Novel Techniques II

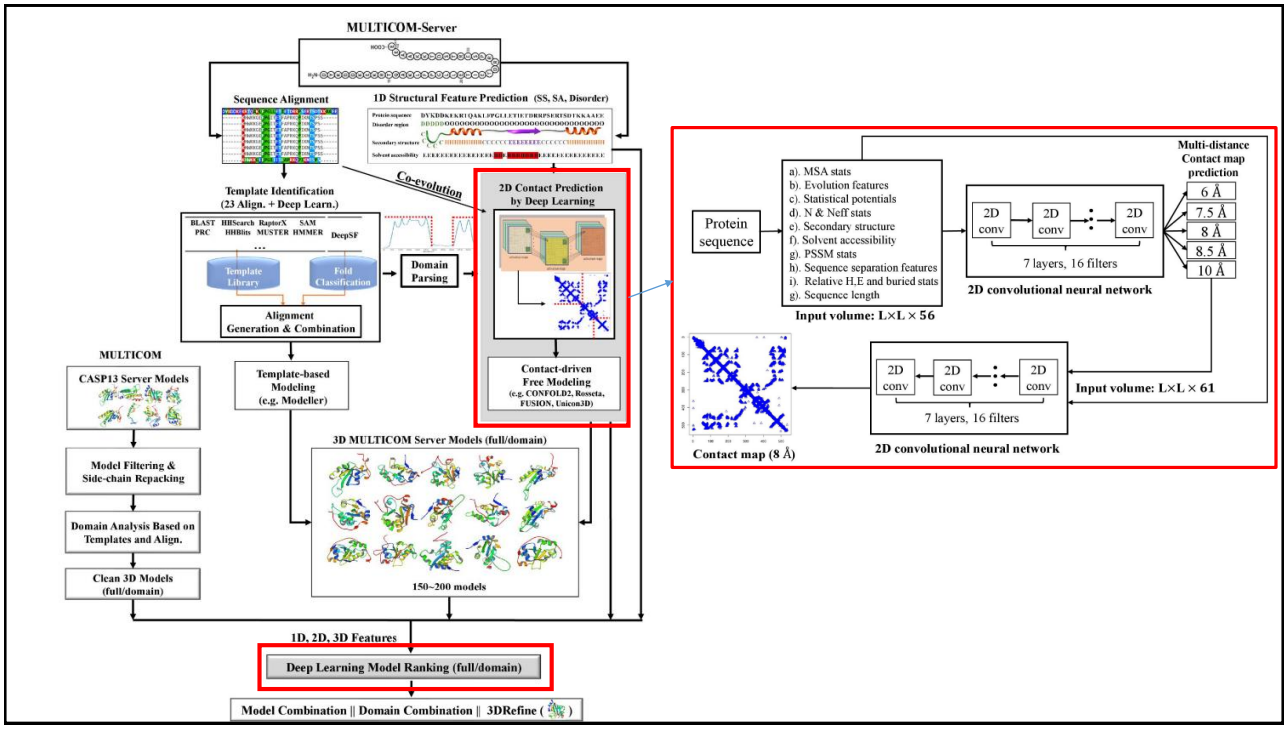
E.M. Bakker

J. Hou, et al. Protein tertiary structure modeling driven by deep learning and contact distance prediction in CASP13, 2019.

- CASP10 (2012) deep learning for contact and distance distribution prediction.
- CASP11 (2014) prediction of residue-residue distance relationships (e.g. contacts) is the key direction to advance protein tertiary structure prediction.
- CASP11 and CASP12: successes of residue-residue co-evolutionary analysis

CASP13 (2018) MULTICOM (3rd place) a protein structure prediction system with three major deep learning components:

- contact distance prediction based on deep convolutional neural networks
- contact distance-driven template-free (ab initio) modeling
- protein model ranking empowered by deep learning and contact prediction
- further components: template library, sequence database, and alignment tools.
- MULTICOM was ranked 3rd out of all 98 predictors in both template-free and template-based protein structure modeling in CASP13.



[11] Jinbo Xu, Distance-based Protein Folding Powered by Deep Learning. Nov. 2018.

- Deep ResNet for distance prediction, several classes of distances
- ***Predict inter-atom distance by deep 1D and 2D deep residual networks (ResNet)***
  - We discretize inter-atom distance into 25 bins: <4Å, 4-4.5Å, 4.5-5Å, 5-5.5Å, …, 15-15.5Å, 15.5-16Å, and >16Å.
  - Predict both C_β-C_β distance distribution, as well as distance distributions for: C_α-C_α, C_α-C_g, C_g-C_g, and N-O. Here C_g is the first CG atom in an amino acid, if CG does not exist, OG or SG is used.
- ***Predict secondary structure and torsion angles by 1D deep residual network***
  - predict 3-state secondary structure and backbone torsion angles ϕ and ψ for each residue.
- **Folding by predicted distance, secondary structure and torsion angles**
  - first predict its inter-atom distance matrix, secondary structure and backbone torsion angles,
  - then convert the predicted information into CNS* restraints
  - finally build its 3D models by CNS28, a software program for experimental protein structure determination.

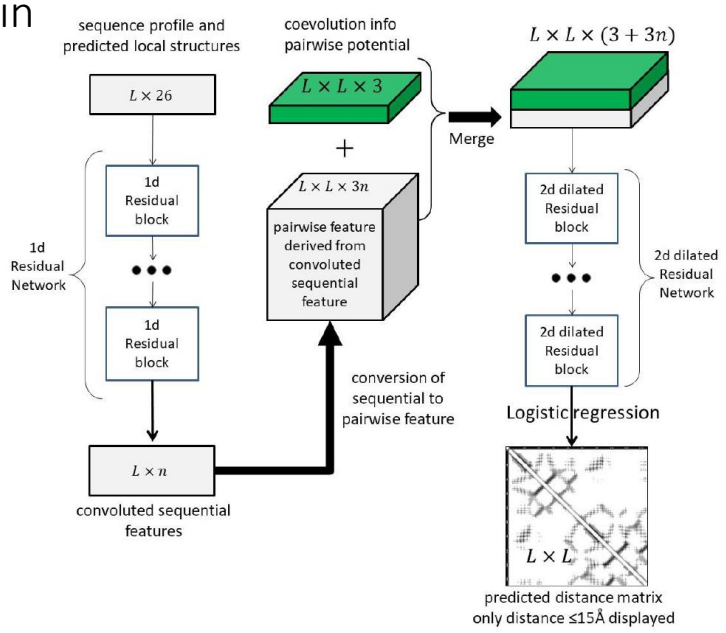
*) CNS (Crystallography and NMR System) is a suite of programs designed for crystallography and NMR.

Distance-based Protein Folding Powered by Deep Learning. [11]

The overall deep network architecture for the prediction of protein distance matrix.

- The left column is a 1D deep residual neural network that transforms sequential features (e.g., sequence profile and predicted secondary structure).
- The right column is a 2D deep dilated residual neural network that transforms pairwise features.
- The middle column converts the convoluted sequential features to pairwise features and combine them with the original pairwise features.

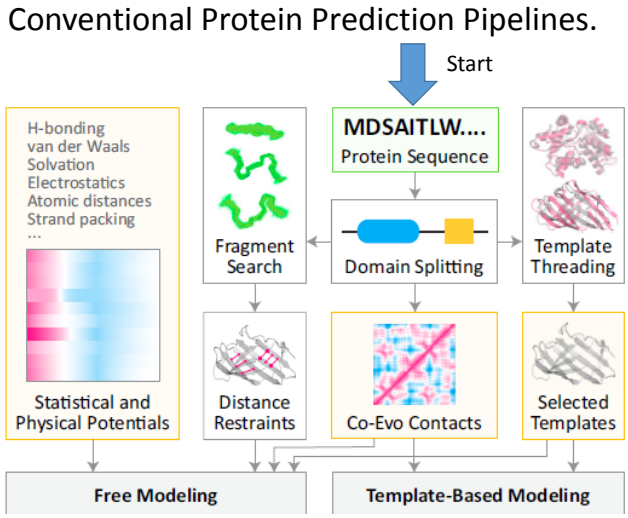
Figure from [11].
RaptorX



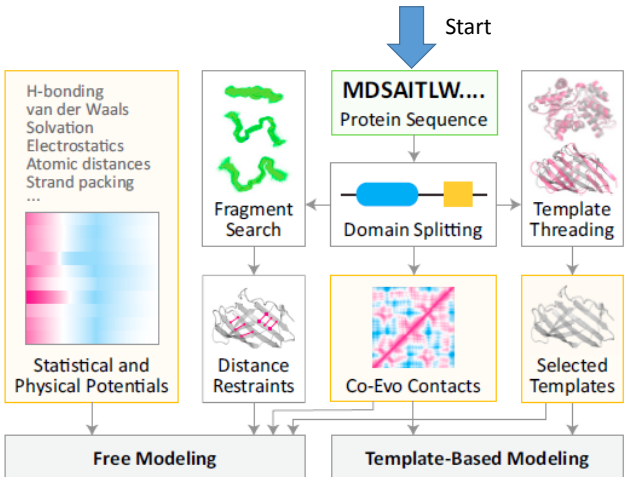
[14] M. AlQuraishi, End-to-End Differentiable Learning of Protein Structure, Cell Systems 8, 292–301, April 24, 2019.

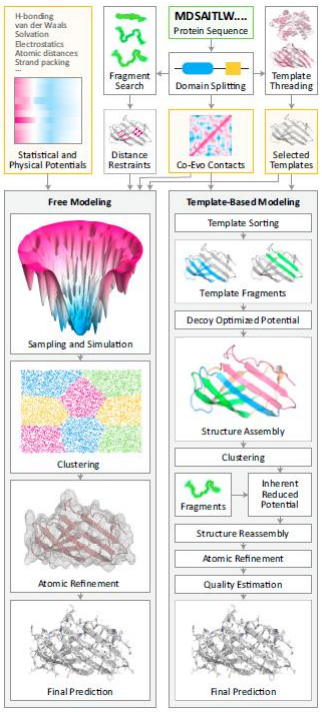
- Neural network predicts protein structure from sequence without using co-evolution
- Model replaces structure prediction pipelines with one mathematical function
- Achieves state-of-the-art performance on novel protein folds
- Learns a low-dimensional representation of protein sequence space

[14] M. AlQuraishi, End-to-End Differentiable Learning of Protein Structure, Cell Systems 8, 292–301, April 24, 2019.



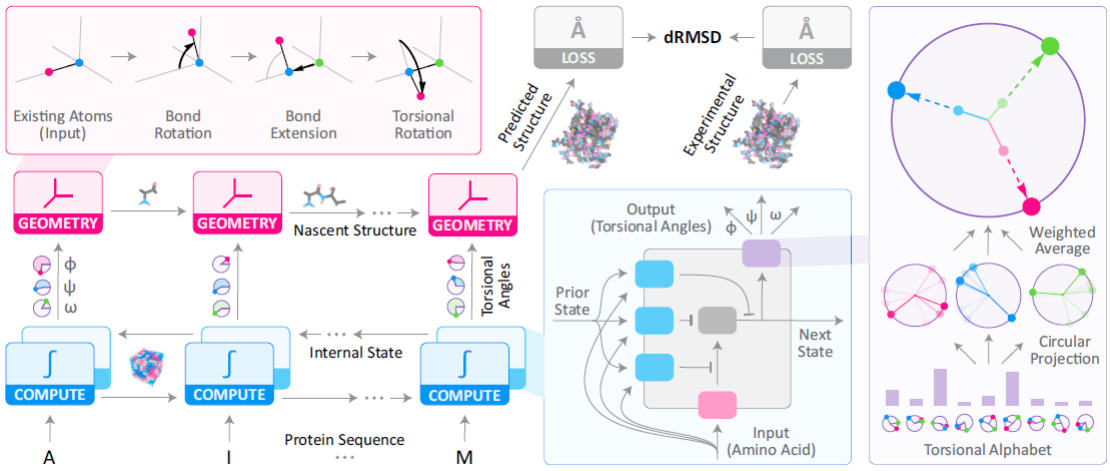
[14] M. AlQuraishi, End-to-End Differentiable Learning of Protein Structure.





[14] M. AlQuraishi, End-to-End Differentiable Learning of Protein Structure:

Recurrent Geometric Networks



[14] M. AlQuraishi, End-to-End Differentiable Learning of Protein Structure:

Recurrent Geometric Networks

Table 1. Comparative Accuracy of RGNs Using dRMSD

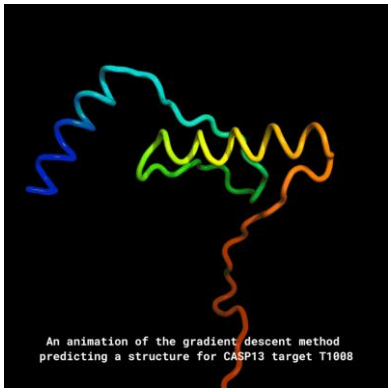
	FM (Novel Folds) Category (Å)						TBM (Known Folds) Category (Å)					
	CASP7	CASP8	CASP9	CASP10	CASP11	CASP12	CASP7	CASP8	CASP9	CASP10	CASP11	CASP12
RGN	9.3*	7.3*	8.7*	10.0*	8.5*	10.7*	5.6	5.9	6.5	6.9	7.4	6.9
1st server	9.3	8.3	9.0	10.3	9.3	11.0	4.0*	4.3*	5.2*	5.3*	5.8*	4.7*
2nd server	9.9	8.6	9.1	10.6	9.6	11.2	4.0	4.6	5.2	5.4	6.0	4.8
3rd server	10.0	9.2	9.7	10.9	11.2	11.3	4.1	4.8	5.4	5.7	6.5	5.6
4th server	10.1	9.9	10.1	11.7	11.7	11.4	4.2	5.0	5.4	5.9	6.8	5.8
5th server	10.4	10.4	13.5	12.0	12.9	13.0	4.8	5.0	5.5	7.2	6.9	5.9

The average dRMSD (lower is better; asterisk indicates best performing method) achieved by RGNs and the top five servers at each CASP is shown for the novel folds (left) and known folds (right) categories. Numbers are based on common set of structures predicted by top 5 servers during each CASP. A different RGN was trained for each CASP, using the corresponding ProteinNet training set containing all sequences and structures available prior to the start of that CASP. See also Tables S1–S3.

dRMSD: root-mean-square deviation between the atoms in two configurations

AlphaFold [9] by DeepMind

- A7D is the best preforming algorithm in the Free Modelling Category of CASP13

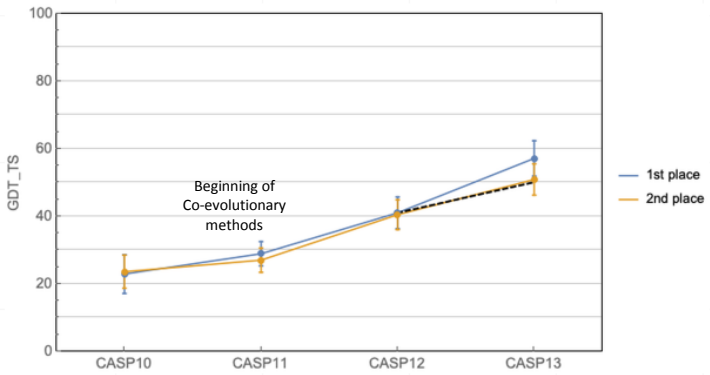


https://deepmind.com/blog/alphafold/

#	GR code	GR name	Domains Count	SUM Zscore (>-2.0)	Rank SUM Zscore (>-2.0)	AVG Zscore (>-2.0)	Rank AVG Zscore (>-2.0)	SUM Zscore (>0.0)	Rank SUM Zscore (>0.0)	AVG Zscore (>0.0)	Rank AVG Zscore (>0.0)
1	043	A7D	104	120.4307	1	1.1580	1	128.0693	1	1.2314	1
2	322	Zhang	104	107.5948	2	1.0346	2	108.1948	2	1.0403	2
3	089	MULTICOM	104	99.4661	3	0.9564	3	99.9886	3	0.9614	3
4	145	QUARK	104	90.9915	4	0.8749	4	91.5625	4	0.8804	4
5	261	Zhang-Server	104	88.9540	5	0.8553	5	89.7597	5	0.8631	5
6	460	McGuffin	104	81.6353	6	0.7850	6	84.4019	6	0.8116	6
7	354	wfAll-Cheng	104	77.7039	7	0.7472	7	80.9951	7	0.7788	8
8	135	SBROD	102	71.5656	9	0.7408	8	78.9792	8	0.7743	9
9	324	RaptorX-DeepModeller	104	75.4891	8	0.7259	9	78.5878	9	0.7557	10
10	197	MESHI	104	70.9761	10	0.6825	11	76.6354	10	0.7369	11

AlphaFold [9] by DeepMind

- Until CASP10 no big improvements for a decade.
- CASP11: co-evolutionary methods.
  - Required MSA's. But Free Modelling targets would benefit only slightly from this.
- CASP11 – CASP13 showed further improvements because of co-evolutionary methods, e.g. Zhang (2nd place)



Curves show the best and second best predictors at each CASP, while the dashed line shows the expected improvement at CASP13 given the average rate of improvement from CASP10 to 12. Ranking is based on CASP assessor's formula, and does not always coincide with highest mean GDT_TS (e.g. CASP10.) Error bars correspond to 95% confidence intervals.

From [10].

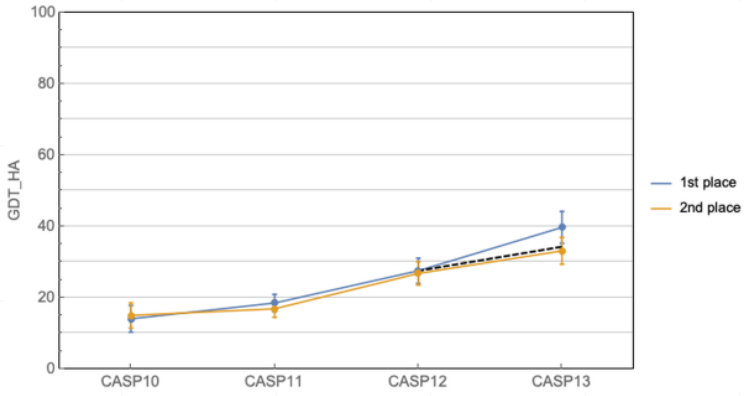
GDT_TS
GDT_TS - GlobalDistanceTest_TotalScore
GDT_TS = (GDT_P1 + GDT_P2 + GDT_P4 + GDT_P8)/4, where GDT_Pn denotes percent of residues under distance cutoff <= nÅ

AlphaFold [9] by DeepMind

- GDT_TS measures the quality of the overall topology gives a distorted image of the problem status.
- GDT_HA measures the quality of the topology in higher resolution, which is more appropriate for further applications using the predicted 3d structure

⇒problem far from solved.

⇒local goodness of fit?



From [10].

AlphaFold

Co-evolution method:

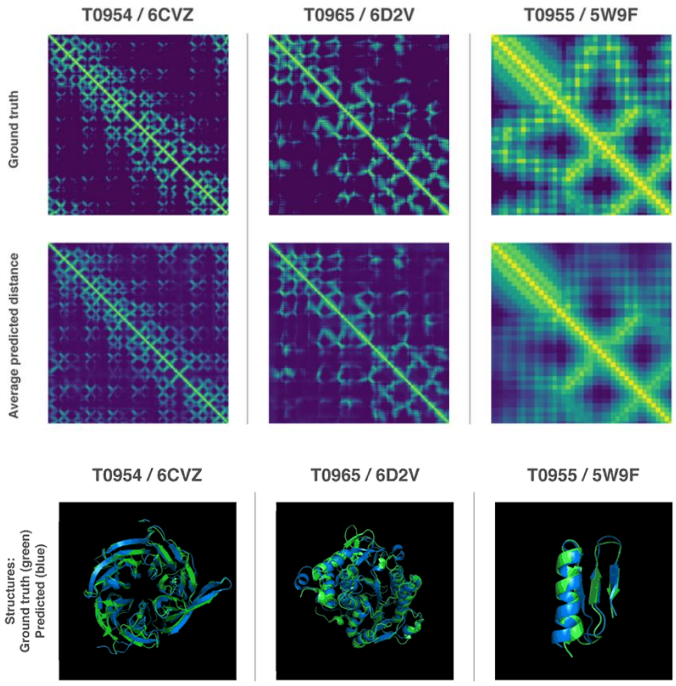
- evolutionary couplings from protein MSAs by detecting residues that co-evolve
- suggesting physical proximity in 3D space.
- predicted binary contact matrices from MSAs, i.e. whether two residues are “in contact” or not (typically defined as being within <8Å),
- Used in geometric constraint satisfaction methods

Exploited by:

- coupling of such binary contacts with folding pipelines such as Rosetta and I-Tasser,
- convolutional networks and deep architectures (residual networks) to integrate information with the matrix of raw couplings to obtain more accurate contacts. Jinbo Xu’s group.
- **inter-residue distance prediction** instead of **binary contacts**: predicted probabilities over a discretized spatial range and then picked the highest probability one for feeding into CNS to fold the protein. (Xu’s preprint before CASP13)
- Is one of the key ingredients of AlphaFold

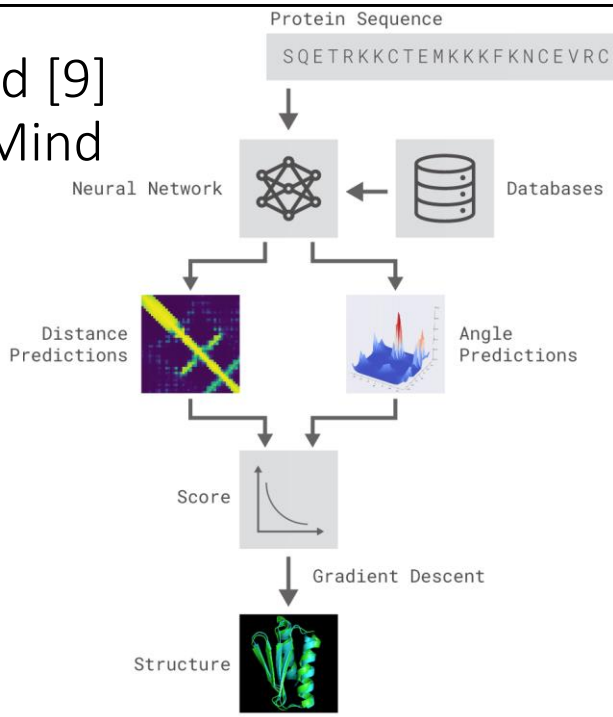
Alpha Fold [9] by DeepMind

- Predict Physical Properties



Alpha Fold [9] by DeepMind

- Predict Physical Properties
- no plans to release the source code



AlphaFold characterized as in [10]

General ideas [10]:

- A softmax over discretized spatial ranges gives a predicted probability distribution over distances

  (Note: value distribution vs Value [12])
- The convolutional ResNet uses the distribution as a (protein-specific) statistical (normalized) potential function
- Normalizing is done using a learned reference state.
- Minimizes the statistical potential function using gradient descent (L-BFGS), to generate the protein fold.

Details:

- The L-BFGS minimizer operates independently from the neural network.
- The energy potential is coupled with a more traditional physics-based potential. The combined energy function is minimized.
- The potentials are a consequence of the MSA (or sequence + PSSM).
- A smooth potential surface for the given protein family is constructed, and whose minimum closely matches that of the native protein (-family average) fold.

AlphaFold characterized as in [10]

In most methods the following paradigm was used when handling co-evolutionary data:

- predict contacts → feed into complex folding algorithm

But:

- More complex approaches were tried, such as fragment assembly using a generative variational autoencoder.
- But the more simple and direct minimization of their predicted energy function was found to be more effective in predicting a high accuracy fold.

AlphaFold Compared [10]

	Zhang	Xu [11]	AlphaFold [2,3]	NEMO	RGN
Inputs	MSA	MSA	MSA	Sequence or PSSM	PSSM
Outputs (pre-folding)	Binary Contacts	Distances	Distributions over distances	Cartesian coordinates (folding internal)	Cartesian coordinates (folding internal)
Folding	I-Tasser	CNS	L-BFGS	Differentiable Langevin dynamics	Implicit
Energy function	Explicit, fixed, and universal	None	Explicit, learned, and MSA-specific	Explicit, learned, and sequence- or PSSM-specific	Implicit, learned, and PSSM-specific
Uses templates	Yes	No	No	No	No
End-to-end differentiable	No	No	No	Yes	Yes

Both AlphaFold and Xu use simple folding engines L-BFGS (L- Broyden–Fletcher–Goldfarb–Shanno (BFGS)) and CNS (Crystallography and NMR System), respectively, i.e., improvements come from a better energy potential using distributional information.

NB: important slide for the Final Assignment.

M. AlQuraishi, ProteinNet: a standardized data set for machine learning of protein structure. Feb 2019

ProteinNet data for training and validating ML Protein Structure Predictors:

- Integrated data: sequence, structure, and evolutionary information
- Multiple sequence alignments of all structurally characterized proteins
- Standardized data to emulate past CASP (Critical Assessment of protein Structure Prediction) experiments by capturing the historical states for CASP7 – CASP12.
- New validation sets constructed using evolution-based distance metrics to segregate distantly related proteins

Availability: Data sets and associated TensorFlow-based parser are available for download at https://github.com/aqlaboratory/proteinnet

M. AlQuraishi, ProteinNet: a standardized data set for machine learning of protein structure. Feb 2019

Data set	Cutoff date	Structures*	Sequences*
ProteinNet 7	2006/5/1	34,557	4,817,827
ProteinNet 8	2008/5/5	48,087	15,756,117
ProteinNet 9	2010/5/3	60,350	24,688,095
ProteinNet 10	2012/5/1	73,116	63,477,198
ProteinNet 11	2014/5/1	87,573	173,908,140
ProteinNet 12	2016/5/1	104,059	332,283,871

* Non-redundant

M. AlQuraishi, ProteinNet. Feb 2019

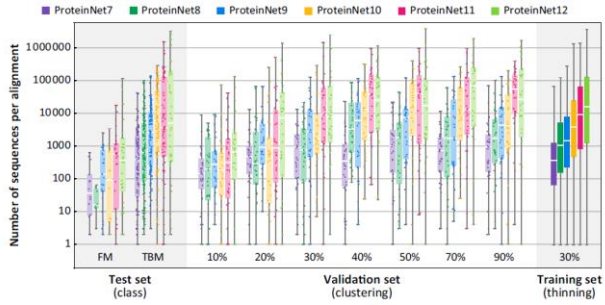
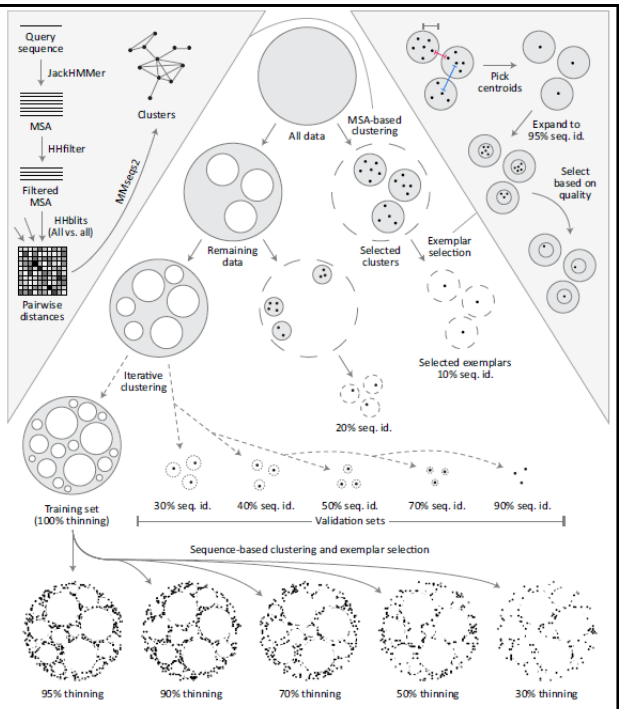


Fig. 3. Alignment size as a function of ProteinNet subset. Box and whisker charts depict the distribution of number of sequences per MSA for ProteinNet training (30% thinning), validation, and test sets. Individual data points for training sets are not shown due to their large size.



References

[1] L.A. Abriata, G.E. Tam, B. Monastyrskyy, A. Kryshtafovych, M. Dal Peraro, Assessment of hard target modeling in CASP12 reveals an emerging role of alignment-based contact prediction methods. Proteins, Vol. 86, pp. 97-112, 2018.

[2] A. Kryshtafovych, B. Monastyrskyy, K. Fidelis, J. Moult, T. Schwede, A. Tramontano Evaluation of the template-based modeling in CASP12. Proteins, Vol. 86, pp. 321-334, 2018.

[3] K. Paliwal, J. Lyons, R. Heffernan, A Short Review of Deep Learning Neural Networks in Protein Structure Prediction Problems, Adv. Tech. Biol. Med. Volume 3, Issue 3, 2015.

[4] J. Schaarschmidt, B. Monastyrskyy, A. Kryshtafovych, A.M.J.J. Bonvin, Assessment of contact predictions in CASP12: Co-evolution and deep learning coming of age. Proteins, Vol. 86, pp. 51-66, 2018.

Not Peer Reviewed

[5] G. Derevyanko, S. Grudininy, Y. Bengiozx, G. Lamoureux, Deep convolutional networks for quality assessment of protein folds. arXiv:1801.06252v1, 18 Jan 2018.

[6] J. Hou, T. Wu, R. Cao, J. Cheng. Protein tertiary structure modeling driven by deep learning and contact distance prediction in CASP13, http://dx.doi.org/10.1101/552422doi: bioRxiv preprint first posted online Feb. 17, 2019.

References

[7] S.K. Sønderby, O. Winther, Protein Secondary Structure Prediction with Long Short Term Memory Networks. arXiv:1412.7828v2 [q-bio.QM] 4 Jan 2015.

[8] B. Zhang, J. Li, Q. Lü, Prediction of 8-state protein secondary structures by a novel deep learning architecture. BMC Bioinformatics, 19:293, 2018.

[9] R.Evans, J.Jumper, J.Kirkpatrick, L.Sifre, T.F.G.Green, C.Qin, A.Zidek, A.Nelson, A.Bridgland, H.Penedones, S.Petersen, K.Simonyan, S.Crossan, D.T.Jones, D.Silver, K.Kavukcuoglu, D.Hassabis, A.W.Senior, De novo structure prediction with deep-learning based scoring. In Thirteenth Critical Assessment of Techniques for Protein Structure Prediction (Abstracts) 1-4 December 2018.

[10] Mohammed AlQuraish,

https://moalquraishi.wordpress.com/2018/12/09/alphafold-casp13-what-just-happened/

[11] Jinbo Xu, Distance-based Protein Folding Powered by Deep Learning. Nov. 2018. (https://arxiv.org/abs/1811.03481)

[12] AMarc G. Bellemare, Will Dabney, Rémi Munos, Distributional Perspective on Reinforcement Learning. July 2017. (https://arxiv.org/abs/1707.06887)

[13] Mohammed AlQuraishi, ProteinNet: a standardized data set for machine learning of protein structure. Feb 2019 (https://arxiv.org/abs/1902.00249)

[14] M. AlQuraishi, End-to-End Differentiable Learning of Protein Structure, Cell Systems 8, 292–301

April 24, 2019.