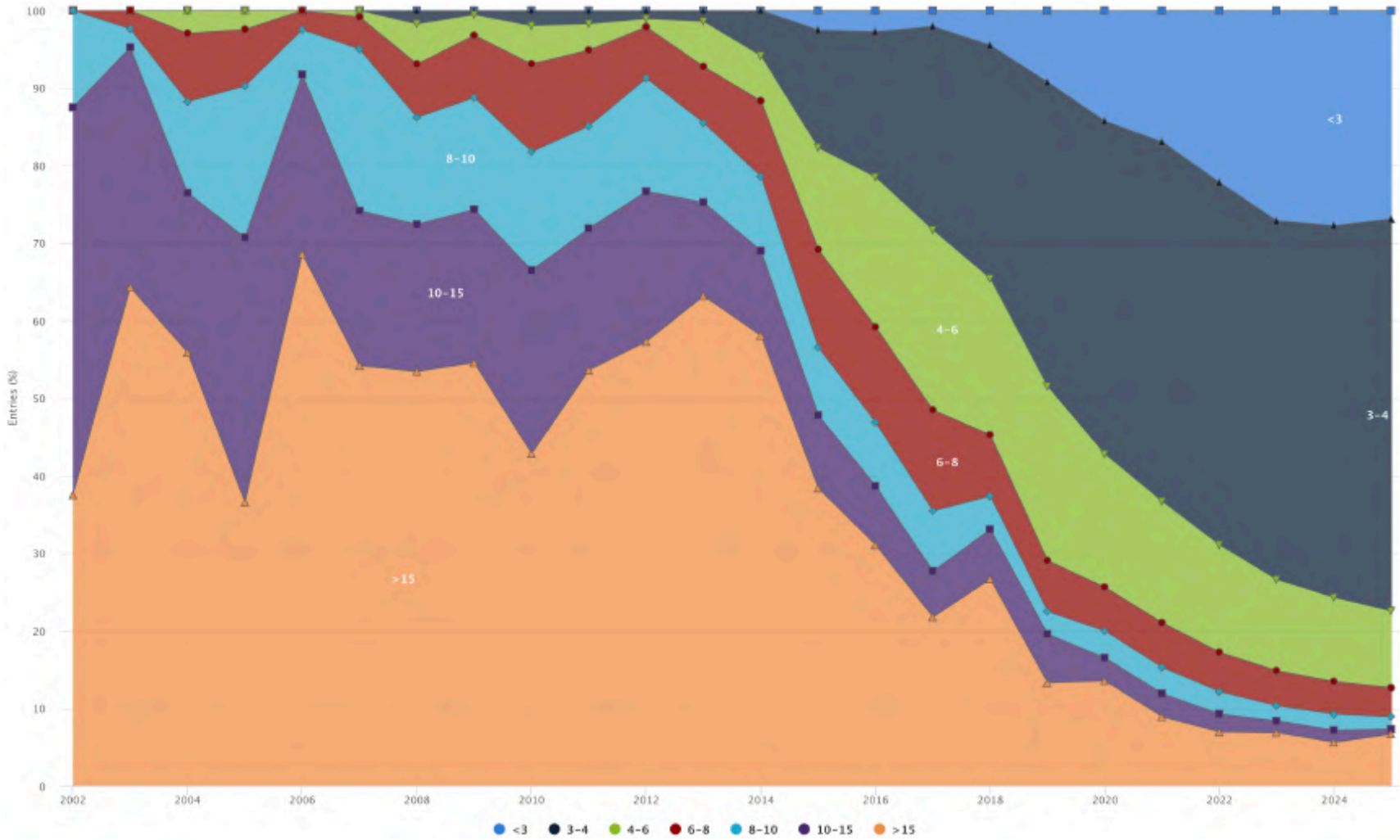


Figure: Resolution Distribution of Cryo-EM Density Maps in EMDB: **77.42%** of deposited maps have a resolution value **up to 4 Å**. Data from `ebi.ac.uk/emdb/`

Cryo2StructData: **Workflow**

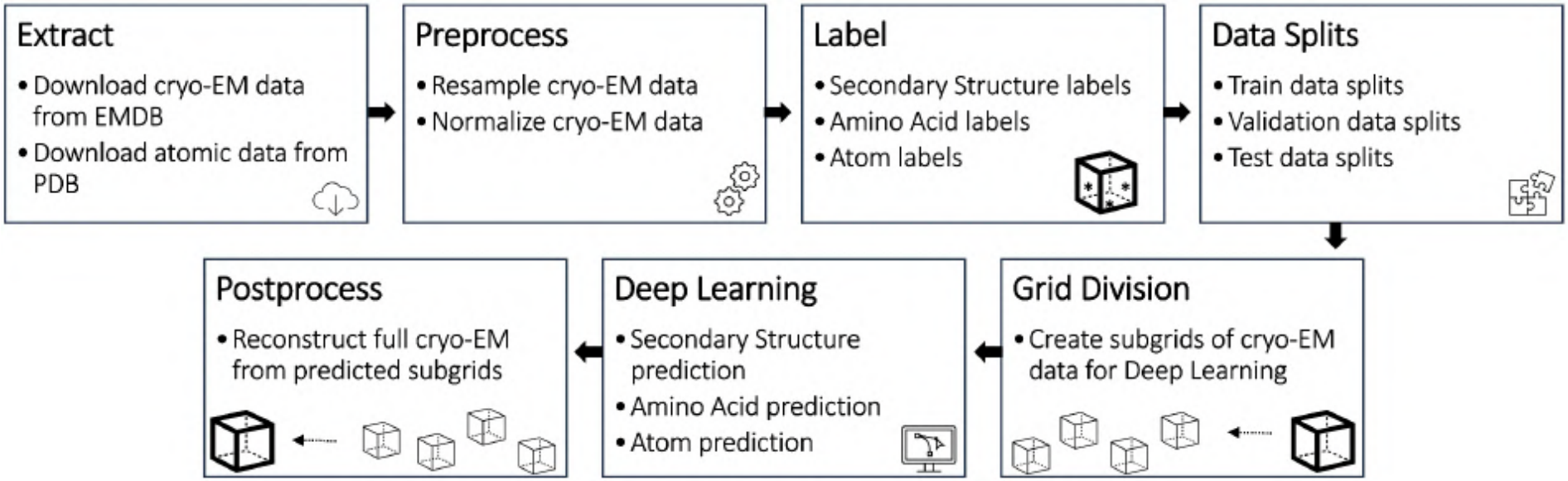
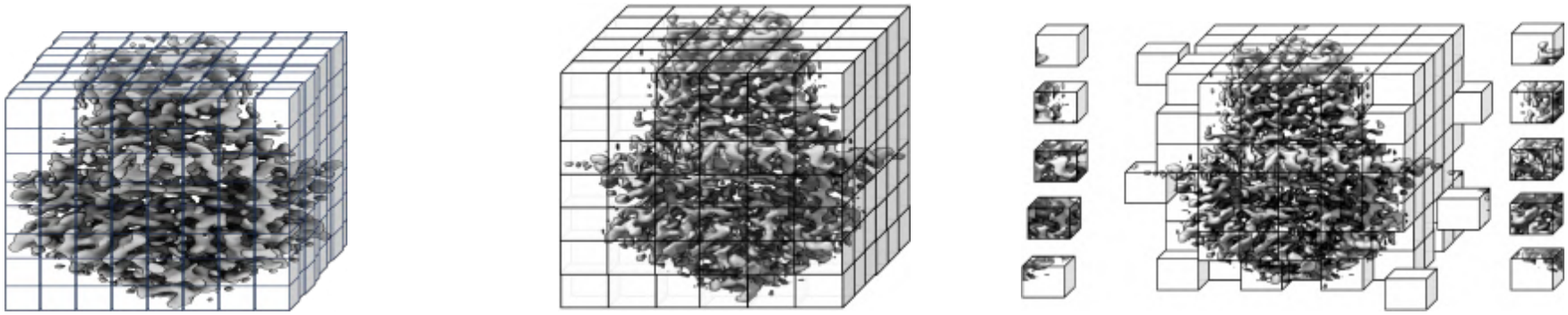


Figure: The data preparation and evaluation pipeline for Cryo2StructData.

Giri, Nabin, Liguo Wang, and Jianlin Cheng. "Cryo2structdata: A large labeled cryo-em density map dataset for ai-based modeling of protein structures." Scientific Data 11.1 (2024): 458.

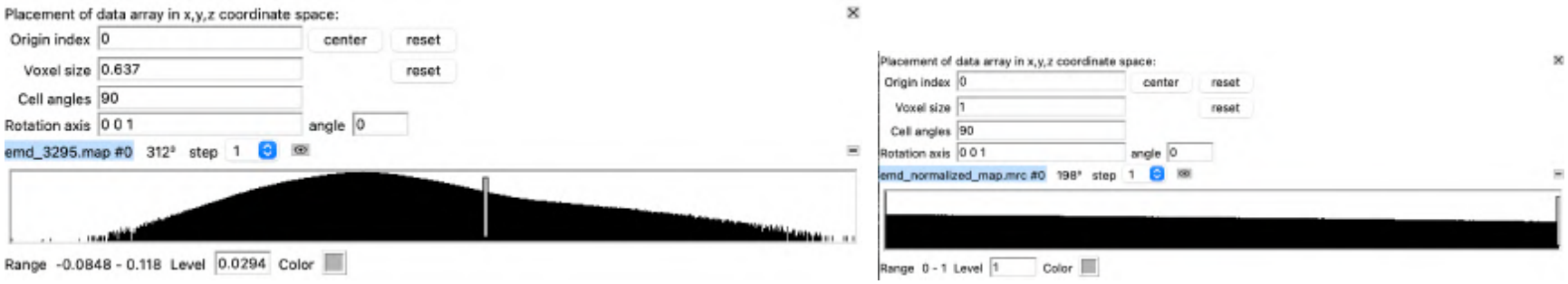
Cryo2StructData: **Resampling and Grid Division**



(a) (b) (c)

Figure: An example of density map grid resampling and division of a cryo-EM density map. (**a**) Density map (EMD-22898) in the original grid of 0.7 Å. (**b**) Resampled density map to the uniform grid size of 1 Å by trilinear interpolation. (**c**) Grid division of the density map where each cube is sub-grid with specific size. In this illustration, it is 1 Å× 1 Å× 1 Å. A coarser grid is shown for the purpose of illustration.

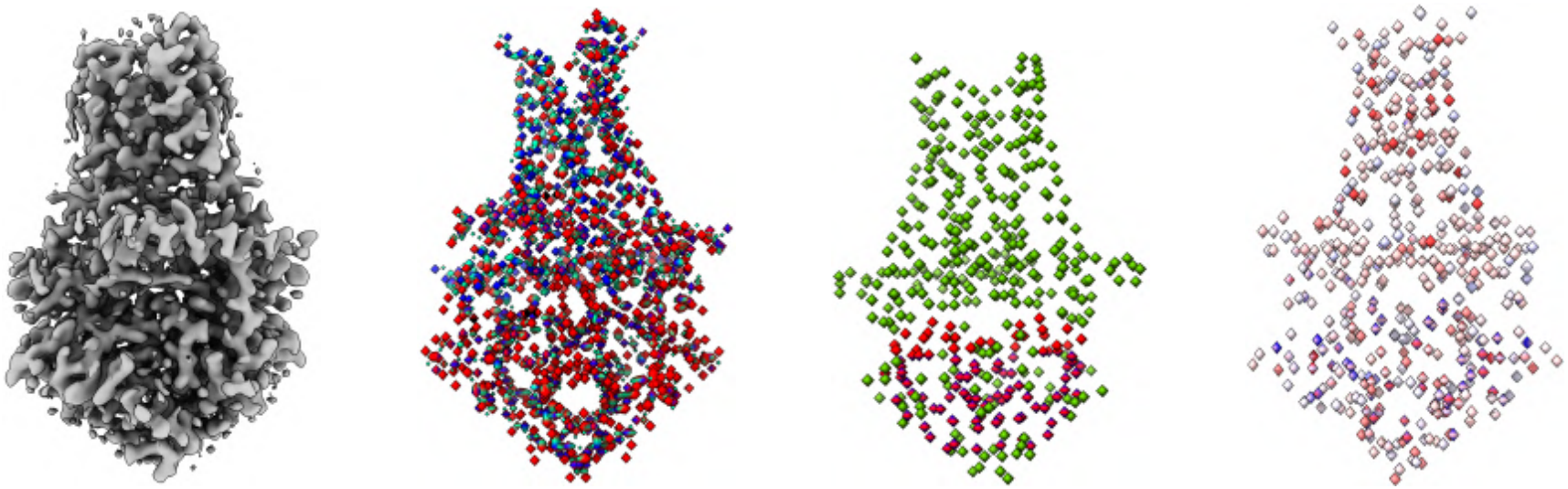
Cryo2StructData: **Normalization**



(a) (b)

Figure: Cryo-EM density map (EMD-3295). **(a)** Density values in experimental cryo-EM density map. **(b)** Density values after preprocessing (resampled and normalized). The **voxel size is 1 Å**, and the **density values are between [0 - 1]**. Percentile normalization by dividing density values by the 95th percentile of positive densities in the map. The image is extracted from volume viewer of UCSF Chimera.

Cryo2StructData: **Labels**



(a) (b) (c) (d)

Figure: (**a**) The density map (EMD-22898) in author recommended contour level of 0.7 (10.3 σ). (**b**) **Three different types** of protein backbone atoms (Cα, N, C) **labeled in different colors.** (**c**) **Three different** secondary structure elements **labeled in different colors.** (**d**) **Twenty different** amino acid **labeled in different colors.**

$$i = \lceil \left(\frac{\lfloor (z - origin_z) \rfloor}{voxel_z} \right); j = \lceil \left(\frac{\lfloor (y - origin_y) \rfloor}{voxel_y} \right); k = \lceil \left(\frac{\lfloor (x - origin_x) \rfloor}{voxel_x} \right) \quad (1)$$

Cryo2StructData: **Data Splits for Train & Validation**

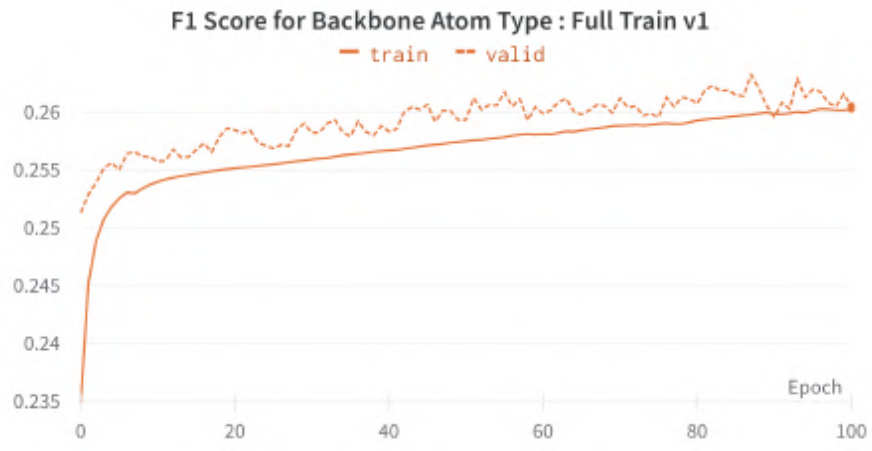
Table: The resolution distribution of the maps in the train and validation datasets. '**Full**' denotes the **entire dataset**, while '**Small**' denotes a smaller subset of the dataset, **25% of the full dataset.**

Low resolution value → **High-resolution map**. (fine atomic details, side chains visible clearly)

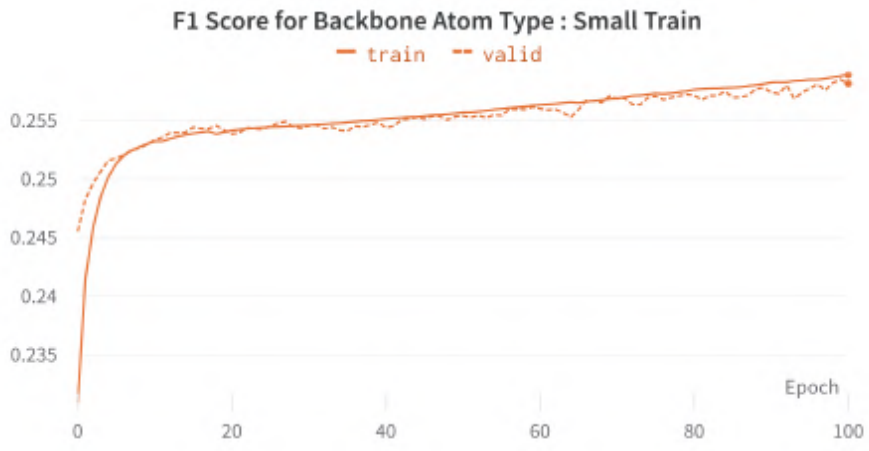
High resolution value → **Low-resolution map**. (blurry, secondary structure visible, atomic details missing)

	Train		Validation	
Reso.(Å)	**Full**	Small	**Full**	Small
1.0 - 2.0	**62 (0.9%)**	34 (2.1%)	**5 (0.7%)**	4 (2.2%)
2.0 - 3.0	**2,147 (32.3%)**	535 (31.8%)	**222 (30%)**	55 (29.4%)
3.0 - 4.0	**4,443 (66.8%)**	1,111 (66.2%)	**513 (69.3%)**	128 (68.5%)
Total	**6,652 (100%)**	1,680 (100%)	**740 (100%)**	187 (100%)

Cryo2StructData: **Validation By DL Training**



(a)



(b)

Figure: The training and validation F1 scores for transformer-model trained on Cryo2StructData. **(a)** Full Train v1. **(b)** Small Train. Same architecture as Cryo2Struct.

Cryo2StructData : A Large Labeled Cryo-EM Density Map Dataset for AI-based Modeling of Protein Structures

- **Preprocessed & Labeled** : High-quality cryo-EM dataset for structure modeling using AI methods.
- **Training, Validation, & Test Data** : Used to develop **Cryo2Struct**, an approach for atomic structure modeling.
- **Purpose-Built** : Designed specifically for modeling atomic structures from cryo-EM density maps.
- **Availability** : Open-sourced and is available in Harvard Dataverse. DOI : https://doi.org/10.7910/DVN/FCDG0W

Giri, Nabin, Liguo Wang, and Jianlin Cheng. "Cryo2structdata: A large labeled cryo-em density map dataset for ai-based modeling of protein structures." Scientific Data 11.1 (2024): 458.

Cryo2Struct

De novo atomic protein structure modeling for cryoEM density maps using 3D transformer and HMM

Nabin Giri and Jianlin Cheng. Nature Communication (2024).

https://doi.org/10.1038/s41467-024-49647-6

Cryo2Struct: Step I

- **Predict** $C\alpha$ voxels and their amino acid types in the cryo-EM density map.

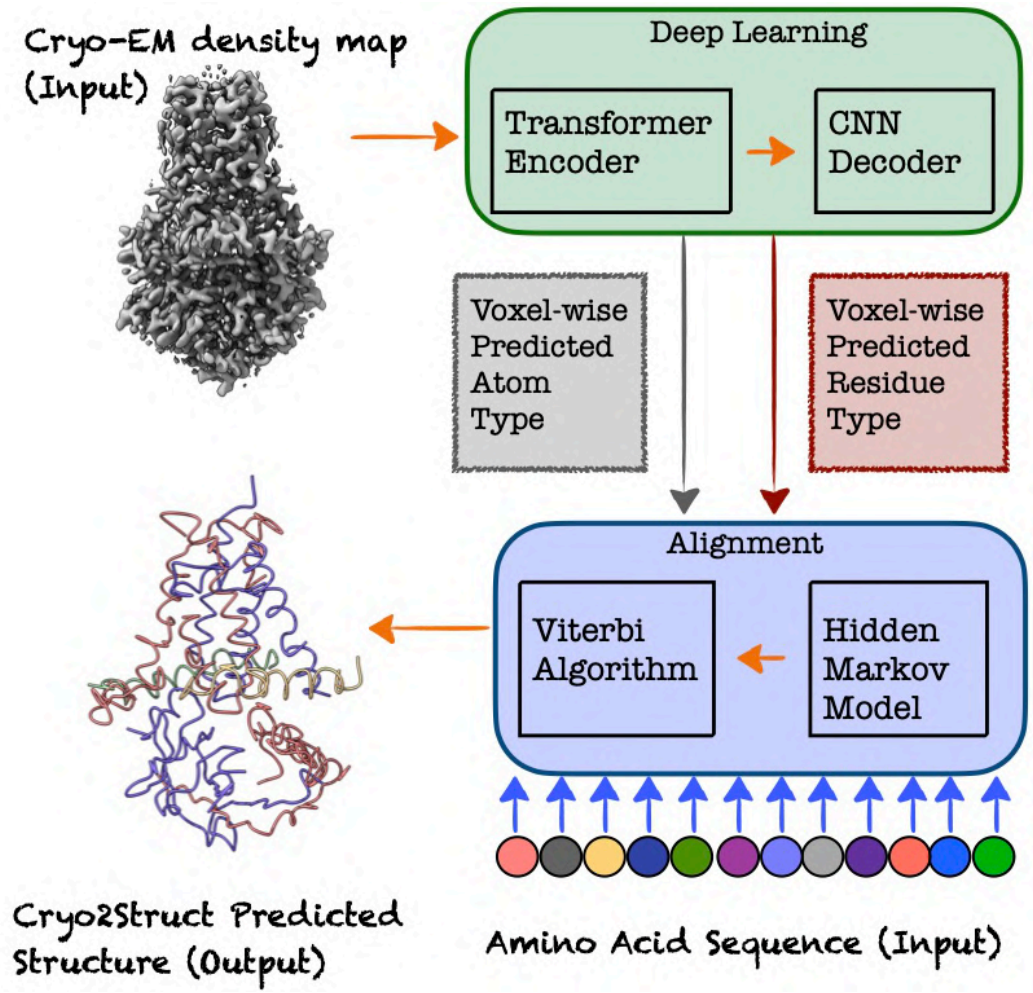


Figure: Cryo2Struct Architecture. Inference.

Cryo2Struct: Step I

- **Predict** Cα voxels and their amino acid types in the cryo-EM density map.

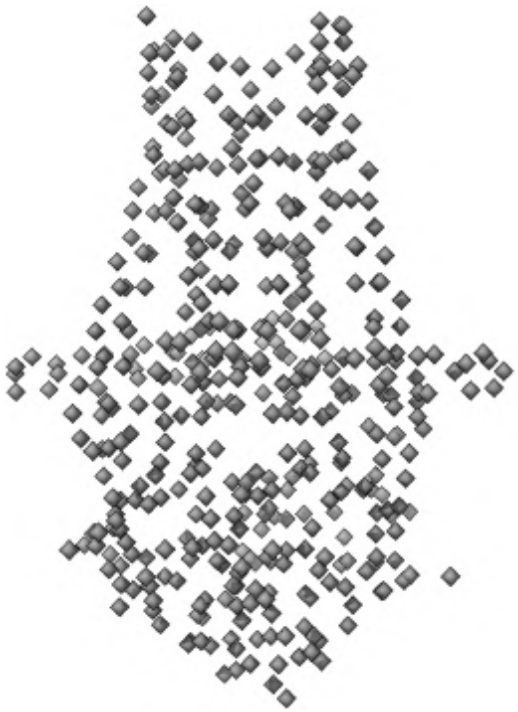


Figure: Cα voxels only.

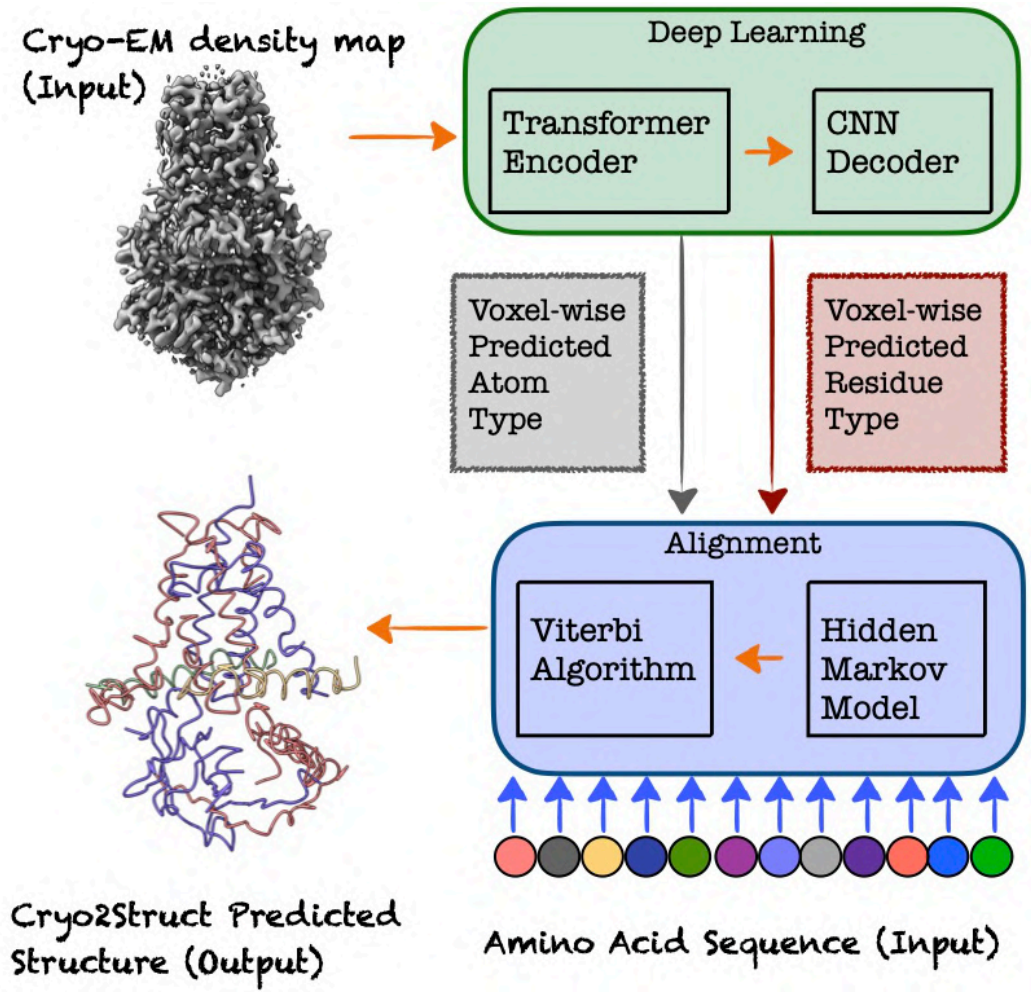


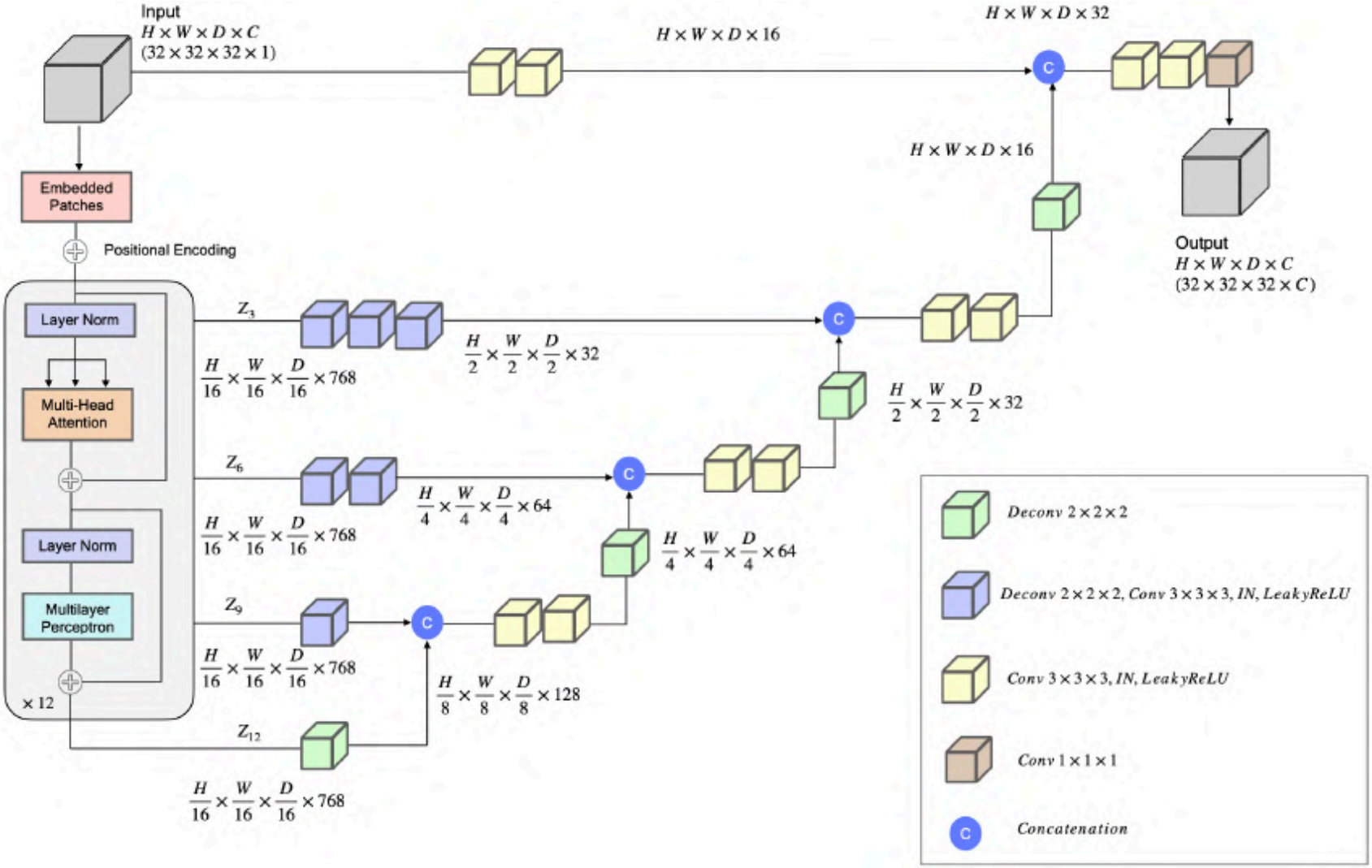
Figure: Cryo2Struct Architecture. Inference.

Cryo2Struct Training Regime

- **Input Processing:**
  Cryo-EM density maps are split into sub-grids of size $32 \times 32 \times 32$ to fit into a GPU memory of 32 GB. Input Shape : $32 \times 32 \times 32 \times 1$
- **Deep Learning Models:**
  - **Atom Type Prediction**: 4 classes (Cα, C, N, no atom presence).
  - **Amino Acid Type Prediction**: 21 classes (20 standard amino acids + 1 for no atom presence).
  - Prediction Shape : $32 \times 32 \times 32 \times C$
- **Addressing Class Imbalance:**
  Applied weighted cross-entropy loss. Weights computed for each sub-grid.
- **Optimization and Training:**
  - **Optimizer**: NADAM with Distributed Data Parallel (DDP) on multi-node, multi-GPU systems.
  - **Computing** : 12 nodes $\times$ 6 V100 GPUs.
  - **Monitoring & Training**: Weights & Biases tool for tracking training progress, with Early Stopping & ReduceLROnPlateau.

Cryo2Struct: Step I: Deep Learning : 92.28 M Parameters.

- **Classify** voxels and their amino acid types in the cryo-EM density map.



Cryo2Struct: Step II: **Alignment**

Hidden Markov Model

A statistical model used to describe the probabilistic relationship between a sequence of observations and a sequence of hidden states.

- Observations.
  - Are the amino acid sequence.
- Transition Probability Matrix.
  - Defines the probability of moving from one hidden state to another.
- Emission Probability Matrix.
  - Defines the probability of emitting an observation from a hidden state.
- Initial Probability.
  - Chance of emitting the first amino acid ÷ Total chance from all states emitting it.

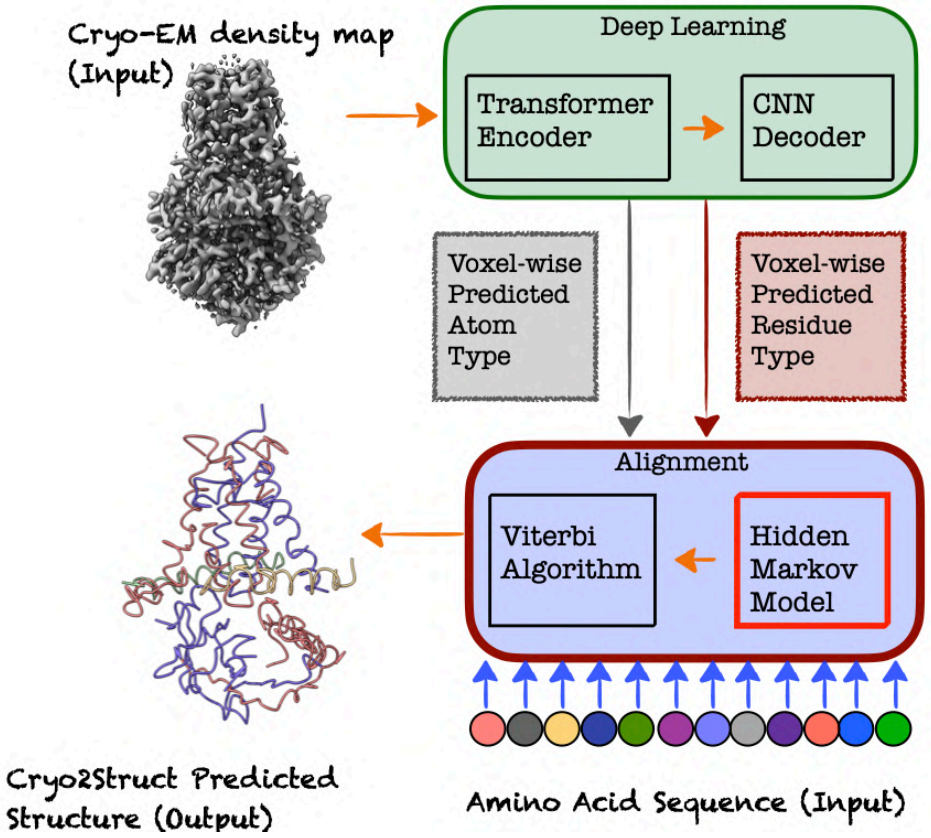


Figure: Cryo2Struct Alignment. Inference.

Cryo2Struct: Step II: **HMM Transition Probability**

- Predicted Cα voxels are used as hidden states. K hidden states.
- Compute pairwise distances using Eq. (2)
- Convert distances to probabilities using a modified Gaussian PDF of Eq. (3):
  - **Mean (μ)** $= 3.8047$ Å
  - **Std Dev (σ)** $= 0.036$ Å
  - **Scaling factor (Λ)** $= 10$
- Shape : $K \times K$. Row sums to 1.

$$\sqrt{(x_2 - x_1)^2 + (y_2 - y_1)^2 + (z_2 - z_1)^2} \tag{2}$$

$$f(x) = \frac{1}{\Lambda\sigma\sqrt{2\pi}} e^{-(x-\mu)^2/2(\Lambda\sigma)^2} \tag{3}$$

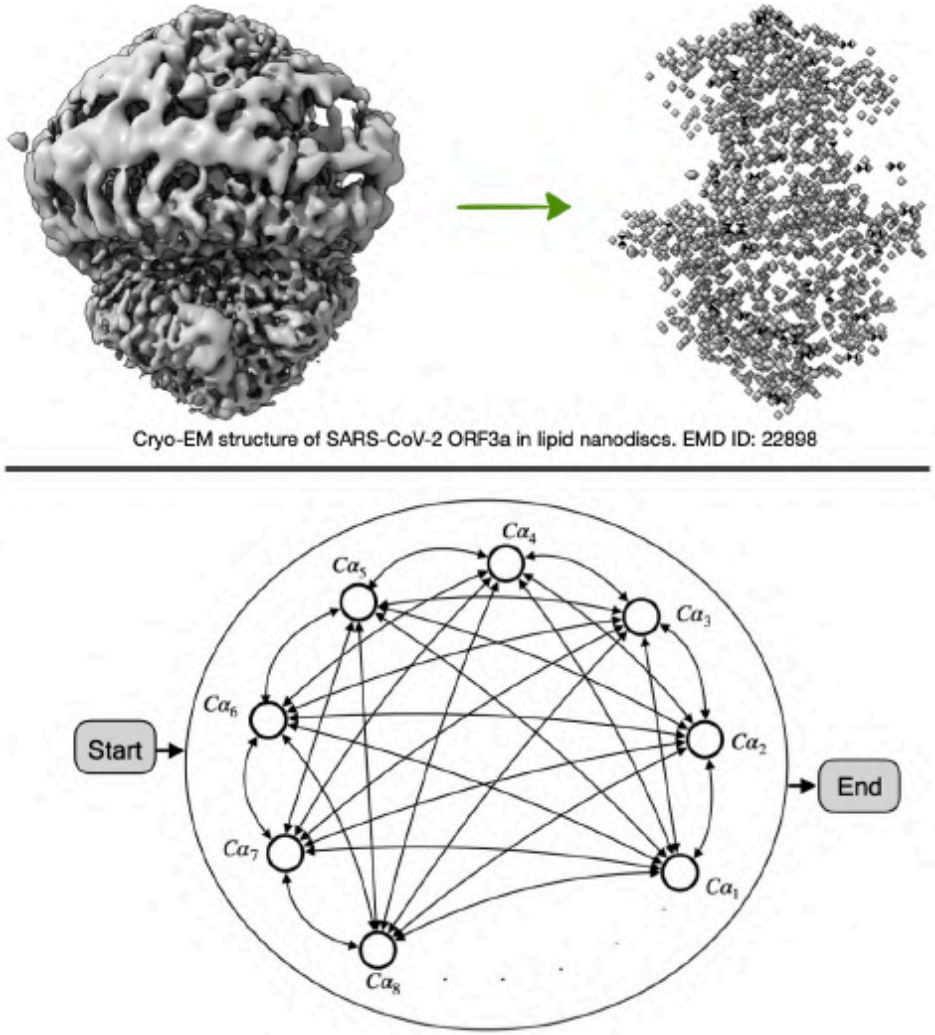


Figure: HMM Transition Probability

Cryo2Struct: Step II: **HMM Emission Probability**

- For each Cα state (voxel), the emission probability is derived from:
  - Predicted amino acid type probability **n**
  - Natural background frequency **m** of 20 amino acids
- Computed using geometric mean : $\sqrt{\mathbf{n} \times \mathbf{m}}$
- Shape : $K \times 20$. Row sums to 1.

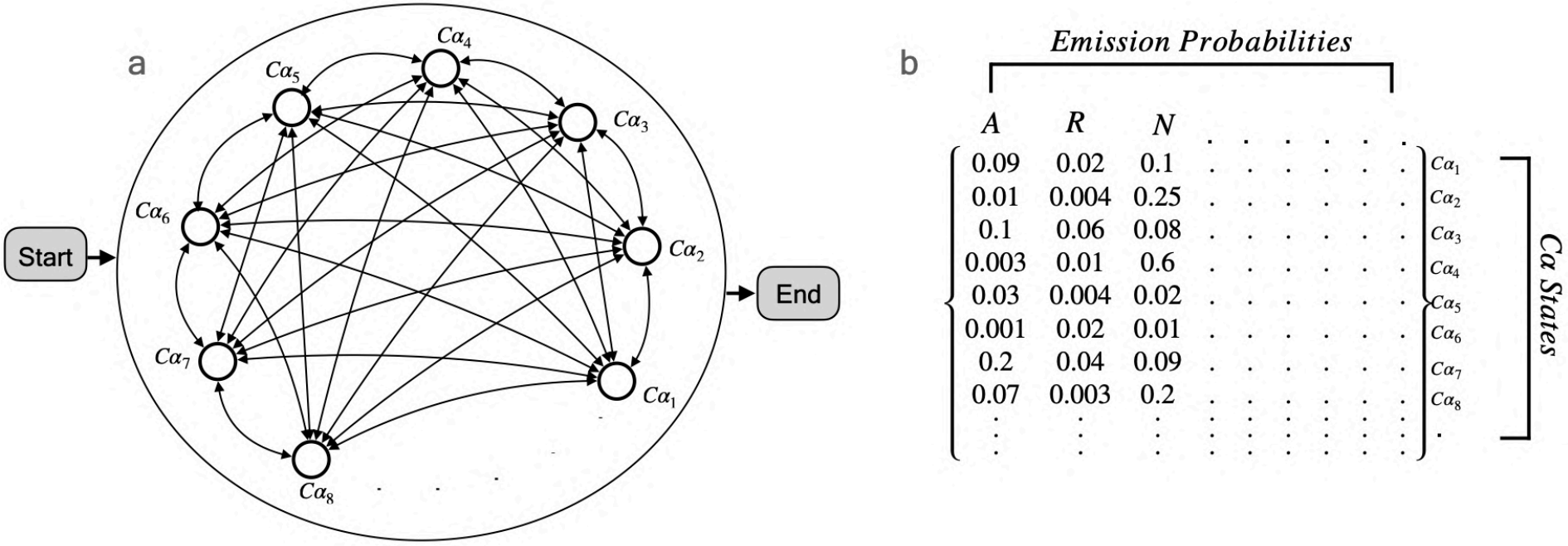


Figure: A HMM used for aligning protein sequences with predicted Cα atoms (voxels) to generate protein backbone traces.

Cryo2Struct: Step II: **HMM Emission Probability**

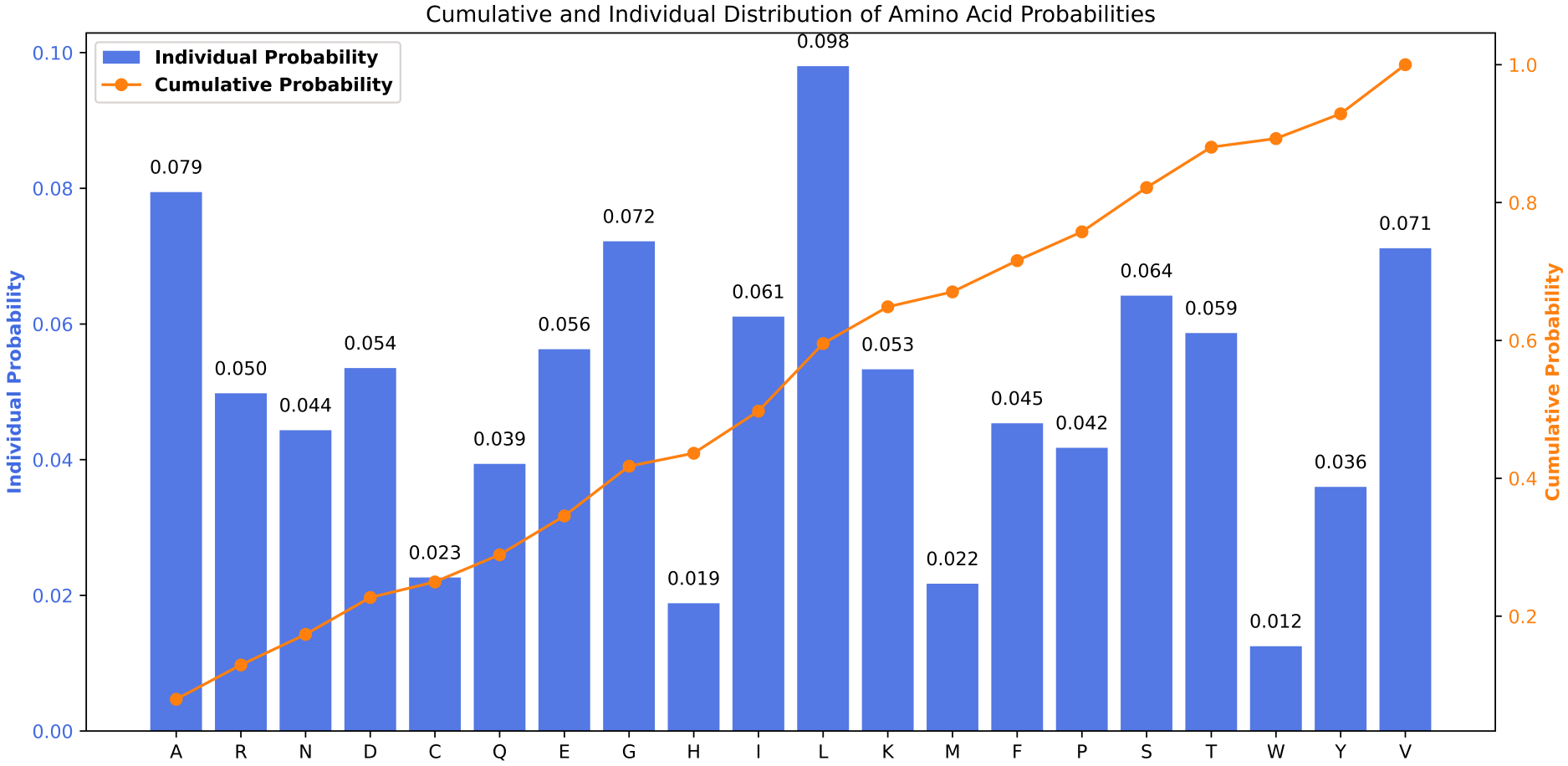


Figure: Prior amino acid probabilities, which are combined with the probabilities of amino acid types predicted by deep learning to construct the emission probabilities of amino acid types in the HMM. **Computed from the Cryo2StructData training dataset.**

Cryo2Struct: Step III: **Customized Viterbi**

- Align **sequence (O)** with the **HMM model** to determine the most likely **C**α **state sequence (X)**
- **O** $= O_1, O_2, .. O_T$, where $O_t \in$ V; V: the set of 20 standard amino acids
- **X** $= x_1, x_2, .. x_T$ where $x_t \in$ S; S: the set of Cα hidden states
- The predicted **C**α **sequence forms the protein backbone structure.**
- For multiple chains, remove all Cα states from HMM and build a new HMM.

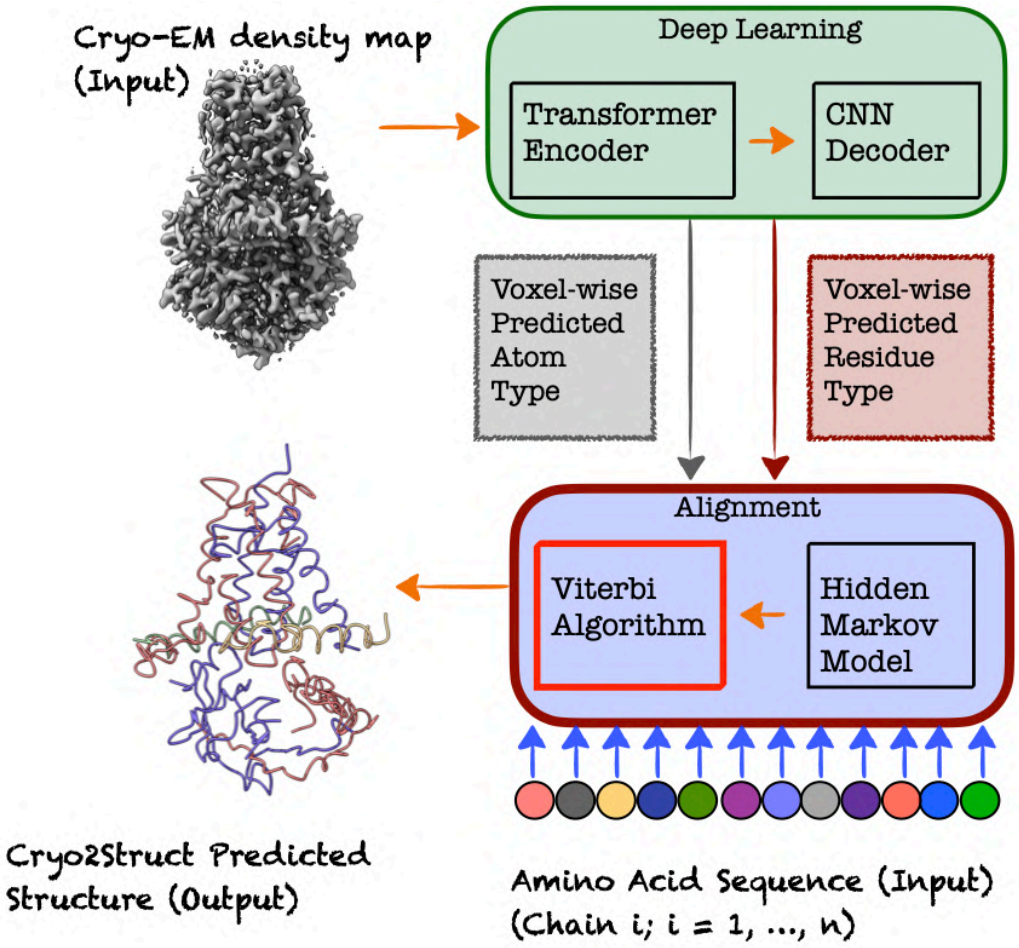


Figure: Cryo2Struct Architecture

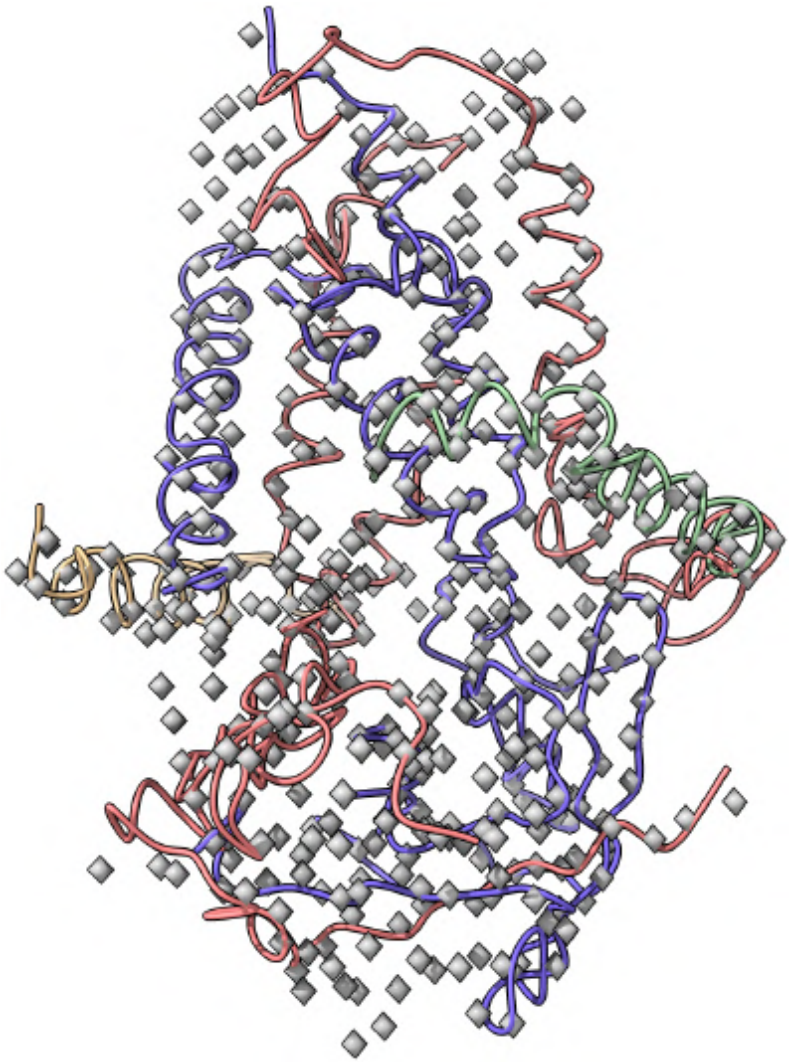
Cryo2Struct: Step III: **Customized Viterbi Algorithm**

Algorithm 1 The customized Viterbi algorithm

```
function MODIFIED-VITERBI(O, S, Π, γ, δ): X
    for each state i = 1, 2, ..., K do
        T_1[i, 1] ← Π_i · δ_{i,O_1}
        T_2[i, 1] ← 0
    end for
    visited_states ← {}
    for each observation j = 2, 3, ..., T do
        for each state i = 1, 2, ..., K do
            if i ∉ visited_states then
                path ← T_1[i, j−1] · γ_i · δ_{i,O_j}
                T_1[i, j] ← max_k(path)
                T_2[i, j] ← arg max_k(path)
            end if
        end for
        z_j ← arg max_k(T_1[:, j])
        visited_states ← visited_states ∪ {z_j}
    end for
    z_T ← arg max_k(T_1[:, T])
    x_T ← S_{z_T}
    for j = T−1, T−2, ..., 1 do
        z_{j−1} ← T_2[z_j, j]
        x_{j−1} ← S_{z_{j−1}}
    end for
    return X
end function
```

Figure: Customized Viterbi algorithm implemented in C++ with -O3 optimization, compiled as a shared library, and linked to Python for HMM construction. Time: $\mathcal{O}(T \times S^2)$, Space: $\mathcal{O}(T \times S)$.

Cryo2Struct Modeled Structure



(a) Known PDB Cα Voxels and Modeled Cα structure overlay.

(b) Known PDB Structure (Green) and Modeled Structure (Blue) overlay.

Cryo2Struct Modeled Structure

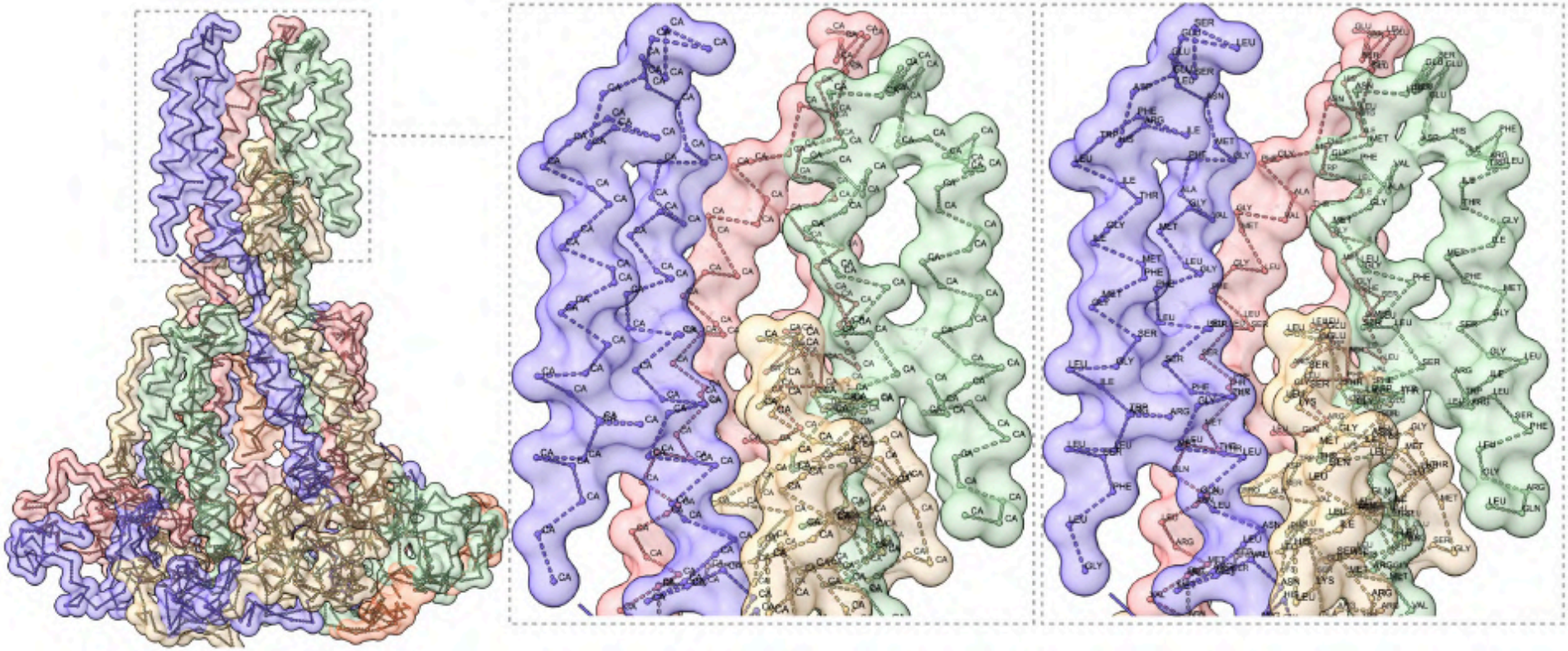


Figure: The skeleton of the Cryo2Struct modeled structure for a test cryoEM density map having <**25% sequence identity** with the training data (EMD ID: 41624; resolution 2.8 Å), where each chain is colored differently. The modeled structure has 1,585 amino acid residues; and the Cα match score is 73.8%.

Confidence Score: Visual Interpretation

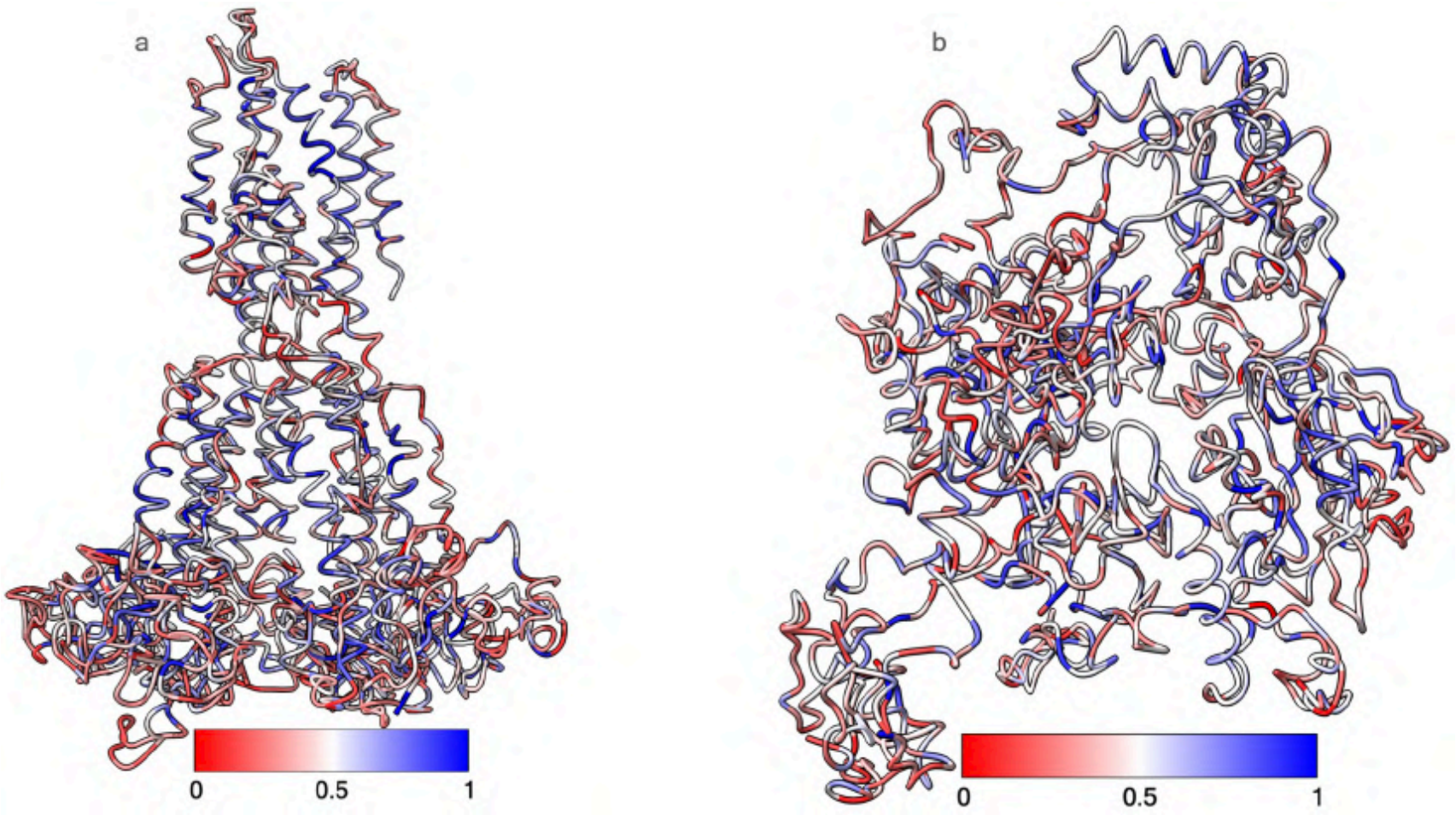


Figure: Residue-wise confidence scores mapped and visualized via a color spectrum. **(a)** EMD ID: 41624, released on 2023-09-13. **(b)** EMD ID: 34402, released on 2023-08-02. Both have resolution 2.8 Å. Both have **<25% sequence identity with training data**.

Confidence Score: Visual Interpretation

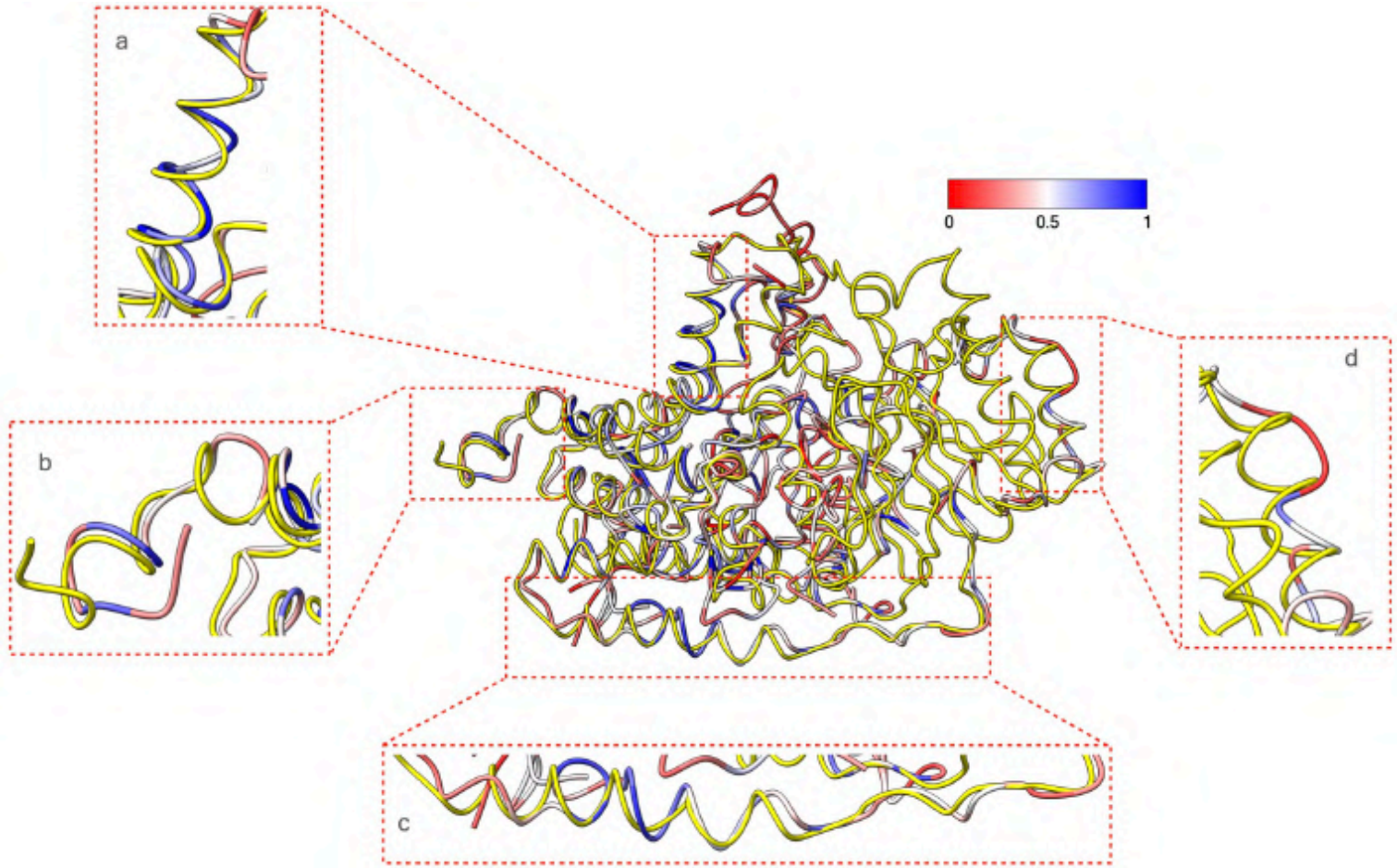


Figure: The modeled structure shares <**25% sequence identity** with the training dataset proteins. PDB ID: 8B0N in Yellow. (**a**) High confidence scores, particularly within helical motifs. (**b**) Region with varying confidence. (**c**) Extended segment with varying confidence. (**d**) Low confidence scores as helix deviates.

(1) Comparison with Phenix. 128 cryo-EM density maps

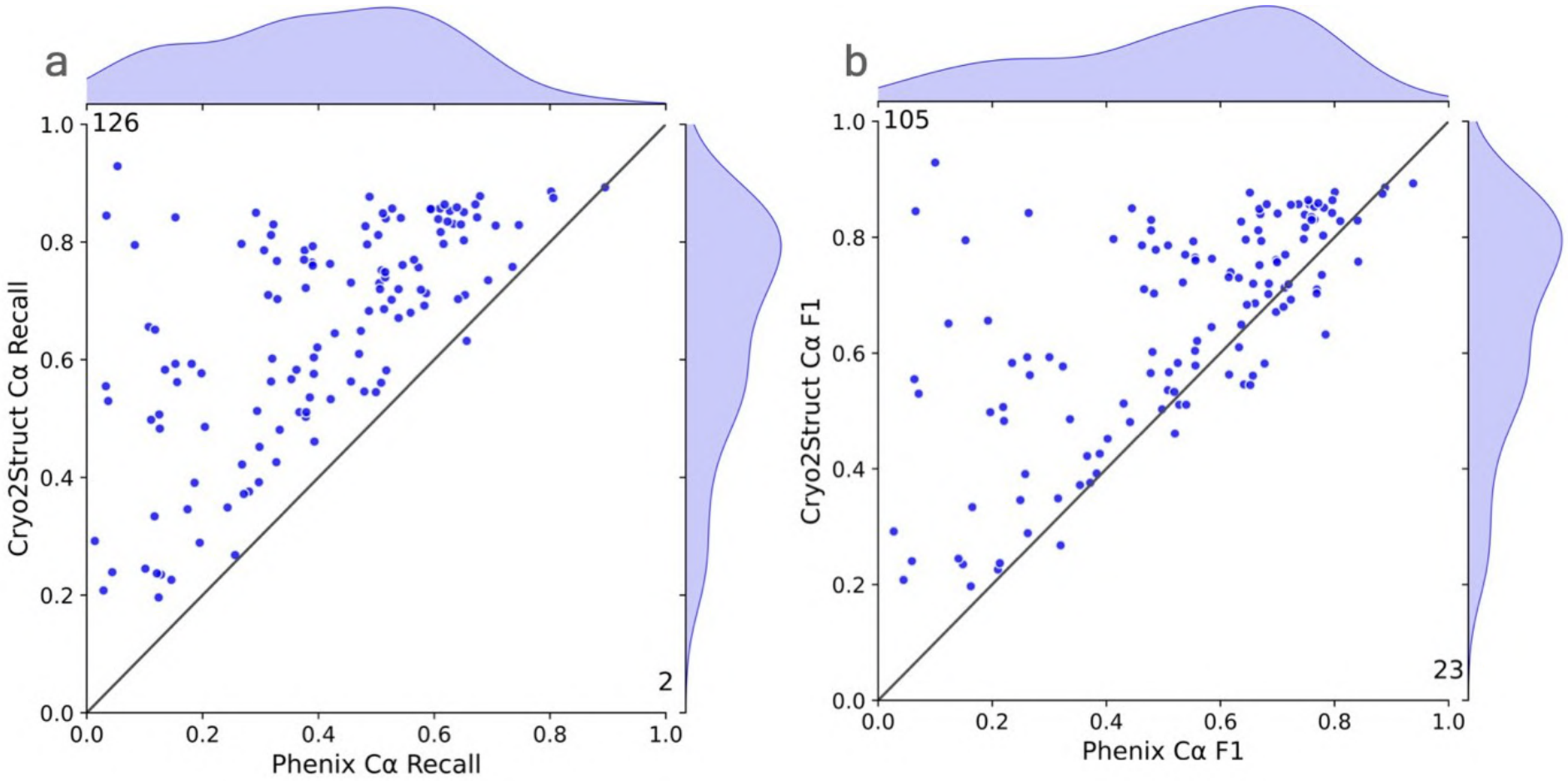
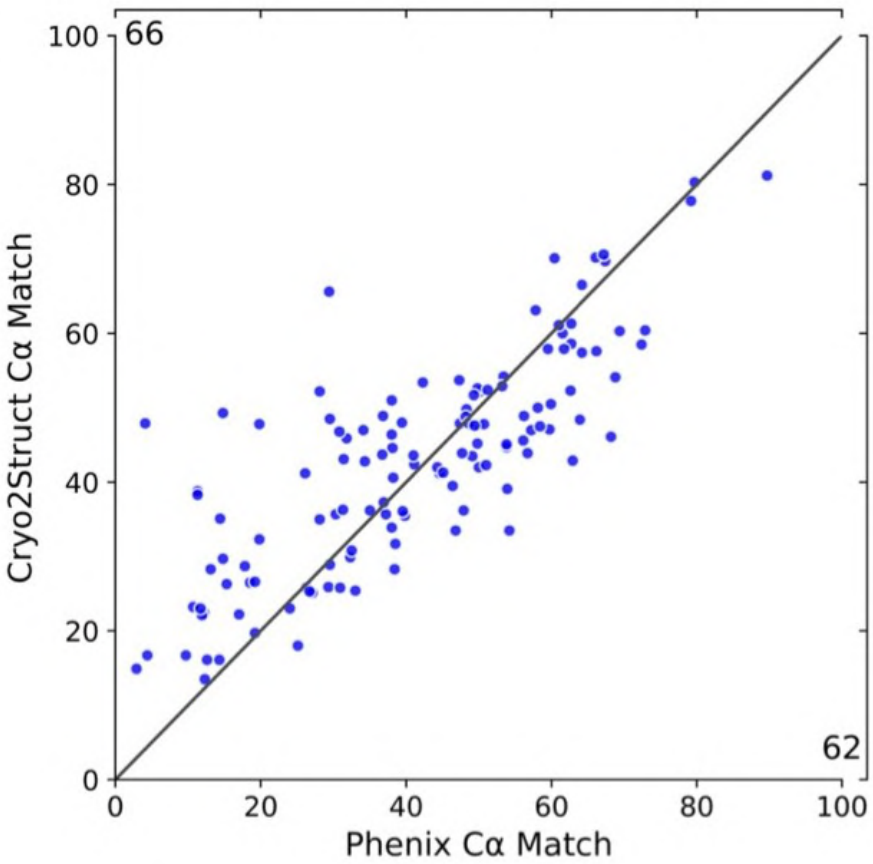


Figure: A dot (structure) above the 45 degree line indicates that Cryo2Struct has higher score than Phenix for the map. **(a) Recall**, **(b) F1-Score**. Map's resolution ranges from **2.08 Å to 5.6 Å**. The average resolution is 3.68 Å. The number of amino acid residues varies from **448 to 8,416.**

(2) Comparison with Phenix. 128 cryo-EM density maps



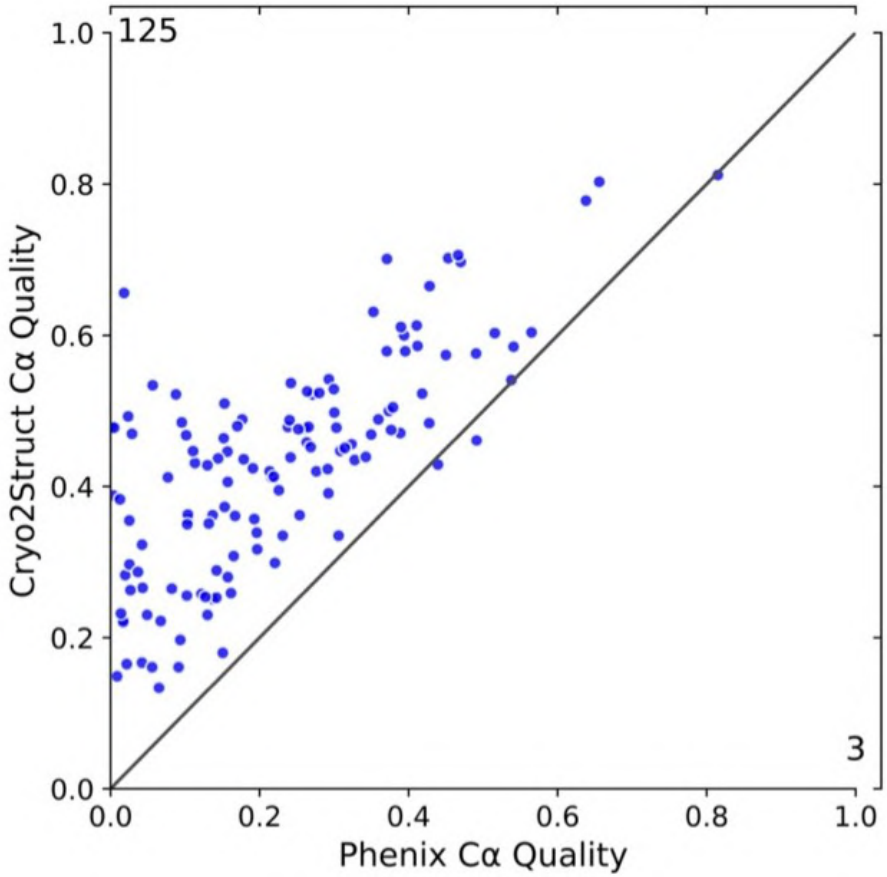


Figure: A dot (structure) above the 45 degree line indicates that Cryo2Struct has higher score than Phenix for the map. **Cα Match** and **Cα Quality**. Map's resolution ranges from **2.08 Å to 5.6 Å**. The average resolution is 3.68 Å. The number of amino acid residues varies from **448 to 8,416.**

Comparison with Phenix. 128 cryo-EM density maps

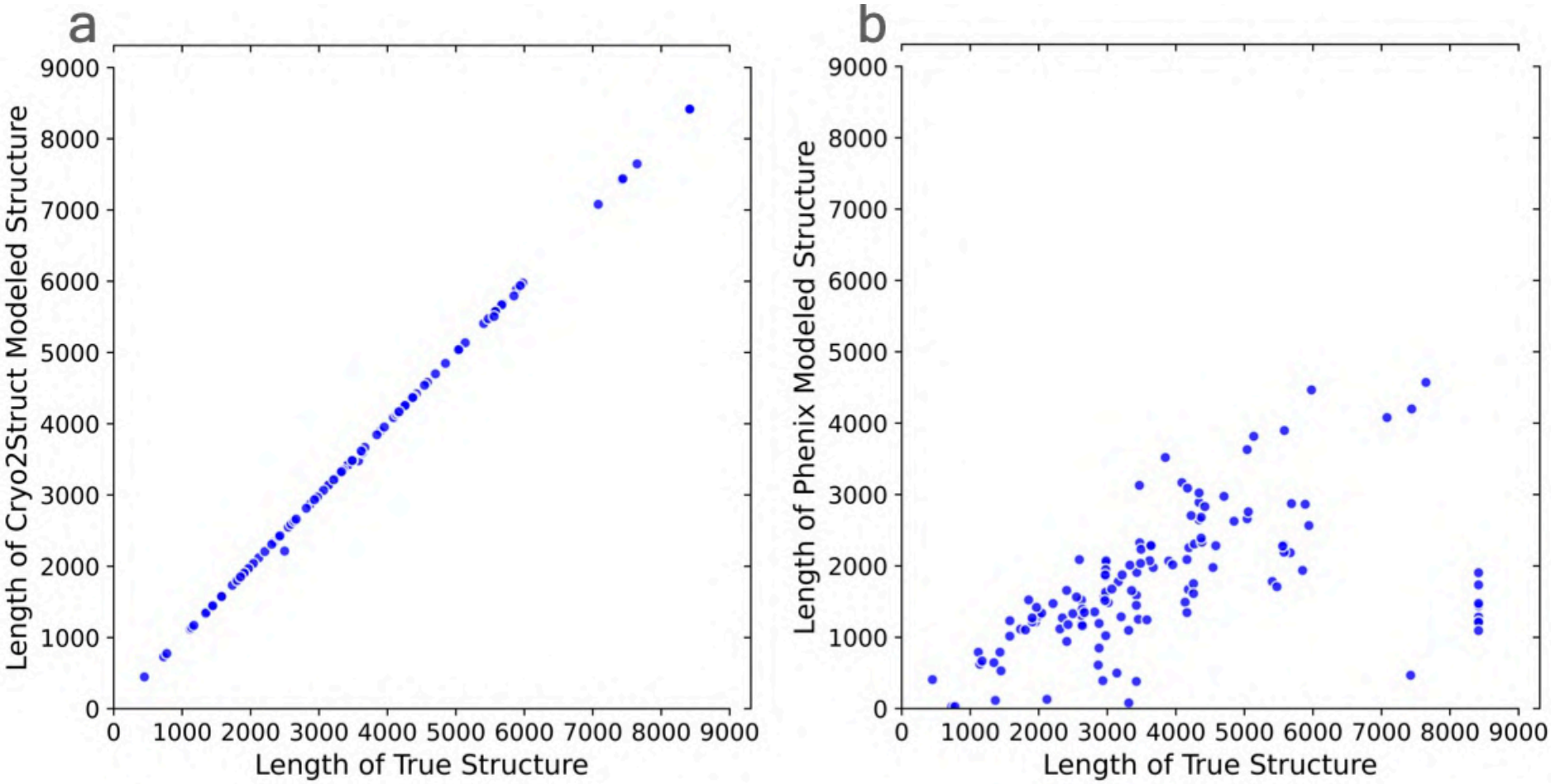
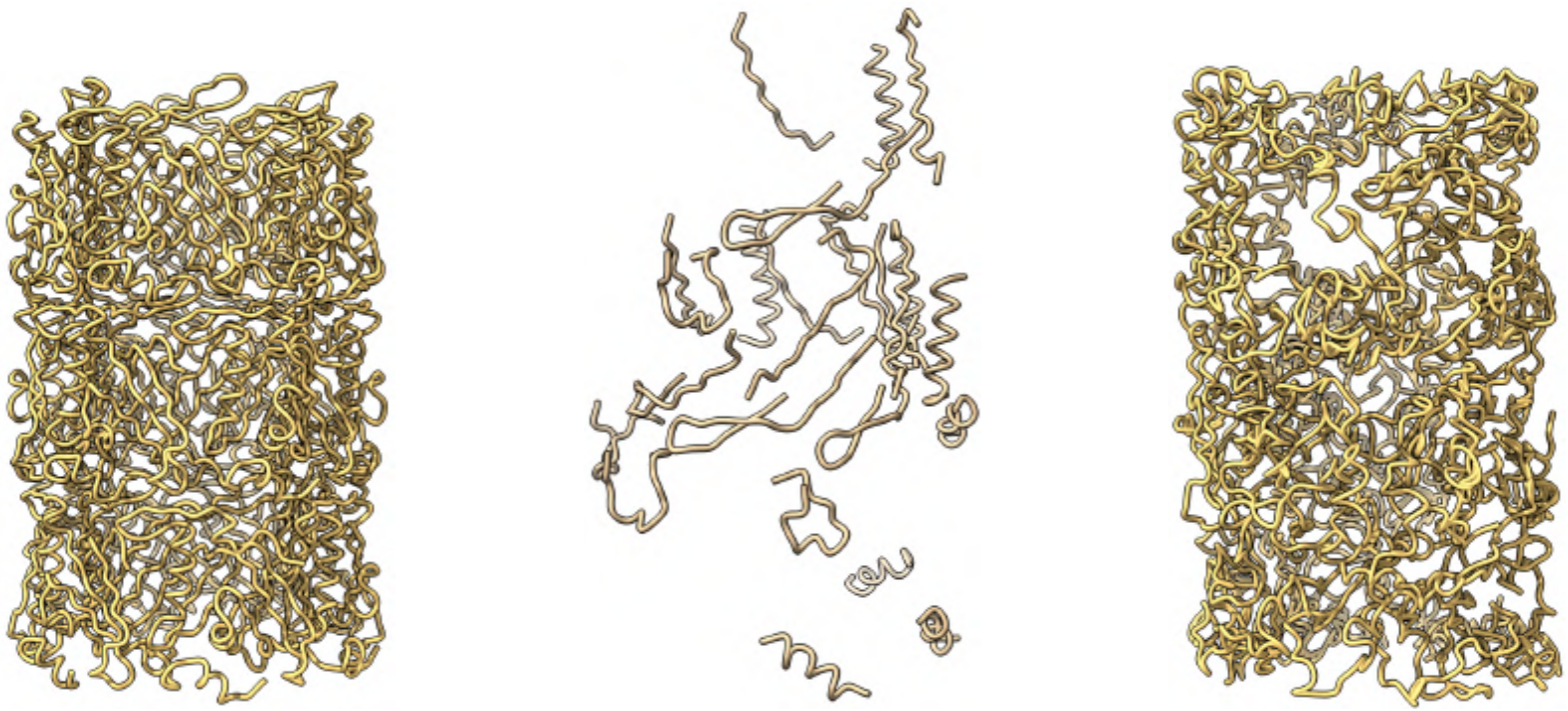


Figure: Completeness of Structures. True Structure is the Known PDB Structure. **(a) Cryo2Struct. (b) Phenix.**

Incomplete & Complete Structures



(a) (b) (c)

Figure: **(a)** Reference PDB structure (ID: 5W5F) with 2,934 Cα atoms. **(b)** Phenix.map_to_model structure with 392 Cα atoms **(Cα match score: 42.3%, Cα quality score: 0.06)**. **(c)** Cryo2Struct modeled structure with 2,932 Cα atoms **(Cα match score: 53.4%, Cα quality score: 0.53)**. EMD ID: 8767, 3.4 Å resolution.

Zheng, Weili et al. "Refined Cryo-EM Structure of the T4 Tail Tube: Exploring the Lowest Dose Limit." Structure.

Comparison with Phenix: In Numbers

Cα Identification			Structure Alignment	
Method/Metric	Recall	F1	Cα Match	Cα Quality
Cryo2Struct	**65%**	**66%**	**43.4%**	**0.43**
Phenix	40%	52%	41.2%	0.23

Table: Cα identification and structure alignment evaluation for Cryo2Struct and Phenix.

Method	EMD-ID	Residues	Weight (kDa)	Unique Protein Chains	Time (hrs)
Phenix	9565	2,395	297.39	1	129
Cryo2Struct	**40492**	**8,828**	**1,013.55**	**10**	**9**

Table: Compute time comparison between Cryo2Struct and Phenix.

Cα Quality:

$$\text{C}\alpha\ \text{Quality} = \frac{\text{C}\alpha\ \text{Match} \times \#\ \text{of modeled residues}}{\#\ \text{of residues in known PDB protein structure}}$$

Comparison with Phenix. 128 cryo-EM density maps

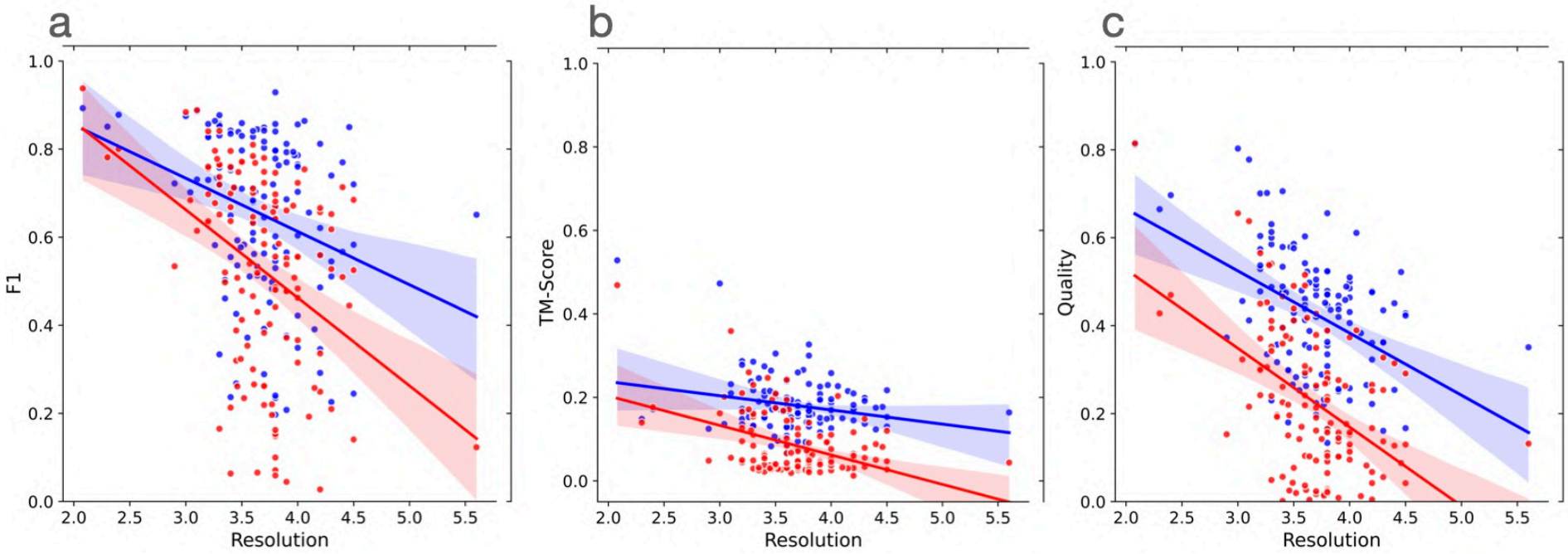


Figure: Blue dots: Cryo2Struct. Red dots: Phenix. Solid line: linear regression line, colored area represents 95% confidence interval. **(a) F1 Score**, **(b) TM-Score**, **(c) Quality Score**. In all scores, higher is better. The confidence interval is narrower in the resolution range [3–4.5 Å] where there are more data points.

Evaluation Cryo2Struct. 500 cryo-EM density maps

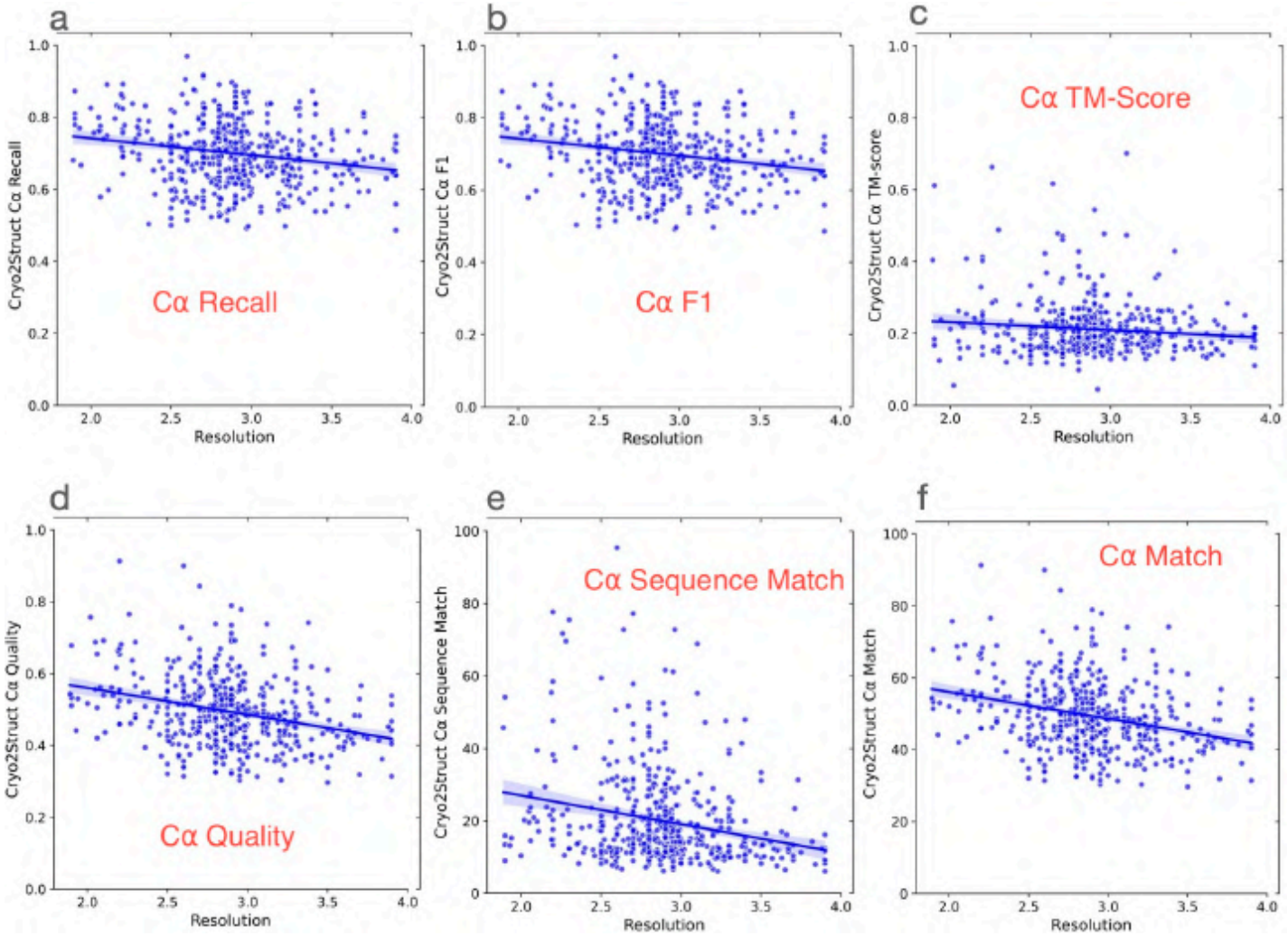
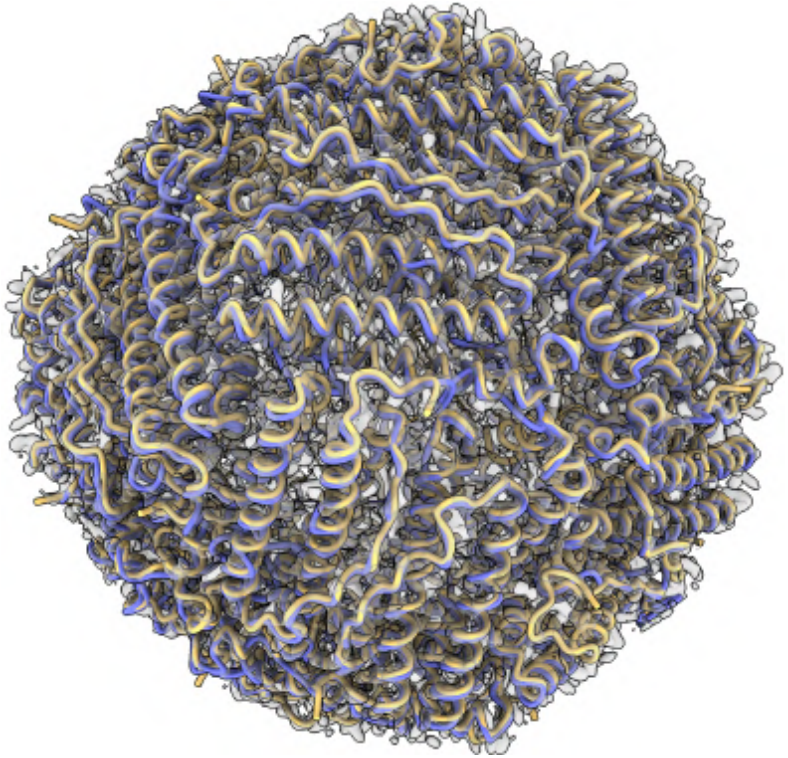
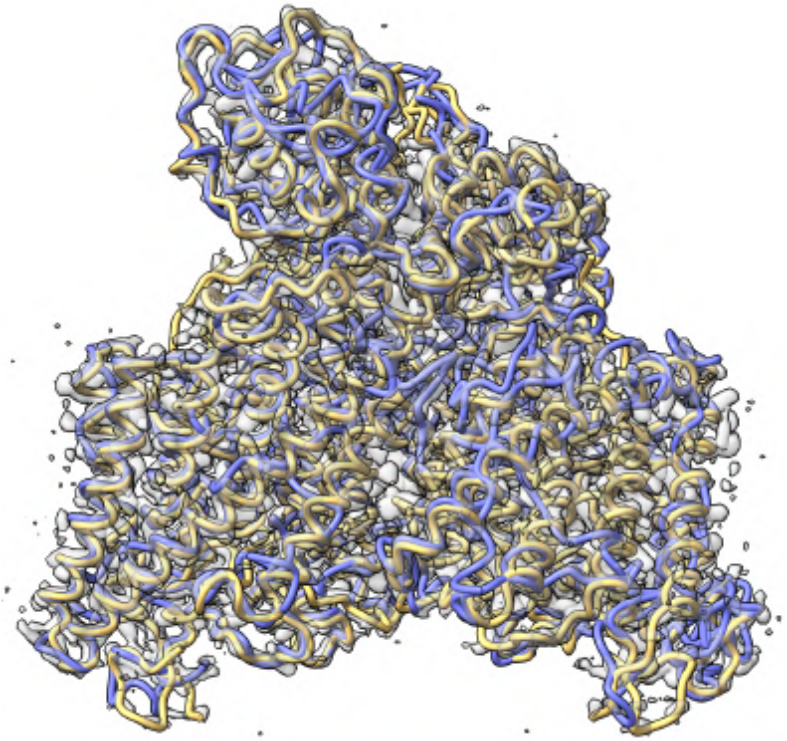


Figure: Map's resolution ranges from **1.9 Å to 4.0 Å**. The average resolution is 2.88 Å. The number of amino acid residues varies from **234 to 8,828.**

Cryo2Struct Modeled Structure Examples



(a) EMD ID: 17961 in transparent. PDB ID: 8PVC in Yellow. Cryo2Struct in Blue.



(b) EMD ID: 35299 in transparent. PDB ID: 8IAB, in Yellow. Cryo2Struct in Blue.

Why? Where? and How?

- **Exponential Growth:** The EMDB cryo-EM structure repository is expanding rapidly, projected to reach 100,000 structures by 2030.
- **Manual Building Challenges:** Manual model building is slow, error-prone, and often leads to overcorrections aimed at perfecting validation scores.
- **Critical Resolution Range:** Intermediate-resolution maps (3–4 Å) make accurate model building particularly difficult.
- **AI for Acceleration:** AI methods generate high-quality backbone traces automatically, reducing modeling time from days–weeks to minutes–hours.
- **Targeted Corrections:** AI tools provide local confidence scores, guiding users to regions that need manual refinement.
- **Improved Model Quality:** AI-based models improve fit-to-density, maintain correct geometry, and ensure overall consistency with the cryo-EM map.

Kiarash Jamali, Lukas Käll, Rui Zhang, Alan Brown, Dari Kimanius, Sjors H W Scheres. Automated model building and protein identification in cryo-EM maps. Nature (2024). https://doi.org/10.1038/s41586-024-07215-4

Vincent Chen, Christopher Williams, Michael Prisant, David Richardson, Jane Richardson. Almost, but not quite, entirely unlike proteins. Struct. Dyn. (2025). https://doi.org/10.1063/4.0000560

Key Takeaways

- Developed and benchmarked **Cryo2Struct**, an AI-based method for accurate and complete atomic structure modeling from cryo-EM maps.
- Demonstrated better performance over existing automated method Phenix.map_to_model across multiple resolutions, especially in the challenging 3–4 Å range.
- Developed fast, consistent, and robust backbone tracing and model quality assessment, reducing dependency on manual efforts.
- Contributed to the growing toolkit of automated structure-building methods, supporting scalable structural biology.

Potential Research Directions

- Integrating side-chain modeling and refinement with end-to-end learning.
- Joint 3D cryo-EM density map reconstruction and model building.
- Utilizing cryo-EM data for studying Protein-ligand interaction.

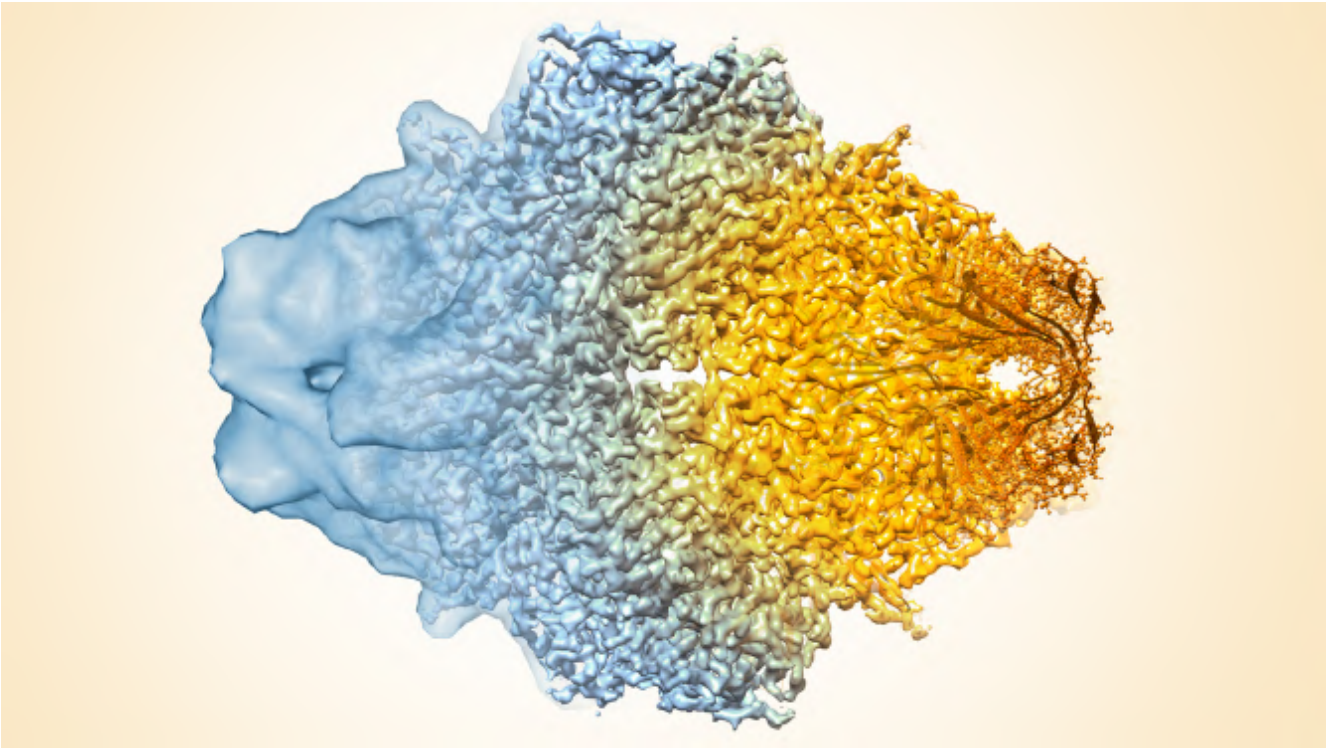


Image Source: National Cancer Institute.

Acknowledgments

- I would like to sincerely thank my advisor, **Dr. Jianlin Cheng**, for his invaluable guidance, support, and encouragement throughout my Ph.D. journey.
- I am deeply grateful to my committee members **Dr. Dong Xu**, **Dr. Yi Shang**, and **Dr. Clarissa Durie** for their valuable feedback and support.
- I also thank the **members of the Bioinformatics and Machine Learning Lab** for their help and collaboration throughout my PhD journey.

Thank You!