Improved protein structure prediction by deep learning irrespective of co-evolution information

Jinbo Xu et al.
Luis Lazcano

Table of contents

01 Introduction

02 Methods

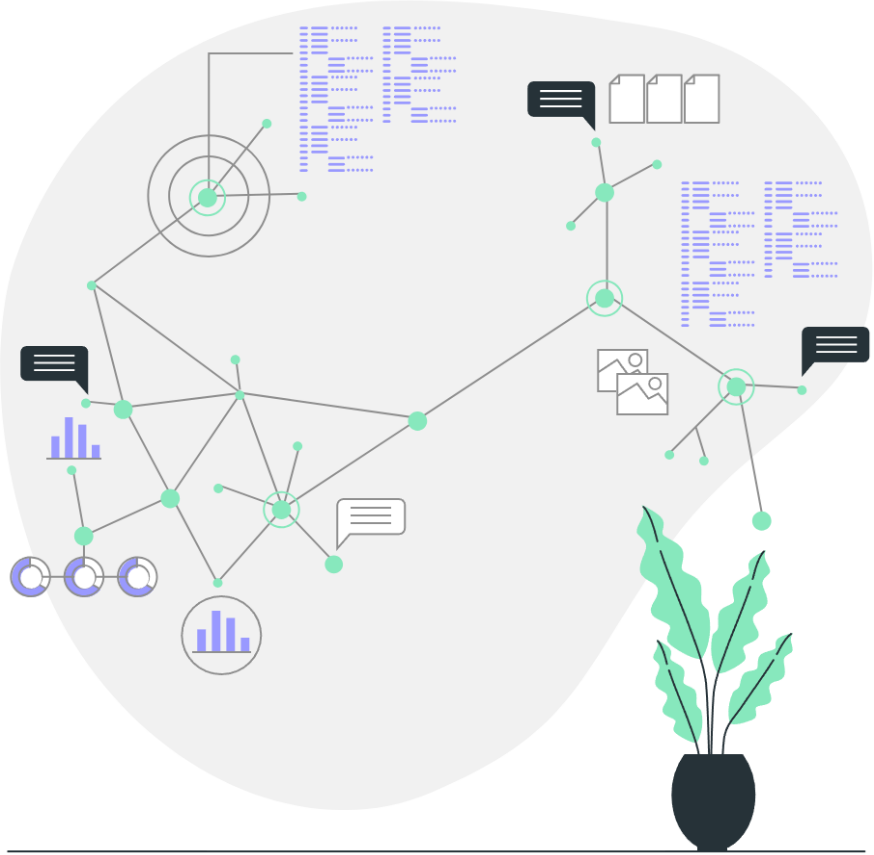
03 Results

04 Conclusion

01

Introduction

The basics.



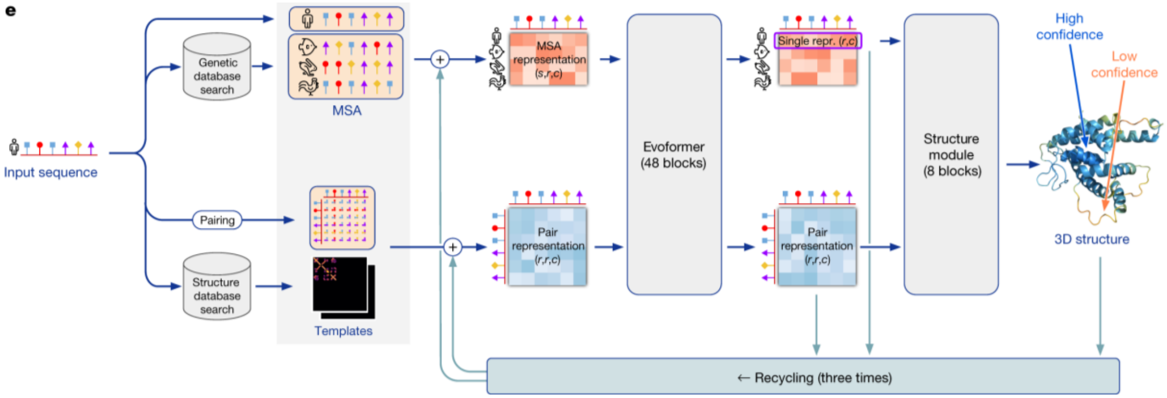
Introduction

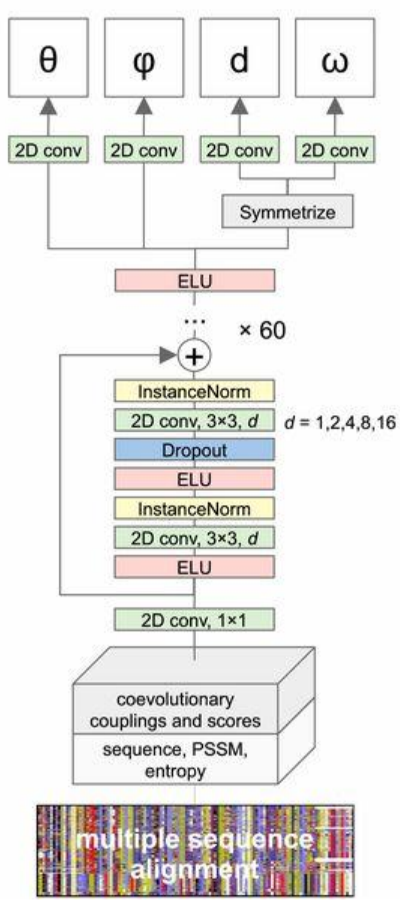
- The structure of a protein is important in determining its function.
- Many structure prediction methods use co-evolution information.
- Human designed proteins there is no evolutionary history.
- Natural proteins have evolution to guide their folding.
- Prediction the folded state of a protein without the need for coevolution data should be possible in principle.

Current landscape

- Current best performing systems:
  - Alpha Fold 2
  - RoseTTAFold
  - trRosetta
- Written before
  - OmegaFold
  - ESMfold

AlphaFold 2 and trRosetta


e
Input sequence
Genetic database search
MSA
Pairing
Structure database search
Templates
MSA representation (s,r,c)
Pair representation (r,r,c)
Evoformer (48 blocks)
Single repr. (r,c)
Pair representation (r,r,c)
Structure module (8 blocks)
High confidence
Low confidence
3D structure
← Recycling (three times)



MSA information is used as part of the input for both AlphaFold 2 and trRosetta

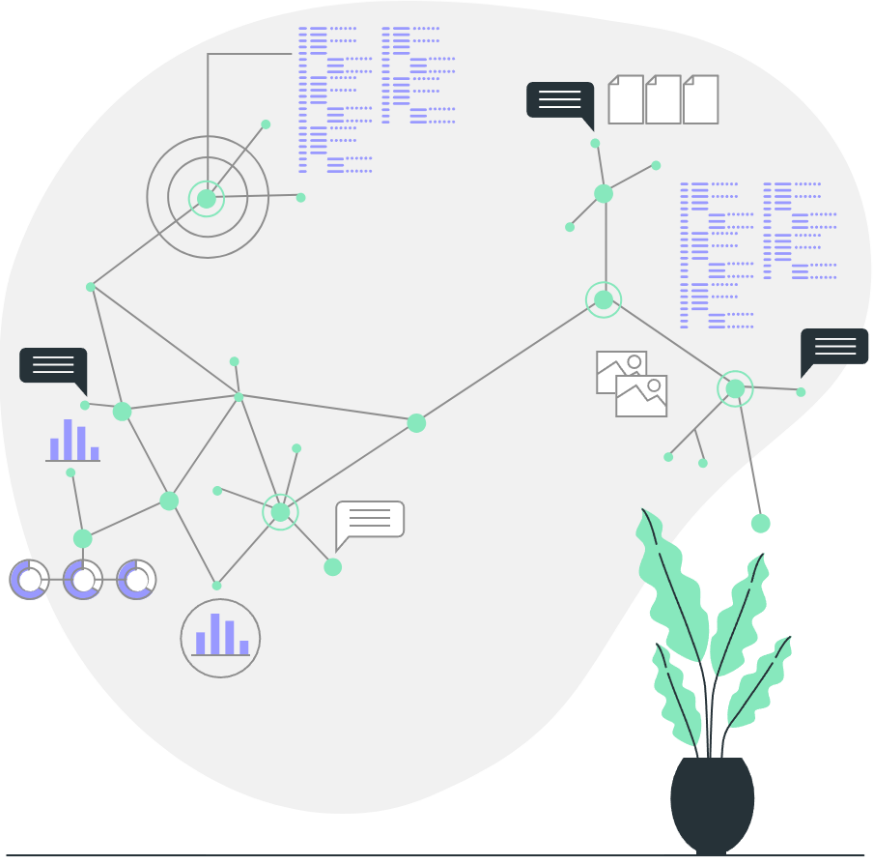
What do we want to do?

- The paper focuses on the impact of components in a ResNet system.
- The impact of co-evolutionary data on a model's performance is important.

02

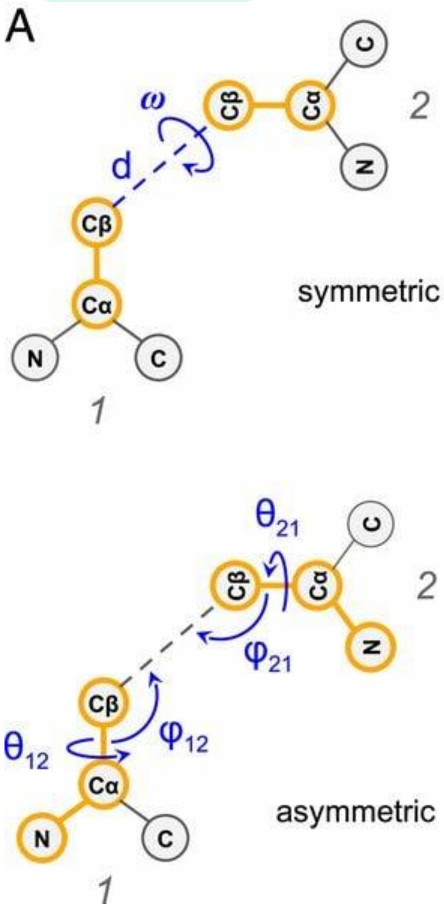
Methods

How are we doing
what were doing?



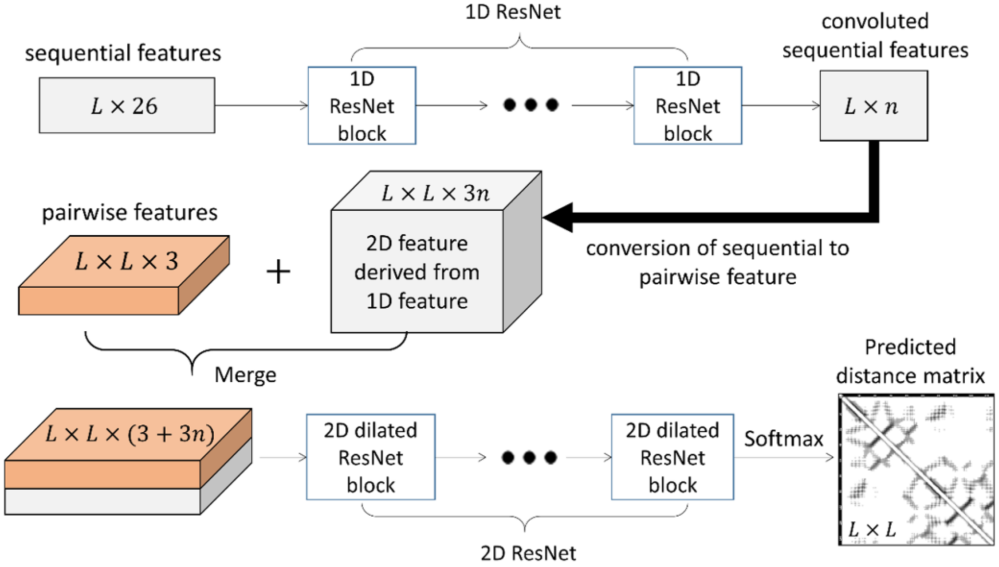
System’s Design

- Inter-residue orientations defined in trRosetta are used
- PyRosetta’s fast relaxation protocol is used for the generation of 3D structures



Network Architecture

- Input features are able to be turned on and off easily.
- 100 2D convolutional layers and, on average, 150 filters per layer.



Model training

- The deep ResNet was trained with the following data:
    - PDB25 was used in CASP13
    - CATH S35 is used for their training and validation process
        - March 2018 and 1 January 2020
        - Not much difference was found in the different versions after training

Coevolutionary data

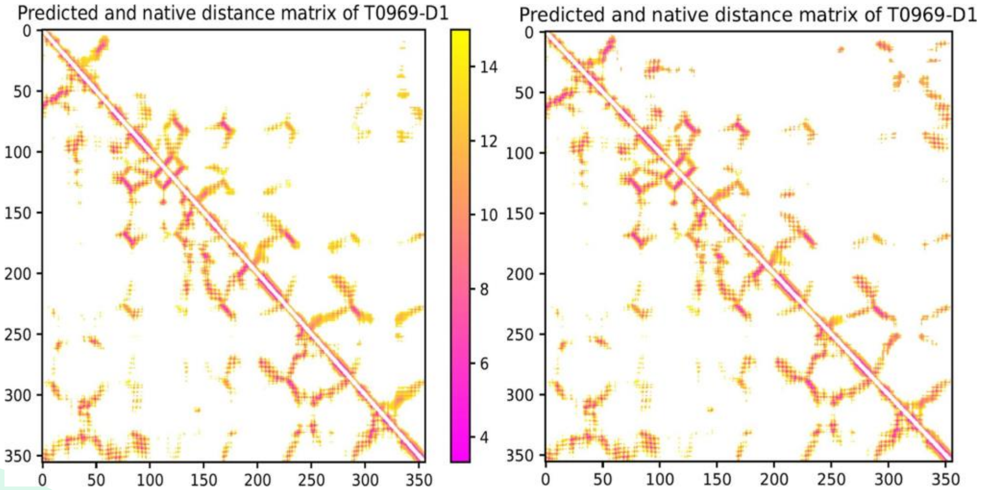
- CCMpred
  - A performance-optimized MSA contact prediction algorithm
- Metagenomic data
  - Metagenomic data was taken from the MetaClust dataset

Ablation study of contact prediction

- Various models were trained using the CATH S35 data
  - The model sizes and input features varied between models
  - Co-evolution
  - CCMpred
  - Metagenomic data
- The contributions of different factors were determined by comparing the resulting models' performance

I/O

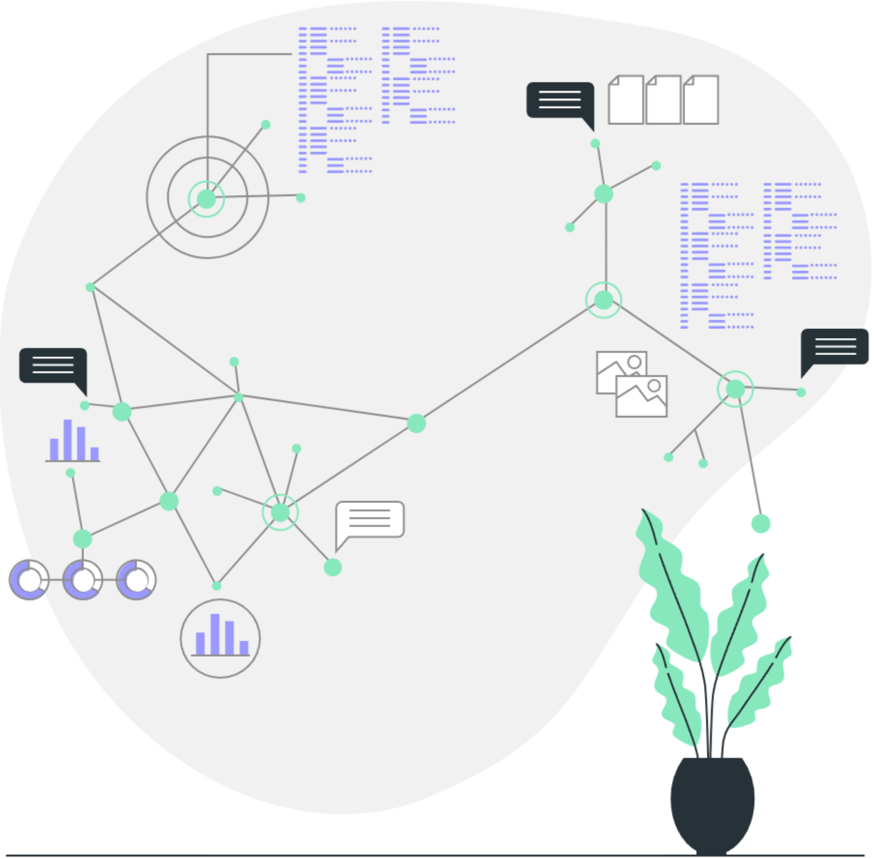
- Input varies between models
  - MSA data
- Output: a 2D distance map predicted by the model the 3D representation is done by the use of pyRosetta.



03

Results

How'd it go?



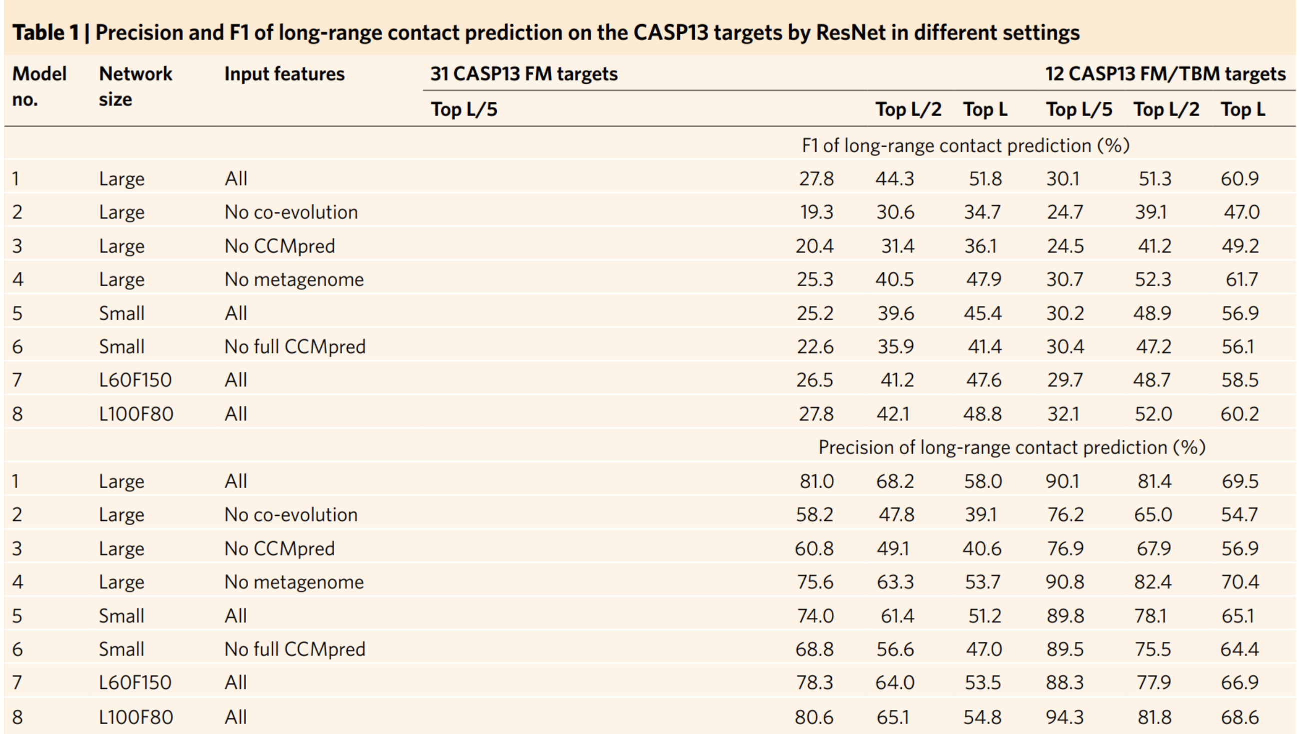
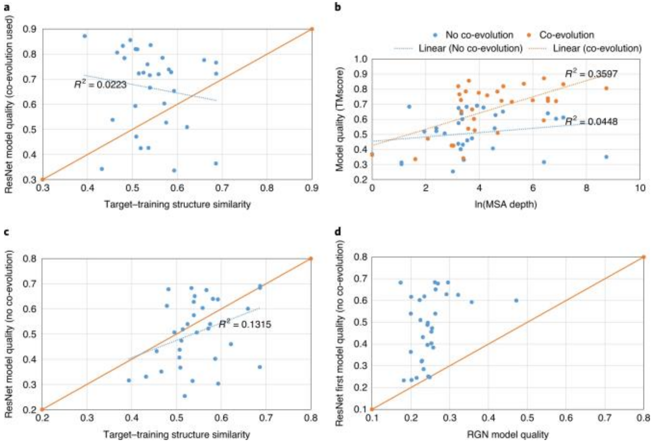
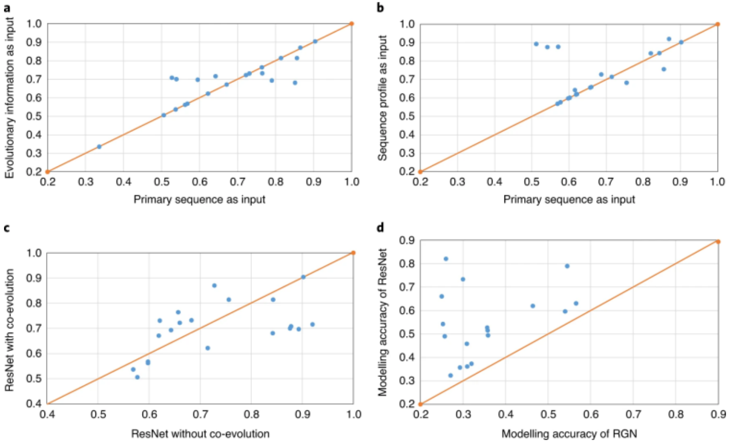


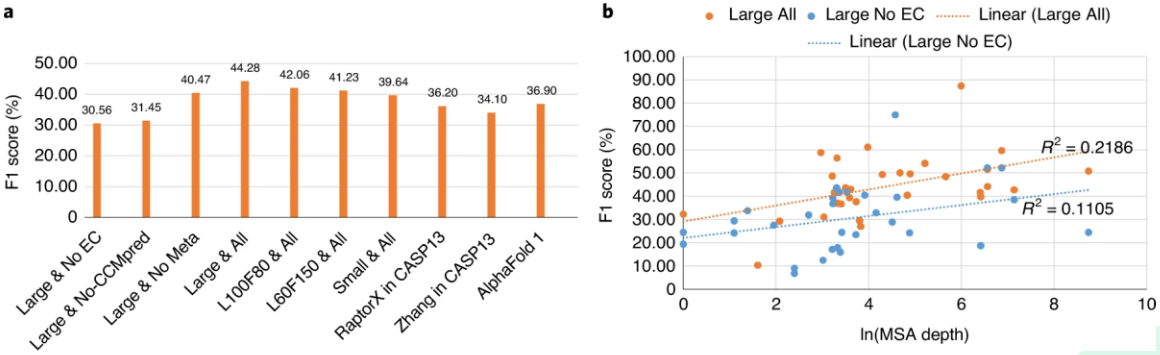
Table 1 | Precision and F1 of long-range contact prediction on the CASP13 targets by ResNet in different settings

Model no.	Network size	Input features	31 CASP13 FM targets			12 CASP13 FM/TBM targets		
			Top L/5	Top L/2	Top L	Top L/5	Top L/2	Top L
			F1 of long-range contact prediction (%)					
1	Large	All	27.8	44.3	51.8	30.1	51.3	60.9
2	Large	No co-evolution	19.3	30.6	34.7	24.7	39.1	47.0
3	Large	No CCMpred	20.4	31.4	36.1	24.5	41.2	49.2
4	Large	No metagenome	25.3	40.5	47.9	30.7	52.3	61.7
5	Small	All	25.2	39.6	45.4	30.2	48.9	56.9
6	Small	No full CCMpred	22.6	35.9	41.4	30.4	47.2	56.1
7	L60F150	All	26.5	41.2	47.6	29.7	48.7	58.5
8	L100F80	All	27.8	42.1	48.8	32.1	52.0	60.2
			Precision of long-range contact prediction (%)					
1	Large	All	81.0	68.2	58.0	90.1	81.4	69.5
2	Large	No co-evolution	58.2	47.8	39.1	76.2	65.0	54.7
3	Large	No CCMpred	60.8	49.1	40.6	76.9	67.9	56.9
4	Large	No metagenome	75.6	63.3	53.7	90.8	82.4	70.4
5	Small	All	74.0	61.4	51.2	89.8	78.1	65.1
6	Small	No full CCMpred	68.8	56.6	47.0	89.5	75.5	64.4
7	L60F150	All	78.3	64.0	53.5	88.3	77.9	66.9
8	L100F80	All	80.6	65.1	54.8	94.3	81.8	68.6

Casp13 FM, human designed, and contact prediction







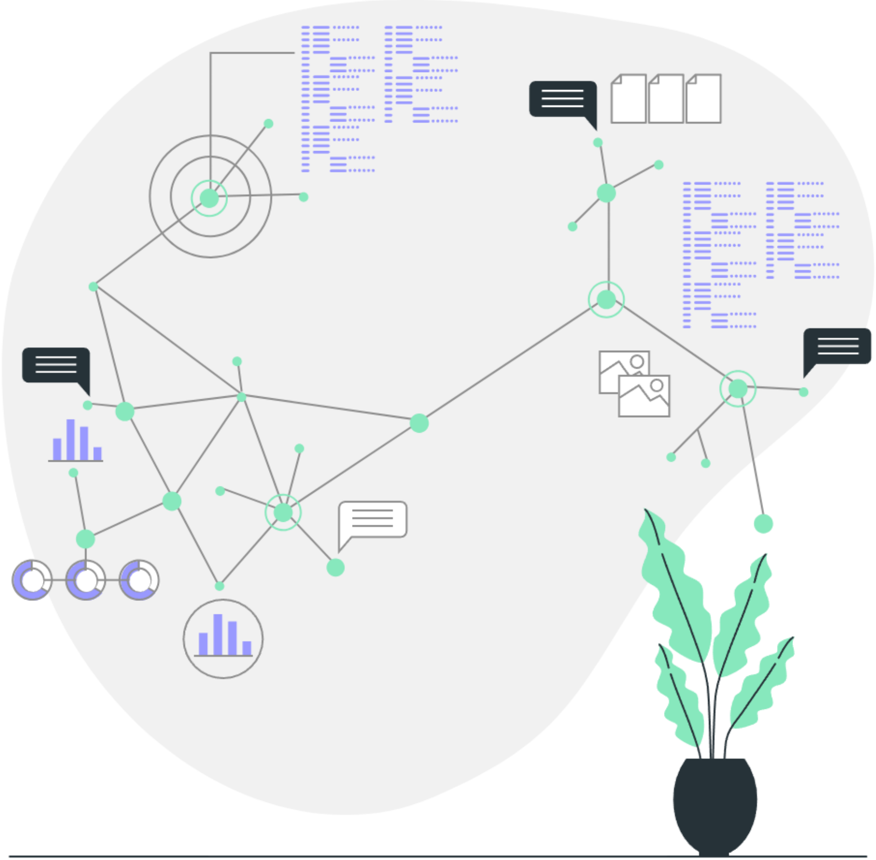
Impact of different settings

- Without co-evolution the model showed a decrease of 13% in the F1 value
- The model had a 4.6% decrease in the F1 when using the smaller model
- Model depth is the main contributing factor not the width
- The metagenomic data had a 3.4% contribution

04

Conclusion

What can we take away?



Key points

- Co-evolutionary data is a large factor in structure prediction
- The size of the model and metagenomic data can boost performance
- Predicting natural proteins without coevolutionary data doesn't work well
- Human-designed proteins work well
    - Probably due to low energy wells
- Their method still needs some sequences to work

What can be Improved

- The systems could be improved by working on the ResNet and its training.
- It has been shown that larger models provide better results
- An improved architecture can help boost performance
- The use of techniques like recycling could help improve the results

References

- Xu, J., McPartlon, M. & Li, J. Improved protein structure prediction by deep learning irrespective of co-evolution information. Nat Mach Intell 3, 601–609 (2021). https://doi.org/10.1038/s42256-021-00348-5
- Yang, J. Y. et al. Improved protein structure prediction using predicted interresidue orientations. Proc. Natl Acad. Sci. USA 117, 1496–1503 (2020).
- Jumper, J., Evans, R., Pritzel, A. et al. Highly accurate protein structure prediction with AlphaFold. Nature 596, 583–589 (2021). https://doi.org/10.1038/s41586-021-03819-2
- Stefan Seemayer, Markus Gruber, Johannes Söding, CCMpred—fast and precise prediction of protein residue–residue contacts from correlated mutations, Bioinformatics, Volume 30, Issue 21, November 2014, Pages 3128–3130, https://doi.org/10.1093/bioinformatics/btu500

Thanks!

Congrats you survived :)

CREDITS: This presentation template was created by **Slidesgo**, and includes icons by **Flaticon**, and infographics & images by **Freepik**