Highly Accurate Macromolecular Structure Modeling by Deep Learning

Xiao Wang

What is Macromolecular Structure?

- Large molecules: DNA, RNA, protein

DNA

Gene Information

RNA

Gene Regulation
Protein Synthesis

Protein

Cellular Functions

Complex

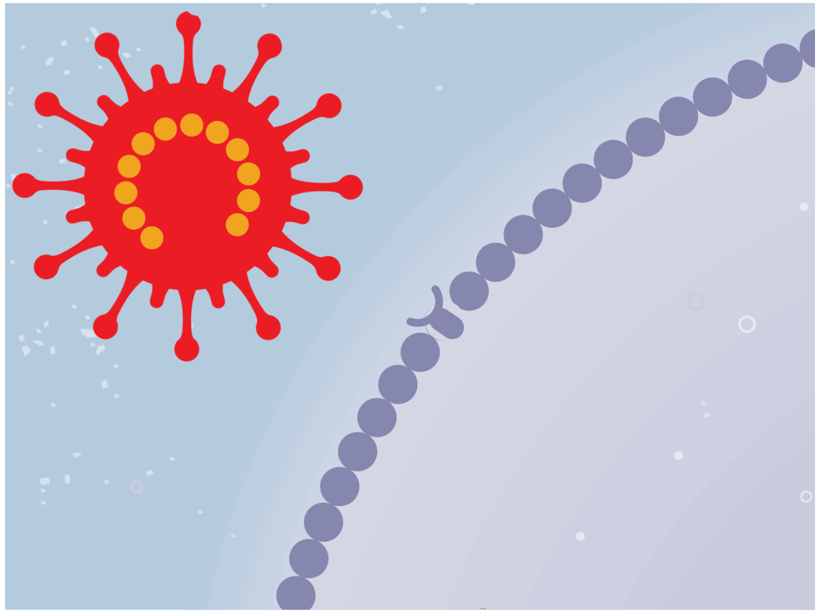
Gene Regulation
Protein Synthesis

https://en.wikipedia.org/wiki/Central_dogma_of_molecular_biology#/media/File:Centraldogma_nodetails.png
https://www.rcsb.org/structure/5BKM https://www.rcsb.org/structure/7MLW https://www.rcsb.org/structure/7XKM https://www.rcsb.org/structure/6V5B https://www.rcsb.org/structure/6CIJ

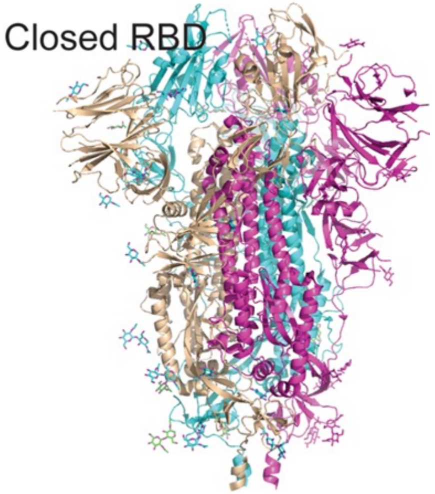
Why we study Macromolecular Structure?

Structure → **Function**

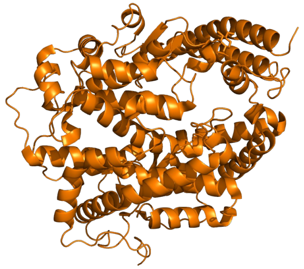
SARS-CoV-2 enters human cells

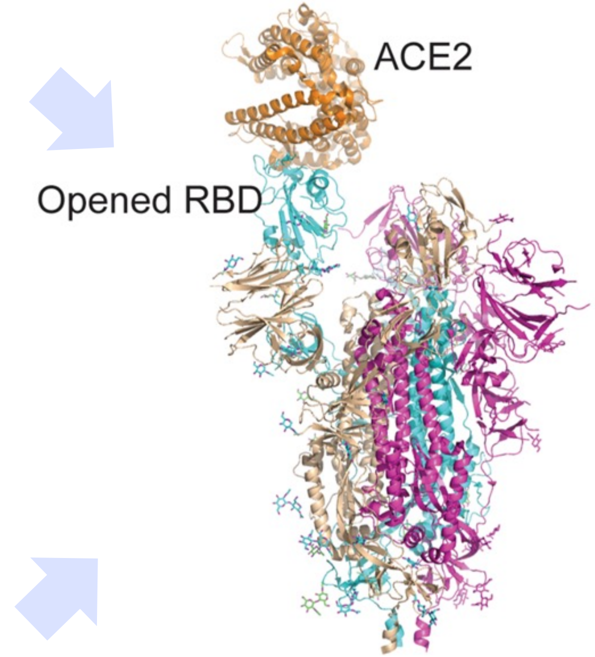


Spike protein



ACE2 protein



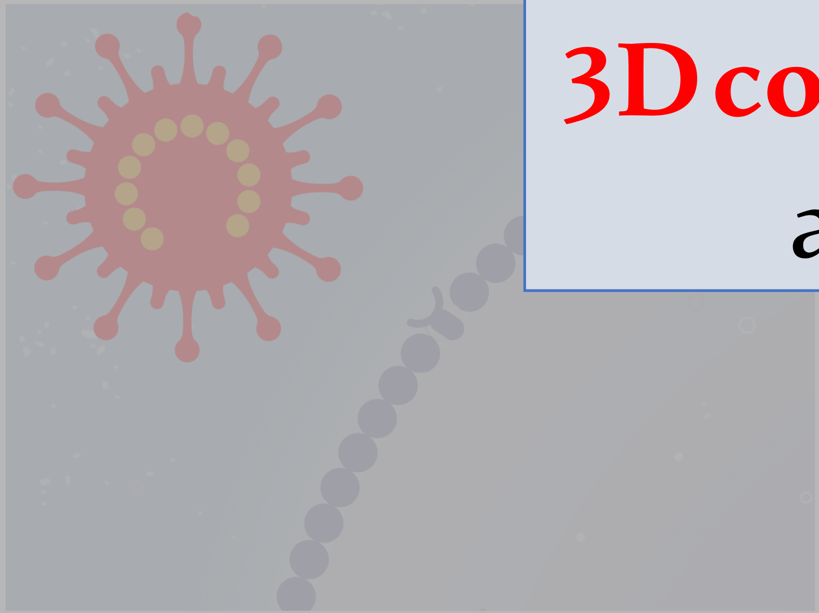


Huang, Yuan, et al. "Structural and functional properties of SARS-CoV-2 spike protein: potential antivirus drug development for COVID-19." *Acta Pharmacologica Sinica* 41.9 (2020): 1141-1149.

Why we study Macromolecular Structure?

Structure → Function

SARS-CoV-2 enters human ce

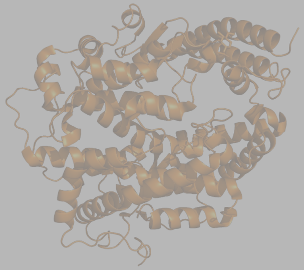


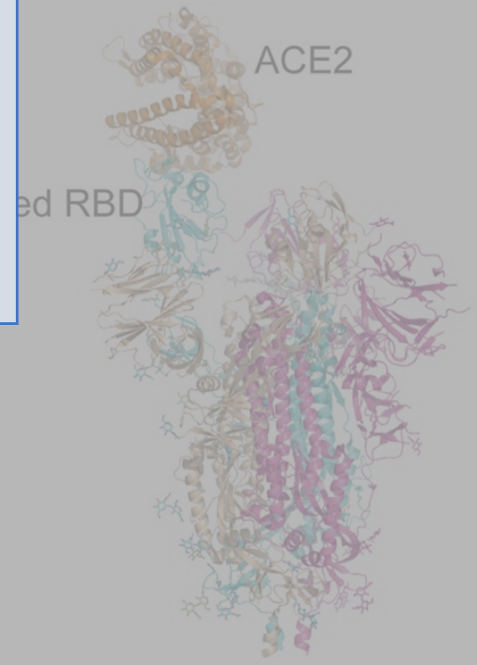
Structure is important!

3D coordinates of atoms are important!

Spike protein

ACE2 protein





Huang, Yuan, et al. "Structural and functional properties of SARS-CoV-2 spike protein: potential antivirus drug development for COVID-19." *Acta Pharmacologica Sinica* 41.9 (2020): 1141-1149.

How to Determine Macromolecular Structure?

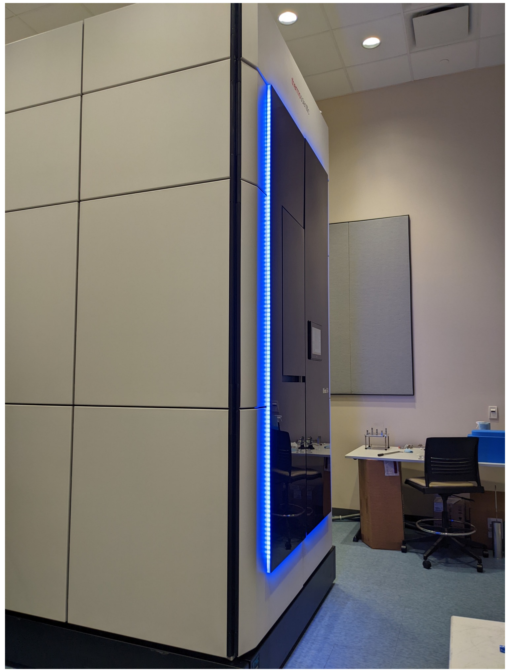
X-ray Crystallography

X-Ray Crystallography Facility
Florida State University

Nuclear Magnetic Resonance (NMR)

Bruker 800MHz NMR
BRWN @ Purdue

Cryo-Electron Microscopy(cryo-EM)



(Photo by Charles Christoffer, 2022)

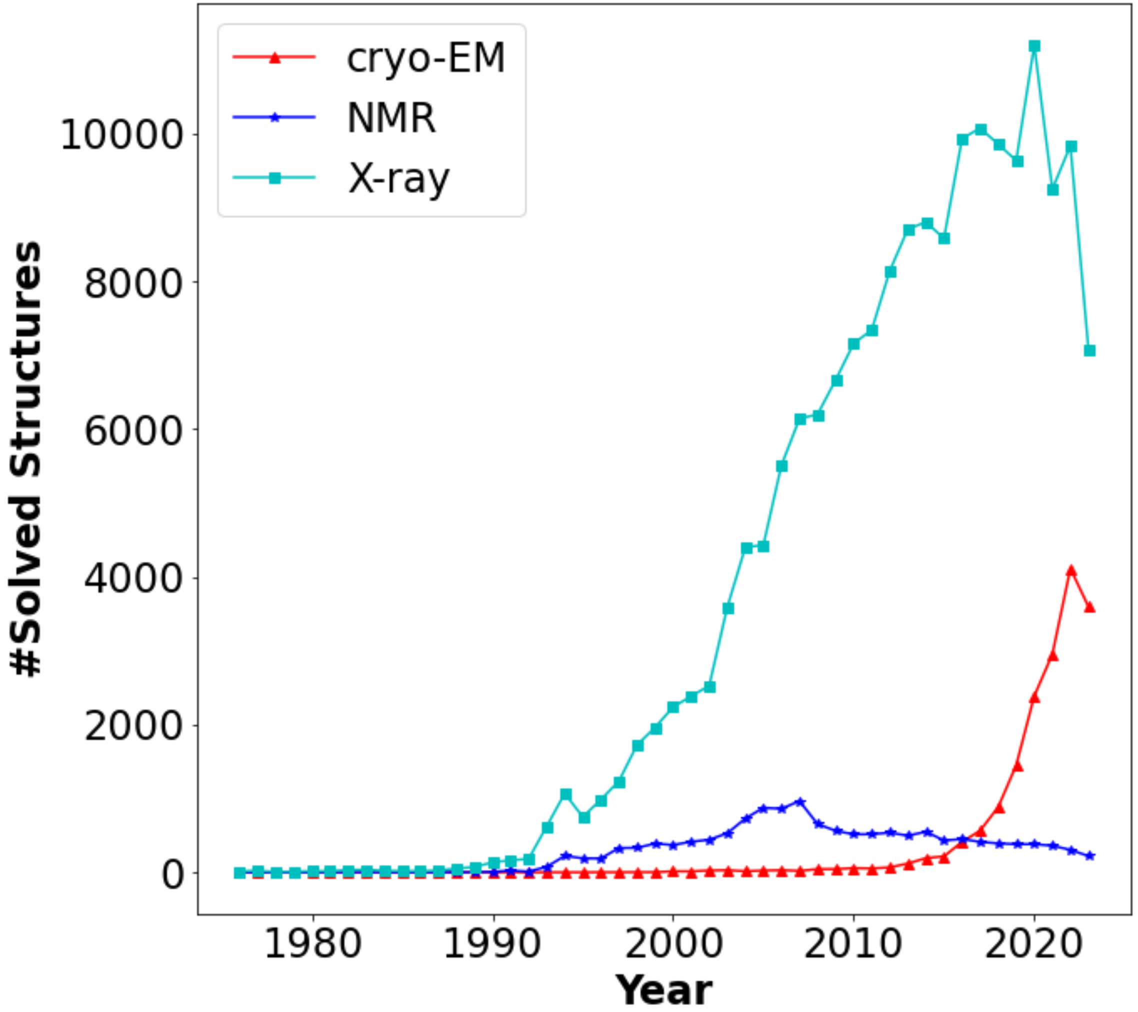
Krios G4 Cryo-EM
HOCK @ Purdue

Statistics – Structure Determination Methods

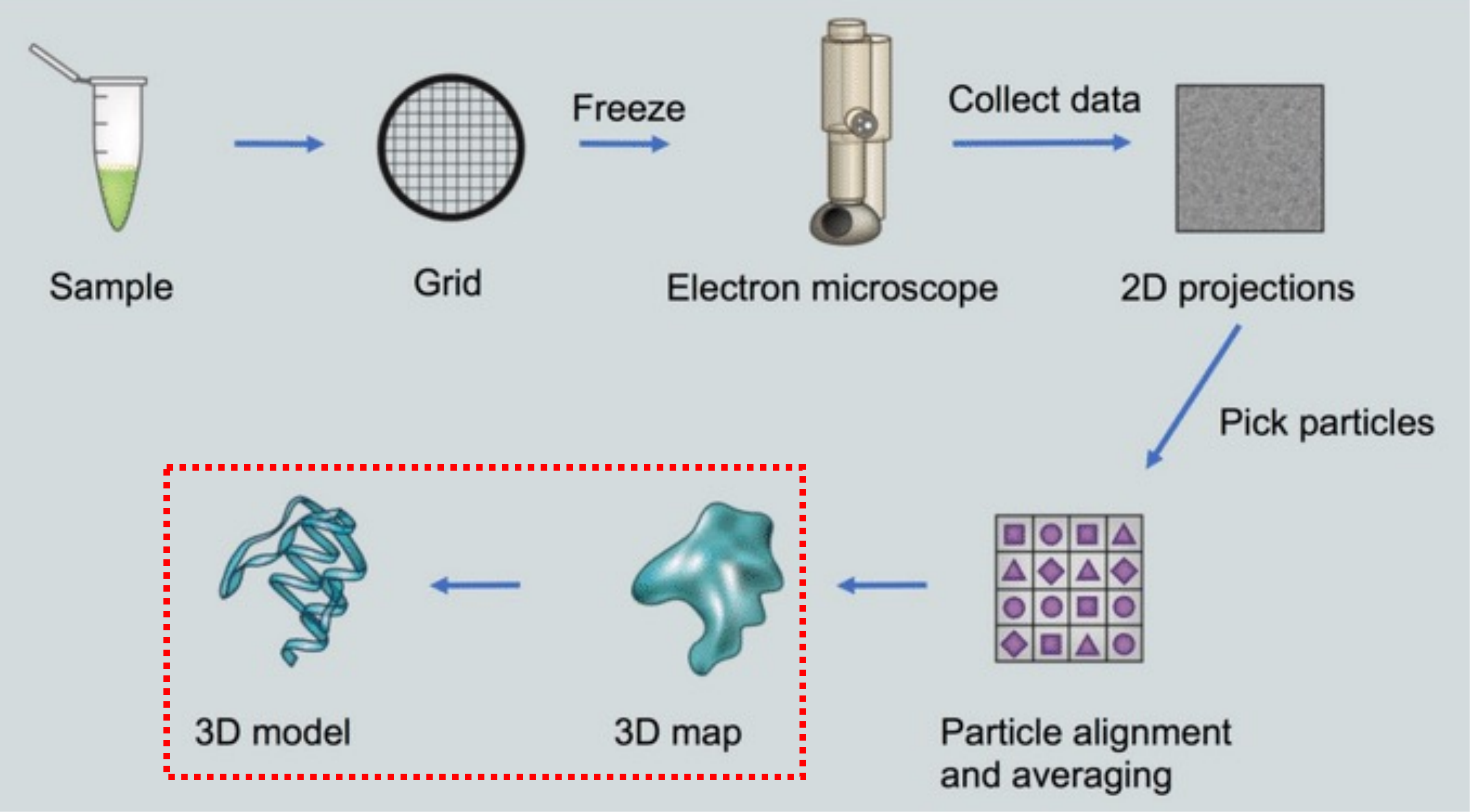
- Number of solved structures per year by different methods.

- Cryo-EM become popular!

- Advantage of cryo-EM:
  - No need to be crystallized.
  - Can determine large macromolecules.



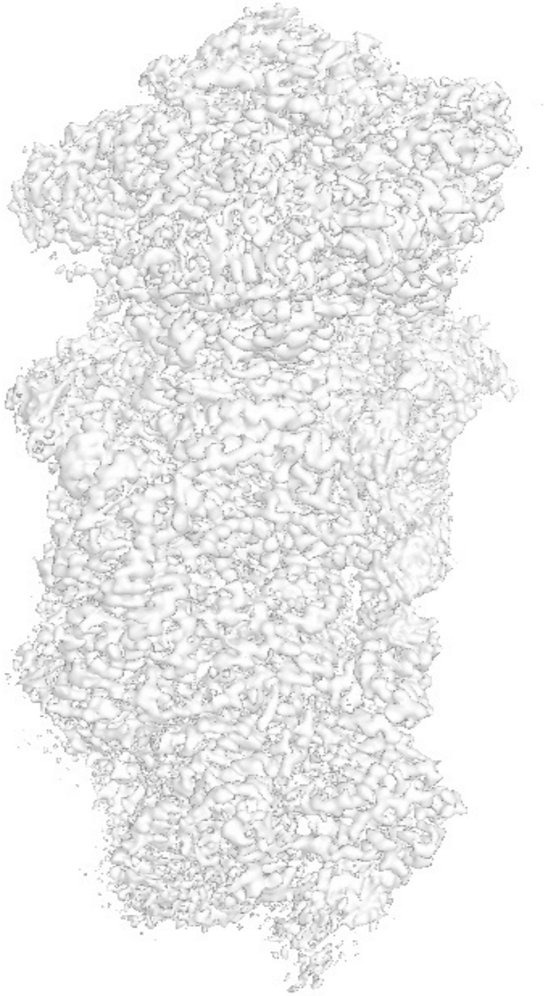
Cryo-EM Structure Determination Pipeline



https://www.researchgate.net/figure/A-schematic-of-the-single-particle-reconstruction-cryoEM-pipeline-Image-thanks-to_fig1_336638894

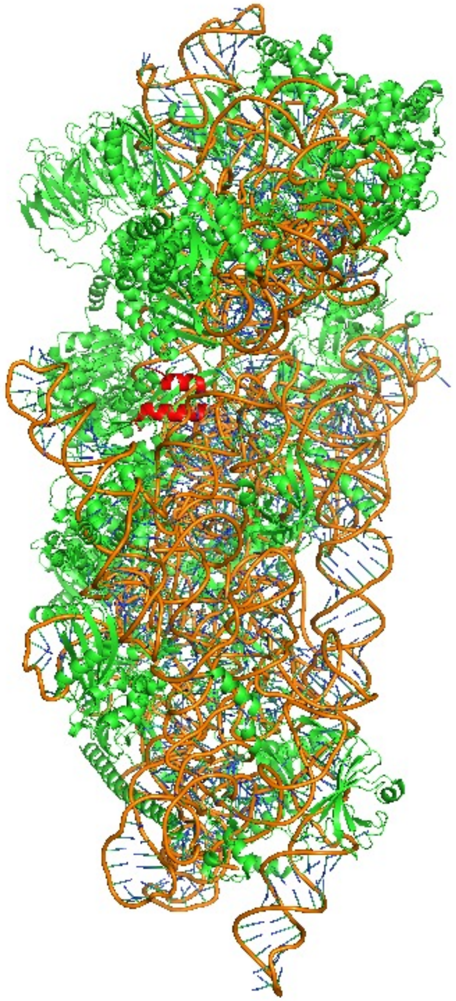
AI for Macromolecular Structure Modeling

Cryo-EM Map



Structure Modeling

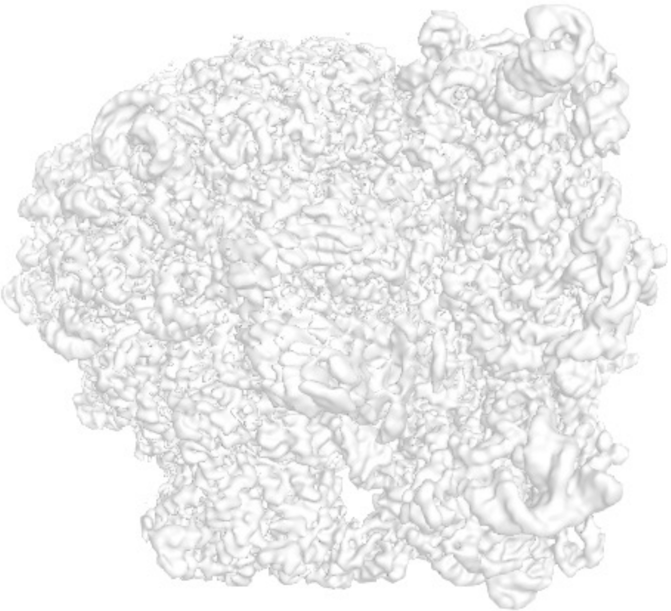
Structure



Xiao Wang, Genki Terashi, & Daisuke Kihara. "Cryo-READ: De novo structure modeling for nucleic acids in cryo-EM maps using deep learning". Nature Methods. (2023)

Method 1: CryoREAD for DNA/RNA structure modeling

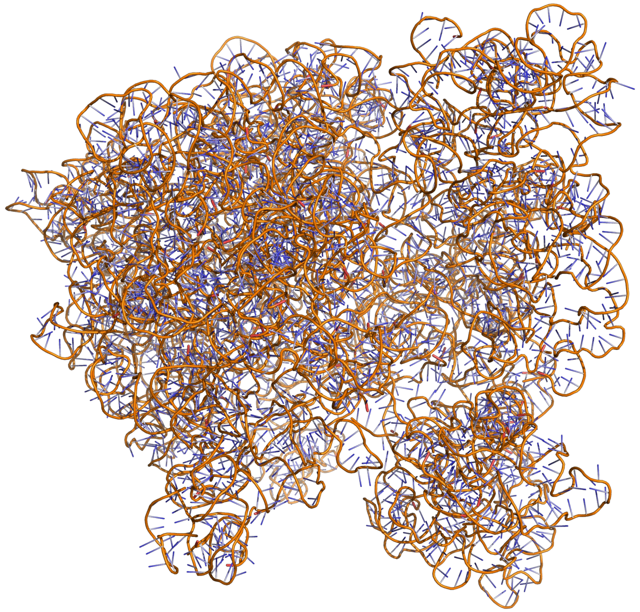
Cryo-EM Map



Structure Modeling

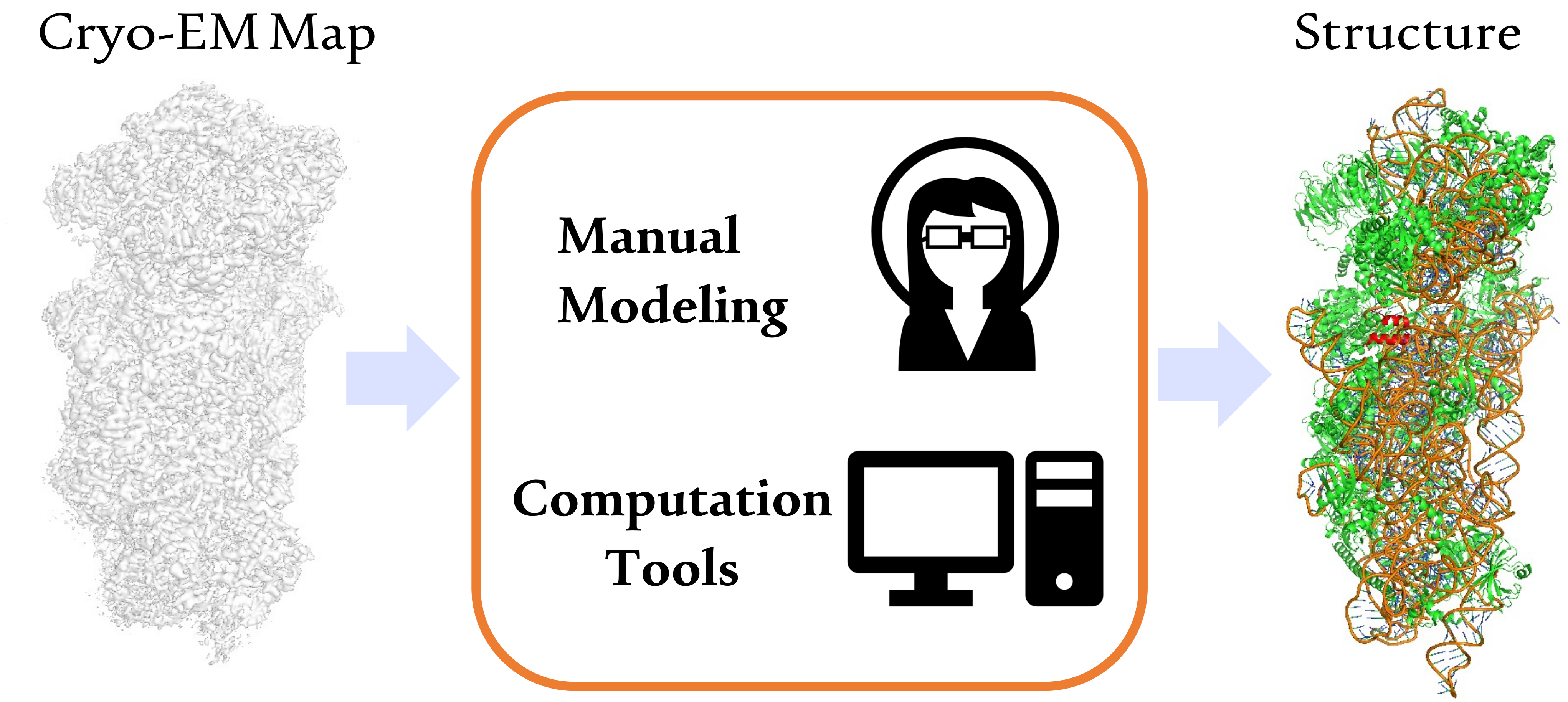
CryoREAD

DNA/RNA Structure



Xiao Wang, Genki Terashi, & Daisuke Kihara. "Cryo-READ: De novo structure modeling for nucleic acids in cryo-EM maps using deep learning". Nature Methods. (2023)

Previous Structure Modeling Methods



Cryo-EM Map

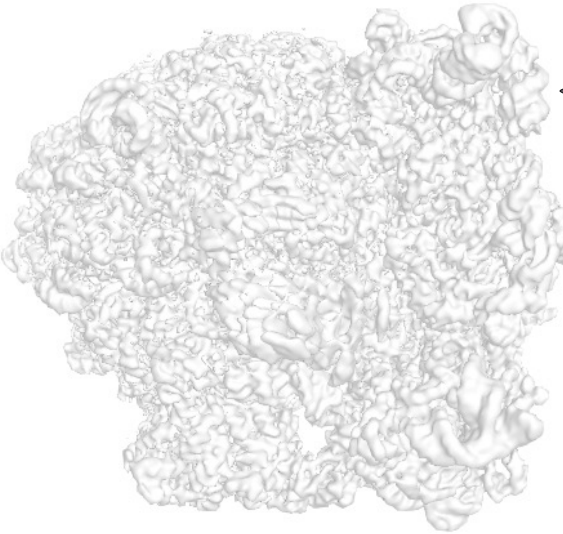
Manual Modeling

Computation Tools

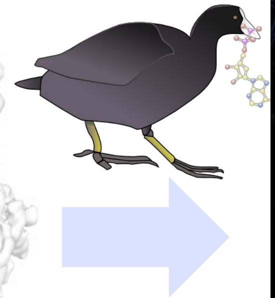
Structure

Manual Modeling: Interactive Structure Modeling

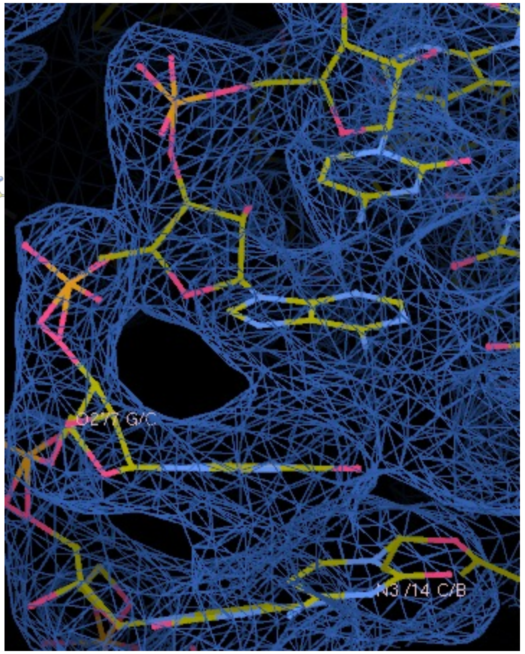
Cryo-EM Map



Coot



Detailed Visualization



Limitations:

- ➢ Time-consuming
- ➢ Local low resolution
- ➢ Human errors

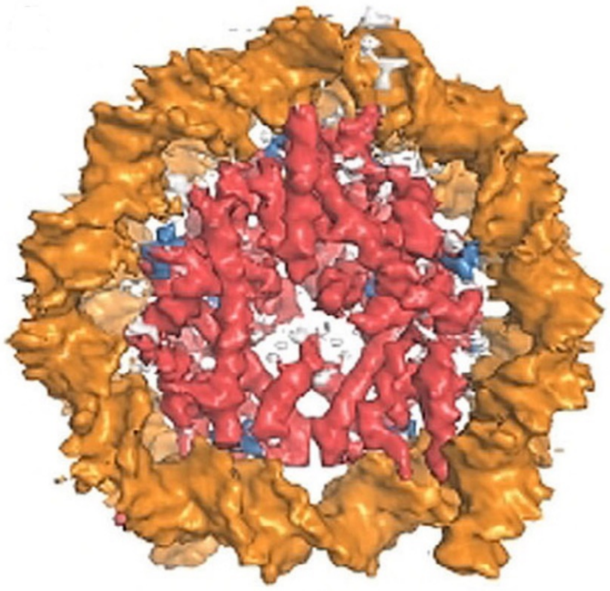
Previous computational modeling

- Haruspex: Only structure detection
- Phenix:

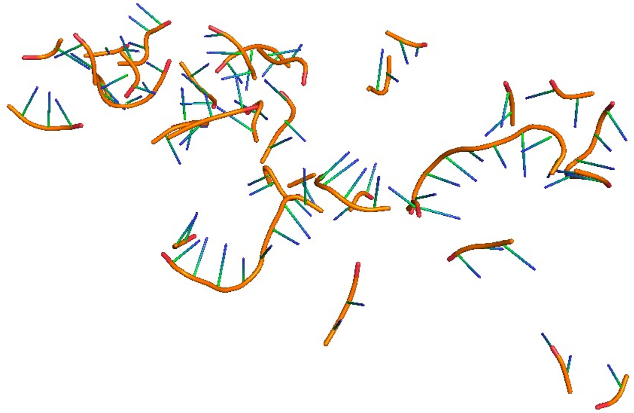
1. Focused on protein
2. DNA/RNA atomic model is not accurate.

No good DNA/RNA computational tools!

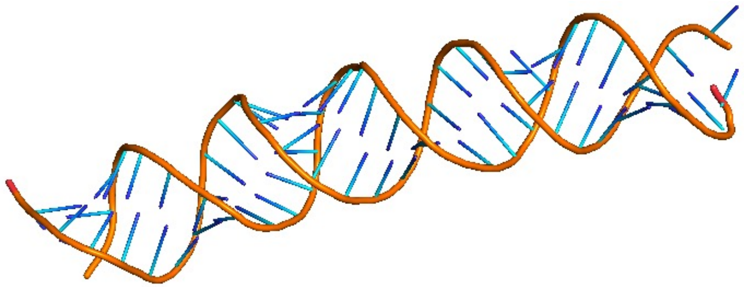
HaruSpex



Phenix



Native Structure (Ground Truth)



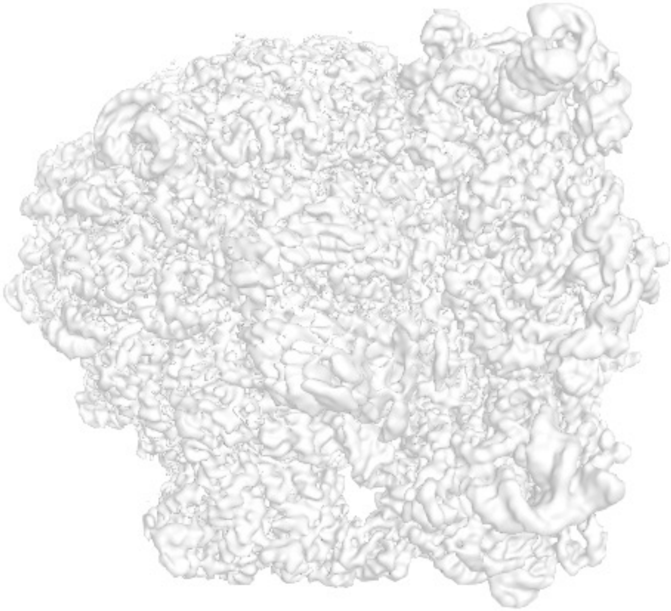
Wang Xiao, et al. "CryoREAD: De novo structure modeling for nucleic acids in cryo-EM maps using deep learning". Nature Methods. (2023)

Mostosi, P., Schindelin, H., Kollmannsberger, P., & Thorn, A. (2020). Haruspex: a neural network for the automatic identification of oligonucleotides and protein secondary structure in cryo-electron microscopy maps. Angewandte Chemie International Edition, 59(35), 14788-14795.

Terwilliger, T. C., Adams, P. D., Afonine, P. V., & Sobolev, O. V. (2018). A fully automatic method yielding initial models from high-resolution cryo-electron microscopy maps. Nature methods, 15(11), 905-908.

CryoREAD: *De Novo* DNA/RNA Structure Modeling

Cryo-EM Map



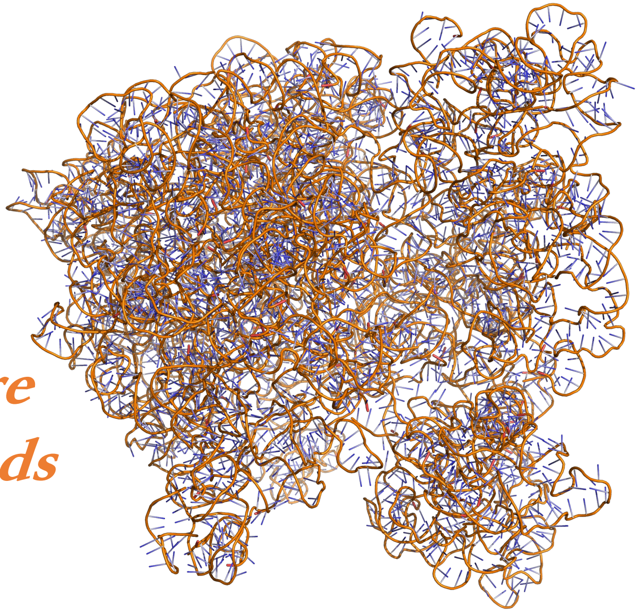
Sequence(Optional)

```
>chain A
UACCUGGUUGAUCCUGCCAGU
AGCAUAUGCUUGUCUCAAAGA
UUAAGCCAUGCAUGUCUAAGU
ACGCACGGCCGGUACAGUGAA
ACUGCGAAUGGCUCAUUAAAU
CAGUUAUGGUUCCUUUGGUCG
>chain B
UAACUGUGGUAAUUCUAGAGC
UAAUACAUGCCGACGGGCGCU
GACCCCCUUCGCGGGGGGAU
GCGUGCAUUUAUCAGAUCAAA
ACCAACCCGGUCAGCCCCUC
```

Cryo READ

Nature Methods

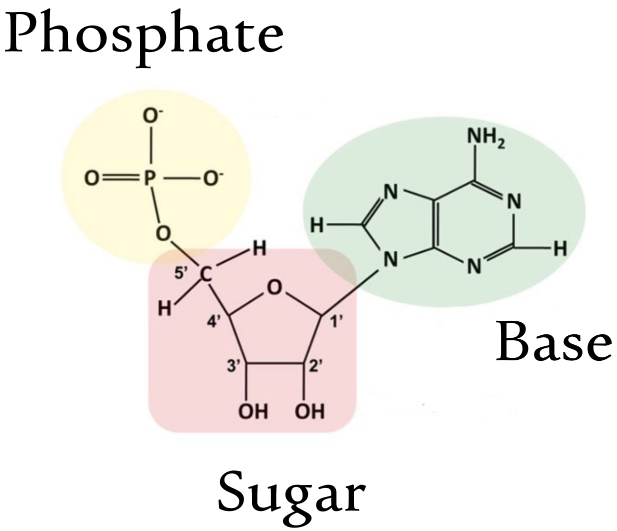
DNA/RNA Structure



Wang Xiao, et al. "Cryo-READ: De novo structure modeling for nucleic acids in cryo-EM maps using deep learning". Nature Methods. (2023)

Background: DNA/RNA Structure

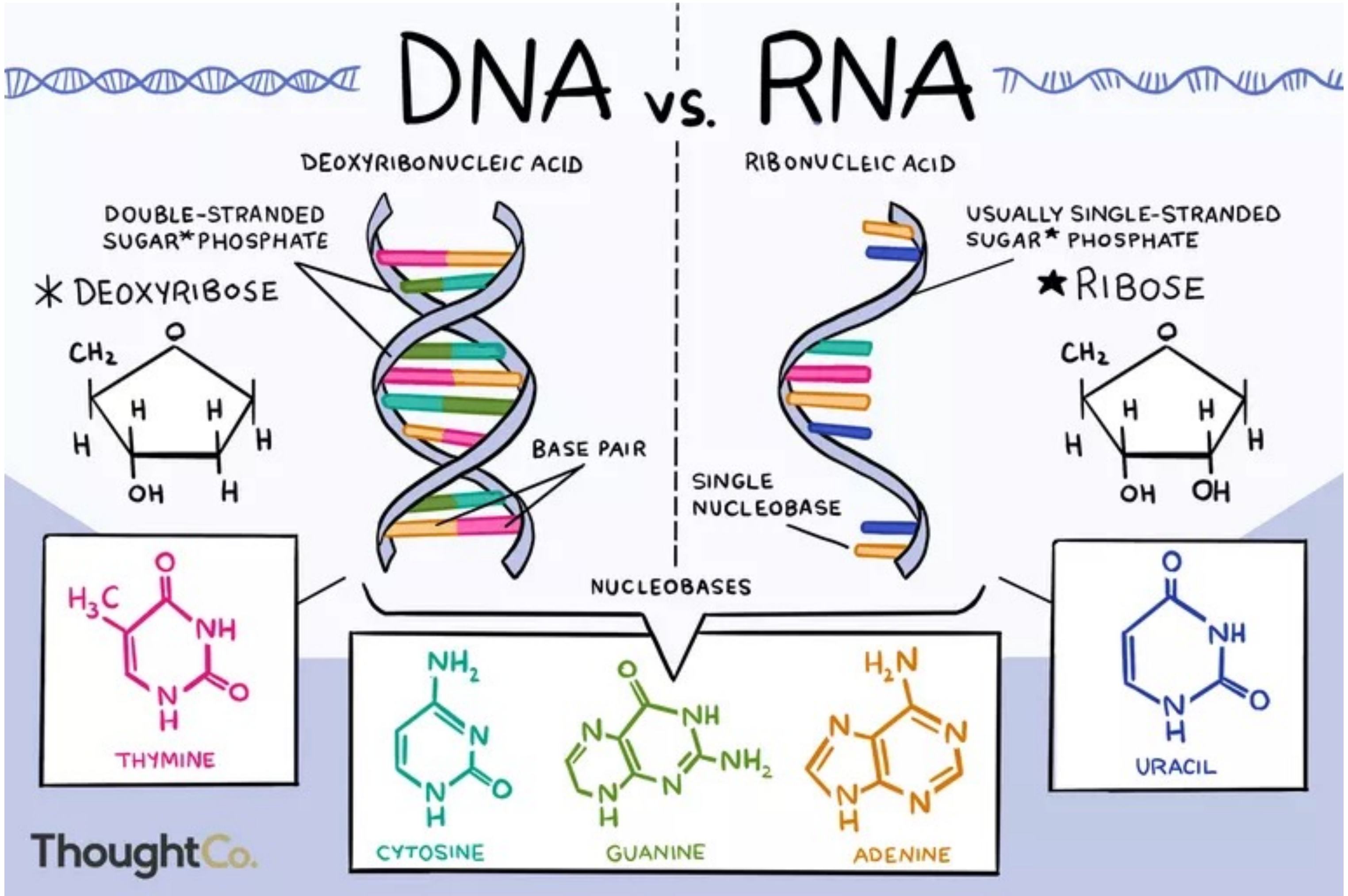
Nucleotides



Phosphate

Base

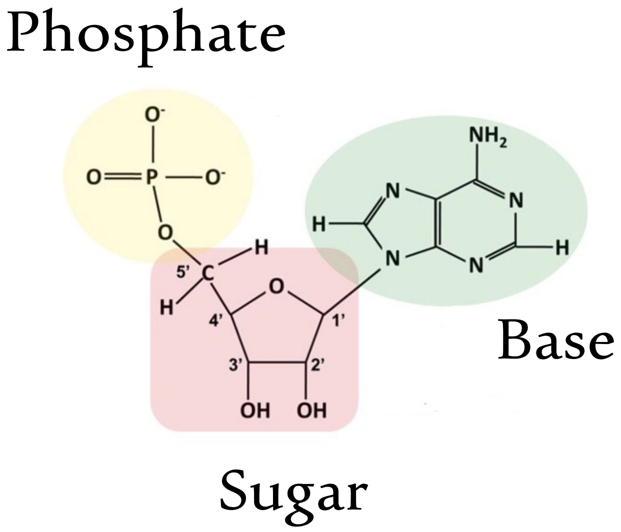
Sugar



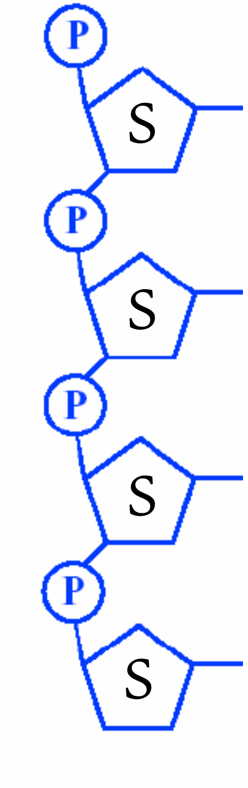
Wang Xiao, et al. "Cryo-READ: De novo structure modeling for nucleic acids in cryo-EM maps using deep learning". Nature Methods. (2023)

Domain Knowledge: DNA/RNA Structure

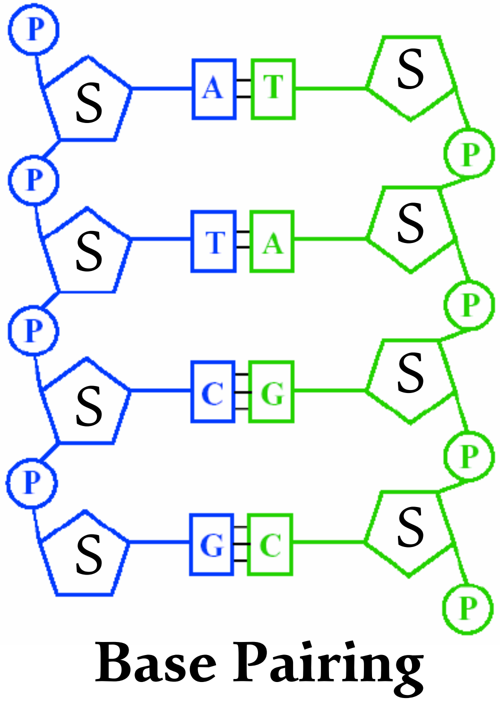
Nucleotides



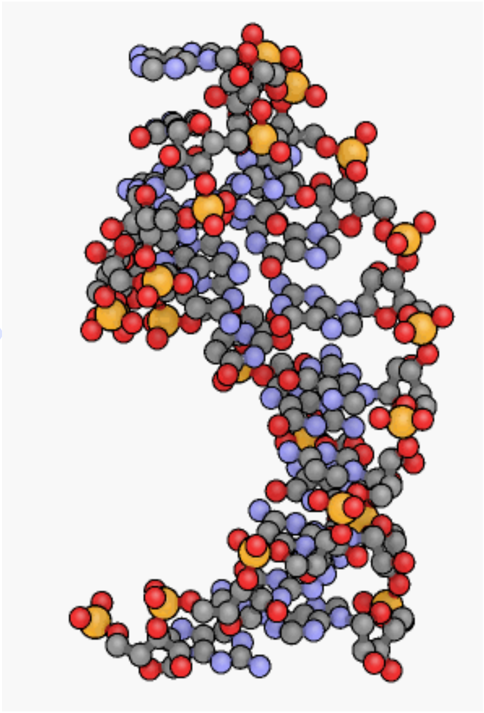
Phosphate-Sugar Backbone



Base Type

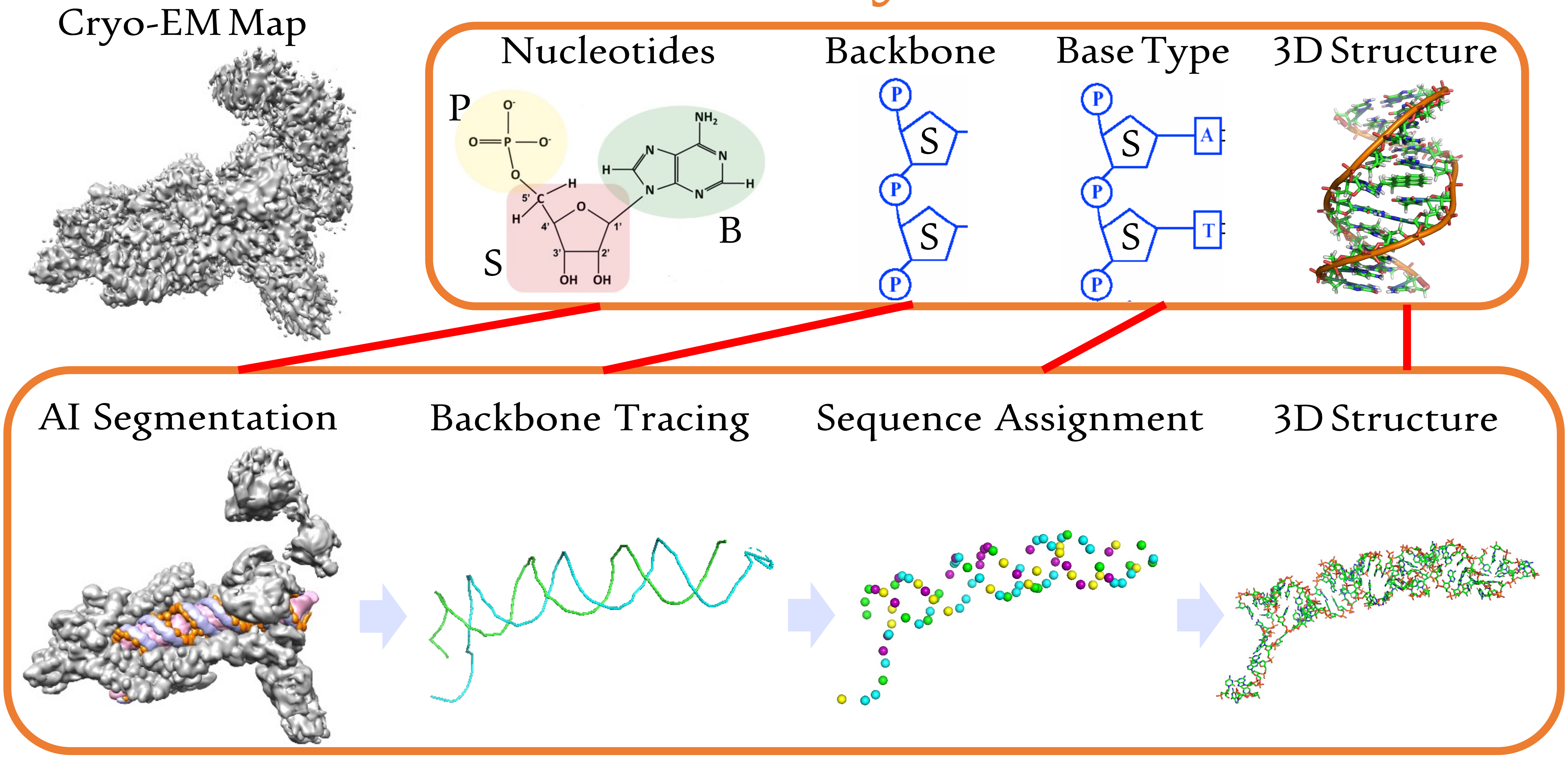


3D Structure



Wang Xiao, et al. “Cryo-READ: De novo structure modeling for nucleic acids in cryo-EM maps using deep learning”. Nature Methods. (2023)

Overview of CryoREAD



Cryo-EM Map

Nucleotides

Backbone

Base Type

3D Structure

AI Segmentation

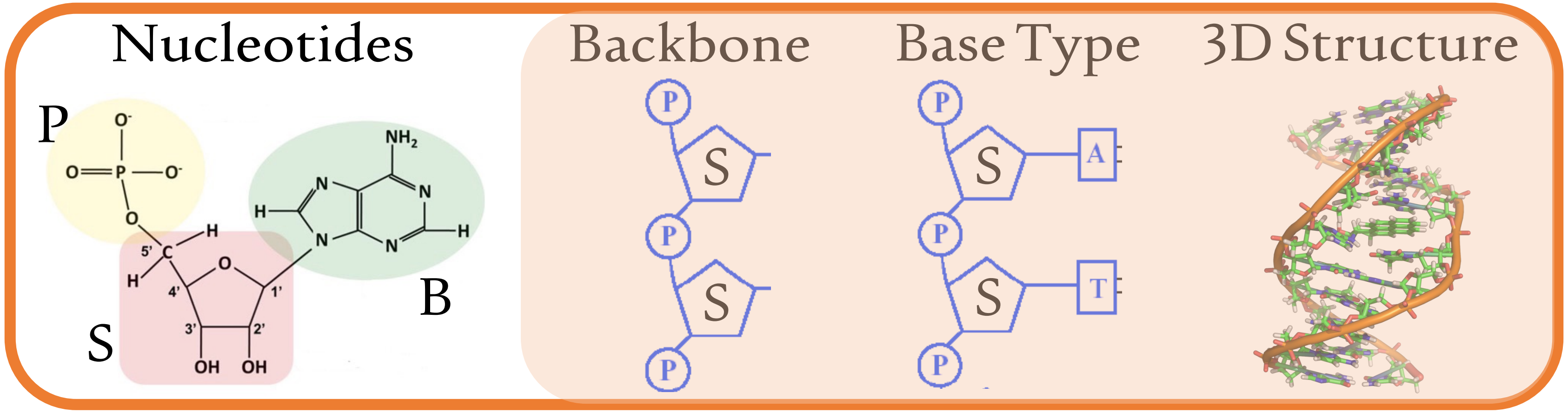
Backbone Tracing

Sequence Assignment

3D Structure

Wang Xiao, et al. "Cryo-READ: De novo structure modeling for nucleic acids in cryo-EM maps using deep learning". Nature Methods. (2023)

Step 1: Phos-Sugar-Base-Protein Segmentation

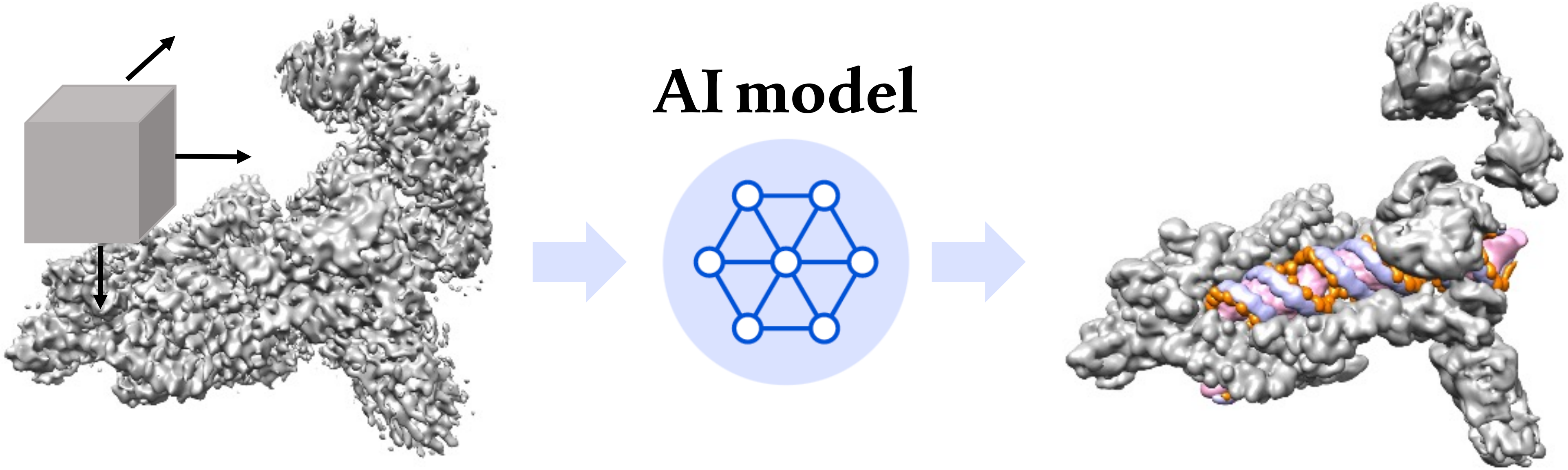


Nucleotides

Backbone

Base Type

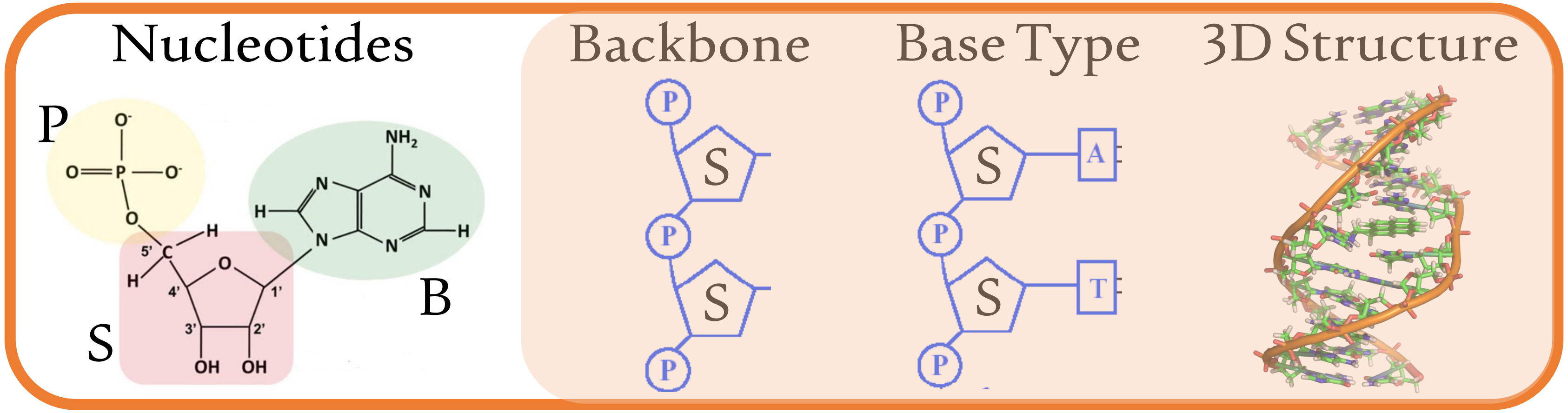
3D Structure



AI model

Wang Xiao, et al. "Cryo-READ: De novo structure modeling for nucleic acids in cryo-EM maps using deep learning". Nature Methods. (2023)

Step 1: Phos-Sugar-Base-Protein Segmentation

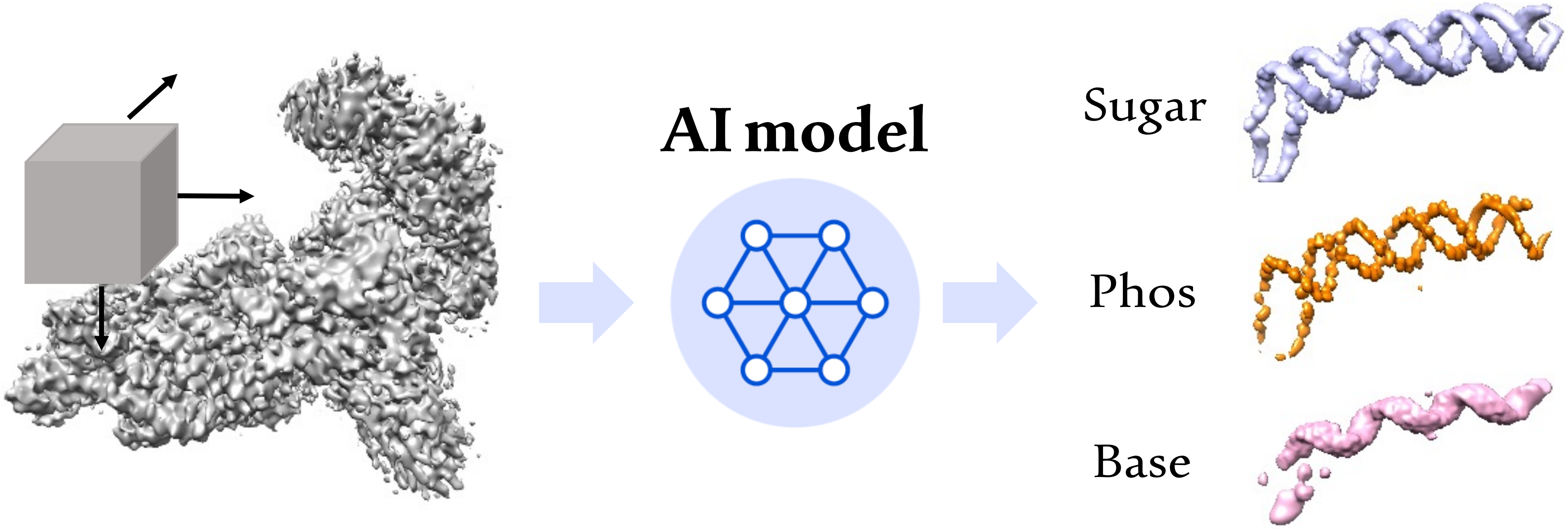


Nucleotides

Backbone

Base Type

3D Structure



AI model

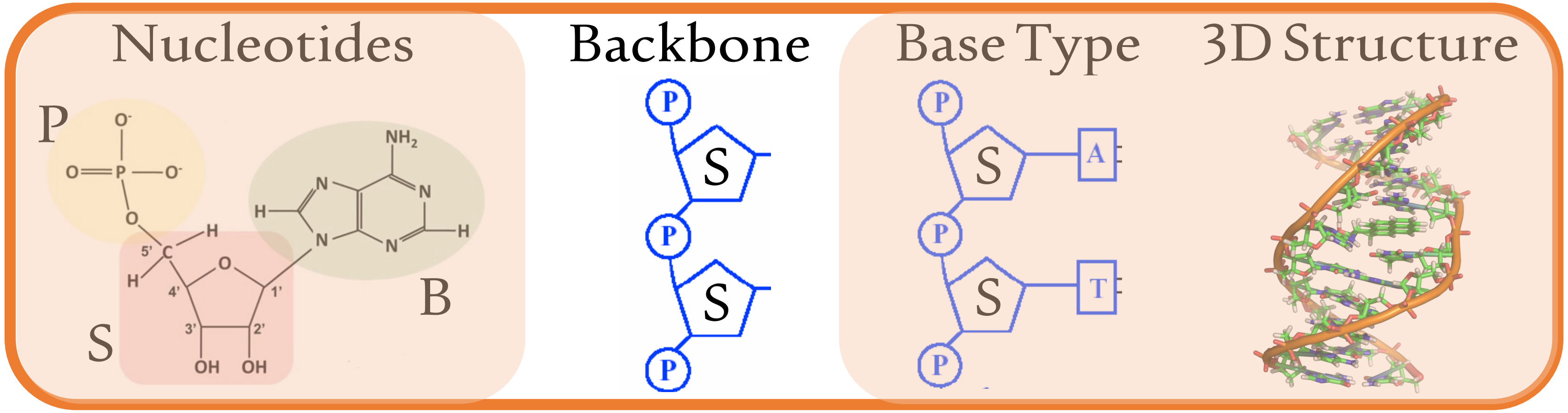
Sugar

Phos

Base

Wang Xiao, et al. "Cryo-READ: De novo structure modeling for nucleic acids in cryo-EM maps using deep learning". Nature Methods. (2023)

Step 2: Backbone Tracing

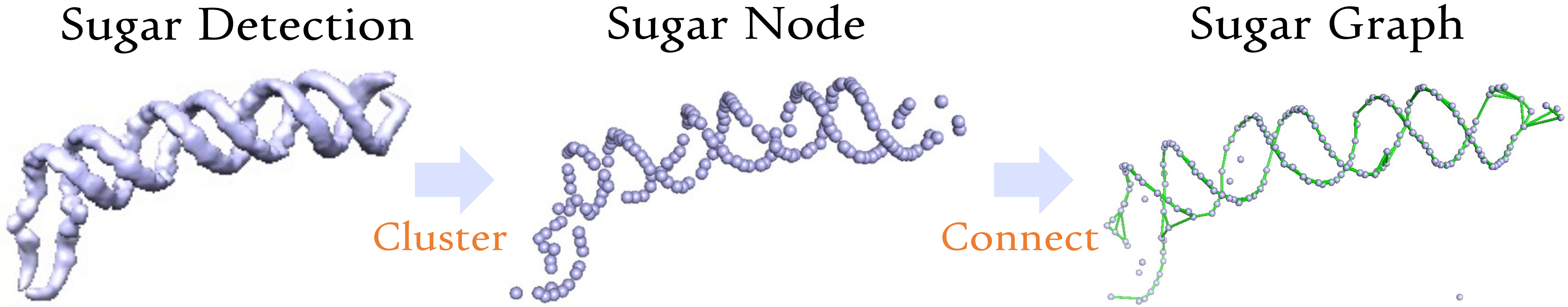


Nucleotides

Backbone

Base Type

3D Structure



Sugar Detection

Sugar Node

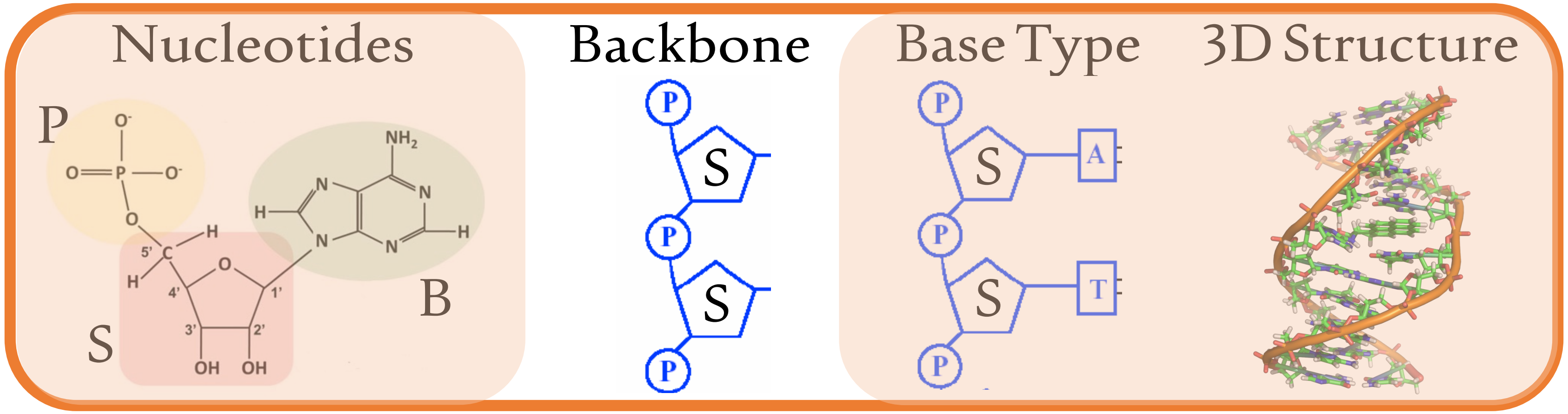
Sugar Graph

Cluster

Connect

Wang Xiao, et al. "Cryo-READ: De novo structure modeling for nucleic acids in cryo-EM maps using deep learning". Nature Methods. (2023)

Step 2: Backbone Tracing



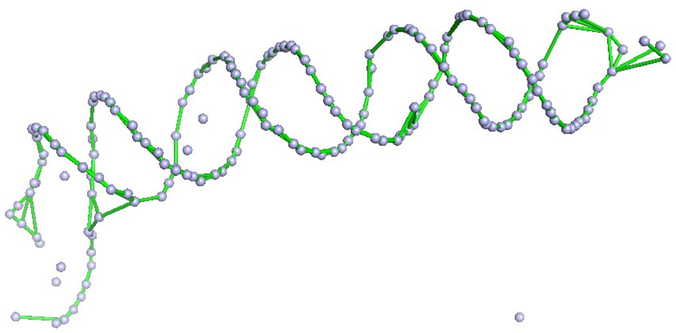
Nucleotides

Backbone

Base Type

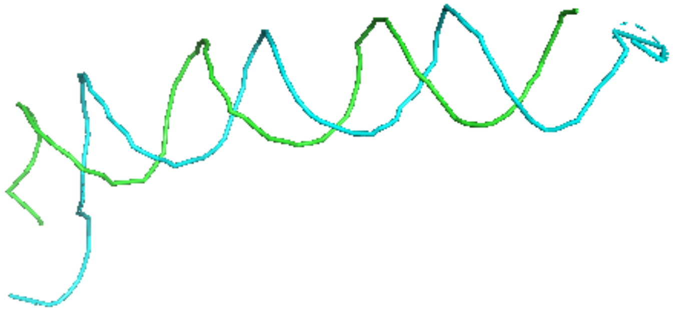
3D Structure

Sugar Graph

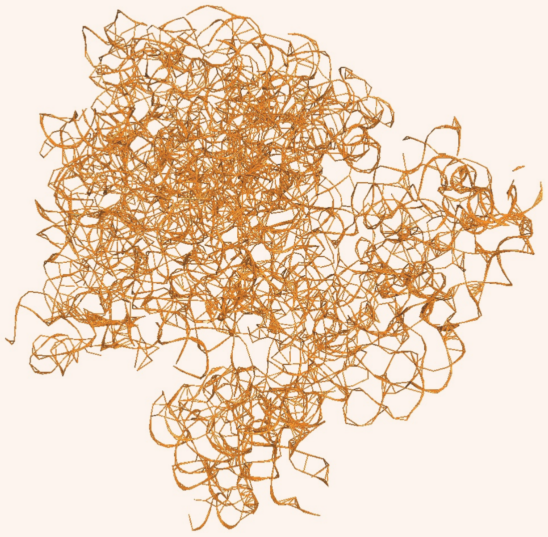


Easy Job?

Sugar Backbone

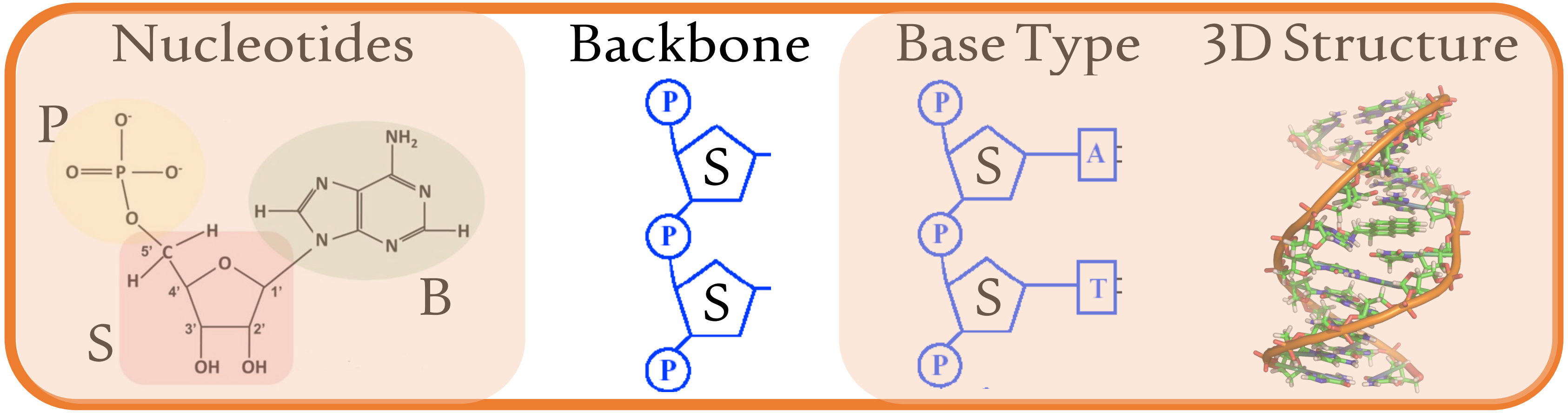


How about This?



Wang Xiao, et al. “Cryo-READ: De novo structure modeling for nucleic acids in cryo-EM maps using deep learning”. Nature Methods. (2023)

Step 2: Backbone Tracing



Nucleotides

Backbone

Base Type

3D Structure

Vehicle Routing Problem

Problem Reformulation

$$Cost = \sum_{i=1}^{N_p} \sum_{j=1}^{L(i)-1} w_{i,j} + P_{drop} \sum_{k=1}^{N} drop_k$$

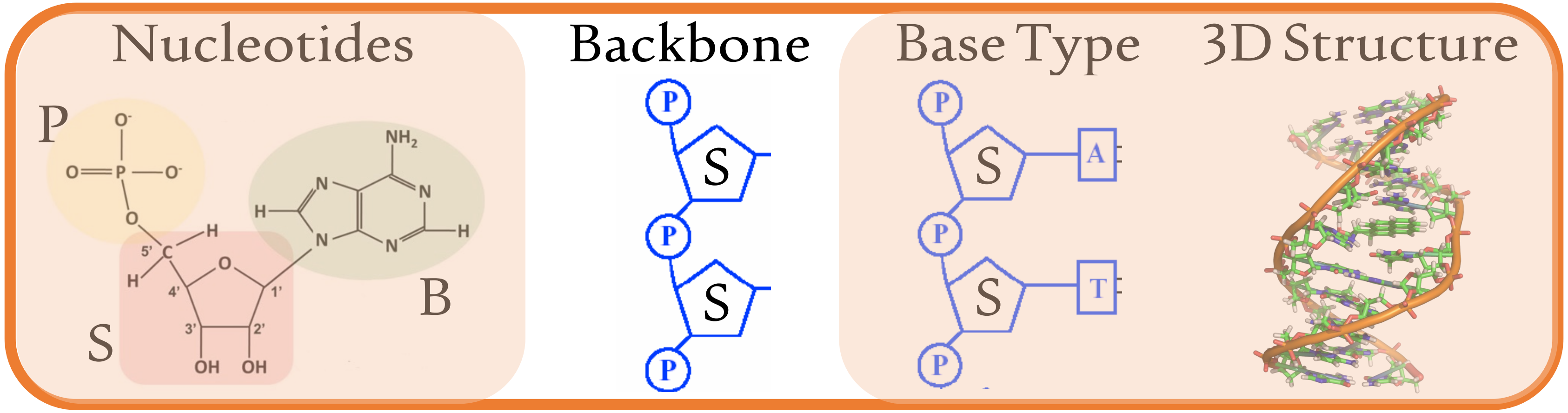
Sugar Node: Locations

N_p Chains: N_p Vehicles

Drop Wrong Node: Drop Penalty

Wang Xiao, et al. "Cryo-READ: De novo structure modeling for nucleic acids in cryo-EM maps using deep learning". Nature Methods. (2023)

Step 2: Backbone Tracing



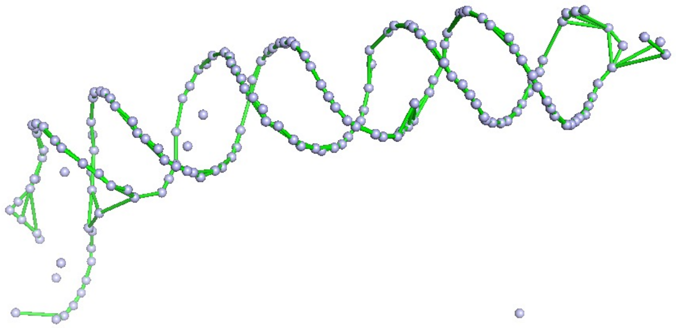
Nucleotides

Backbone

Base Type

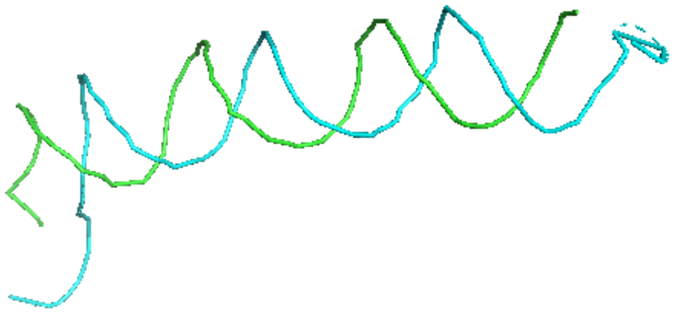
3D Structure

Sugar Graph



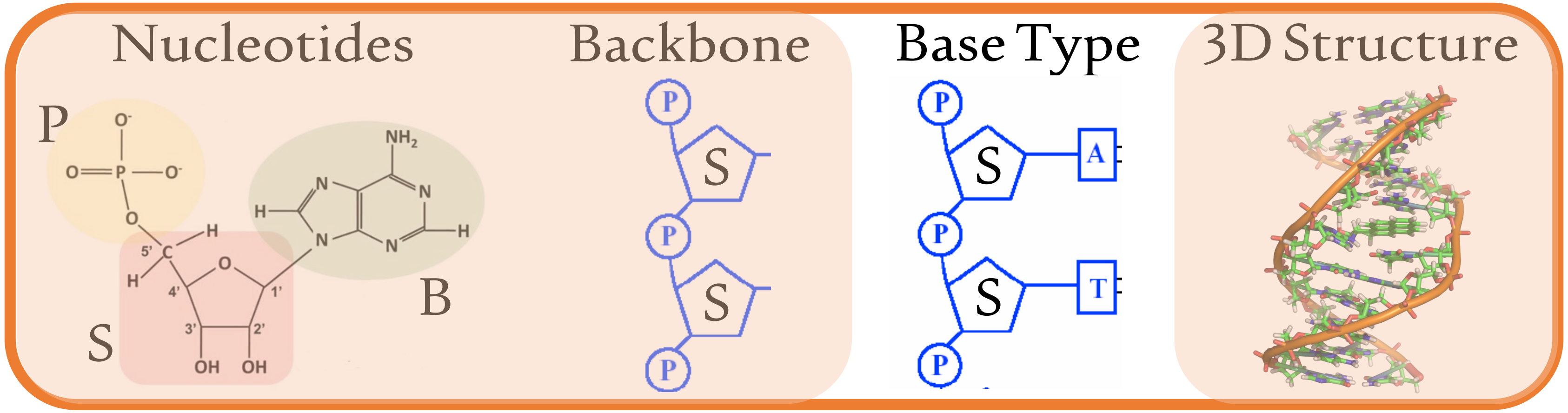
Vehicle Routing Problem

Sugar Backbone



Wang Xiao, et al. "Cryo-READ: De novo structure modeling for nucleic acids in cryo-EM maps using deep learning". Nature Methods. (2023)

Step 3: Sequence Assignment

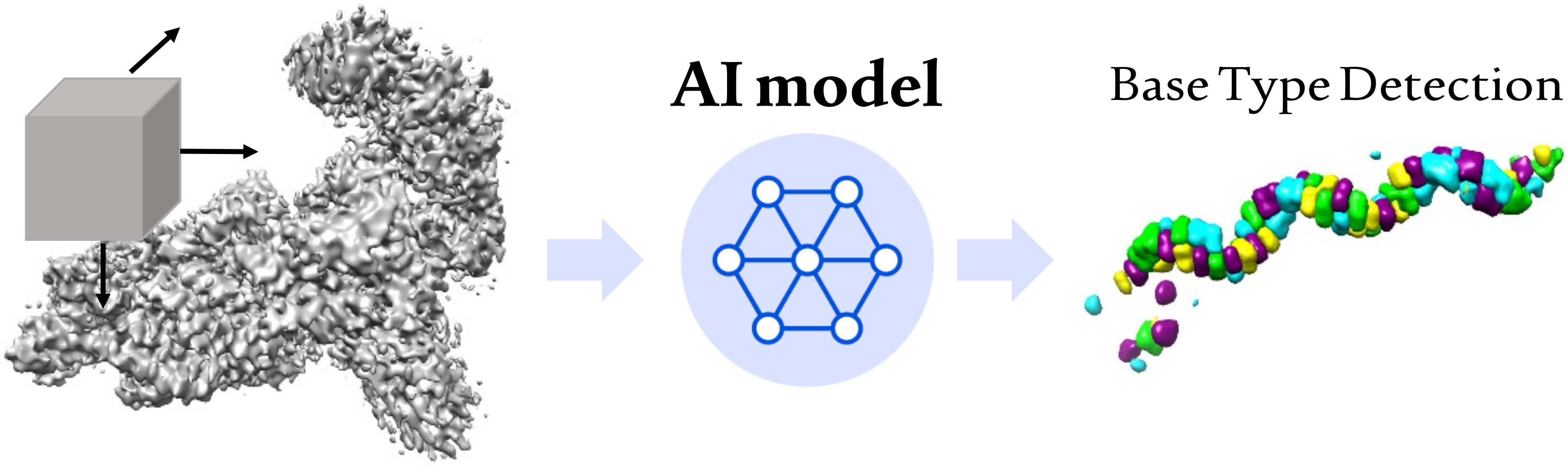


Nucleotides

Backbone

Base Type

3D Structure

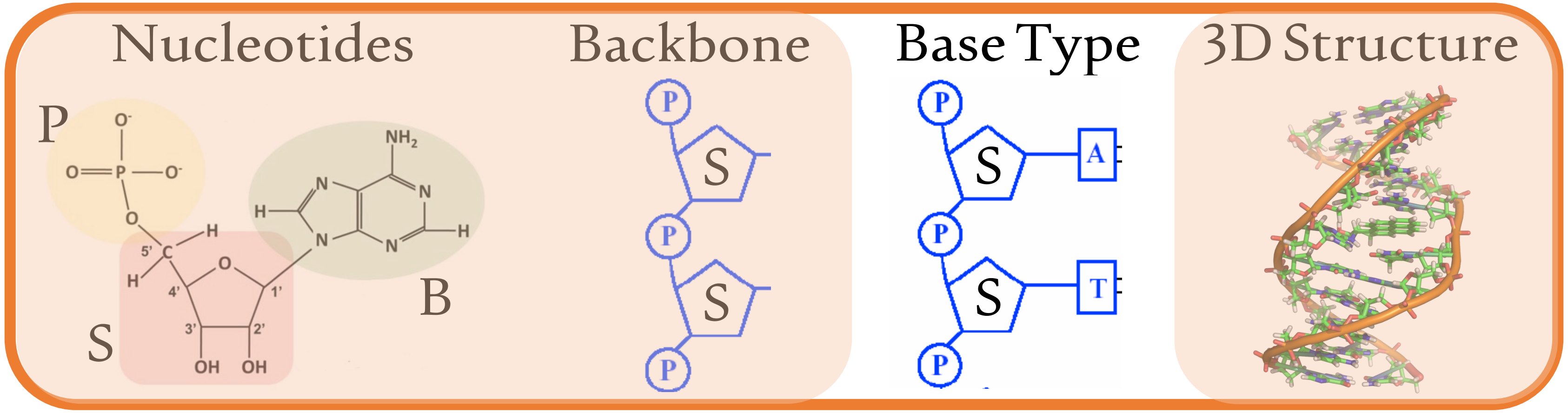


AI model

Base Type Detection

Wang Xiao, et al. "Cryo-READ: De novo structure modeling for nucleic acids in cryo-EM maps using deep learning". Nature Methods. (2023)

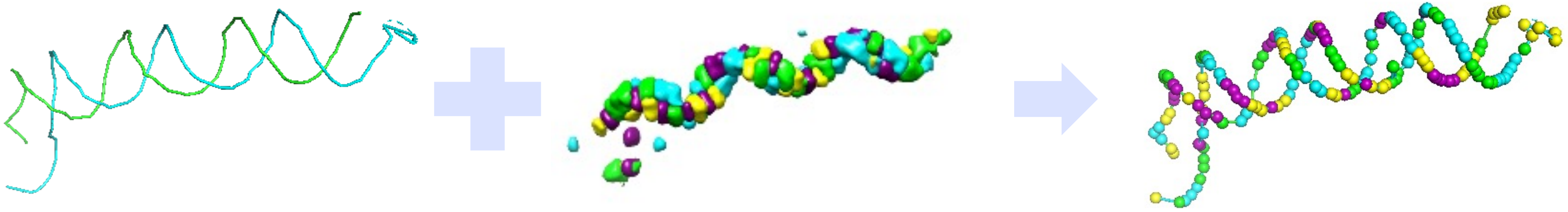
Step 3: Sequence Assignment



Sugar Backbone

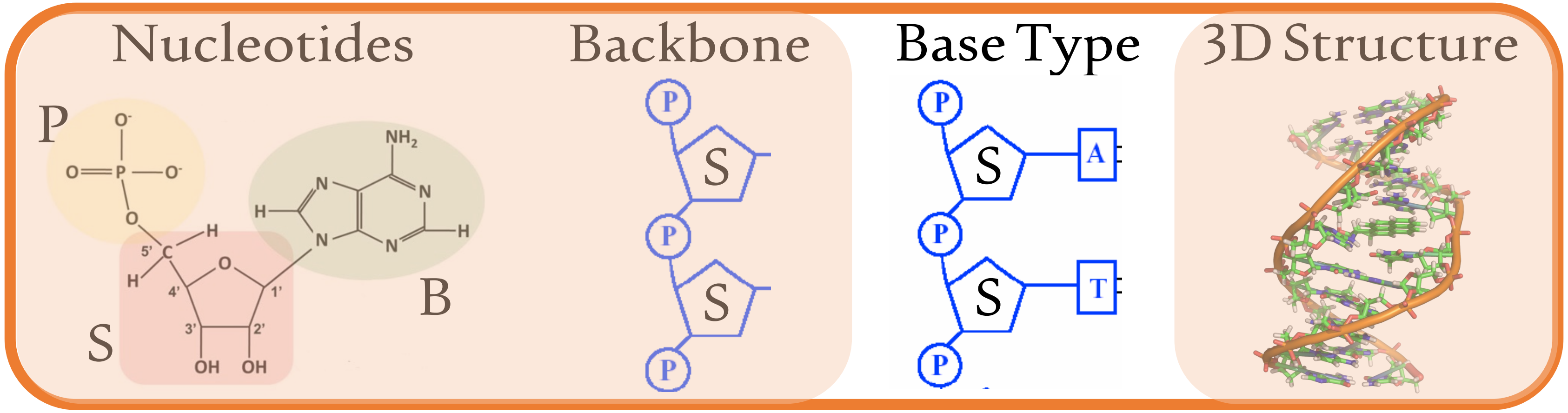
Base Type Detection

Initial Sequence Assignment

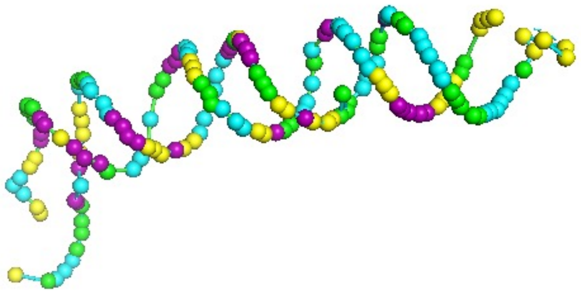


Wang Xiao, et al. "Cryo-READ: De novo structure modeling for nucleic acids in cryo-EM maps using deep learning". Nature Methods. (2023)

Step 3: Sequence Assignment

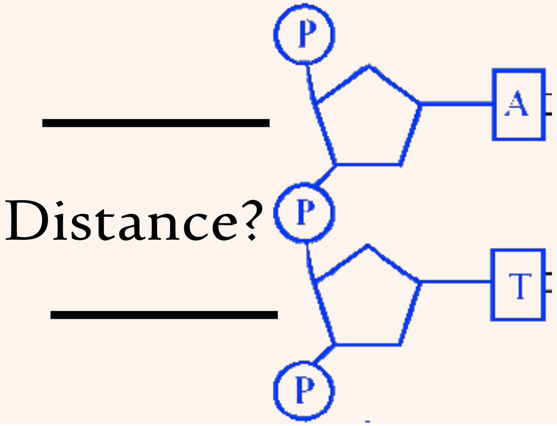


Initial Sequence Assignment



Limitations

- Geometry Constraint

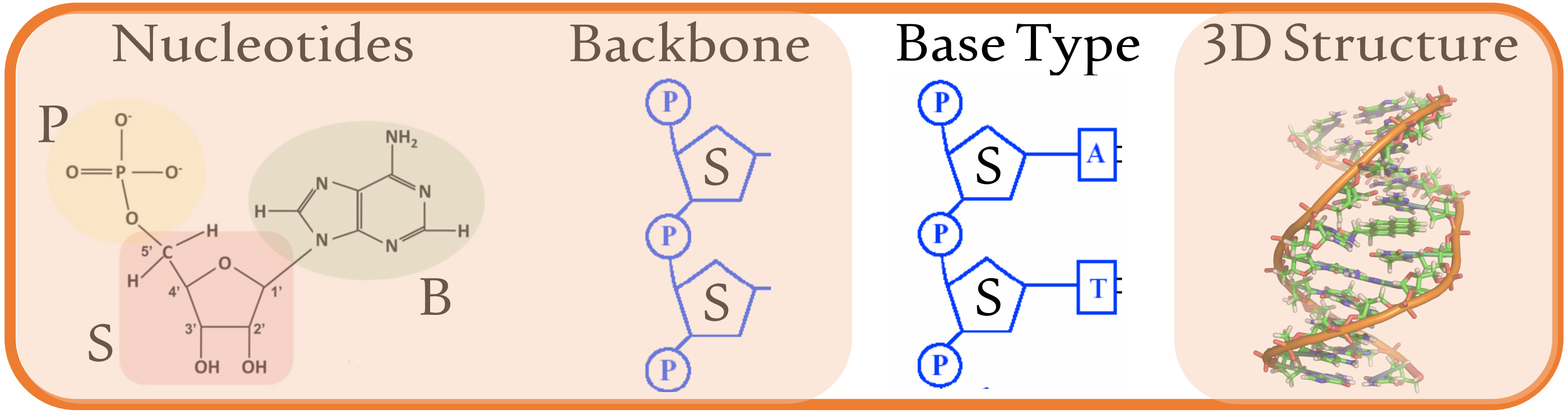


- Sequence Information

>chain A
UACCUGGUUGAUCC
UGCCAGUAGCAUAU
GCUUGUCUCAAAGA
UUAAGCCA

Wang Xiao, et al. "Cryo-READ: De novo structure modeling for nucleic acids in cryo-EM maps using deep learning". Nature Methods. (2023)

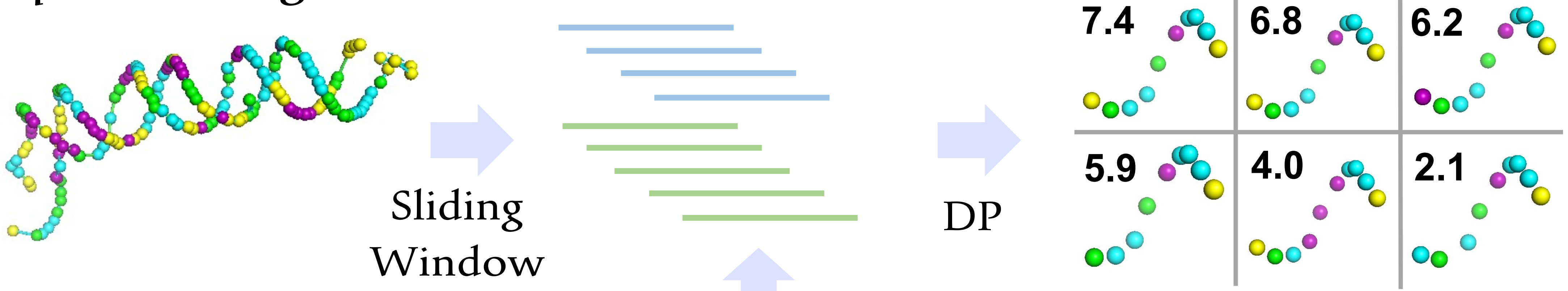
Step 3: Sequence Assignment



Initial Sequence Assignment

Scanned Fragments

Fragment Assignment

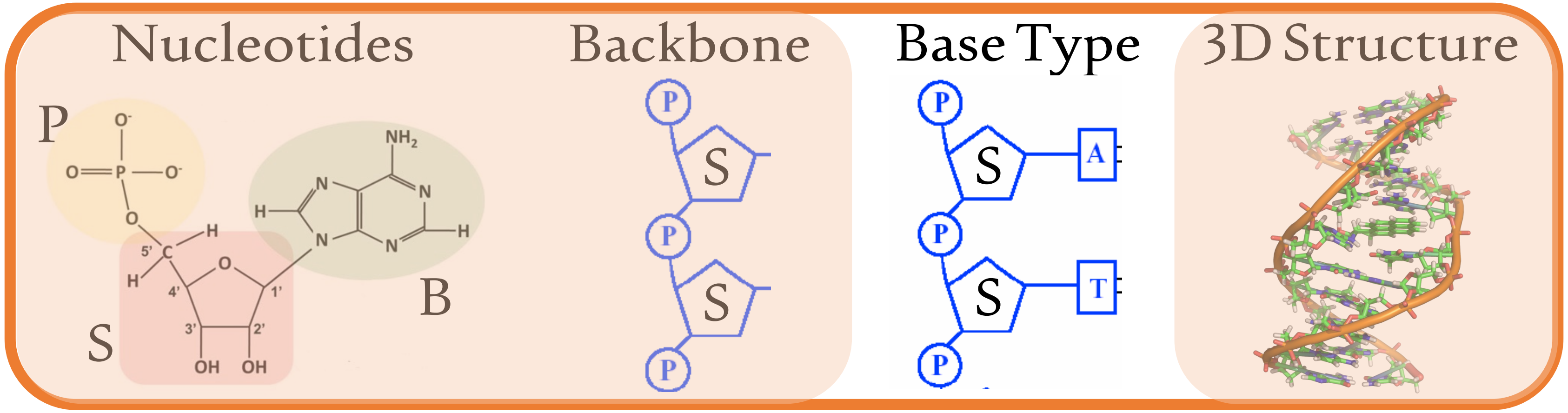


Input Sequence

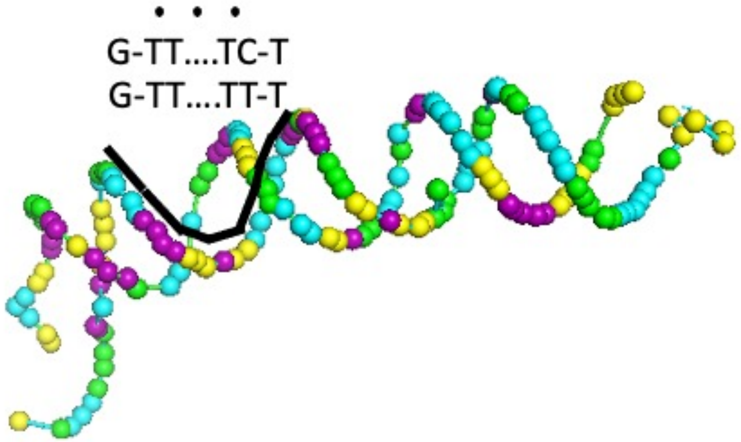
....UACCUGGUUGAUCCUGCCAGUAGCA........

Wang Xiao, et al. “Cryo-READ: De novo structure modeling for nucleic acids in cryo-EM maps using deep learning”. Nature Methods. (2023)

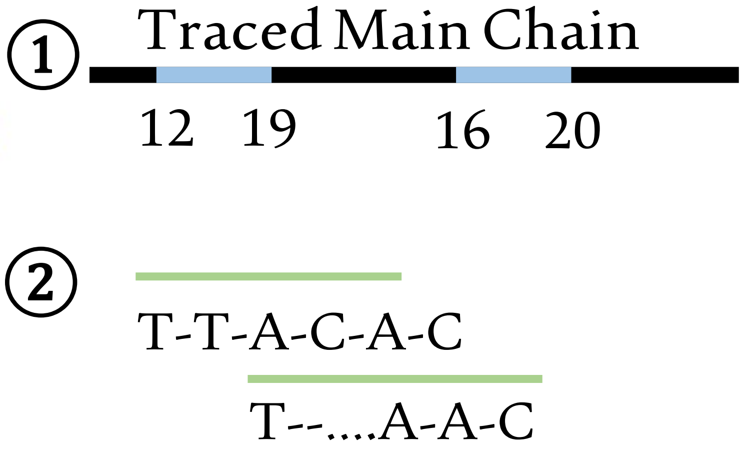
Step 3: Sequence Assignment



Multiple Assignments 😭



Conflict Assignments 😭



Constraint Programming

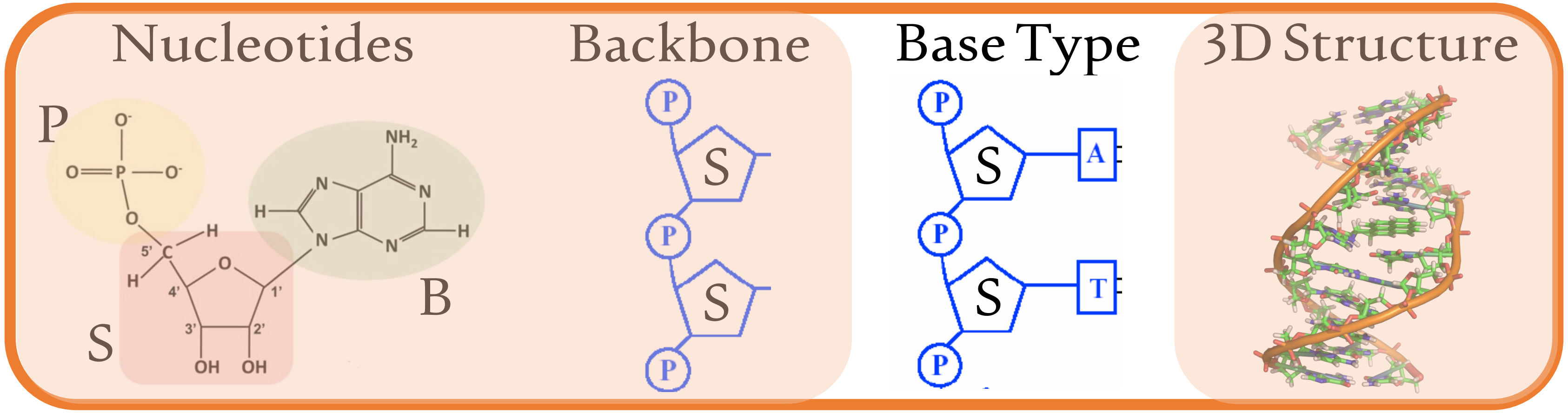
$$\max \sum_{i=1}^{N} x_i r_i$$

with $x_i \in \{0,1\}$

s.t. $x_i + x_j + x_k + \cdots + x_l = 1$

Wang Xiao, et al. "Cryo-READ: De novo structure modeling for nucleic acids in cryo-EM maps using deep learning". Nature Methods. (2023)

Step 3: Sequence Assignment



Nucleotides

Backbone

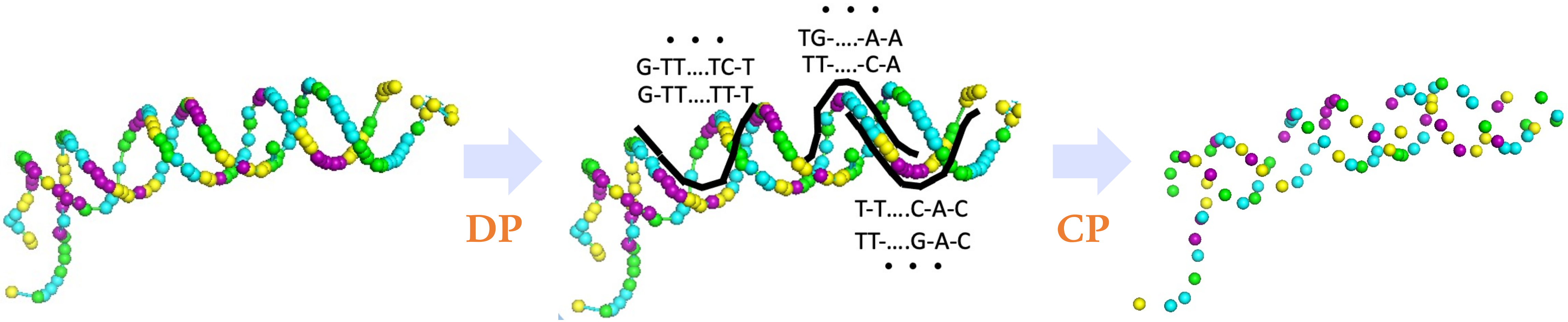
Base Type

3D Structure

Initial Assignment

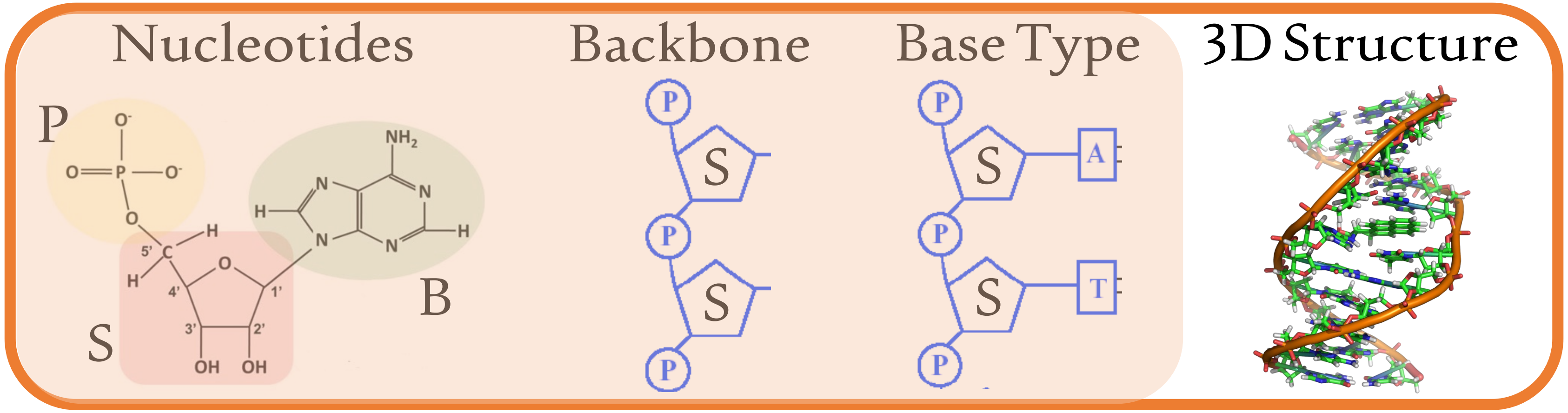
Fragment-based Assignment

Final Assignment



Wang Xiao, et al. “Cryo-READ: De novo structure modeling for nucleic acids in cryo-EM maps using deep learning”. Nature Methods. (2023)

Step 4: Atomic Structure Modeling

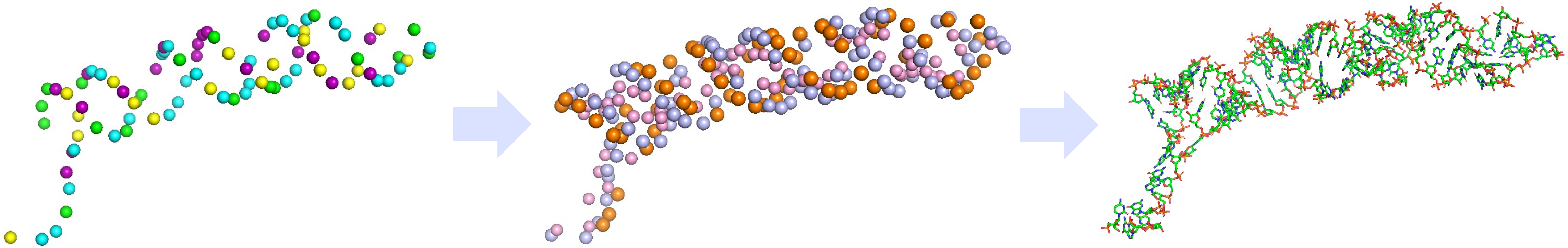


Nucleotides

Backbone

Base Type

3D Structure



Sequence Assignment

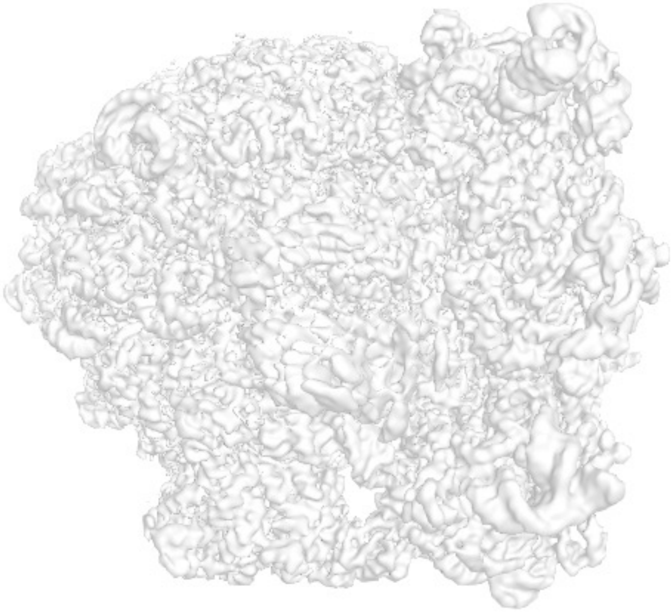
S-P-B Backbone

Atomic Structure

Wang Xiao, et al. "Cryo-READ: De novo structure modeling for nucleic acids in cryo-EM maps using deep learning". Nature Methods. (2023)

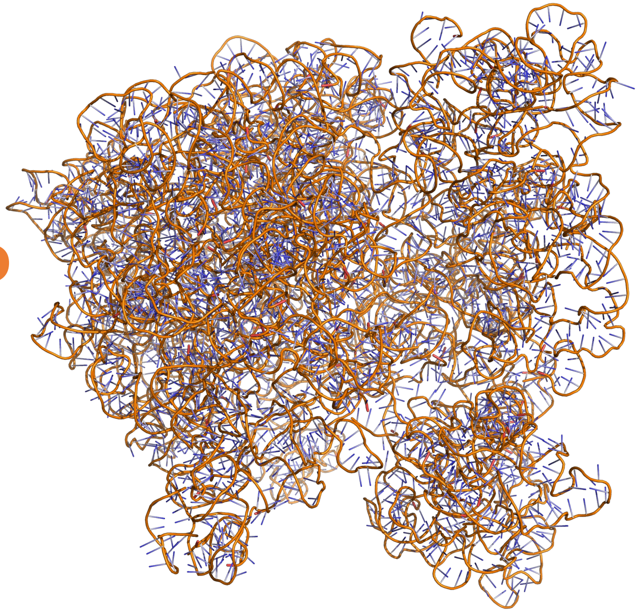
RNA-protein Complex Example by CryoREAD

Cryo-EM Map

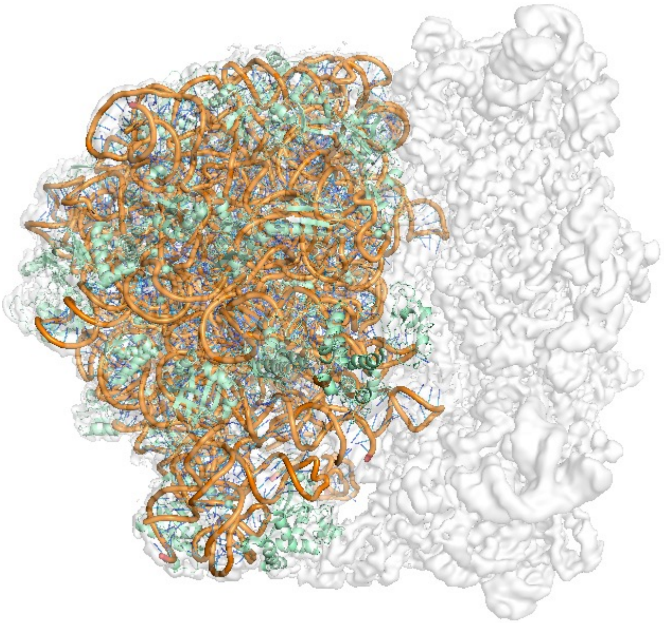


Cryo READ

Modeled Structure



Native Structure (Ground Truth)

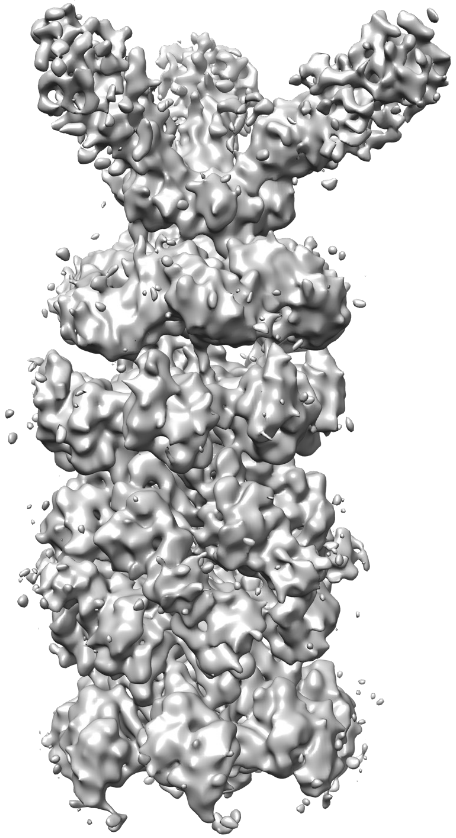


Automatically model **full** structure

Wang Xiao, et al. "Cryo-READ: De novo structure modeling for nucleic acids in cryo-EM maps using deep learning". Nature Methods. (2023)
Thoms, Matthias, et al. "Structural basis for translational shutdown and immune evasion by the Nsp1 protein of SARS-CoV-2." *Science* 369.6508 (2020): 1249-1255.

Method 2: DiffModeler: Protein Complex Structure Modeling

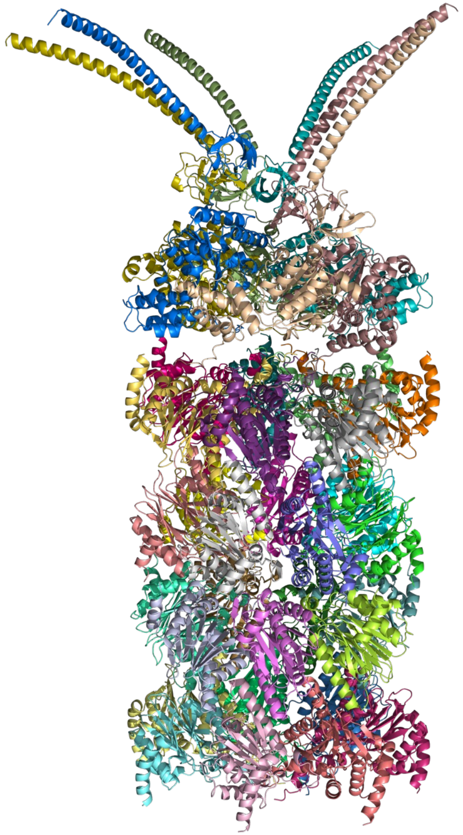
Cryo-EM Map



Structure Modeling

DiffModeler

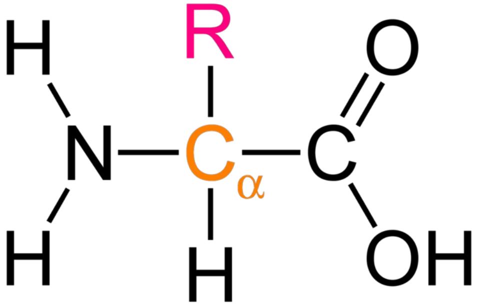
Protein Structure



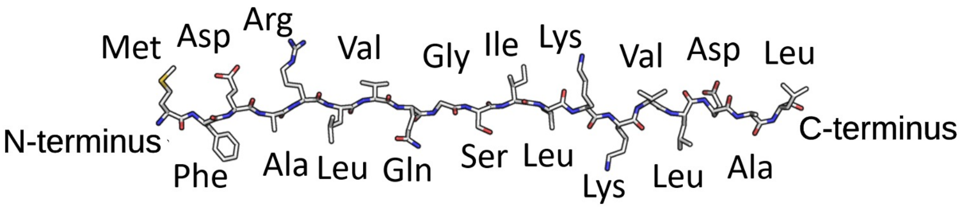
Wang Xiao, et al. “DiffModeler: Protein Complex Structure Modeling with Diffusion Model and AlphaFold in cryo-EM maps”. In preparation. (2023)

Background: Protein Structure

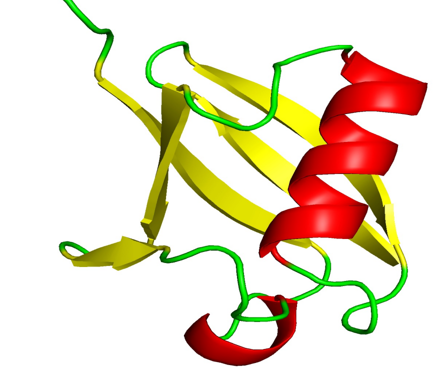
Amino Acid



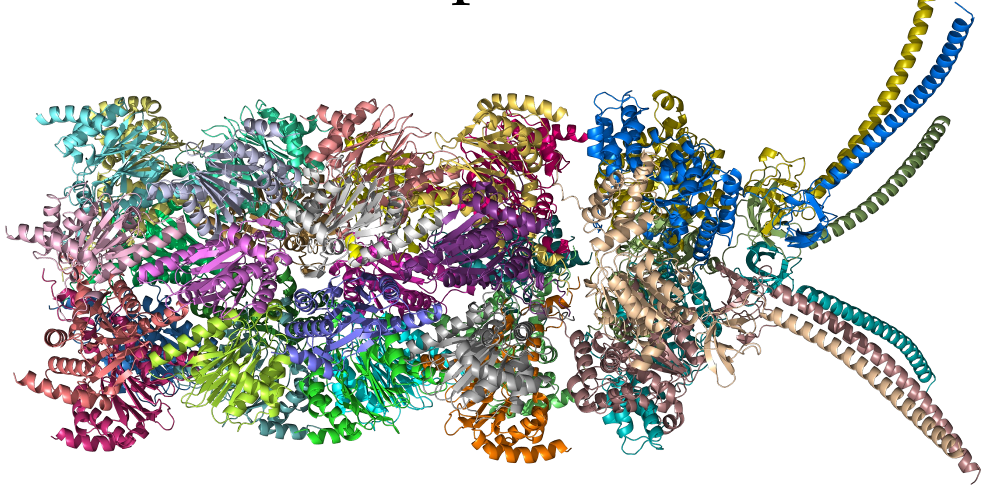
Sequence



Structure



Protein Complex Structure

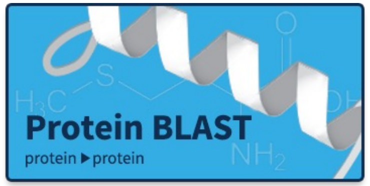


Wang Xiao, et al. “DiffModeler: Protein Complex Structure Modeling with Diffusion Model and AlphaFold in cryo-EM maps”. In preparation. (2023)

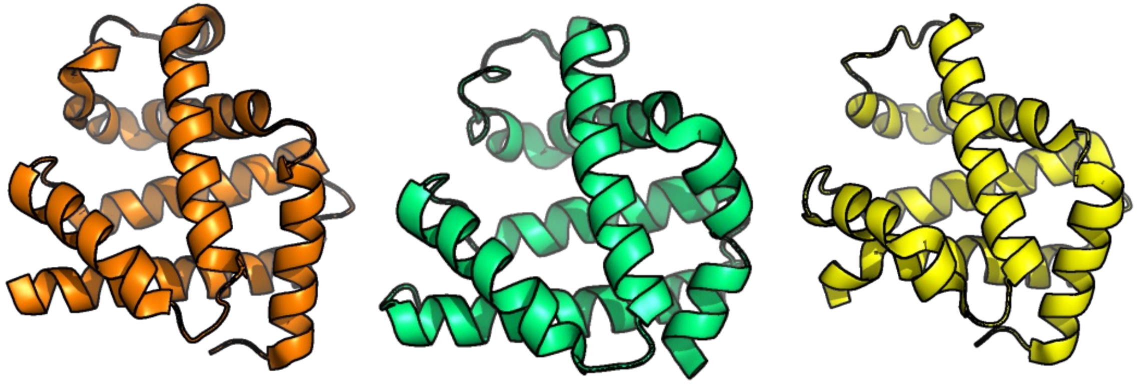
Background: Template-based Structure Modeling

Sequence SQETRKKCTEMKKKFKNCEVRCDESNHCVEVRCSDTKYTLC

Sequence Similarity Search

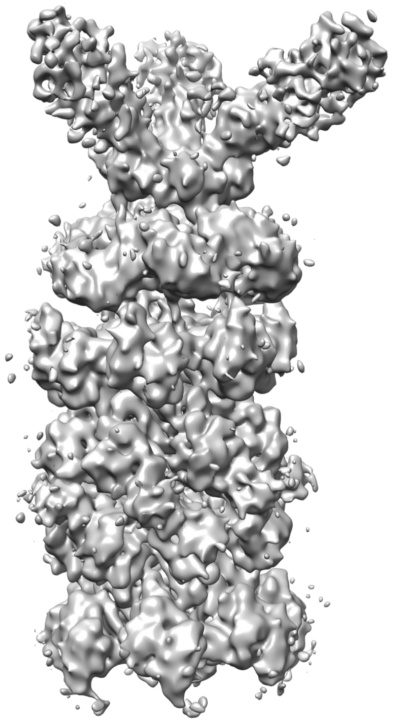


Similar Structures



Fitting

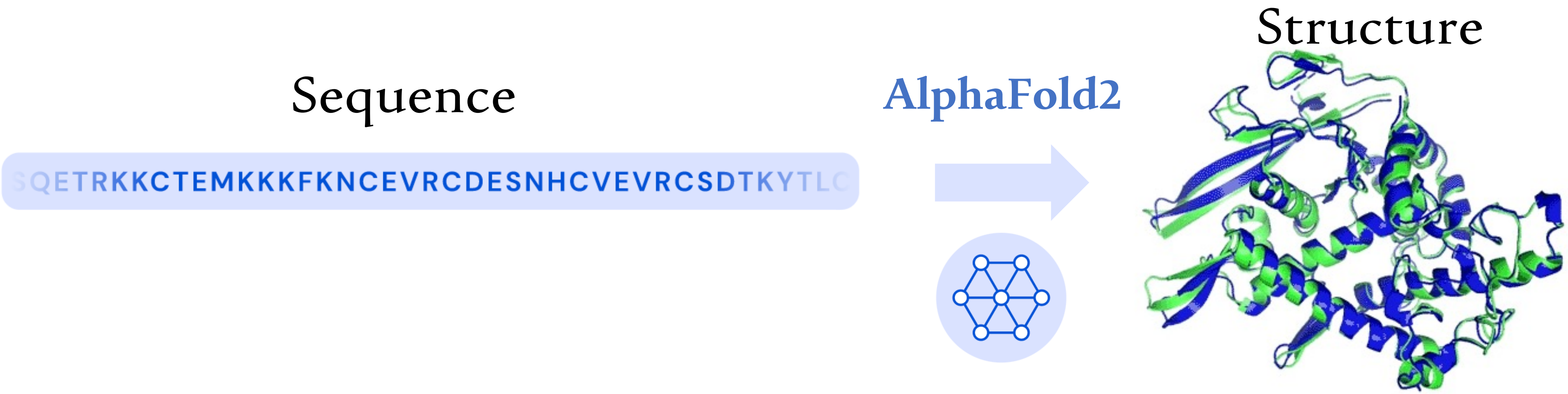
Cryo-EM Map



Challenges: 1. *Where are the templates(candidates)*? 2. *Where to fit*?

Wang Xiao, et al. "DiffModeler: Protein Complex Structure Modeling with Diffusion Model and AlphaFold in cryo-EM maps". In preparation. (2023)

Where are the templates(candidates): AlphaFold2



Sequence

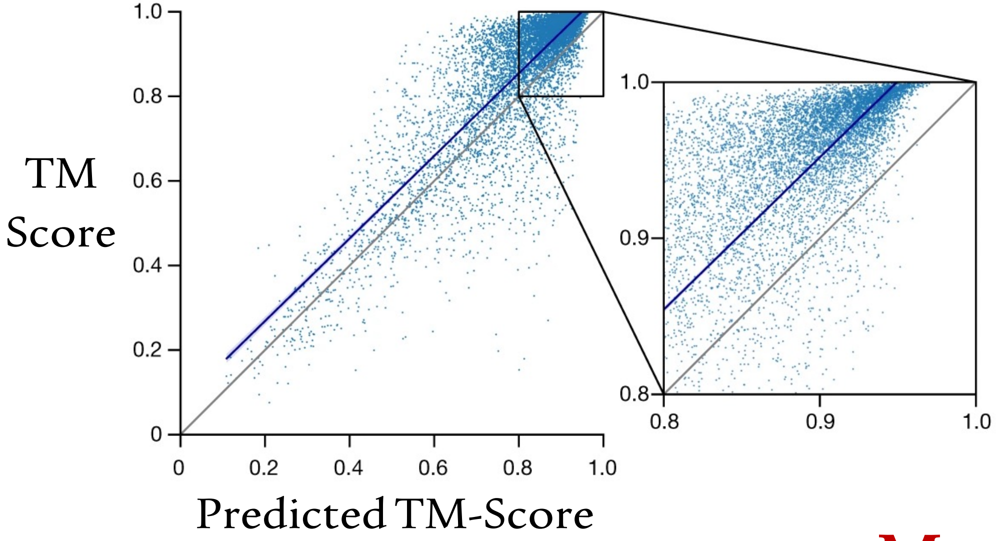
SQETRKKCTEMKKKFKNCEVRCDESNHCVEVRCSDTKYTLC

AlphaFold2

Structure

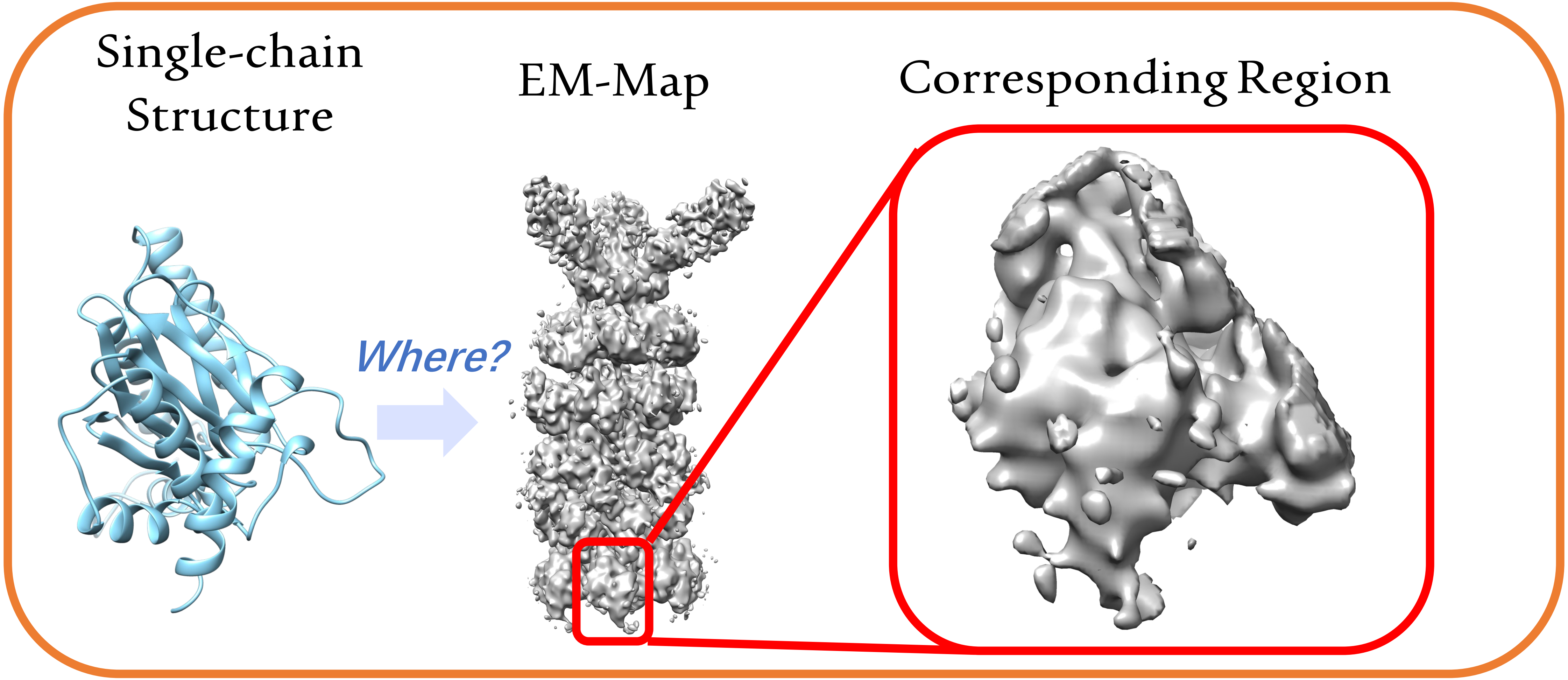
Highly accurate structure by AlphaFold2 can serve as **reliable templates** for structure modeling in cryo-EM maps

Performance



Wang Xiao, et al. "DiffModeler: Protein Complex Structure Modeling with Diffusion Model and AlphaFold in cryo-EM maps". In preparation. (2023)

Where to fit?



Single-chain Structure

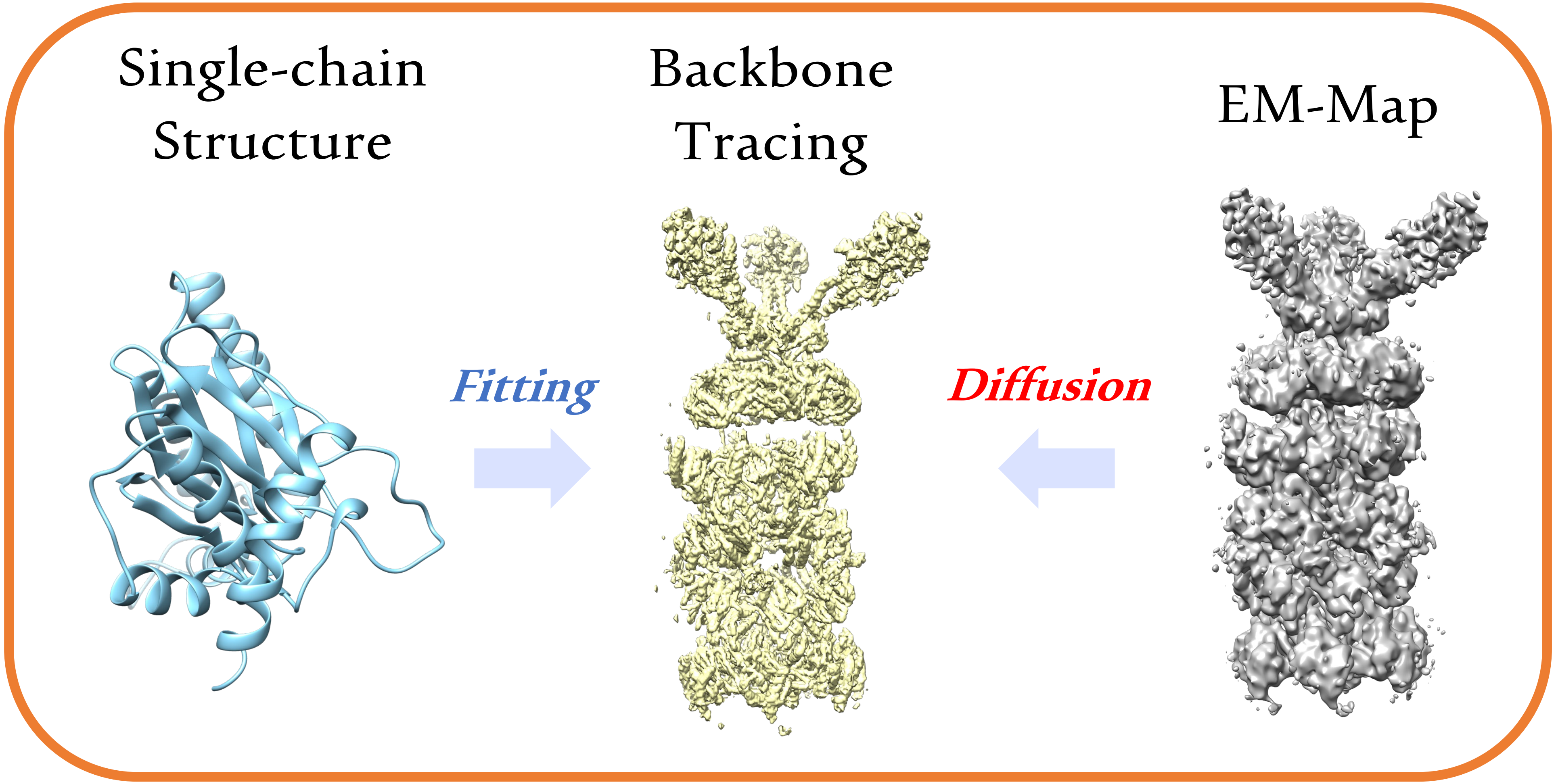
Where?

EM-Map

Corresponding Region

Wang Xiao, et al. "DiffModeler: Protein Complex Structure Modeling with Diffusion Model and AlphaFold in cryo-EM maps". In preparation. (2023)

Where to fit? Backbone tracing via diffusion model



Single-chain Structure

Fitting

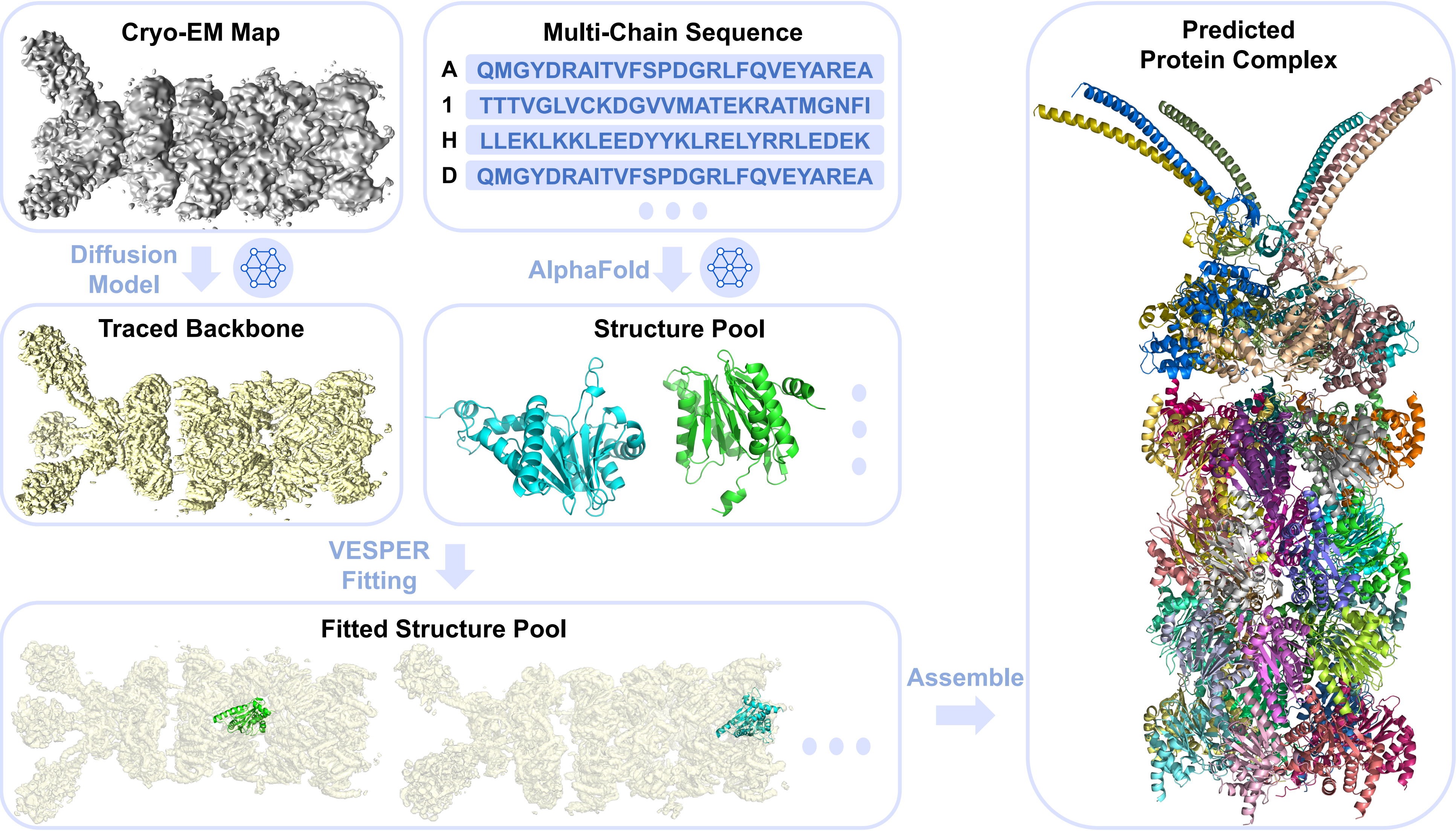
Backbone Tracing

Diffusion

EM-Map

Wang Xiao, et al. "DiffModeler: Protein Complex Structure Modeling with Diffusion Model and AlphaFold in cryo-EM maps". In preparation. (2023)

Overview of DiffModeler



Cryo-EM Map

Diffusion Model

Traced Backbone

Multi-Chain Sequence

A QMGYDRAITVFSPDGRLFQVEYAREA

1 TTTVGLVCKDGVVMATEKRATMGNFI

H LLEKLKKLEEDYYKLRELYRRLEDEK

D QMGYDRAITVFSPDGRLFQVEYAREA

• • •

AlphaFold

Structure Pool

VESPER Fitting

Fitted Structure Pool

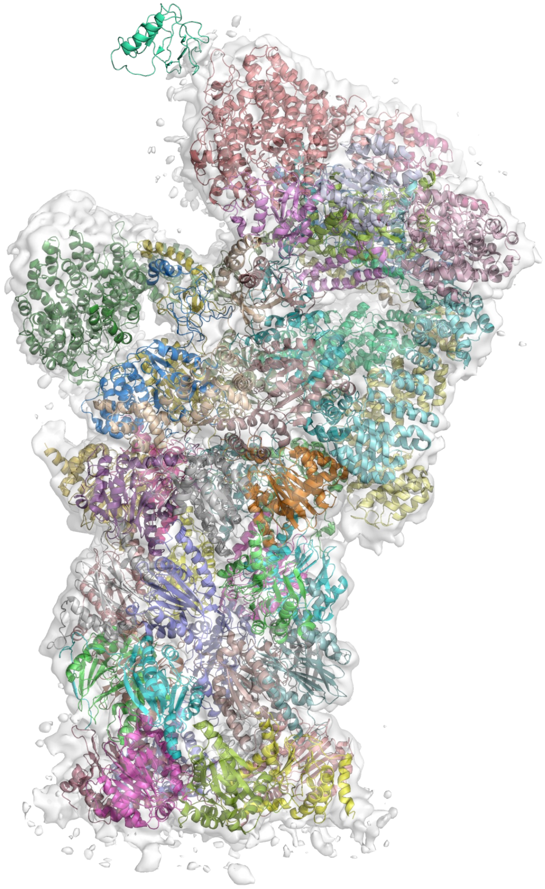
Assemble

Predicted Protein Complex

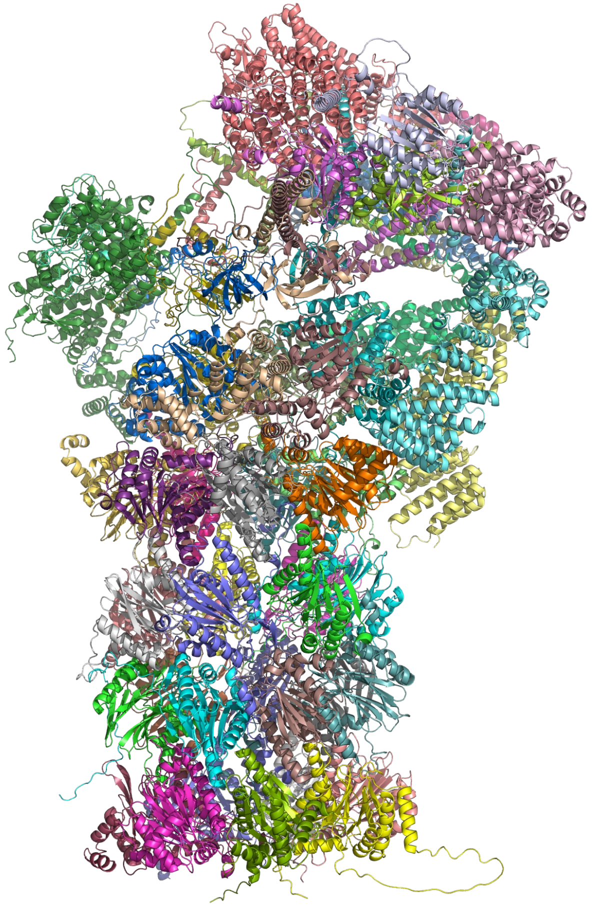
Wang Xiao, et al. "DiffModeler: Protein Complex Structure Modeling with Diffusion Model and AlphaFold in cryo-EM maps". In preparation. (2023)

DiffModeler Modeled Structure

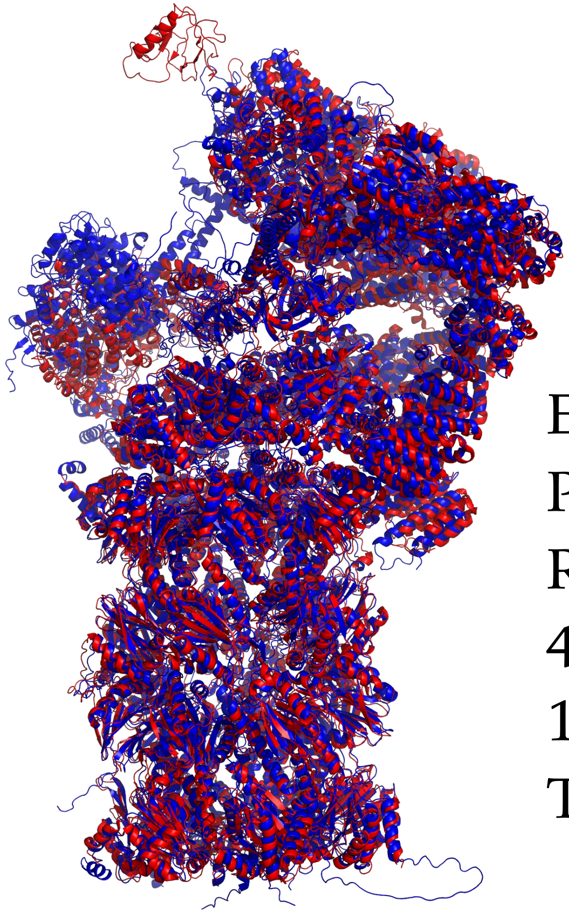
EM-Map and Native Structure



Structure by DiffModeler



Structure Comparison



🟥 **Ground Truth**

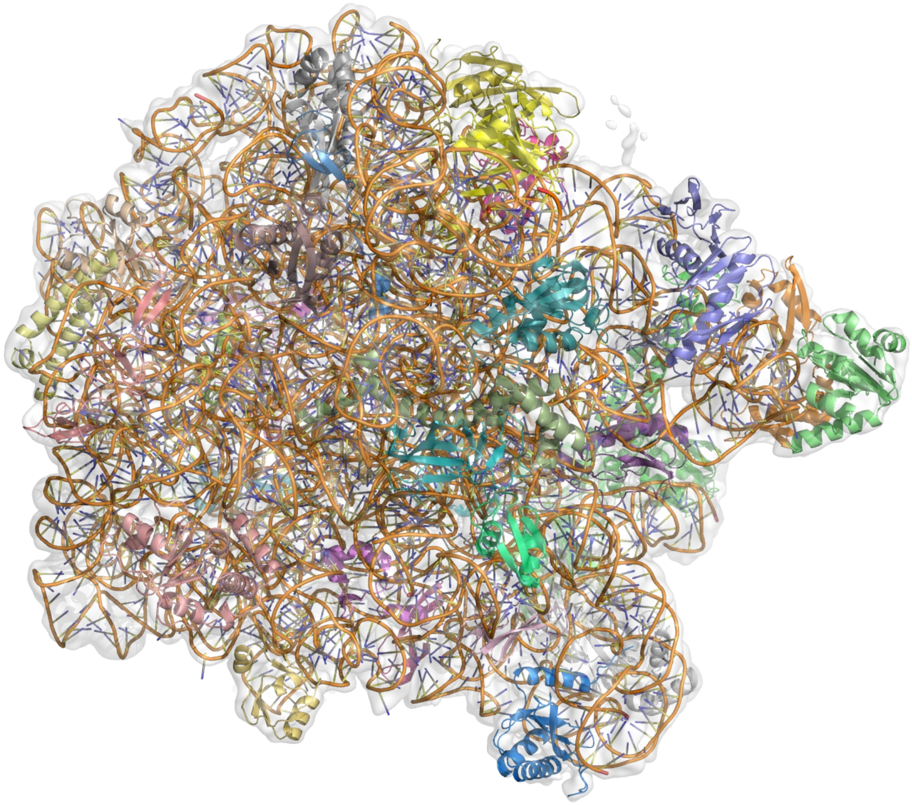
🟦 **Predicted Structure**

EMD-6693
PDB-ID: 5WVI
Resolution: 6.30 Å
47 chains
13,462 residues
TM-Score: **0.94**

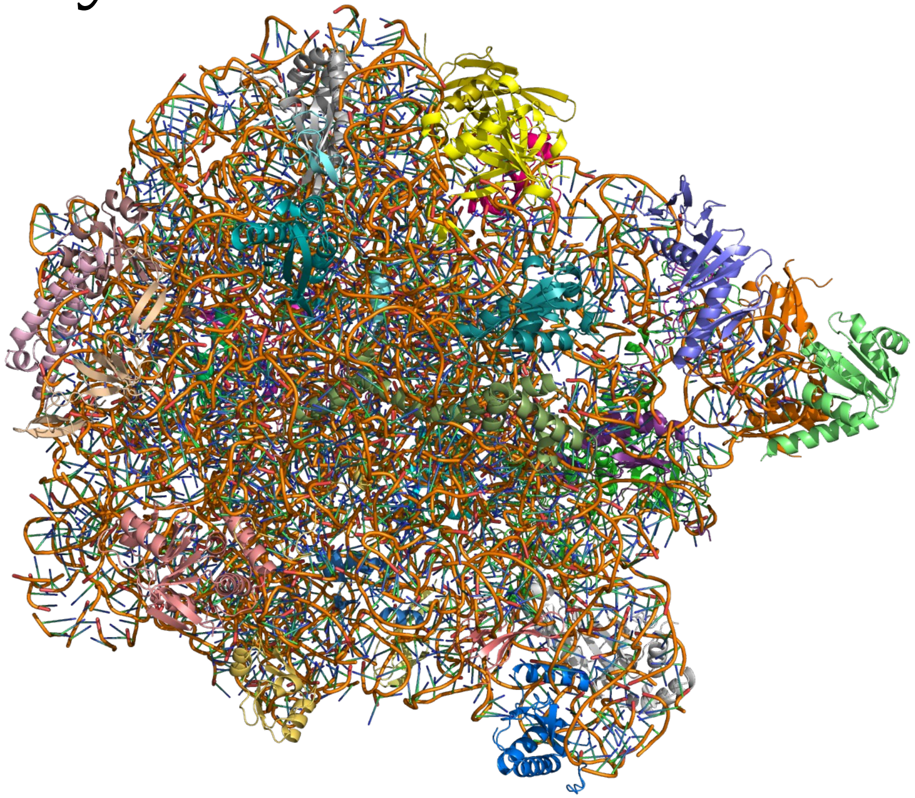
Wang Xiao, et al. "DiffModeler: Protein Complex Structure Modeling with Diffusion Model and AlphaFold in cryo-EM maps". In preparation. (2023)

Protein-RNA complex by DiffModeler+CryoREAD

Cryo-EM Map and Ground Truth Structure



Modeled Structure by CryoREAD and DiffModeler



EMD-13017
PDB-ID: 7OPE
Resolution: 3.20 Å
3,818 residues
2,996 nucleotides
TM-Score(protein): **0.92**
Backbone Recall (RNA): **0.94**

Wang Xiao, et al. "DiffModeler: Protein Complex Structure Modeling with Diffusion Model and AlphaFold in cryo-EM maps". In preparation. (2023)

Acknowledgement