

# Modeling 3D Structures of Covid-19 Proteins Using Deep Learning

Jie Hou, PhD, Project Advisor and Client

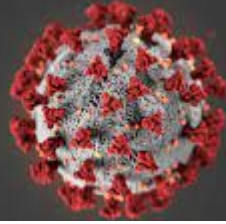
Justin Dulay

Jack Do





Coronavirus  
(COVID-19)



INDIA DISPATCH

## *'This Is a Catastrophe.' In India, Illness Is Everywhere.*

As India suffers the world's worst coronavirus crisis, our New Delhi bureau chief describes the fear of living amid a disease spreading at such scale and speed.

## **EU pivots to Pfizer with world's biggest Covid-19 vaccine deal as it sues AstraZeneca**

By [Angela Dewan](#), [Stephanie Halasz](#) and [Chris Liakos](#), CNN Business

Updated 10:53 AM ET, Wed April 28, 2021

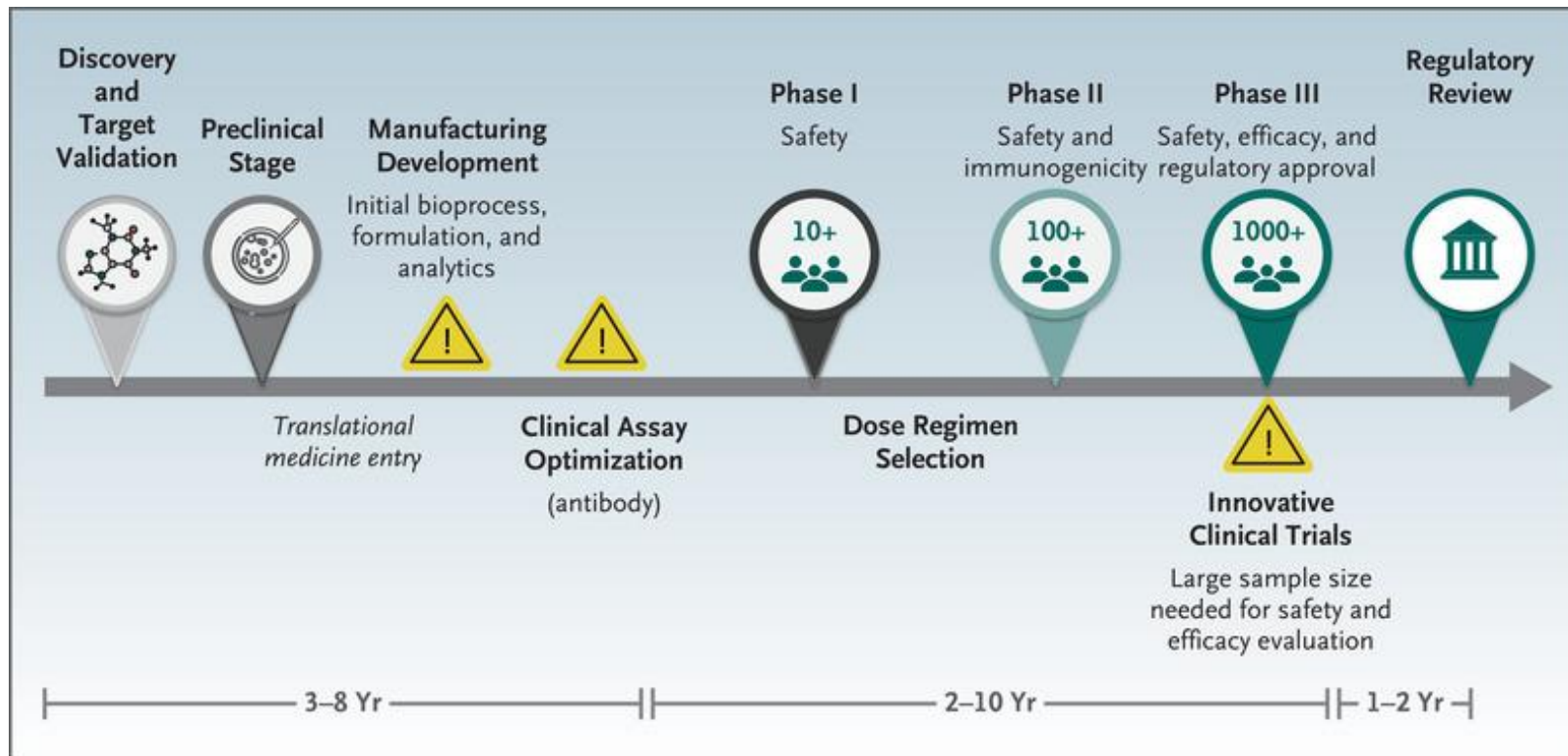
HEALTH AND SCIENCE

## **Pfizer's new at-home pill to treat Covid could be available by end of the year, CEO hopes**

PUBLISHED TUE, APR 27 2021·11:02 AM EDT | UPDATED TUE, APR 27 2021·12:14 PM EDT

# Problem

- Pressing need to disseminate information to researchers
- Slow process to obtain structure of complete proteins from experiments
- Knowing the protein structures aids to vaccine development



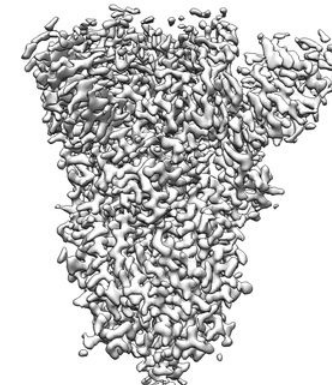
# SARS-CoV-2 proteins with unknown structures

**EMD-21374** 2019-nCoV spike glycoprotein with C3 symmetry imposed **SARS-CoV-2** **No Deposited Model**

[View EMD-21374 in EMDB](#)

**Publication** Cryo-EM structure of the 2019-nCoV spike in the prefusion conformation.  
**Authors** Wrapp D, Wang N, Corbett KS, Goldsmith JA, Hsieh CL, Abiona O, Graham BS, McLellan JS  
**Release Date** 2020-02-26

Note: No available modeled structure

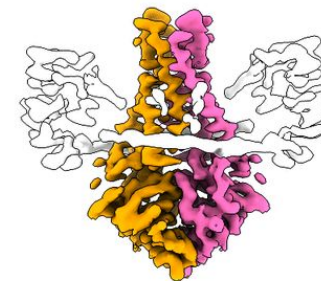


**EMD-22139** SARS-CoV-2 ORF3a with Emodin in a MSP1E3D1 lipid nanodisc **SARS-CoV-2** **No Deposited Model**

[View EMD-22139 in EMDB](#)

**Publication** Cryo-EM structure of the SARS-CoV-2 3a ion channel in lipid nanodiscs.  
**Authors** Kern DM, Sorum B, Hoel CM, Sridharan S, Remis JP, Toso DB, Brohawn SG  
**Release Date** 2020-06-17

Note: No available modeled structure



**EMD-22613** SARS-CoV-2 Nsp15 H235A APO-state dataset **SARS-CoV-2** **No Deposited Model**

[View EMD-22613 in EMDB](#)

**Publication** Cryo-EM structures of the SARS-CoV-2 endoribonuclease Nsp15 reveal insight into nuclea  
**Authors** Pillon MC, Frazier MN, Dillard LB, Williams JG, Kocaman S, Krahn JM, Perera L, Hayne CK, C  
**Release Date** 2020-12-09

Note: No available modeled structure





# Deep Tracer for atom identification



Jobs

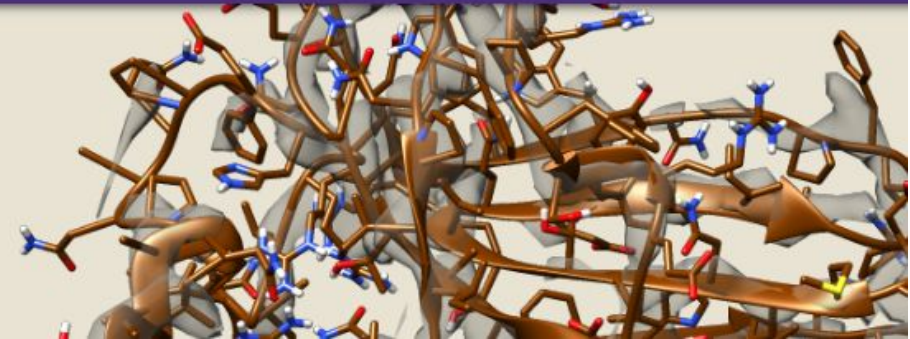
Datasets

Coronavirus

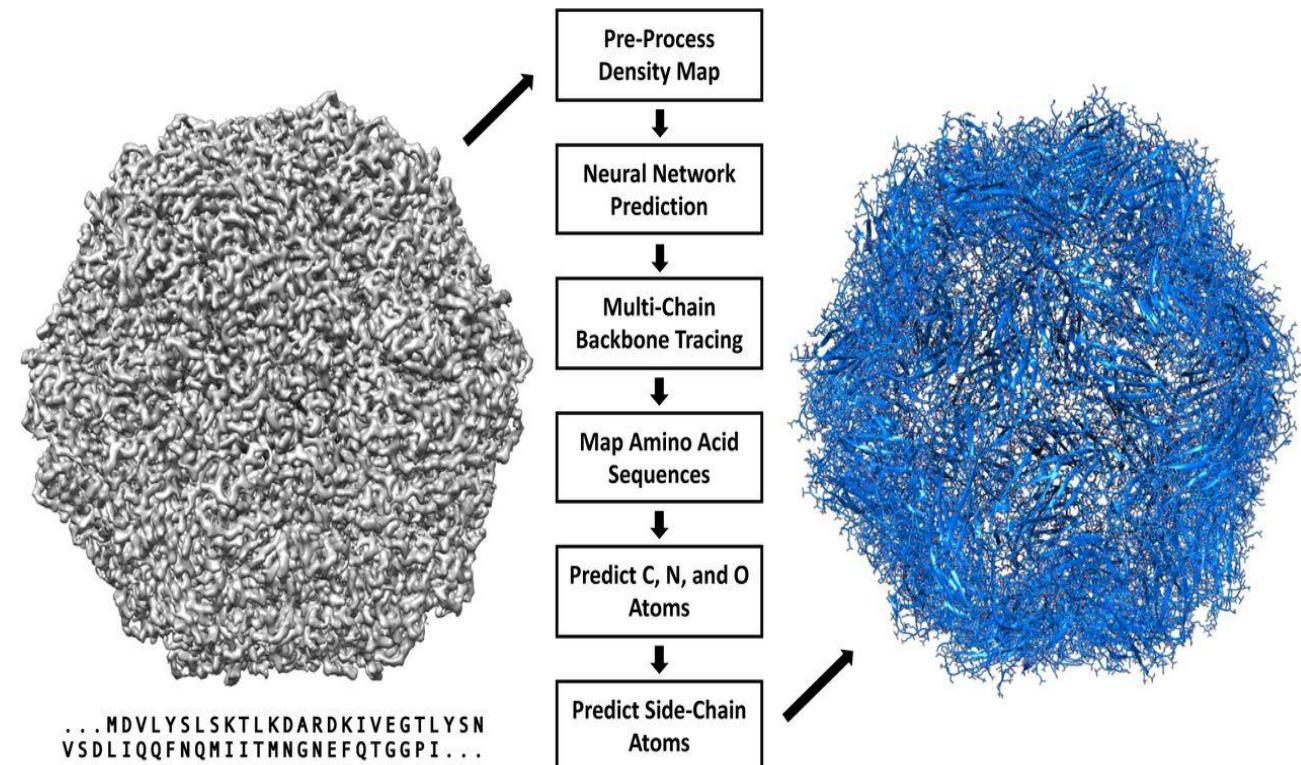
Hot

About Us

## Protein Complex Structure Prediction from Cryo-EM Density Maps

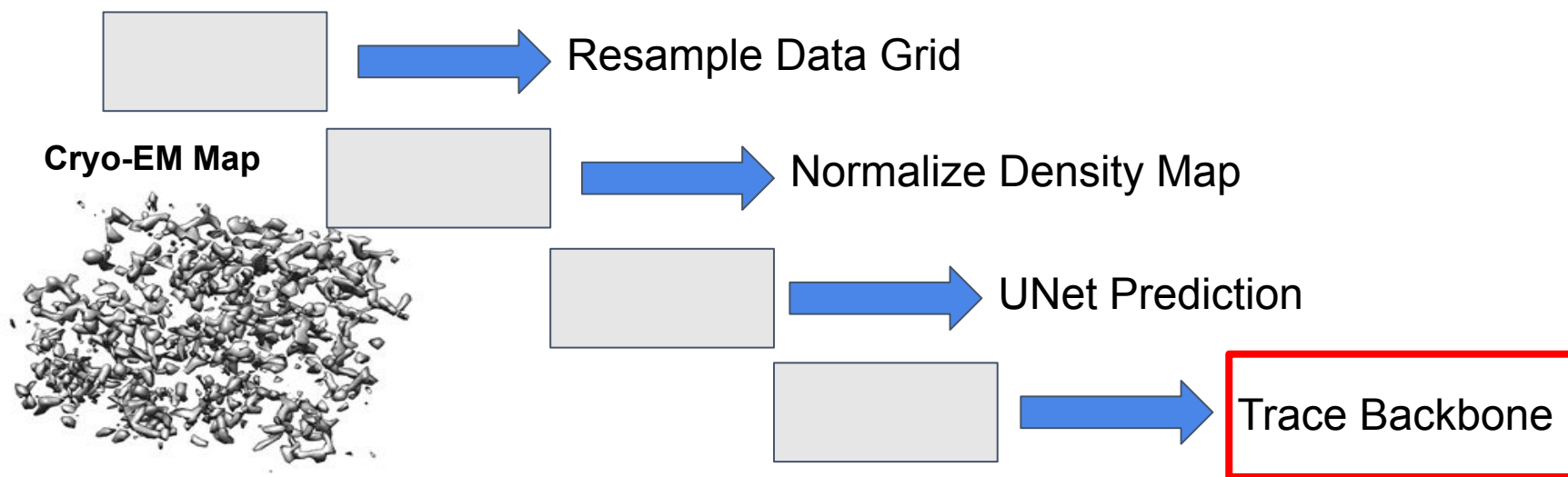


→ Deep Tracer is a web application that determines the structure of a protein complex and allows the user to recognize carbon- $\alpha$  atoms from cryo-EM density map.



<https://deepttracer.uw.edu/home>

# Deep Tracer for atom identification



✓ EMD-11173

emd\_11173.map Cryo-EM Density Map



Trace Backbone

**Input** Density Map, Ca Atoms Pred, Backbone Pred

**Output** [Predicted Structure](#)

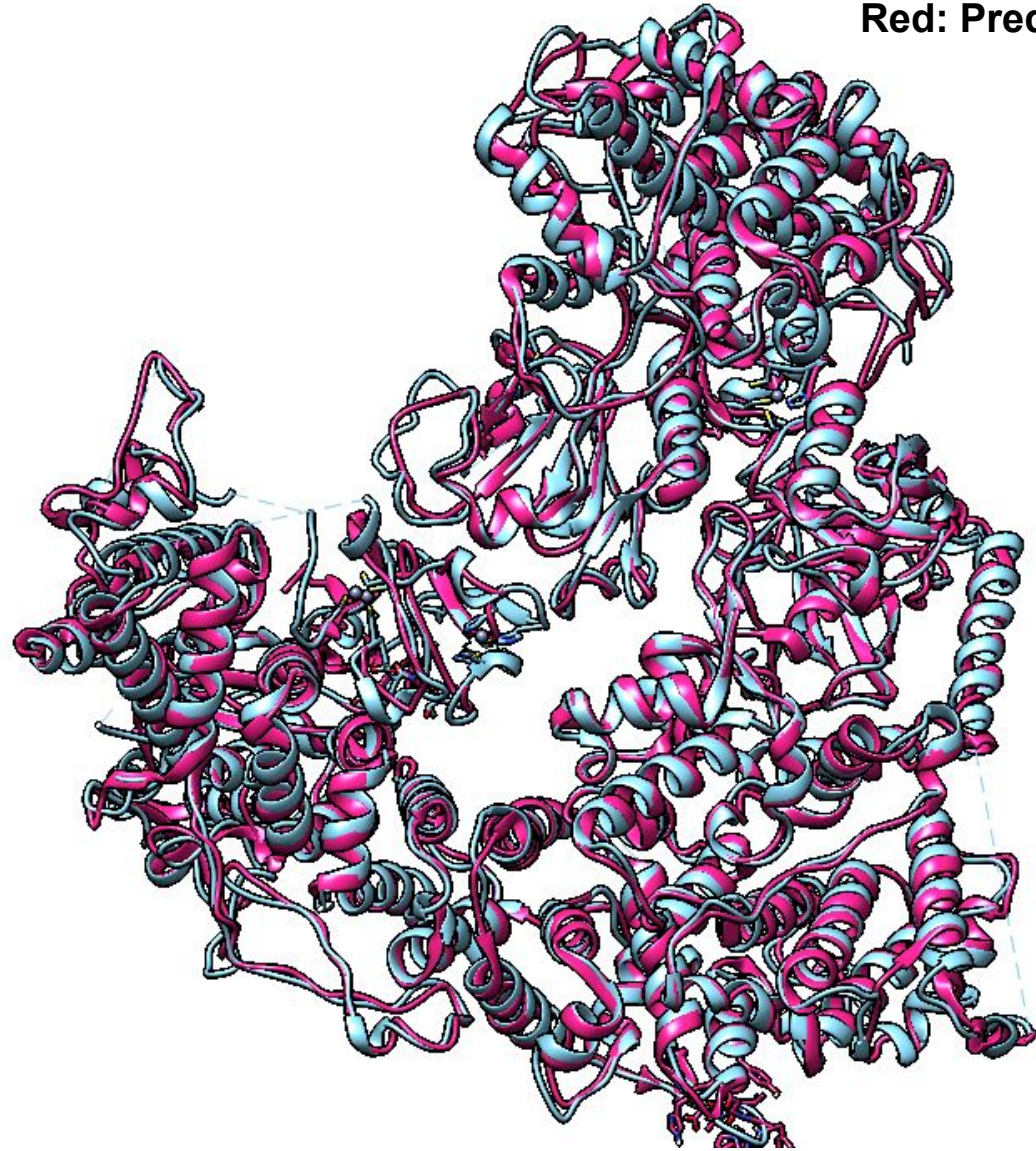
```
[Sep 30, 2020, 8:53:13 PM] Starting new prediction step
[Sep 30, 2020, 8:53:18 PM] Tracing 589 amino acids in 1 chain(s)
[Sep 30, 2020, 8:53:19 PM] Tracing chain with 589 amino acids
[Sep 30, 2020, 8:53:25 PM] Finished computing confidence matrix
[Sep 30, 2020, 8:53:50 PM] Connected 589 amino acids to 39 chains
```



# Current Limitation for Deep Tracer

**Blue:** True structure

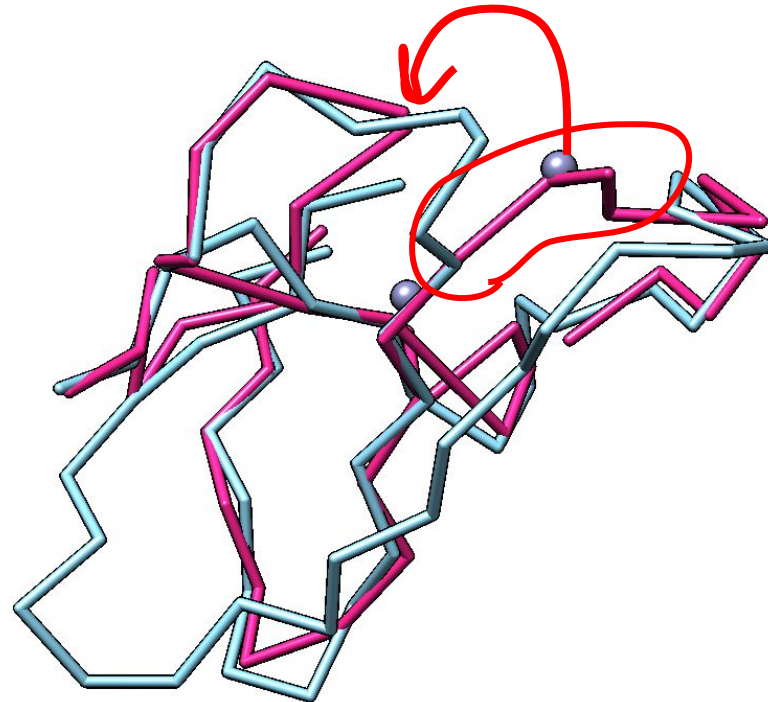
**Red:** Predicted structure by Deep Tracer



**Good:** Accurate prediction on Ca locations

**Limit:** local connections between Ca atoms are not always good

This points connection is incorrectly placed

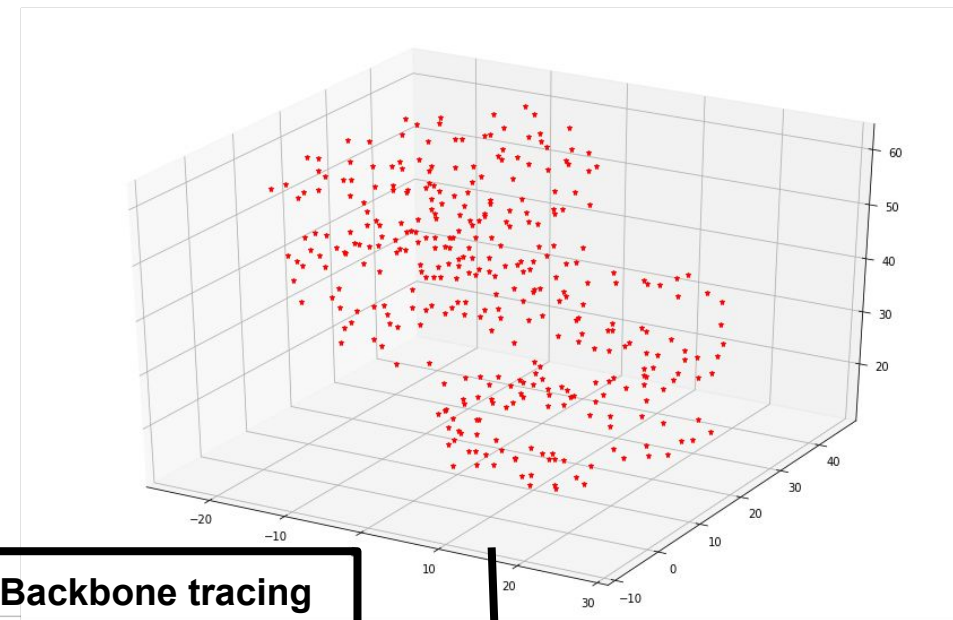


# Backbone tracing in Protein structure prediction

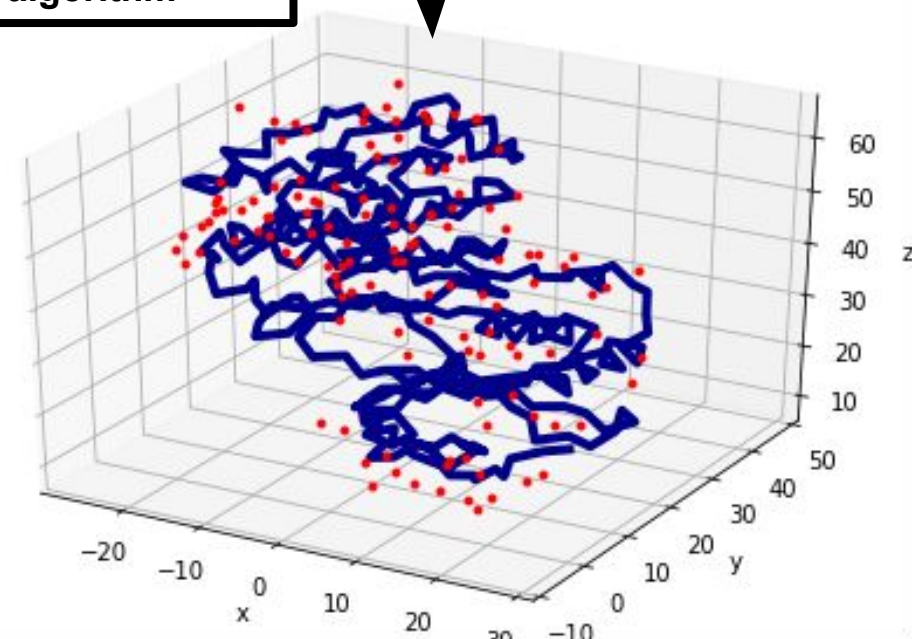
Cryo-EM Map



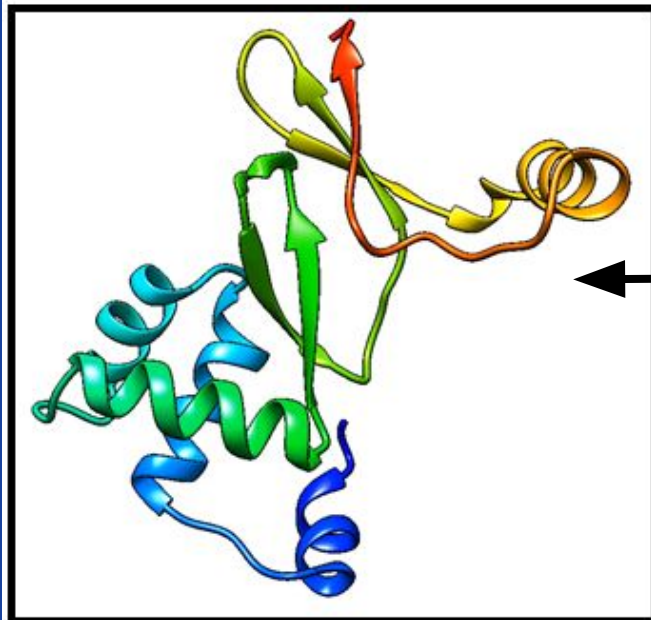
Atom identification by deep learning



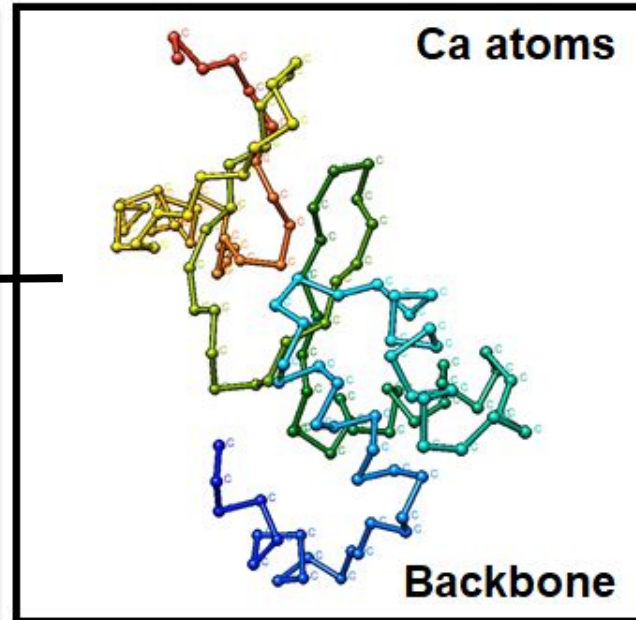
Backbone tracing algorithm



Protein tertiary structure



Protein backbone



# Similar problems in Computer Science



WIKIPEDIA  
The Free Encyclopedia

[Main page](#)  
[Contents](#)  
[Current events](#)  
[Random article](#)

Article [Talk](#)

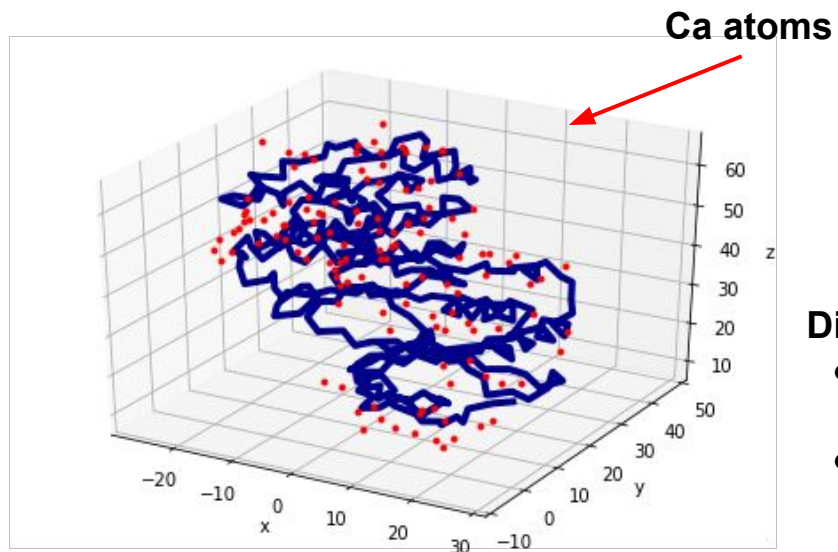
[Read](#) [Edit](#) [View history](#)

Not logged in

## Travelling salesman problem

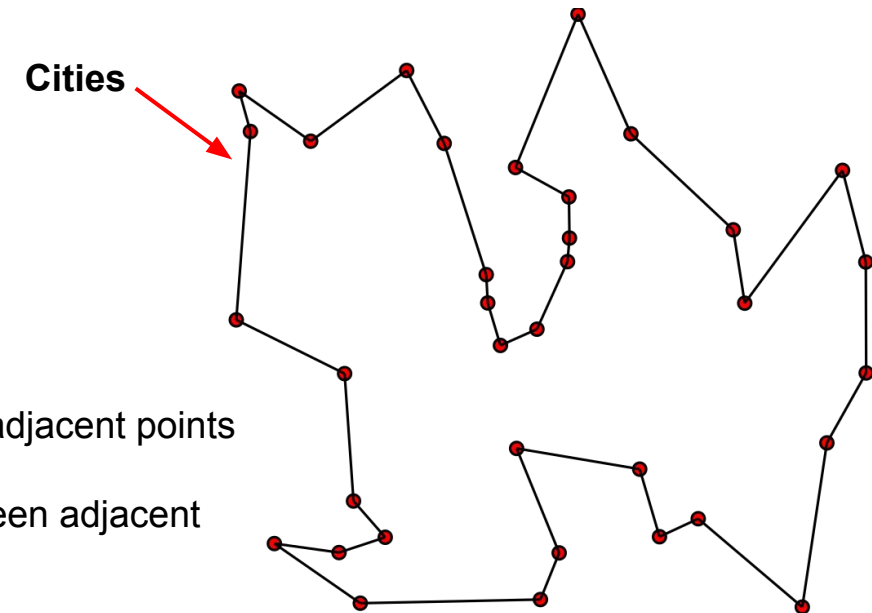
From Wikipedia, the free encyclopedia

The **travelling salesman problem** (also called the **traveling salesperson problem**<sup>[1]</sup> or **TSP**) asks the following question: "Given a list of cities and the distances between each pair of cities, what is the shortest possible route that visits each city exactly once and returns to the origin city?" It is an NP-hard problem in [combinatorial optimization](#), important in [theoretical computer science](#) and [operations research](#).



### Difference:

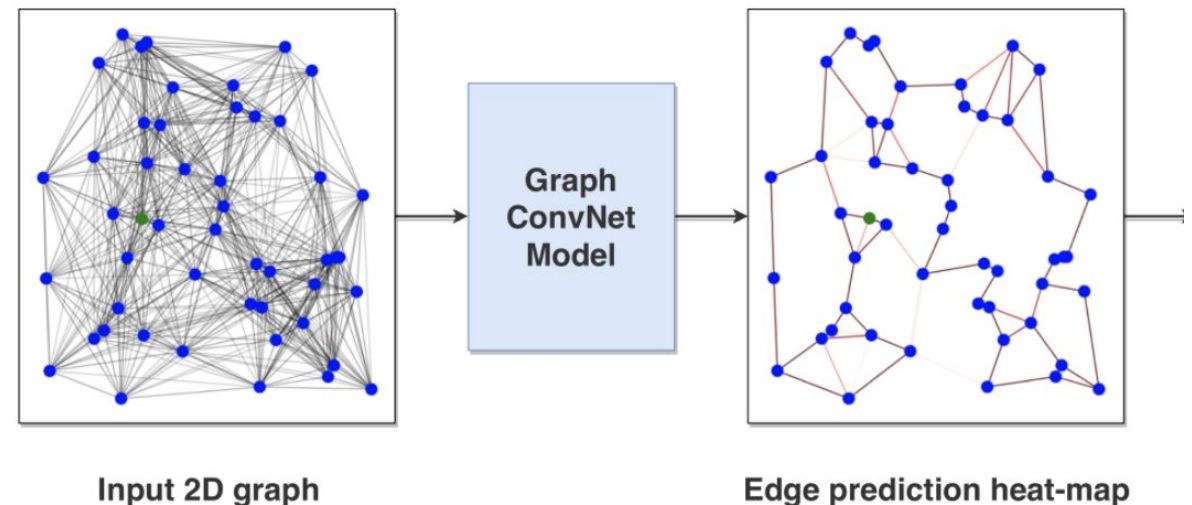
- Among cities, the distance between adjacent points **may differ**
- Among Ca atoms, the distance between adjacent residues is **around 3.8**



# Introducing Traveling Salesman in a Graph Neural Network

Our goal/contribution:

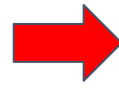
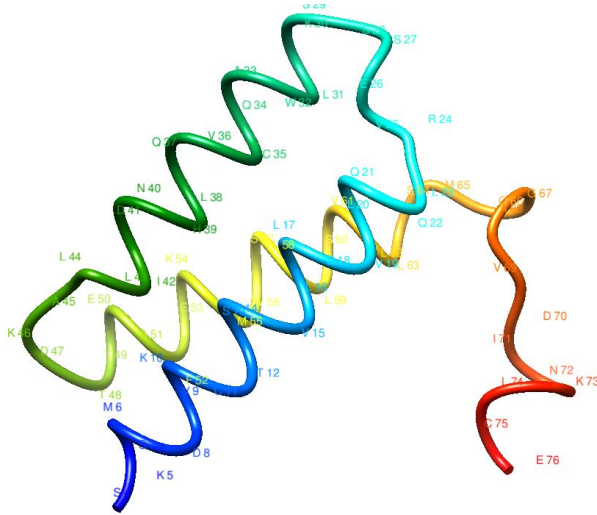
- Algorithm for the connections between atoms
- Practicing traditional traveling salesman algorithm
- Generalize from 2D to 3D training
- Improve computational efficiency, and reduce running time
- Utilize deep learning to form connections



Nazari, Mohammadreza, et al. "Deep Reinforcement Learning for Solving the Vehicle Routing Problem." arXiv preprint arXiv:1802.04240 (2018).

# Data Formatting

Training sample

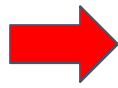
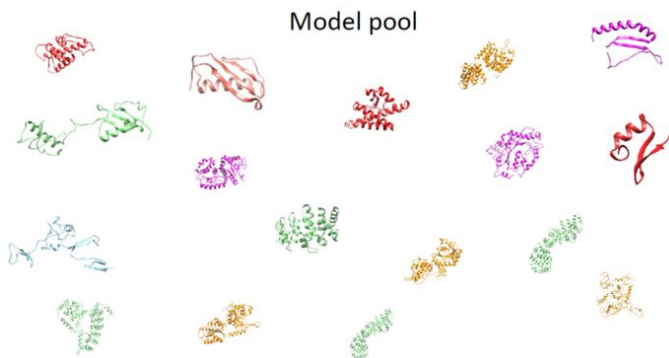


	ResId	AA	x	y	z
0	1	V	46.980000	39.907001	20.481001
1	2	S	46.825001	39.145000	16.837000
2	3	Y	44.955002	36.115002	15.677000
3	4	S	47.201000	33.519001	14.235000
4	5	D	45.662998	31.371000	11.612000

True routes

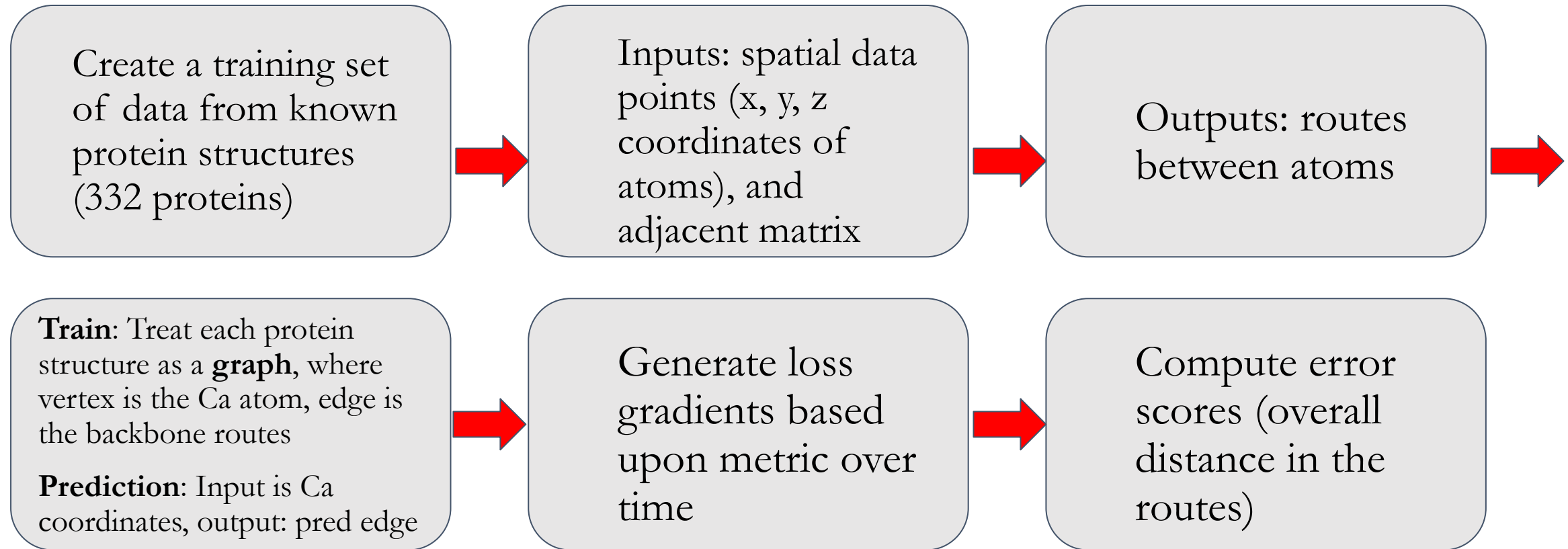
CA coordinates

Training set



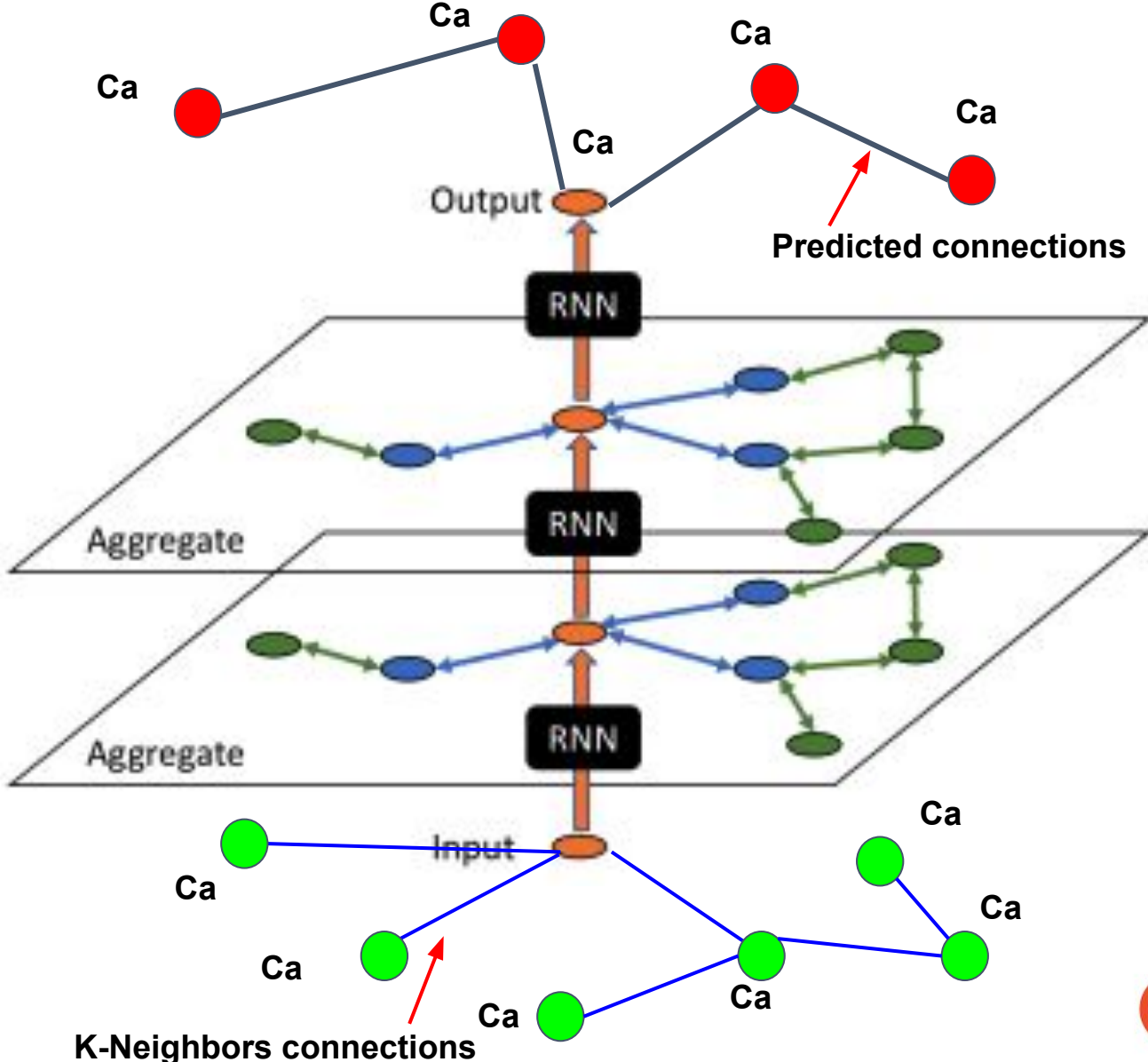
- List of 3D data points
- Connections between point sets
- Implement changes into our previous network
- Separate train, validation, and test data

# How we use deep learning for TSP?



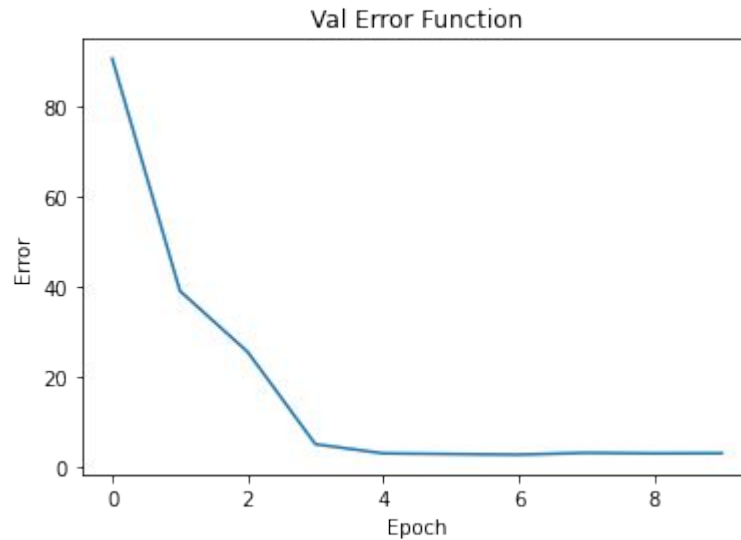
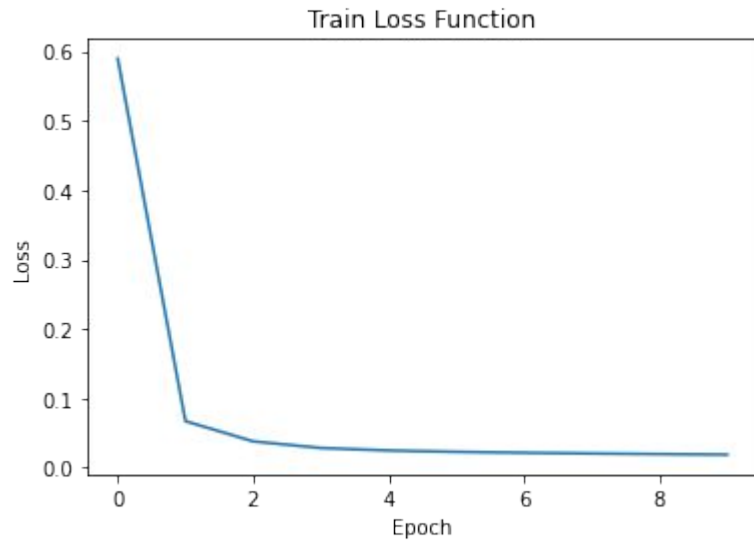
- The network acts a decoder network. Since the dataset is not in an initial order, it directly embeds each batch.
- The saved optimized model from this point onward can be used on test data without routes.

# Graph neural network for link prediction



# Results from GNN

- 92% training accuracy
  - Modest but noticeable improvement from last model
- 81% testing accuracy





	<b>PDB_id</b>	<b>Length</b>	<b>Predicted_Length</b>	<b>True_Length</b>	<b>Error</b>
<b>14</b>	T0845_filtered.pdb	426	572.503174	1690.935791	-1118.432617
<b>57</b>	T0821_filtered.pdb	255	326.697784	1046.697754	-719.999969
<b>12</b>	T0629_filtered.pdb	216	114.406349	832.452148	-718.045799
<b>47</b>	T0494_filtered.pdb	347	686.485901	1379.212646	-692.726746
<b>59</b>	T0848_filtered.pdb	321	611.973633	1292.772949	-680.799316
<b>39</b>	T0835_filtered.pdb	404	882.363525	1540.365479	-658.001953

# Problem: Results on test set

## EMDB › EMD-11207

### Furin Cleaved Spike Protein of SARS-CoV-2 in Closed Conformation


**Source organism:** *Severe acute respiratory syndrome coronavirus 2* [2697049]

**Fitted atomic model:** [6zgj](#)

**Related EM entries by publication:** [EMD-11203](#), [EMD-11204](#), [EMD-11205](#), [EMD-11206](#)

**3Dbionotes:** [available for this entry](#) 

#### Primary publication:

 SARS-CoV-2 and bat RaTG13 spike glycoprotein structures inform on virus evolution and furin-cleavage effects.

Wrobel AG, Benton DJ, Xu P, Roustan C, Martin SR, Rosenthal PB, Skehel JJ, Gamblin SJ

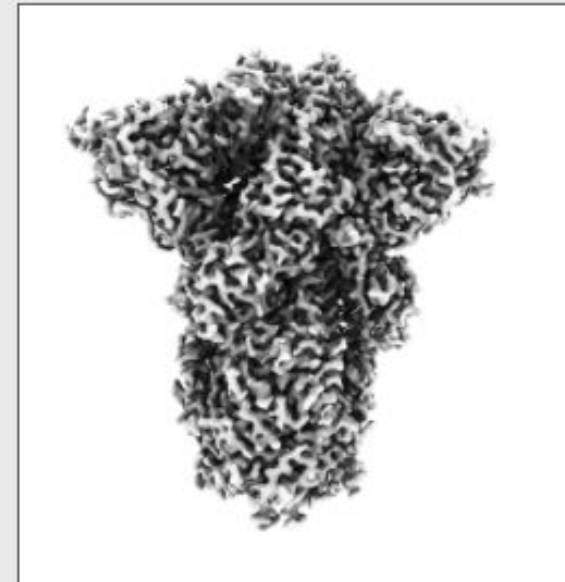
*Nat.Struct.Mol.Biol.* **27** 763-767 (2020)

PMID: [32647346](#)

**Single particle reconstruction  
2.9Å resolution**

**Map released:** 2020-07-01

**Last modified:** 2020-09-16



# Problem: Results on Predicted Length from GNN

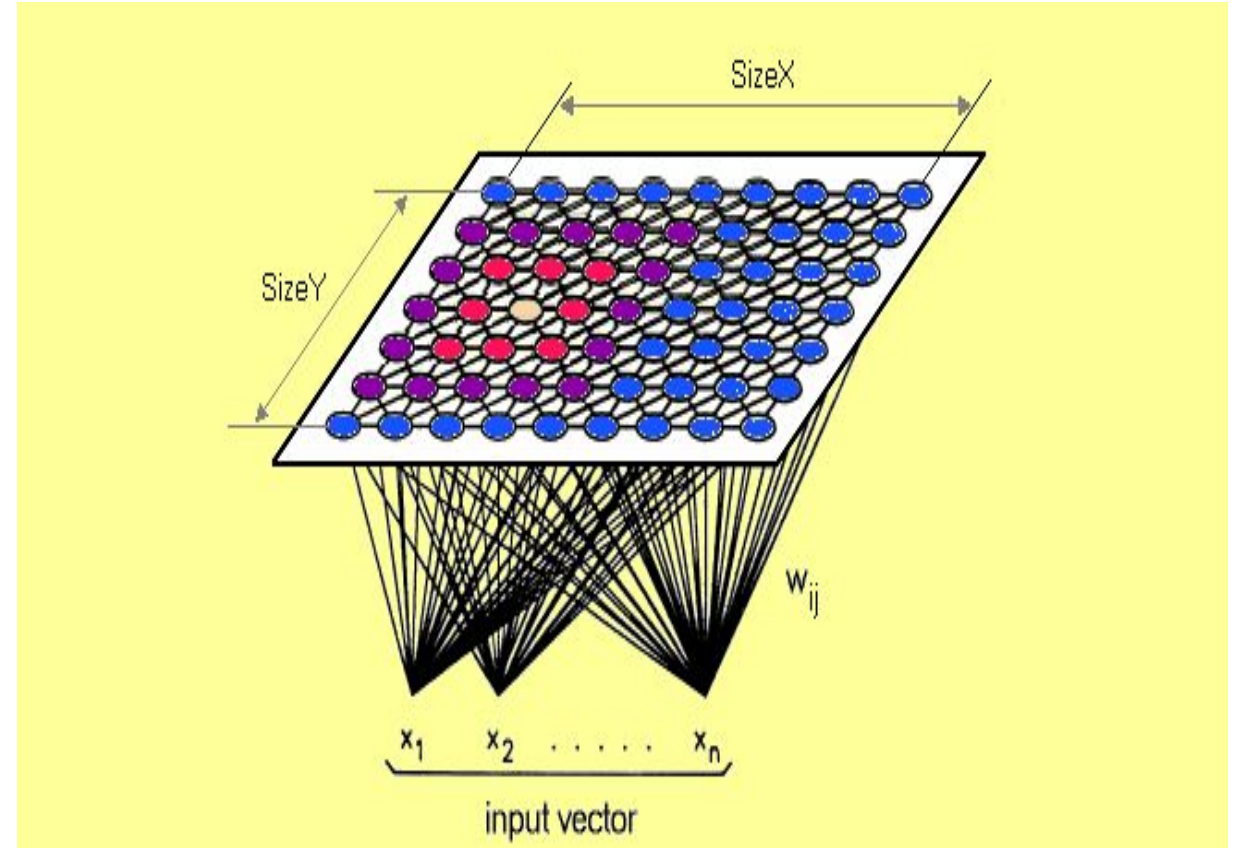
---

- 912 true length vs 1112 predicted length on single protein for example
- Next step, improve mode to match predictions on learned proteins
- This would require obtaining more training data to create a better estimated model
- It was faster



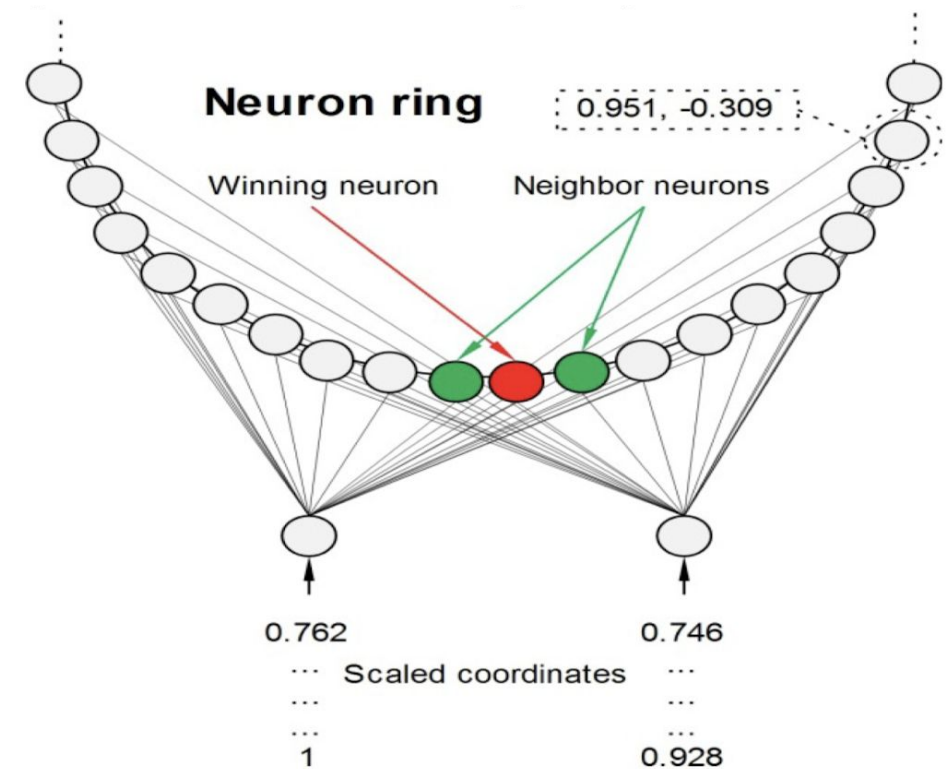
# Self Organizing Map

- Invented by *Teuvo Kohonen* around 1980
- Maps input vectors of any dimension onto map with one, two or more dimensions
- Unsupervised learning ANN (artificial neural network)



# Use SOM to solve TSP

- Given 2 dimensional input (coordinates)
- Create a network with an adequate amount of neurons
- Choose a random city and calculate the winning neuron (minimum euclidean distance)
- $f(\sigma, d) = e^{(-d^2/\sigma^2)}$  (Neighborhood function)
- $W_i^{new} = W_i^{old} + \alpha \cdot f(\sigma, d) \cdot (x_i - W_i^{old})$
- Decay the learning rate
- Finally, calculate the distance of the route we just found

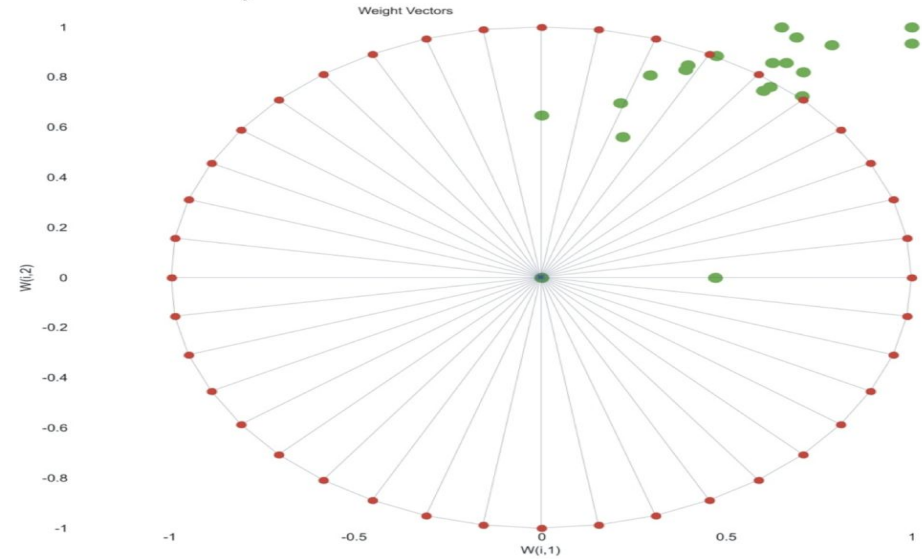


**SOLVING TRAVELLING SALESMAN PROBLEM BY USE OF KOHONEN SELF-ORGANIZING MAPS**

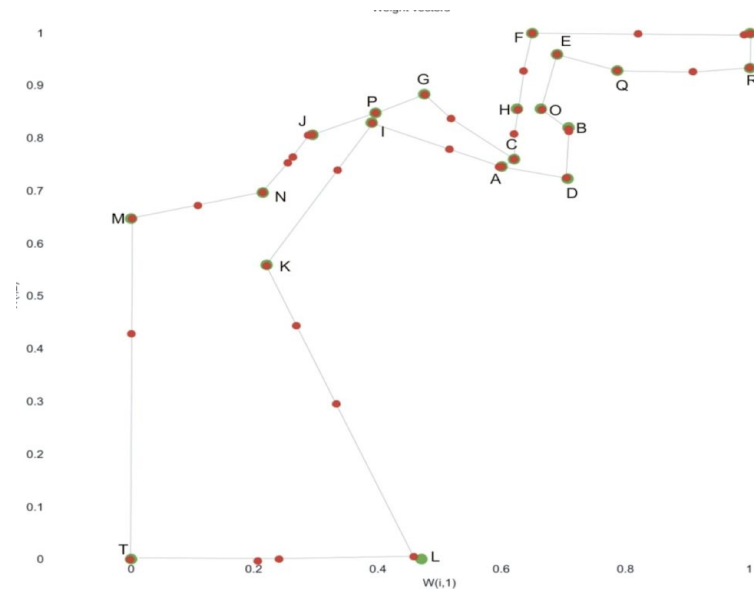
TABLE 1. POSITION OF WDS

Coordinate of WDS			Coordinate of WDS		
WDS	latitude	longitude	WDS	latitude	longitude
A	53.214	19.155	K	52.349	18.924
B	53.560	19.221	L	49.763	19.076
C	53.280	19.167	M	52.758	18.790
D	53.111	19.220	N	52.988	18.920
E	54.200	19.210	O	53.719	19.194
F	54.390	19.185	P	53.684	19.031
G	53.848	19.079	Q	54.058	19.269
H	53.721	19.170	R	54.076	19.399
I	53.603	19.027	S	54.390	19.399
J	53.494	18.969	T	49.763	18.790

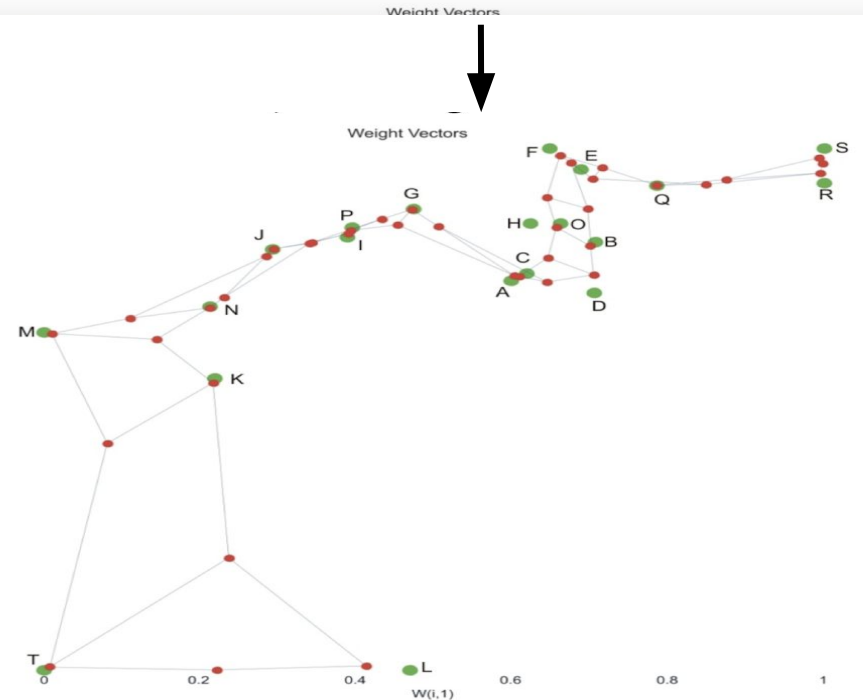
Routes are the same length only difference is that it does not start from the same node.



a) Initial stage



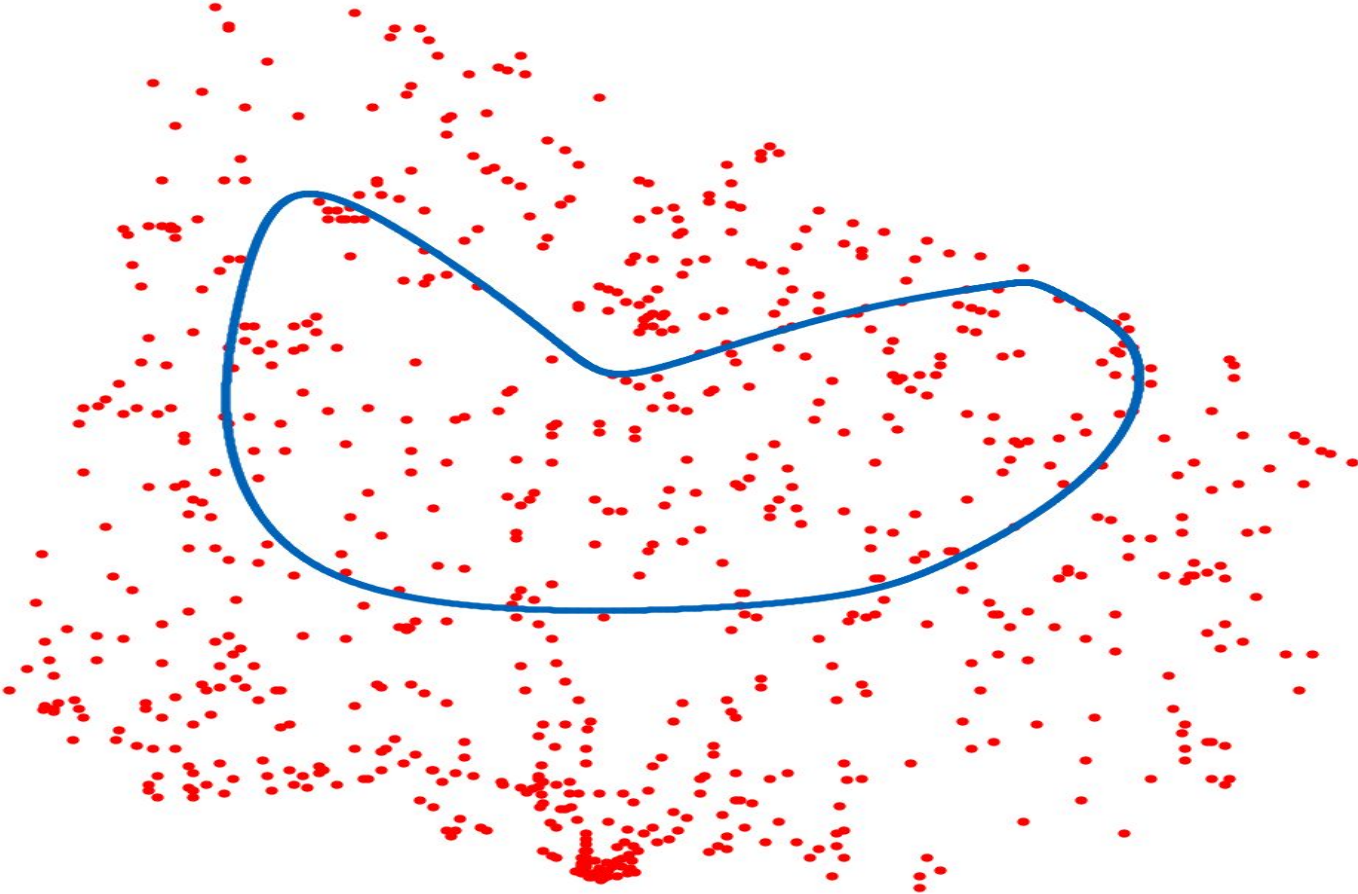
c) Final stage



b) Intermediate stage

Figure 4. Evolution of Kohonen's SOM for TSP

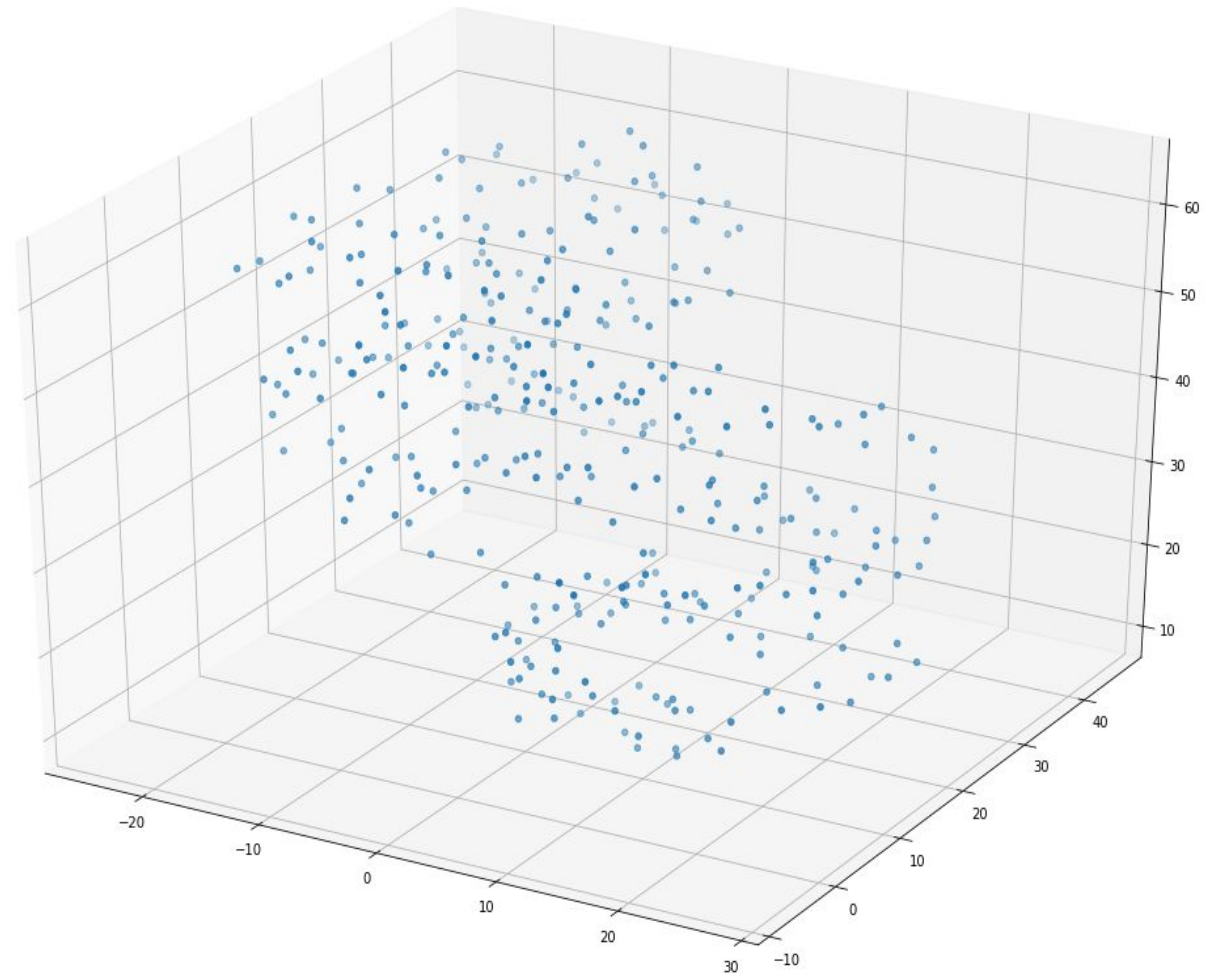
# Demo of SOM in 2D data



# How can we apply the existing algorithm to 3D protein structure?

**Input: Atom coordinates in protein structure**

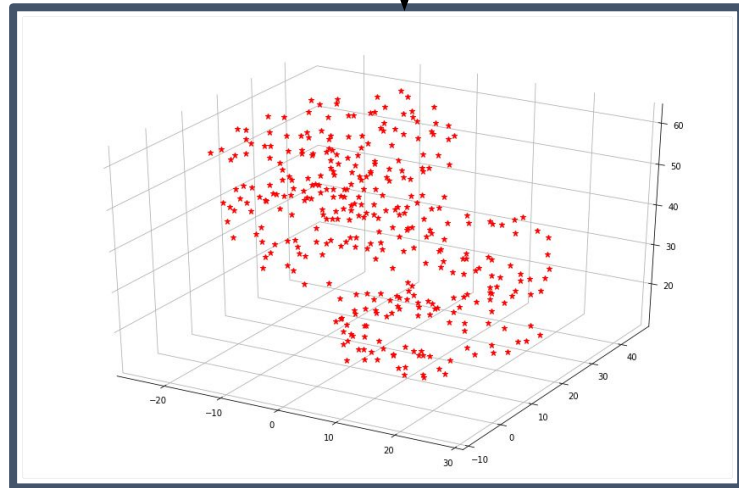
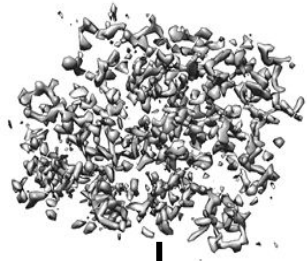
	x	y	z
0	27.552	4.354	23.629
1	24.179	4.807	21.907
2	21.218	2.742	20.697
3	20.409	2.806	16.978
4	17.867	5.477	16.127
..	...	...	...
346	16.970	3.518	33.655
347	14.622	1.905	36.176
348	14.865	-1.931	36.779
349	12.787	-5.145	36.901
350	13.090	-7.723	39.782



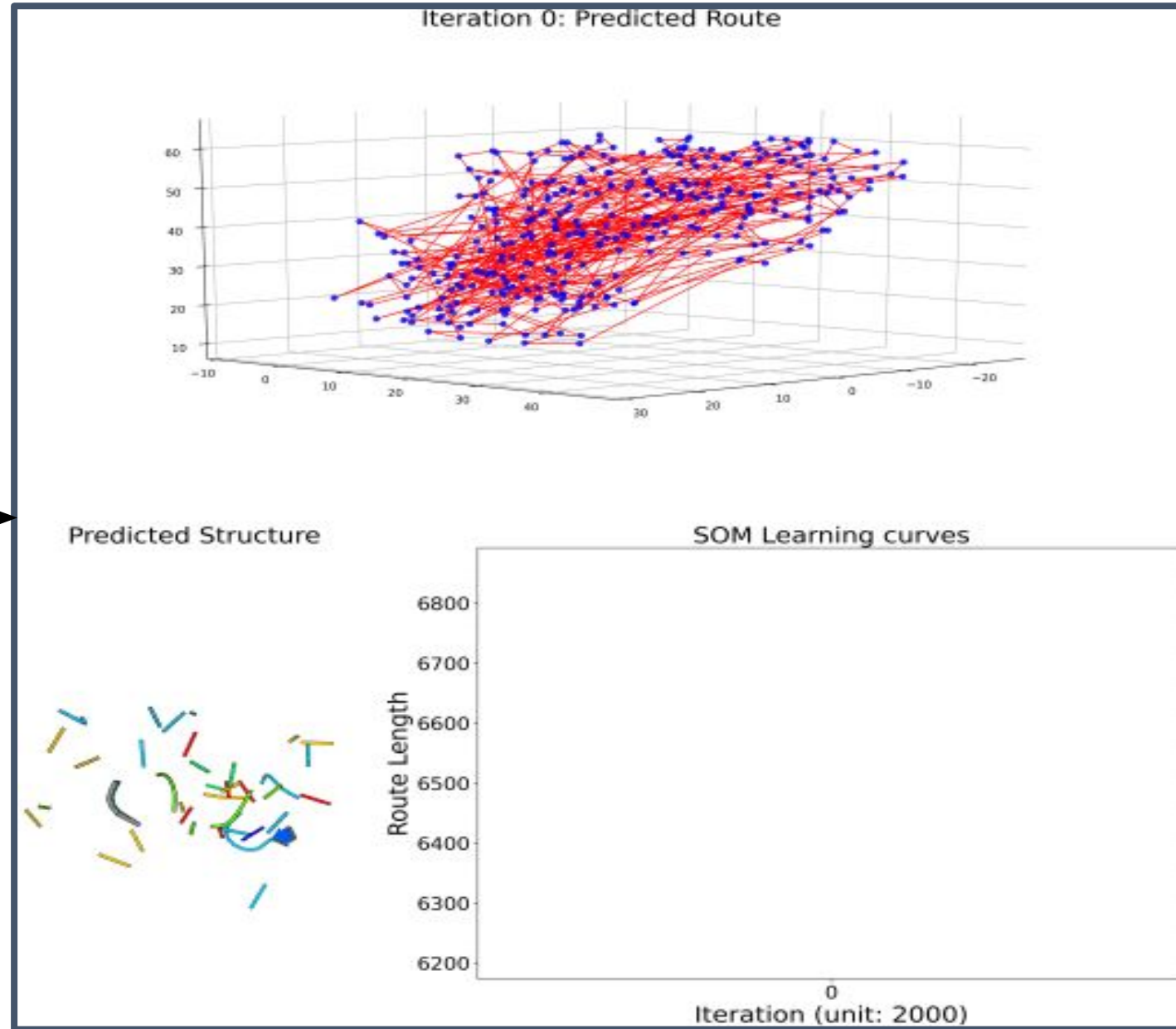
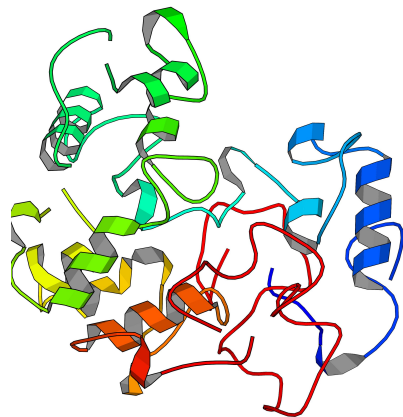


# Demo of SOM in 3D data

Cryo-EM Map

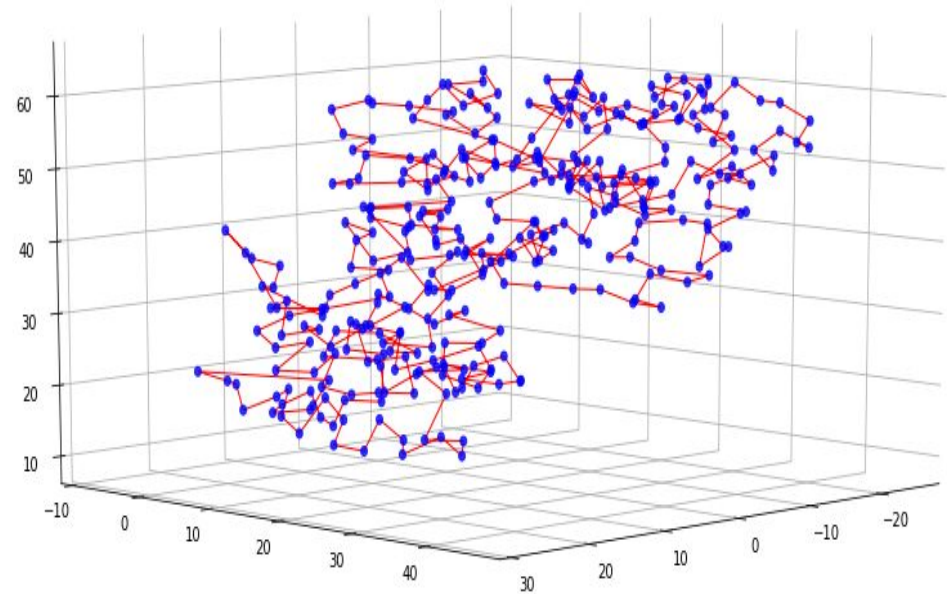
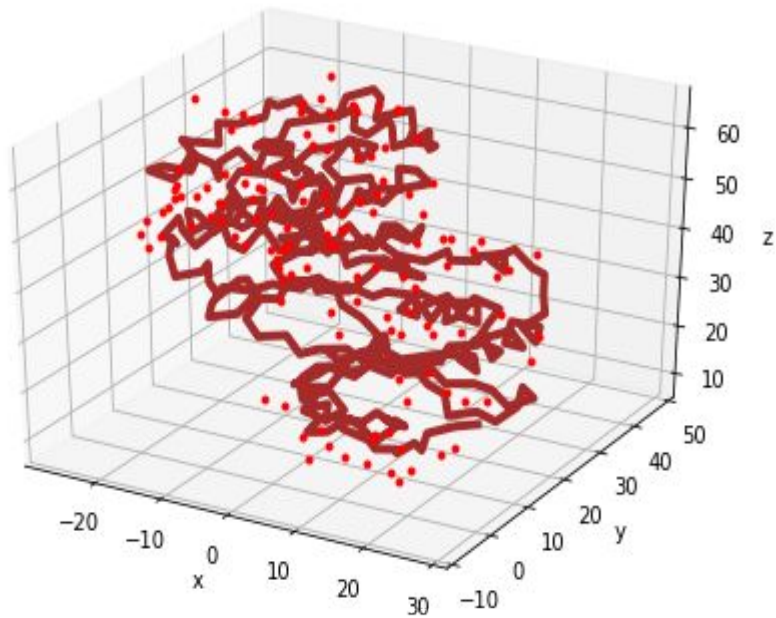


Predicted structure



# Evaluation from Deliverable 2

- Given 351 alpha carbon coordinates
- True backbone length in protein structure: 1332.0114119494347
- Our result: 1547.749479666934



# Subtour elimination in Deliverable 3

$$\text{MIN } C_{i,j} x_{i,j} \quad (9)$$

s.t

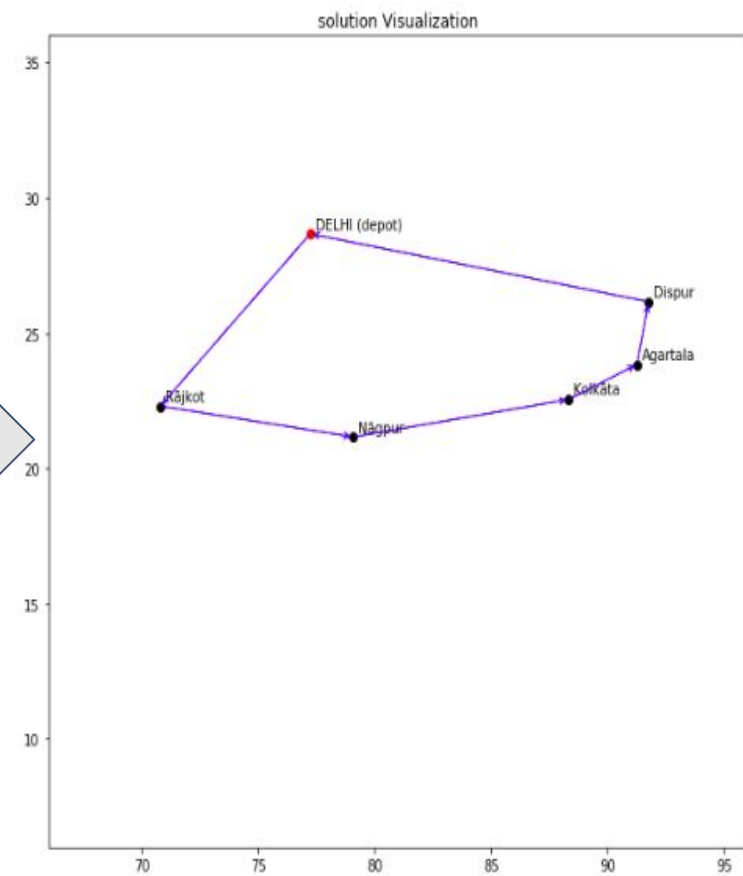
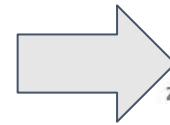
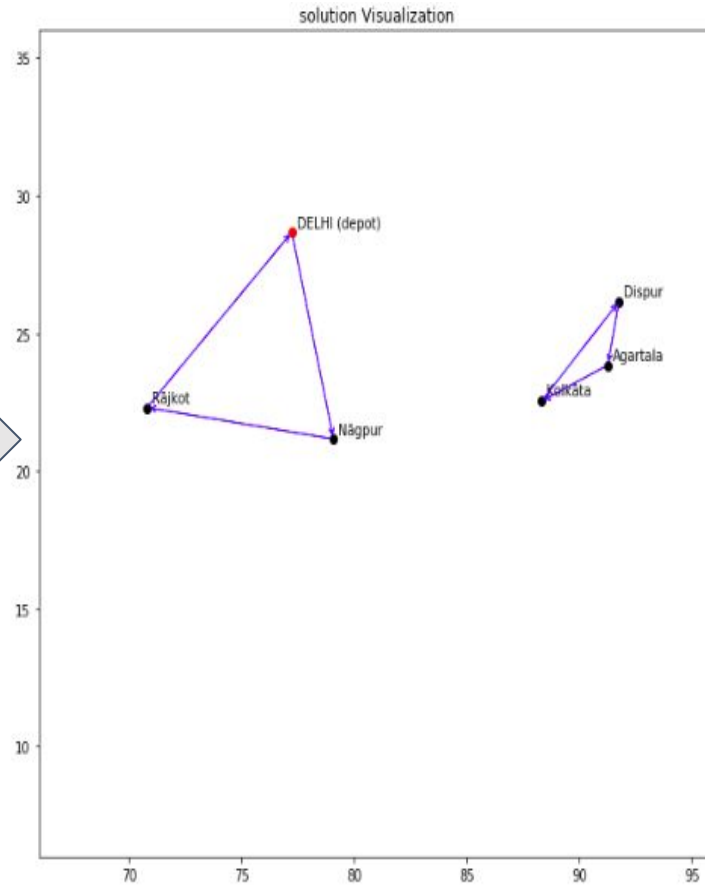
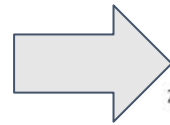
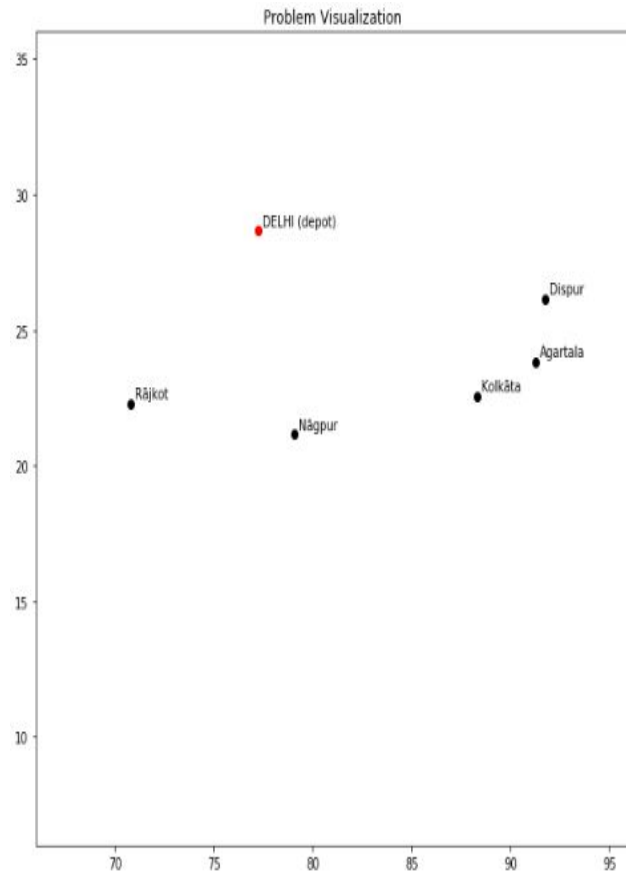
$$\sum_{j=1}^N x_{i,j} = 1 \quad i = 1 \dots N \quad (10)$$

$$\sum_{i=1}^N x_{i,j} = 1 \quad j = 1 \dots N \quad (11)$$

$$x_{i,i} = 0 \quad i = 1 \dots N \quad (12)$$

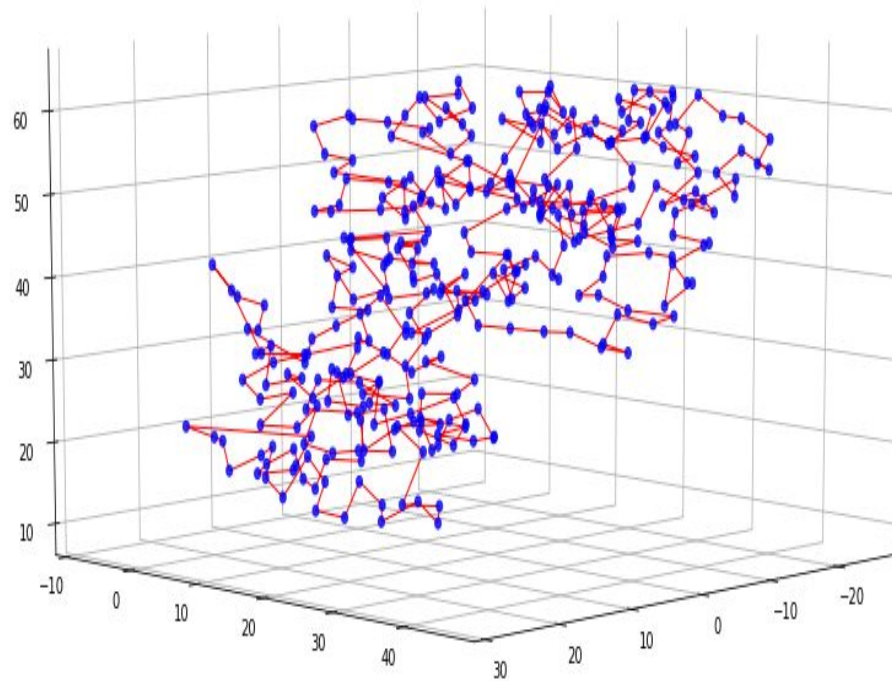
$$\sum_{i \in S} \sum_{j \notin S} x_{i,j} \geq 1 \quad S \subseteq \{1, 2, \dots, n\}, \quad 1 \leq |S| \leq N - 1 \quad (15)$$

# Subtour elimination

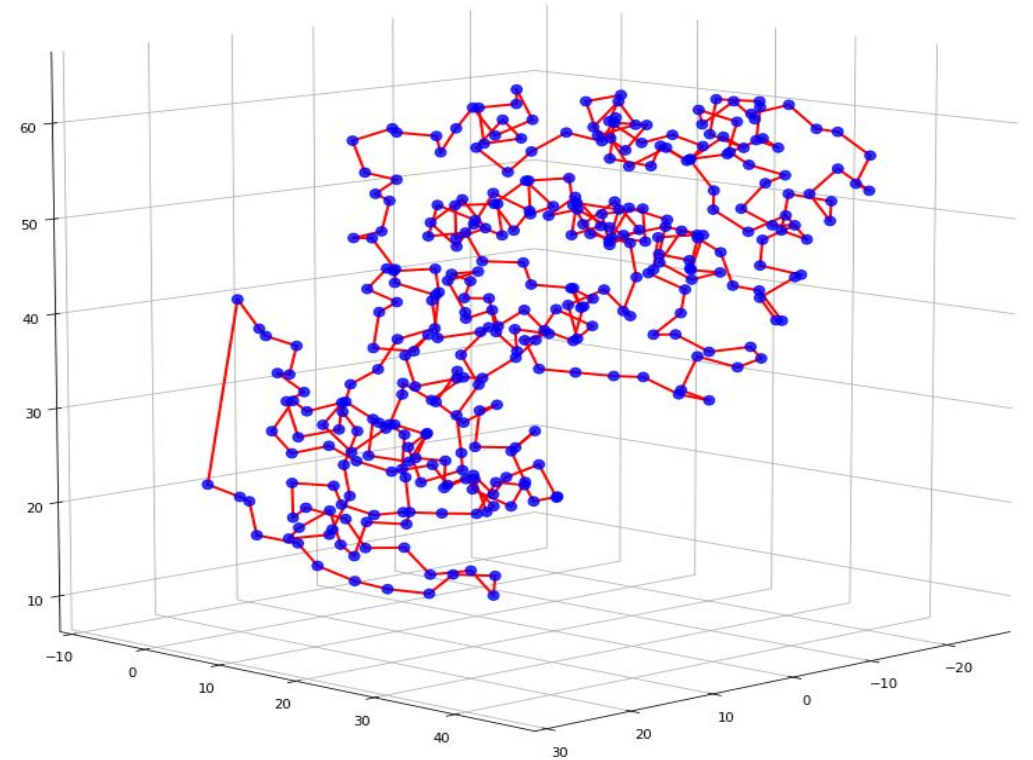


# SOM vs Subtour elimination

SOM's plot



Subtour elimination's plot



# SOM vs Subtour elimination

Protein true length: 1332.0114119494347

Technique	Result	Accuracy	Elapsed time
SOM	1547.749479666934	~ 86%	~ 6 minutes
Subtour elimination	1356.8292035917223	~ 98%	~ 1 hour and a half

# SOM vs Subtour elimination (predicted lengths)

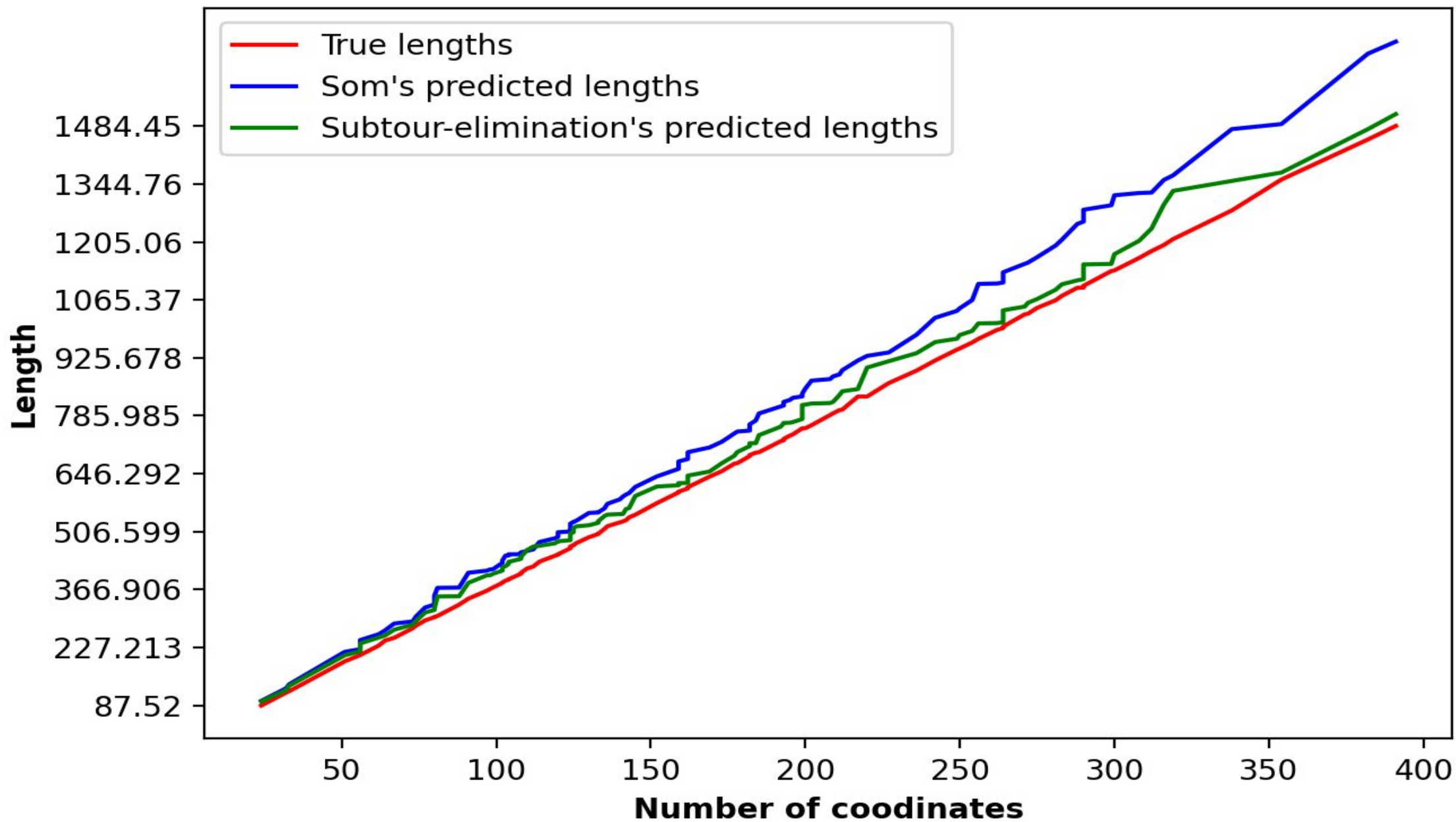
	A	B	C
80	T0849_filtered.pdb	894.866357	981.4506269
81	T0893_filtered.pdb	919.2197821	1022.169751
82	T0738_filtered.pdb	945.1522523	1038.687702
83	T0608_filtered.pdb	948.3916742	1045.226231
84	T0686_filtered.pdb	962.6224003	1103.833194
85	T0486_filtered.pdb	971.4049172	1064.840619
86	T0426_filtered.pdb	992.8523277	1131.810979
87	T0445_filtered.pdb	997.3373902	1107.12205
88	T0478_filtered.pdb	1000.980809	1104.757842
89	T0511_filtered.pdb	1030.511194	1152.240949
90	T0703_filtered.pdb	1031.822802	1155.025808
91	T0505_filtered.pdb	1046.050336	1168.249952
92	T0405_filtered.pdb	1064.444077	1196.439233
93	T0626_filtered.pdb	1074.983735	1210.286789
94	T0421_filtered.pdb	1094.495566	1253.876714
95	T0398_filtered.pdb	1094.719129	1248.385041
96	T0526_filtered.pdb	1099.175785	1282.557461
97	T0721_filtered.pdb	1134.774168	1293.401962
98	T0449_filtered.pdb	1136.36588	1322.882666
99	T0905_filtered.pdb	1166.35049	1317.483954
100	T0861_filtered.pdb	1182.892397	1323.981476
101	T0457_filtered.pdb	1197.592113	1354.450558
102	T0693_filtered.pdb	1211.843761	1365.481524
103	T0609_filtered.pdb	1280.917543	1476.915407
104	T0534_filtered.pdb	1355.243523	1488.999652
105	T0781_filtered.pdb	1451.881008	1658.702151
106	T0917_filtered.pdb	1484.448605	1687.787466
107	<b>Avg length</b>	<b>659.1024615</b>	<b>738.3291995</b>
108			

+ ☰ som-records ▾ subtour-elimiation ▾ som2 ▾

	A	B	C
80	T0849_filtered.pdb	894.866357	917.5067816
81	T0893_filtered.pdb	919.2197821	936.8937661
82	T0738_filtered.pdb	945.1522523	963.6283949
83	T0608_filtered.pdb	948.3916742	971.6163277
84	T0686_filtered.pdb	962.6224003	980.7149631
85	T0486_filtered.pdb	971.4049172	990.4596607
86	T0426_filtered.pdb	992.8523277	1008.582415
87	T0445_filtered.pdb	997.3373902	1009.317532
88	T0478_filtered.pdb	1000.980809	1011.49998
89	T0511_filtered.pdb	1030.511194	1049.212176
90	T0703_filtered.pdb	1031.822802	1040.096196
91	T0505_filtered.pdb	1046.050336	1058.139273
92	T0405_filtered.pdb	1064.444077	1328.032752
93	T0626_filtered.pdb	1074.983735	1089.769749
94	T0421_filtered.pdb	1094.495566	1112.33508
95	T0398_filtered.pdb	1094.719129	1102.871106
96	T0526_filtered.pdb	1099.175785	1115.13053
97	T0721_filtered.pdb	1134.774168	1150.709334
98	T0449_filtered.pdb	1136.36588	1152.055116
99	T0905_filtered.pdb	1166.35049	1175.437903
100	T0861_filtered.pdb	1182.892397	1207.763343
101	T0457_filtered.pdb	1197.592113	1237.209066
102	T0693_filtered.pdb	1211.843761	1351.966435
103	T0609_filtered.pdb	1280.917543	1295.144644
104	T0534_filtered.pdb	1355.243523	1372.154247
105	T0781_filtered.pdb	1451.881008	1476.822202
106	T0917_filtered.pdb	1484.448605	1513.010468
107	<b>Avg length</b>	<b>659.1024615</b>	<b>690.6663401</b>
108			

+ ☰ som-records ▾ subtour-elimiation ▾ som2 ▾

## Subtour elimination results vs Som's results



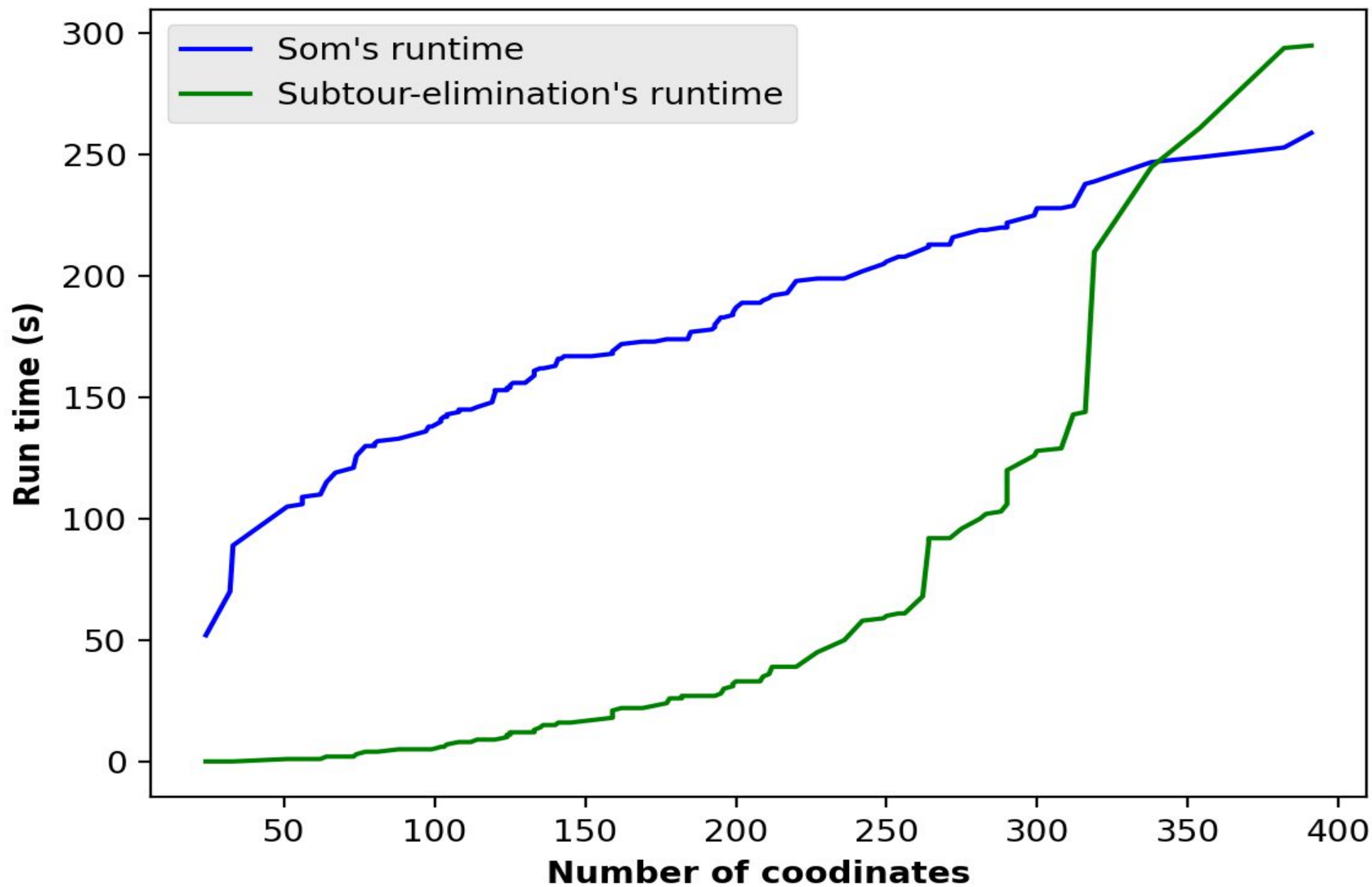


# SOM vs Subtour elimination (run time)

	A	B
89	T0398_filtered.pdb	213
90	T0861_filtered.pdb	216
91	T0526_filtered.pdb	217
92	T0738_filtered.pdb	219
93	T0421_filtered.pdb	219
94	T0445_filtered.pdb	220
95	T0534_filtered.pdb	220
96	T0703_filtered.pdb	222
97	T0405_filtered.pdb	225
98	T0457_filtered.pdb	228
99	T0426_filtered.pdb	228
100	T0905_filtered.pdb	229
101	T0626_filtered.pdb	238
102	T0721_filtered.pdb	239
103	T0693_filtered.pdb	247
104	T0609_filtered.pdb	249
105	T0781_filtered.pdb	253
106	T0917_filtered.pdb	259
107	<b>Avg runtime</b>	<b>171.53333333</b>

	A	B
85	T0708_filtered.pdb	61
86	T0486_filtered.pdb	68
87	T0608_filtered.pdb	90
88	T0635_filtered.pdb	92
89	T0457_filtered.pdb	92
90	T0756_filtered.pdb	93
91	T0526_filtered.pdb	96
92	T0679_filtered.pdb	100
93	T0892_filtered.pdb	102
94	T0880_filtered.pdb	103
95	T0609_filtered.pdb	106
96	T0511_filtered.pdb	120
97	T0421_filtered.pdb	126
98	T0861_filtered.pdb	128
99	T0449_filtered.pdb	129
100	T0516_filtered.pdb	143
101	T0738_filtered.pdb	144
102	T0405_filtered.pdb	210
103	T0781_filtered.pdb	245
104	T0534_filtered.pdb	261
105	T0693_filtered.pdb	294
106	T0917_filtered.pdb	295
107	<b>Avg runtime</b>	<b>43.6952381</b>

## Subtour-elimination's runtime vs Som's runtime



# Questions?



# Subtour elimination

**Coordinates**

	x	y	z
0	27.552	4.354	23.629
1	24.179	4.807	21.907
2	21.218	2.742	20.697
3	20.409	2.806	16.978
4	17.867	5.477	16.127
..	...	...	...
346	16.970	3.518	33.655
347	14.622	1.905	36.176
348	14.865	-1.931	36.779
349	12.787	-5.145	36.901
350	13.090	-7.723	39.782



**Distance matrix**

	0	1	2	...	348	349	350
0	0.000000	3.814135	7.163430	...	19.323139	22.008685	24.817792
1	3.814135	0.000000	3.807341	...	18.797011	21.298826	24.484331
2	7.163430	3.807341	0.000000	...	17.911679	19.896134	23.233980
3	9.882032	6.520118	3.806513	...	21.101059	22.764868	26.161996
4	12.302047	8.584791	6.292402	...	22.144877	23.878701	27.506704
..	...	...	...	...	...	...	...
346	14.601311	13.843652	13.658603	...	6.624349	10.152910	13.377392
347	18.182676	17.417296	16.846579	...	3.890701	7.320885	10.394645
348	19.323139	18.797011	17.911679	...	0.000000	3.829199	6.761353
349	22.008685	21.298826	19.896134	...	3.829199	0.000000	3.877893
350	24.817792	24.484331	23.233980	...	6.761353	3.877893	0.000000

# Evaluation

- Given 351 alpha carbon coordinates
- True backbone length in protein structure: 1332.0114119494347
- Our result: 1356.8292035917223

## Next steps-Deliverable 4

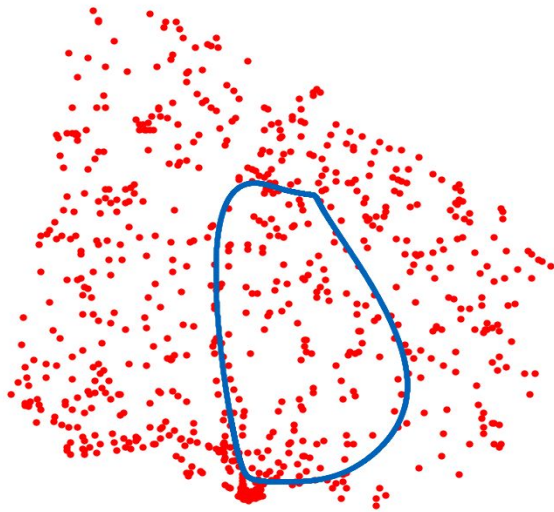
- Incorporate SOM technique's speed and subtour elimination technique's accuracy
- Improve deep learning training using larger datasets
- Evaluate all algorithms on real coronavirus protein datasets
- Create an independent package for Deep Tracer to use
- Implement web-based front end for users to input data



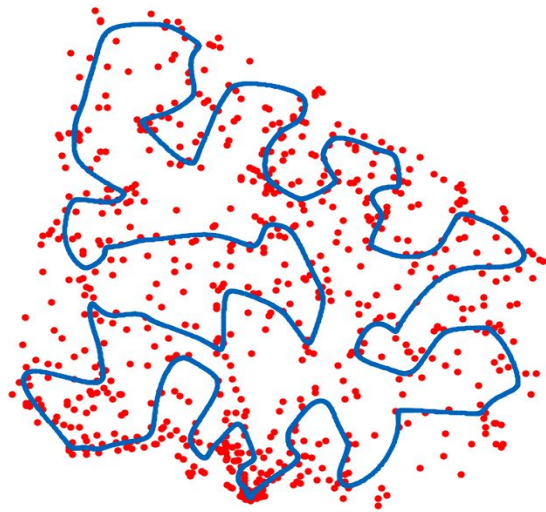
# Existing algorithm

- Self Organizing Map as heuristic
- Uruguay, containing 734 cities with an optimal tour of **79114**.
- 17351 iterations, 23.4s, length: **85072.35**, only **7.5% longer than the optimal tour**

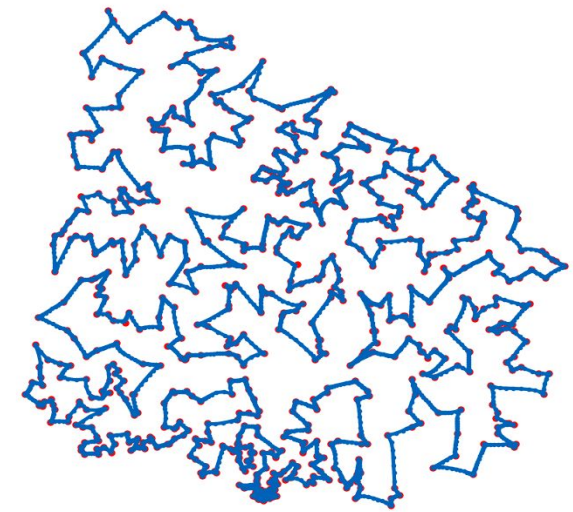
Iterations = 100



Iterations = 6000



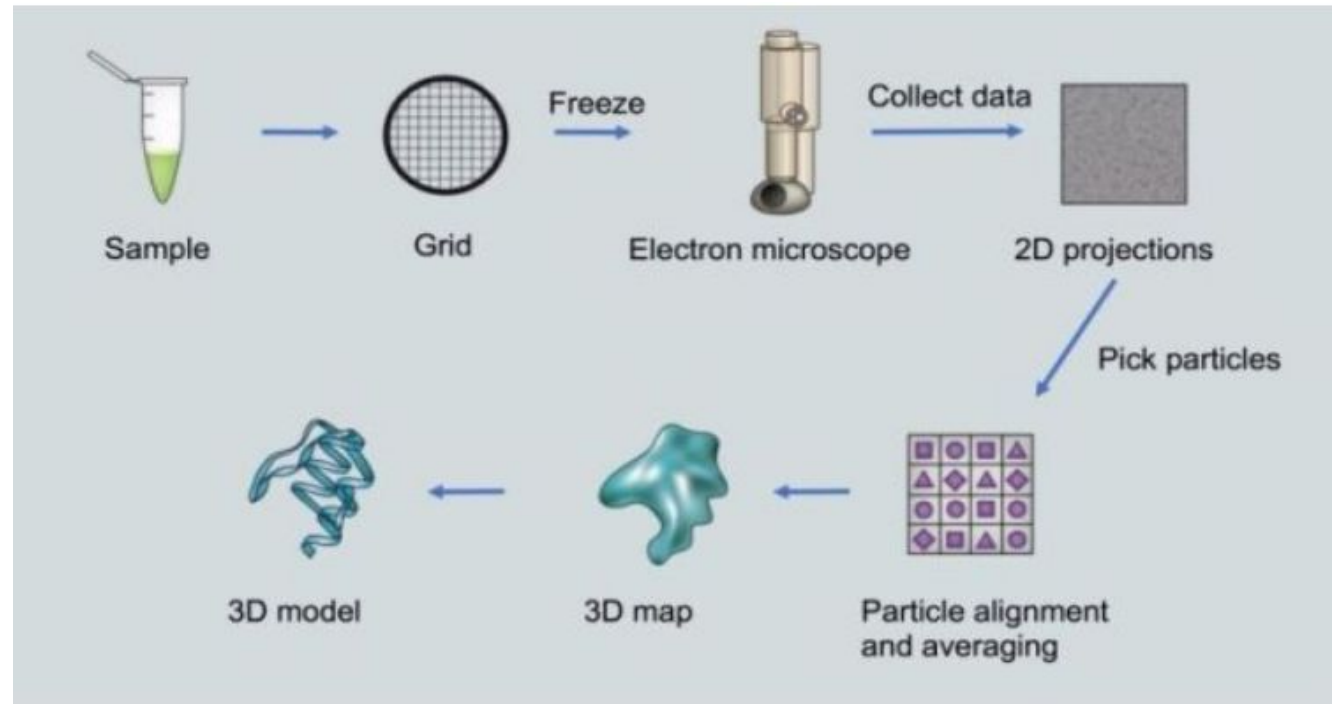
Iterations = 17000





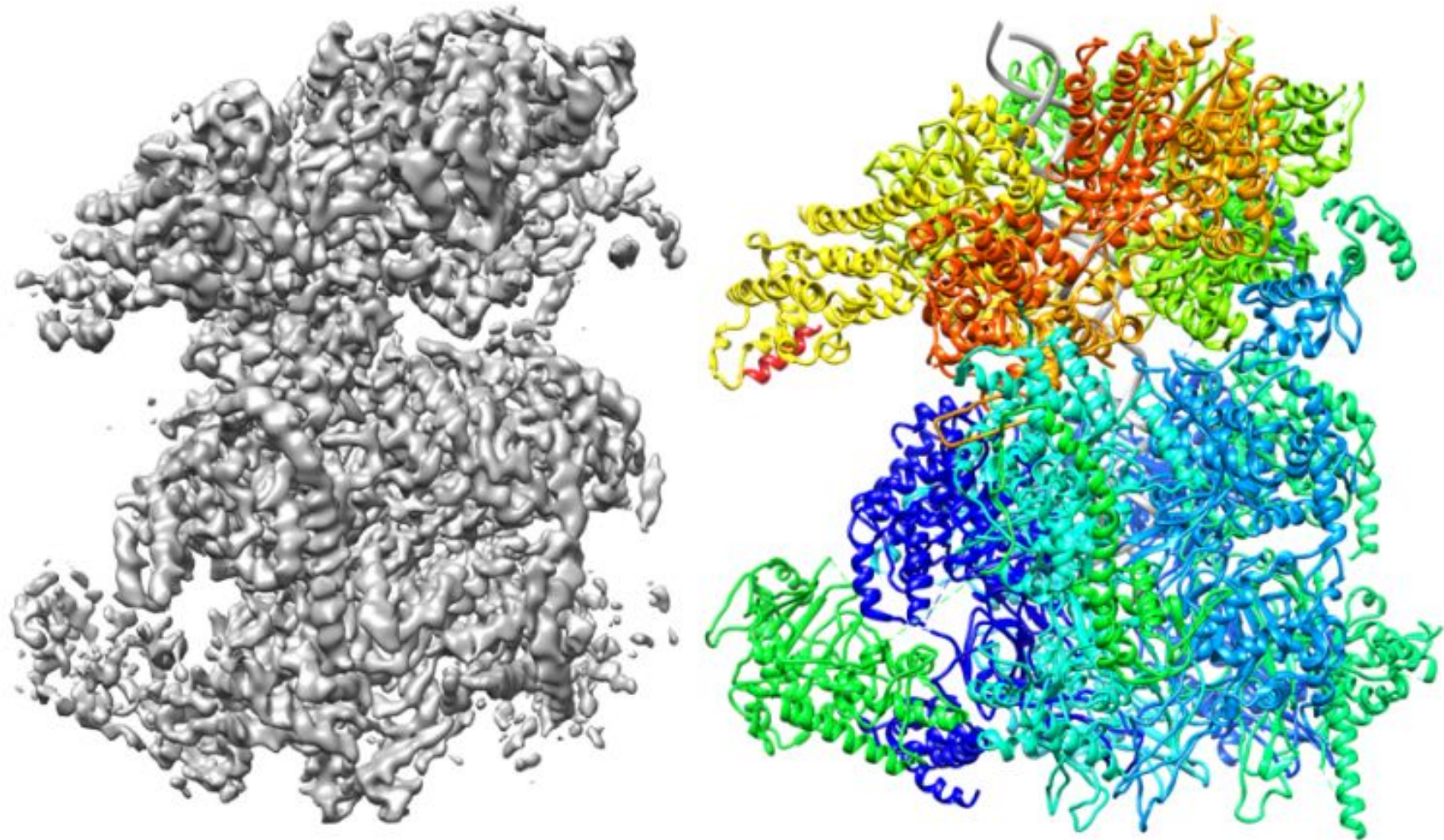
# Cryo-EM

- Cryogenic Electron Microscopy
- Flash freezes proteins for highly accuracy structures
- Data is stored as a density map of the entire protein



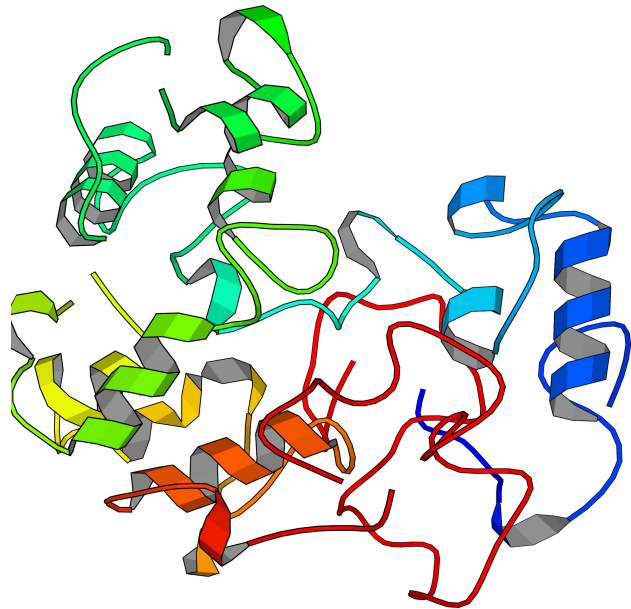
# Cryo-EM

---



# SOM vs Subtour elimination

SOM's predicted structure

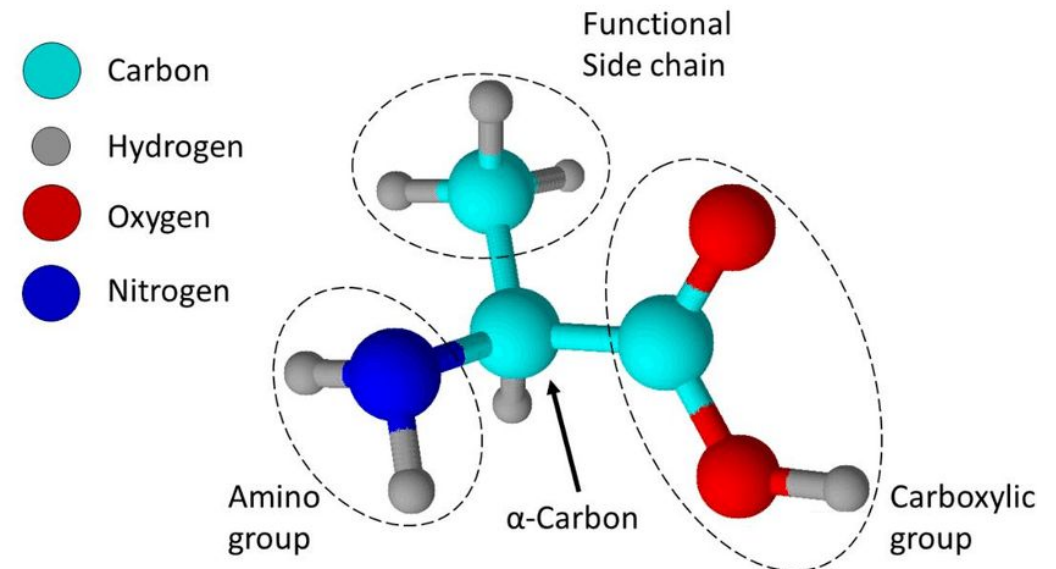
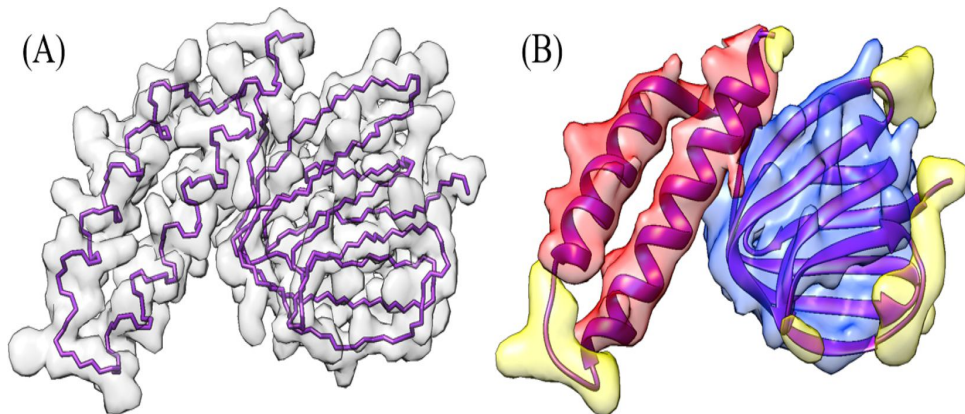


Subtour elimination's predicted structure

# Graph Neural Networks for protein structure prediction

- GNN's abstract tasks into vertices (amino acids) and edges ( $C\alpha$  connections)
- Widely-applied, especially useful for capturing the structural properties of a graph
- Our goal is to create a connected graph of carbon- $\alpha$  in protein backbone structure

The Alpha Carbon is the key identifier for *relational* structures of amino acids



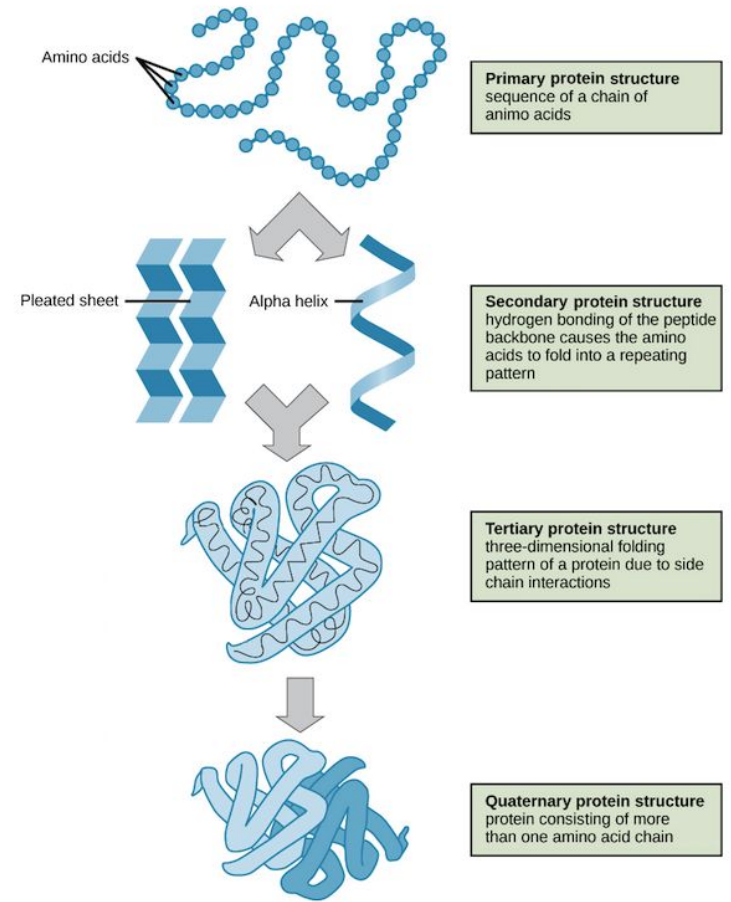
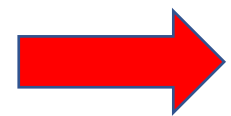
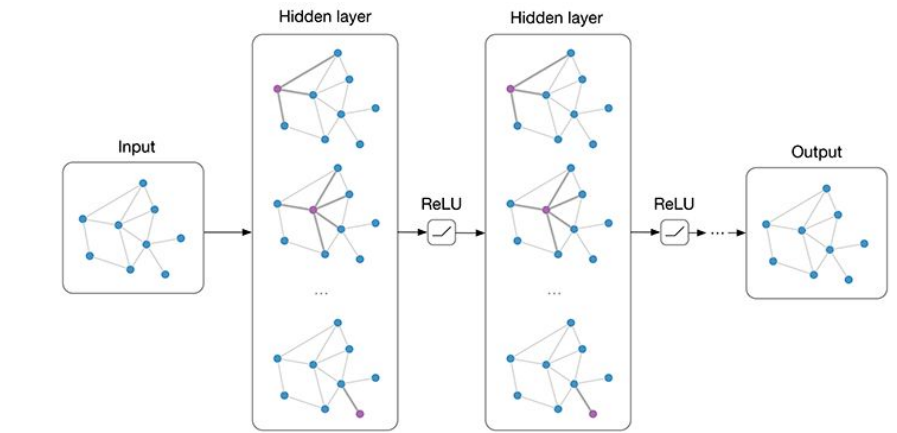
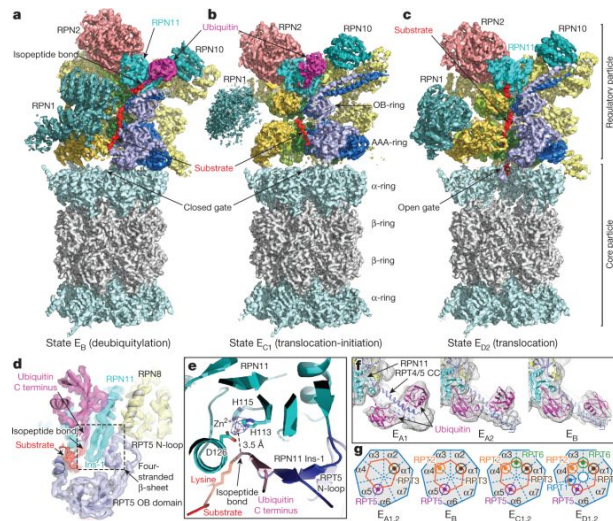
# Our Objective

---

- We hope to build off of this technology in our solution
- Goal: Improve Accuracy of their model and build tertiary protein structures.
  - ★ Efficient backbone tracing (Connecting the backbone  $C\alpha$  atoms using GNN)
  - ★ Amino acids assignment (Mapping the amino acids from the protein sequences onto the backbone traces)
  - ★ Structure refinement (Reconstruct the missing regions, improve the model's stereochemical quality)

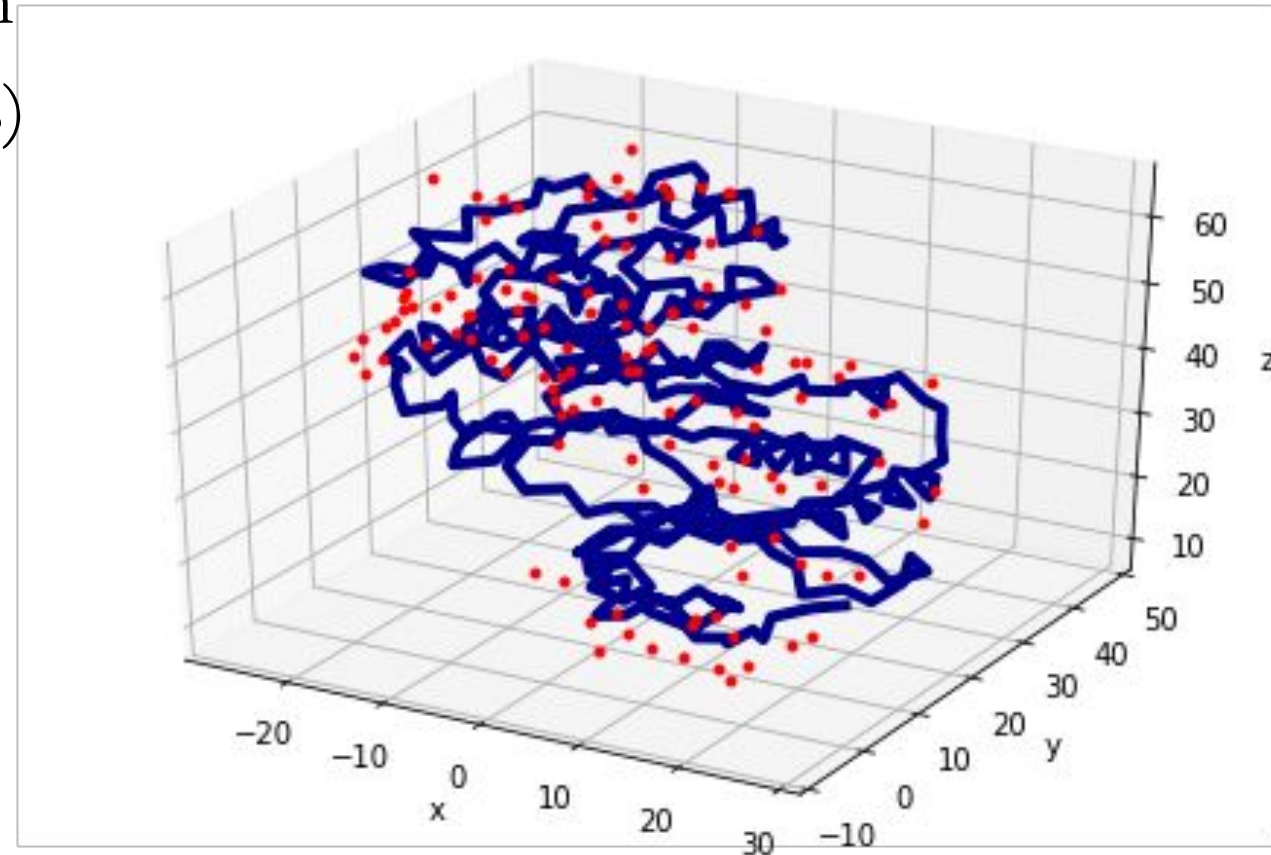
# Solution

- Integrated graph neural network to predict Sars-Cov-2 protein structures from Cryo-EM Data



# 3D Visualization of Structures

- Our algorithm generates routes given a new set of points:
- Our algorithm is only 12% off the true backbone structure when given ample training ( $>20$  epochs)



# Next Steps

---

- Accumulate training and testing data for project development
- Design and implement backbone graph neural network to achieve our goals
- Solicit feedback to finetune and improve the method
- Create a web interface to interact with the method as a client



# Accountability

---

- Weekly Team Meetings
- Weekly Reports
- Agile management and changing of stories and requirements (if necessary)
  - We recognize this is a research-based project, exempt from some standard Agile practices
- Knowledge shares and literature reviews to further our understanding