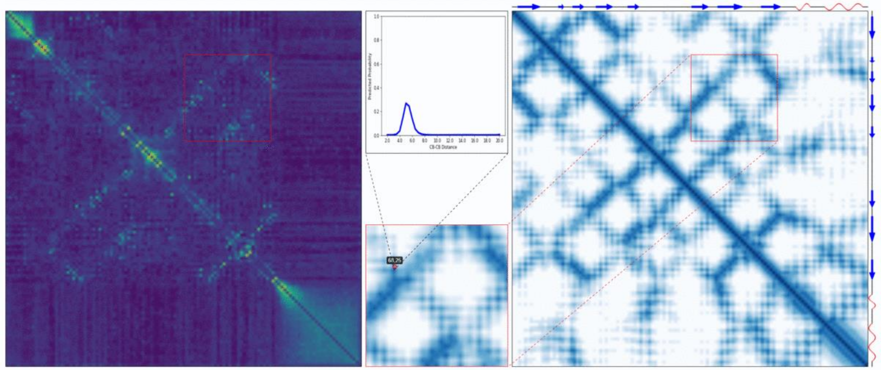
Tencent AI Lab

https://drug.ai.tencent.com/console/en/tfold



Accurate Contact/Distance Prediction by tFold

tFold develop team:

Sheng Wang*#, Haidong Lan*, Tao Shen*, Jiaxiang Wu*, Liangzhen Zheng*, Jianguo Pei*, Yuyi Liu, Junhong Huang, Ningqiao Huang, Zhenlei Xu, Wei Liu#, and Junzhou Huang#

CASP14 Conference

2020.12.03

Acknowledgement to CASP14 Organizers

CASP Organizers

John MOULT
President,
University of Maryland,
USA

Krzysztof FIDELIS
University of California,
Davis,
USA

Andriy KRYSHTAFOVYCH
University of California, Davis,
USA

Torsten SCHWEDE
University of Basel
SIB Swiss Institute of
Bioinformatics
Switzerland

Maya TOPF
Birkbeck, University of London, UK
CSSB (HPI and UKE) Hamburg,
Germany

Contact-prediction Section Chair

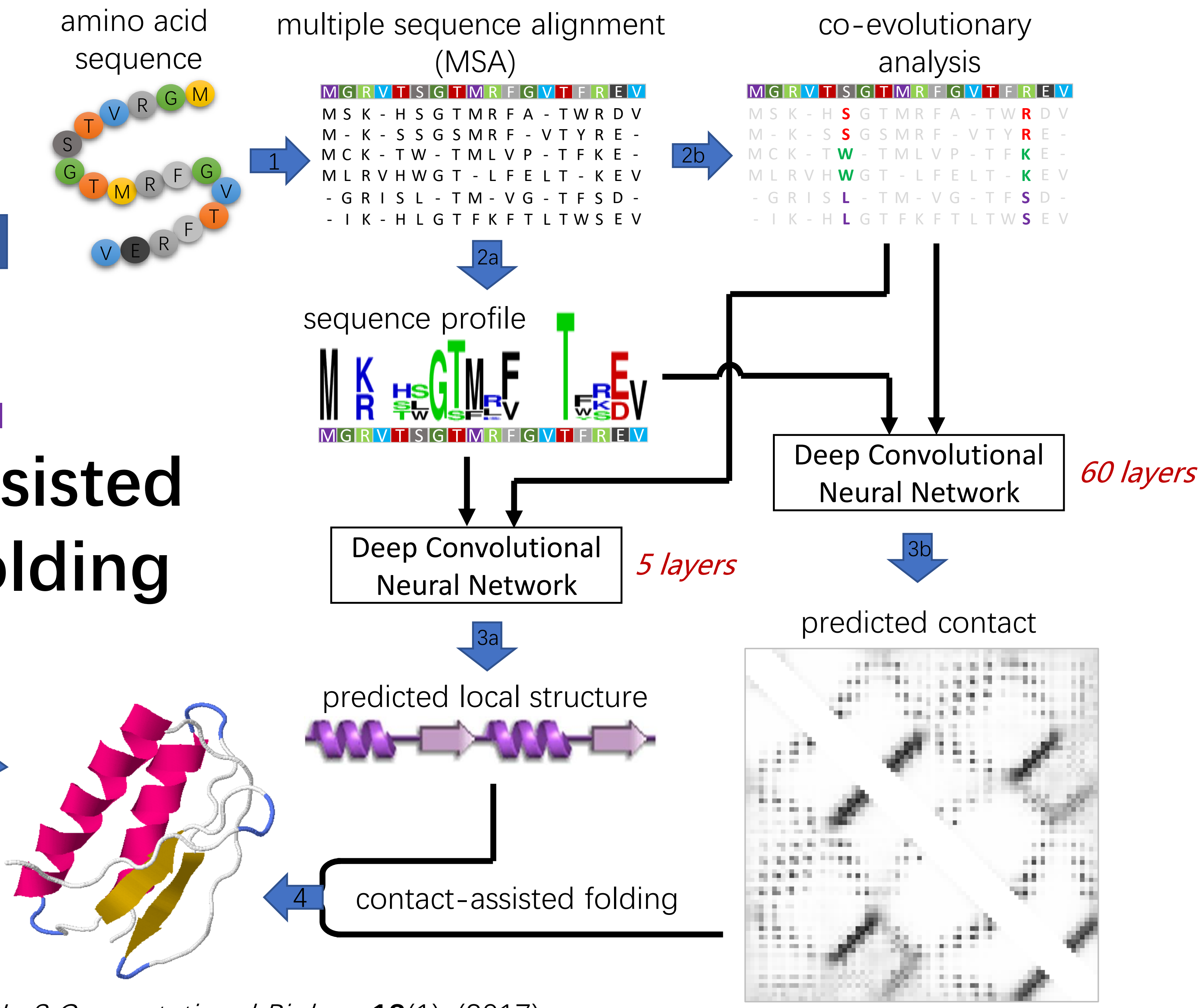
Jinbo Xu

Affiliation: Toyota Technological Institute at Chicago

1) Seq -> MSA
2) MSA -> Feat
3) Feat -> Cont
4) Cont -> Decoy

1) Original model

Contact-assisted de novo folding

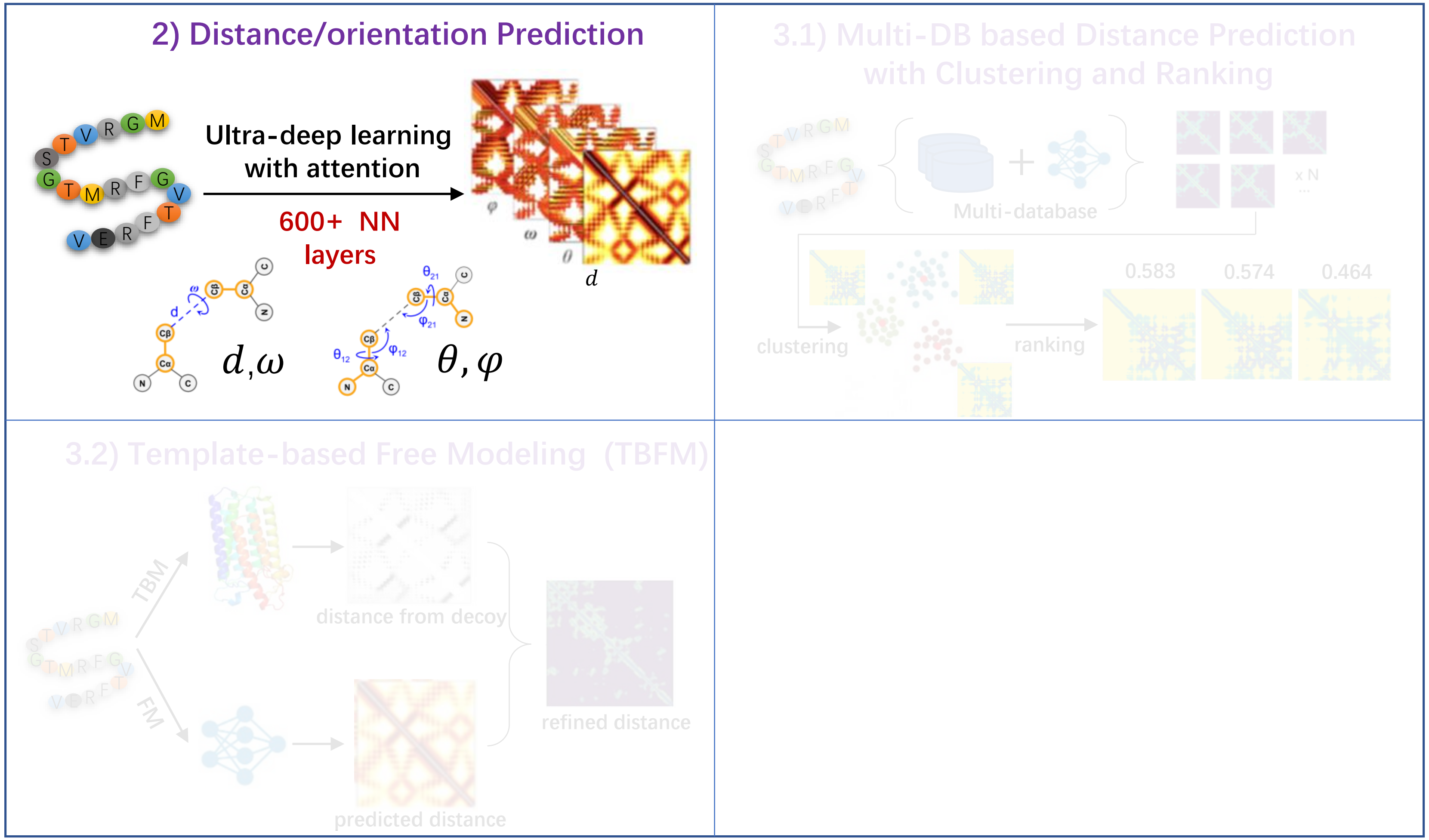


S Wang, ···, J Xu#. *PLoS Computational Biology* **13**(1), (2017)

Candidate issues of my previous work

- Contact V.S. distance/orientation
- Shallow network architecture
- Insufficient data usage
  a. More input MSAs
  b. More input decoys

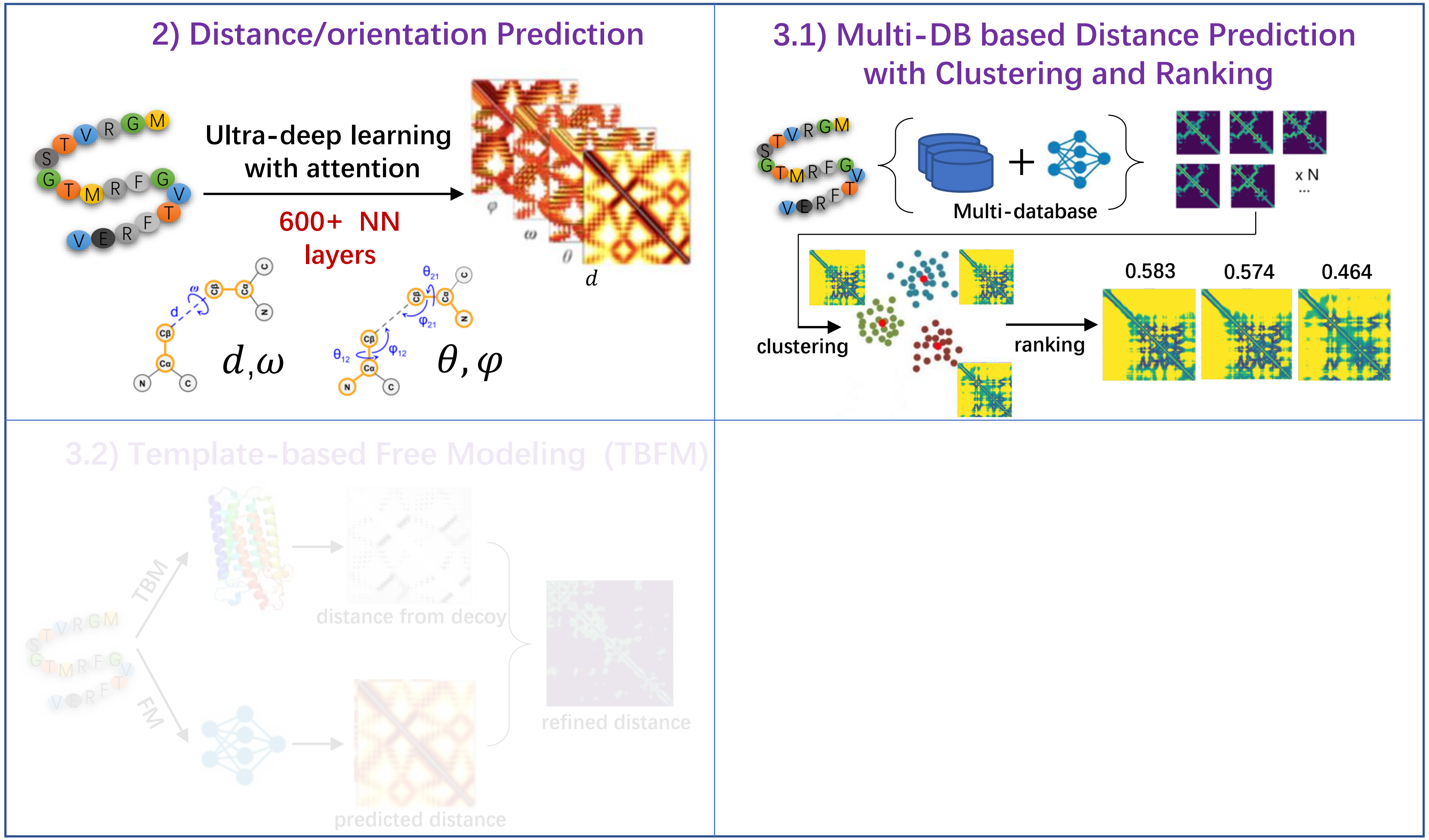
New developments in tFold contact prediction



Distance -> J Zhu, **S Wang**,···, J Xu#. *Bioinformatics* **34**(13), (2018)

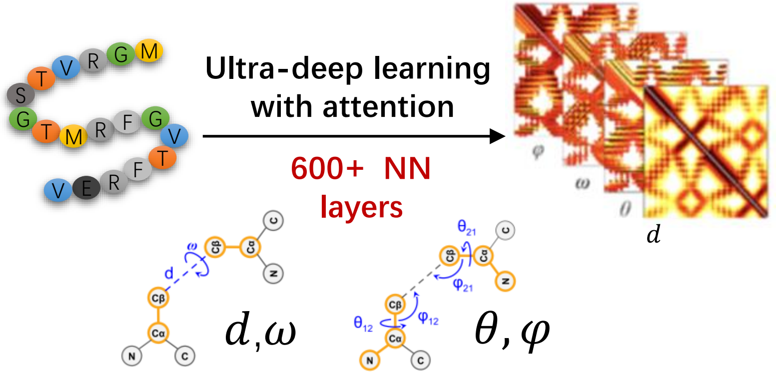
Orientation -> J Yang, ···, D Baker#. *PNAS* **117**(3), (2020)

New developments in tFold contact prediction

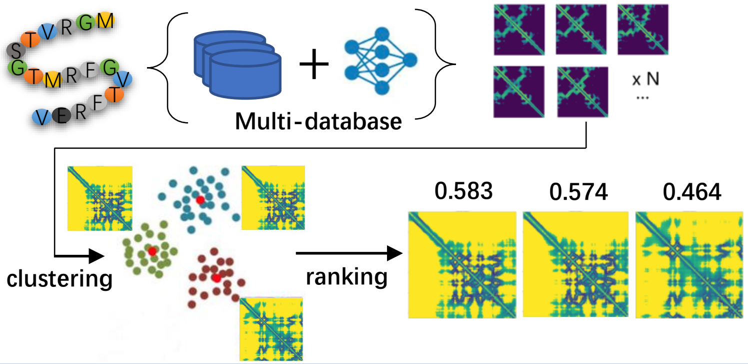


New developments in tFold contact prediction

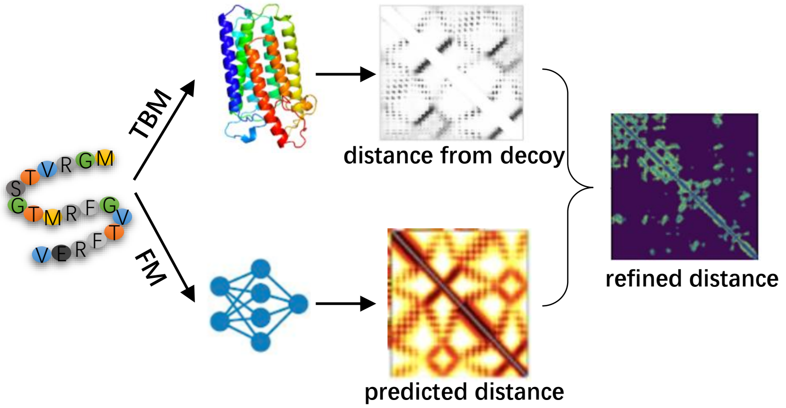
2) Distance/orientation Prediction



3.1) Multi-DB based Distance Prediction with Clustering and Ranking

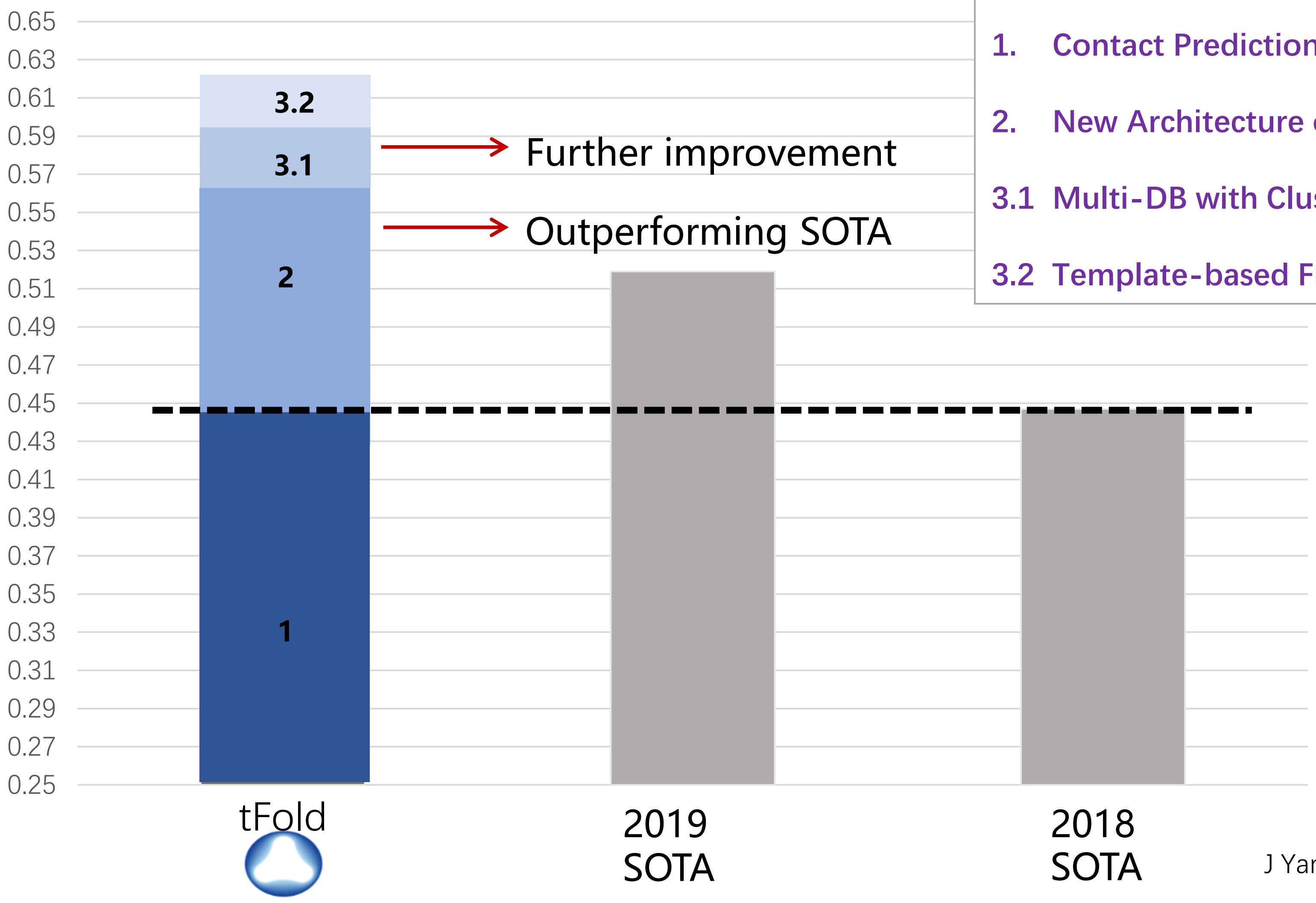


3.2) Template-based Free Modeling (TBFM)



[note]: the similar idea of TBFM first appears in J Xu[#], **S Wang**. *Proteins* **87**(12), (2019)

TopL Long Range Contact Precision



1. Contact Prediction (2018 model)
2. New Architecture of the Network

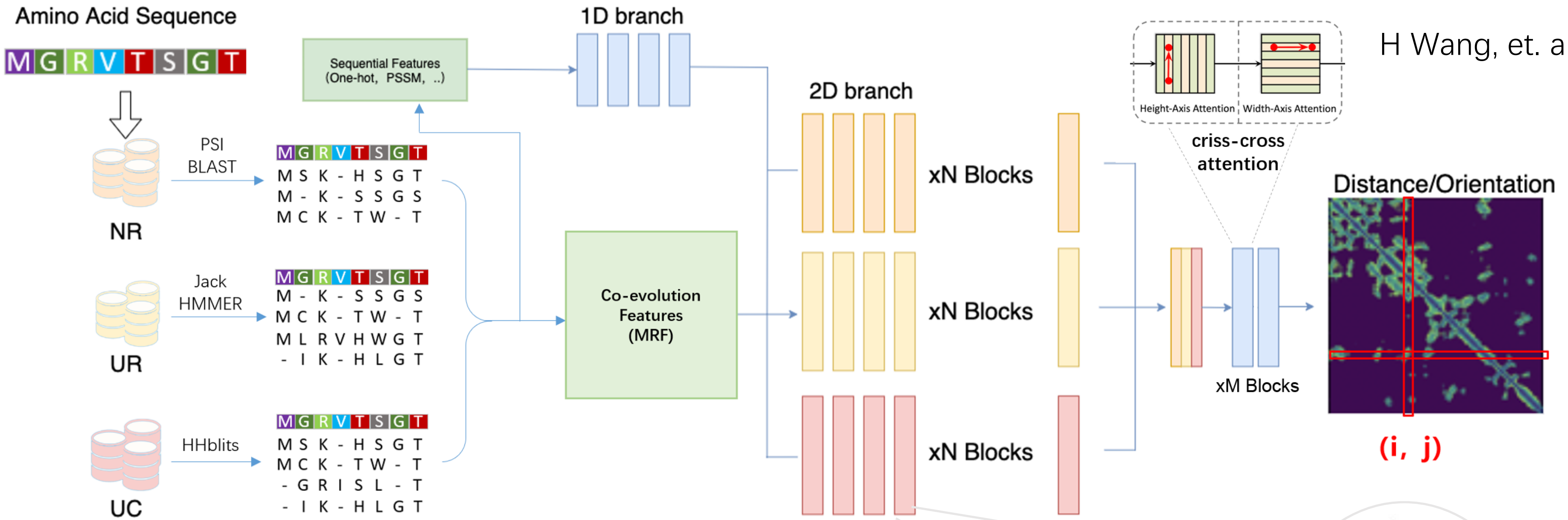
3.1 Multi-DB with Cluster and Rank

3.2 Template-based Free Modeling

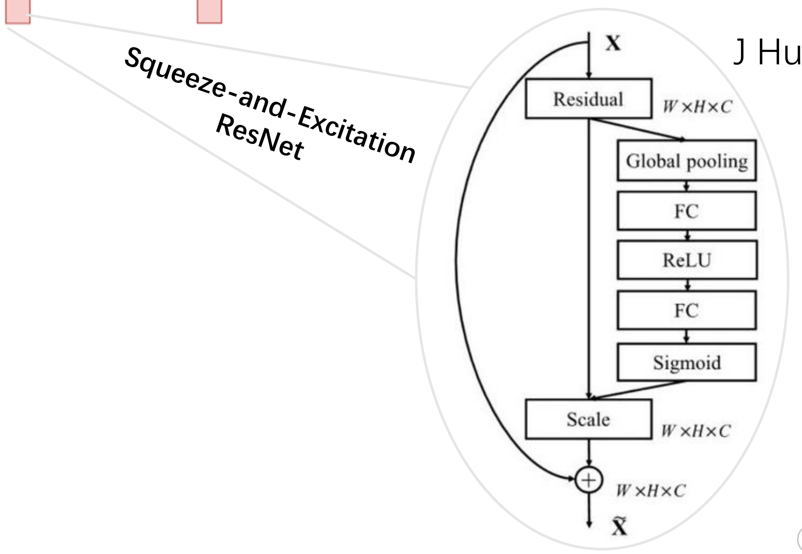
J Yang, ···, D Baker#. *PNAS* **117**(3), (2020)

Data draw using CASP13 targets

tFold: Distance/Orientation Prediction with SEResNet+2DAttention with a Multi-input Multi-task Scheme



H Wang, et. al. *ECCV* (2020)



J Hu, et. al. *CVPR* (2018)

Comparison of trRosetta and tFold

	trRosetta	tFold
Architecture	ResNet	SEResNet+2DAttention
Size	120 layers	600+ layers
Time	9 days (1 * RTX Titan)	2 days (16 * V100)

Ablation study of the Deep Learning model

a) Data construction:

- Construct MSAs from multi databases

…

b) Network architecture and loss design:

- SE ResNet module
- 2D Attention module
- Multi-task learning

…

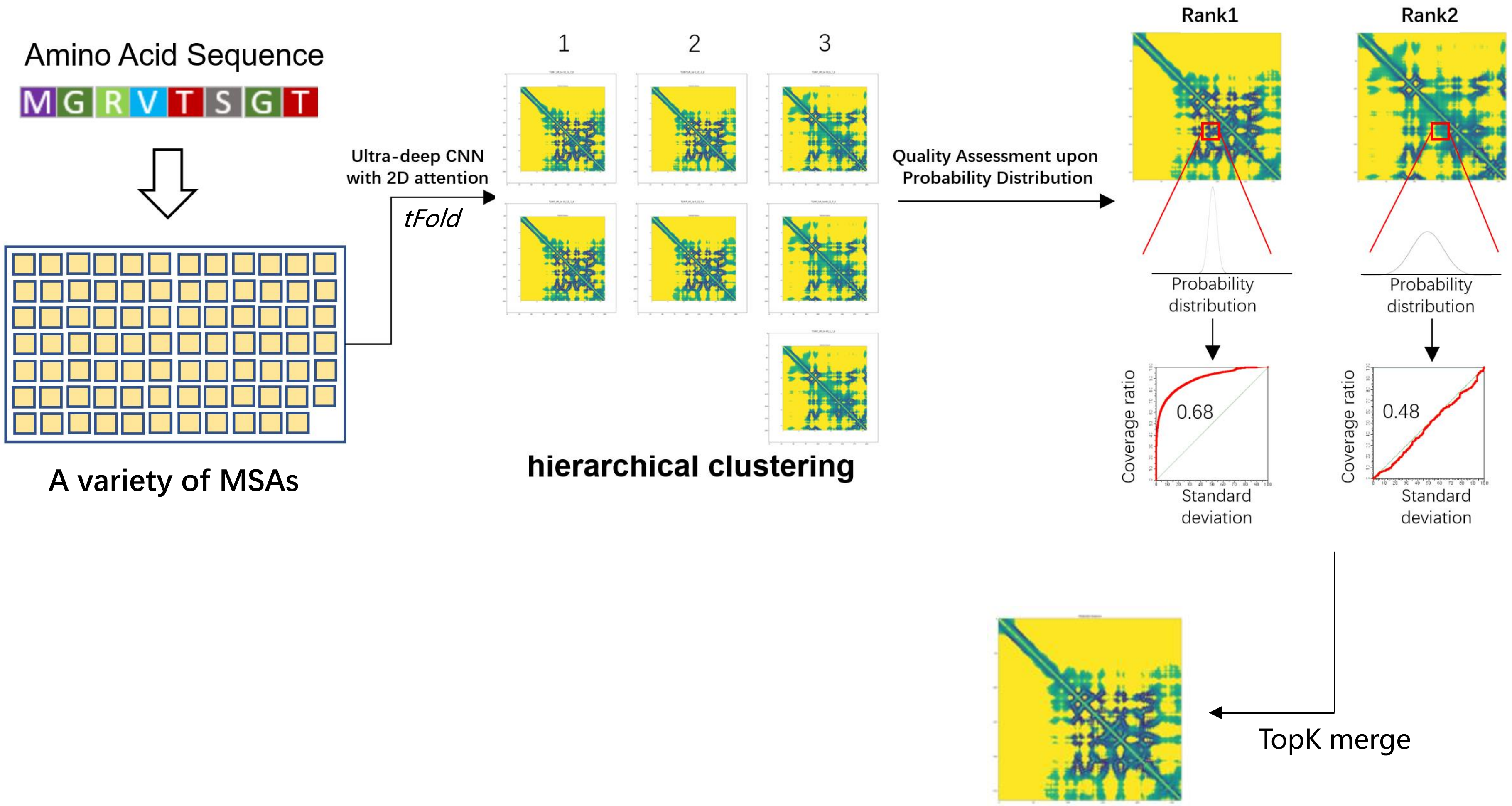
c) Training strategy:

- Progressive training strategy
- 600+ layer ultra-deep network

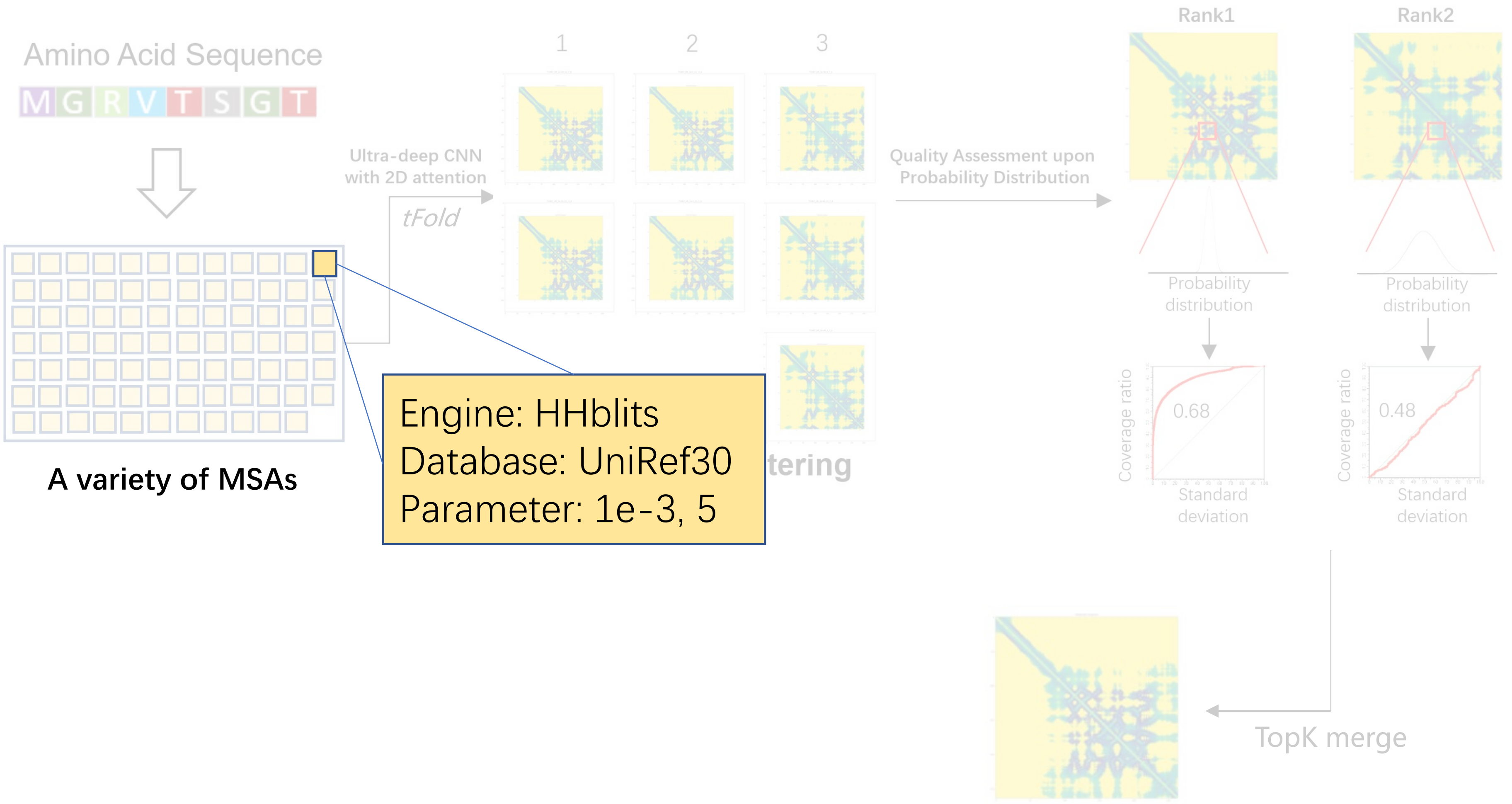
…

Model	CASP13 TopL long range contact
Baseline model	51.32% ~ 2019 SOTA
Baseline + a)	53.67%
Baseline + a) + b)	55.15%
Baseline + a) + b) + c)	56.37%

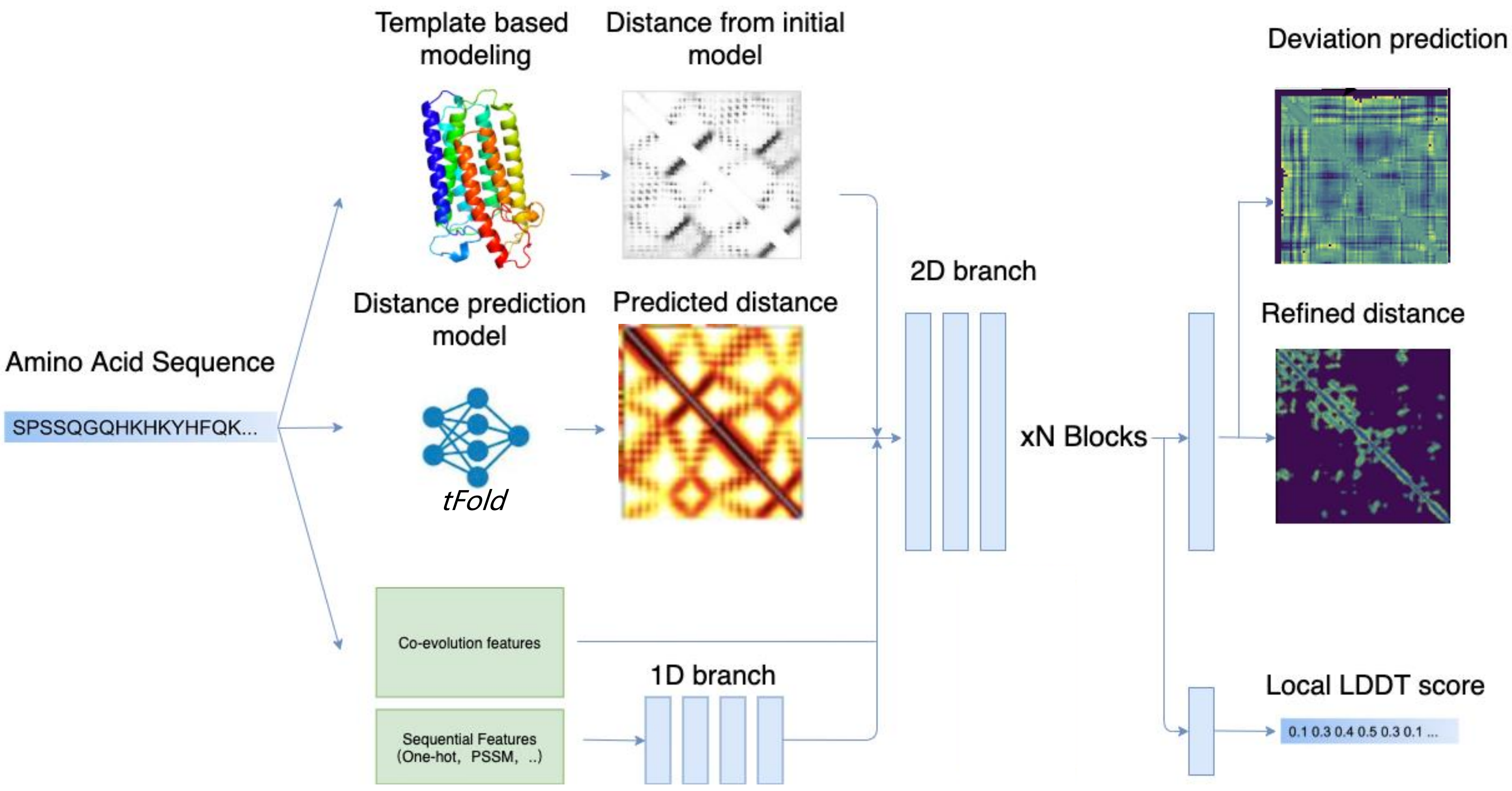
tFold-CaT: Multi-DB based Distance Prediction with Clustering and Ranking



tFold-CaT: Multi-DB based Distance Prediction with Clustering and Ranking

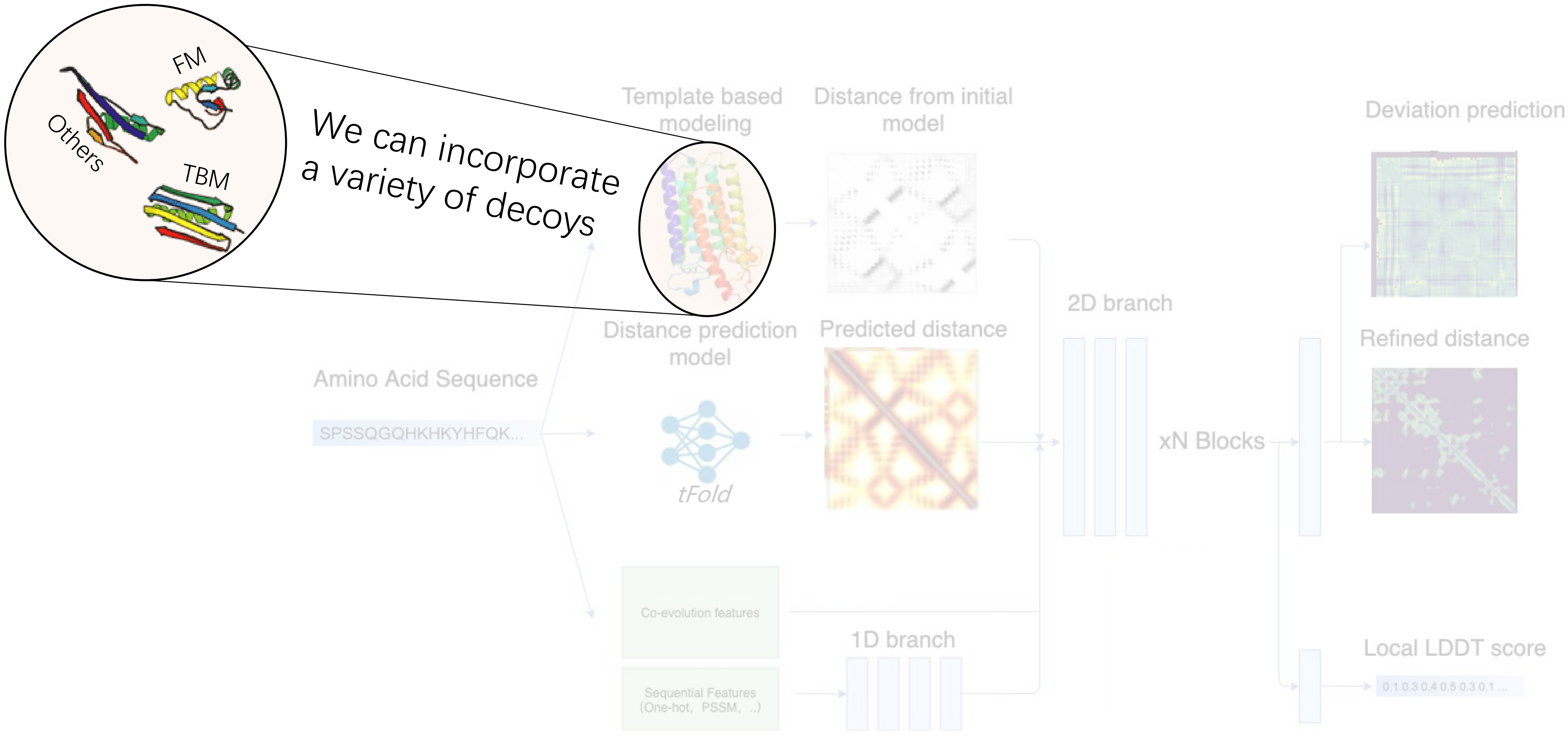


tFold-IDT: Template-based Free Modeling（TBFM)



[note]: the similar idea first appears in J Xu#, **S Wang**. *Proteins* **87**(12), (2019)

tFold-IDT: Template-based Free Modeling (TBFM)



The relationship between the decoy quality and the distance prediction enhancement

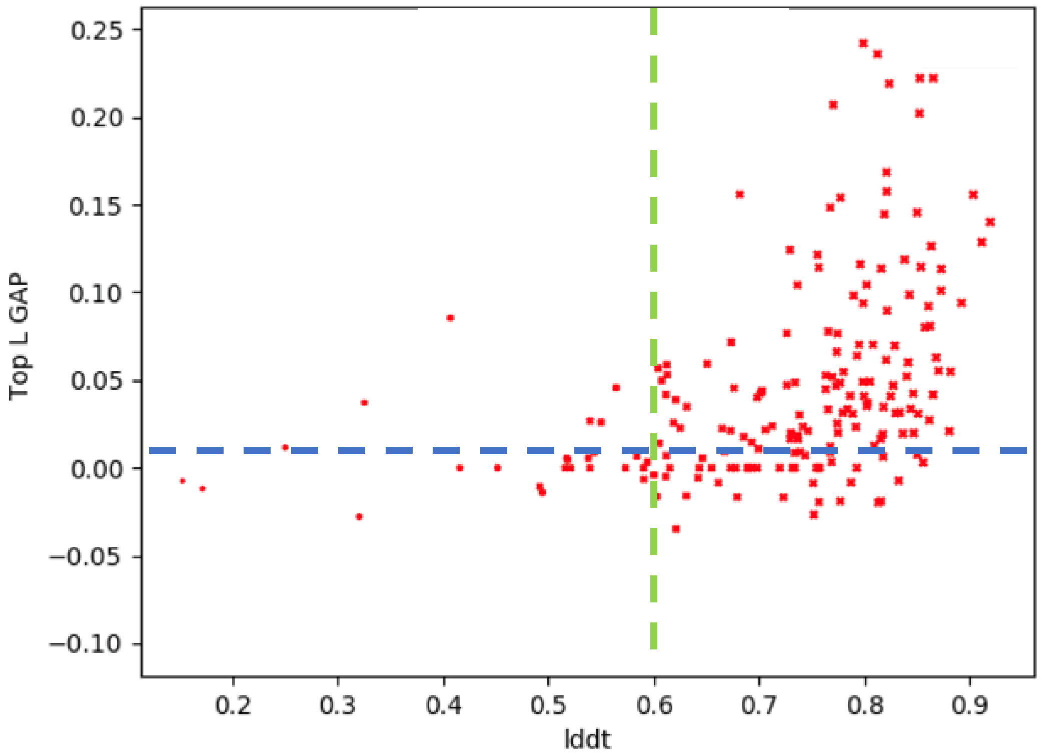
- High quality decoy

  significant enhance

- Low quality decoy

  won’t influence much

The robustness of our algorithm with respect of the decoy quality



Data draw using CAMEO targets from 2020-02-01 to 2020-05-02

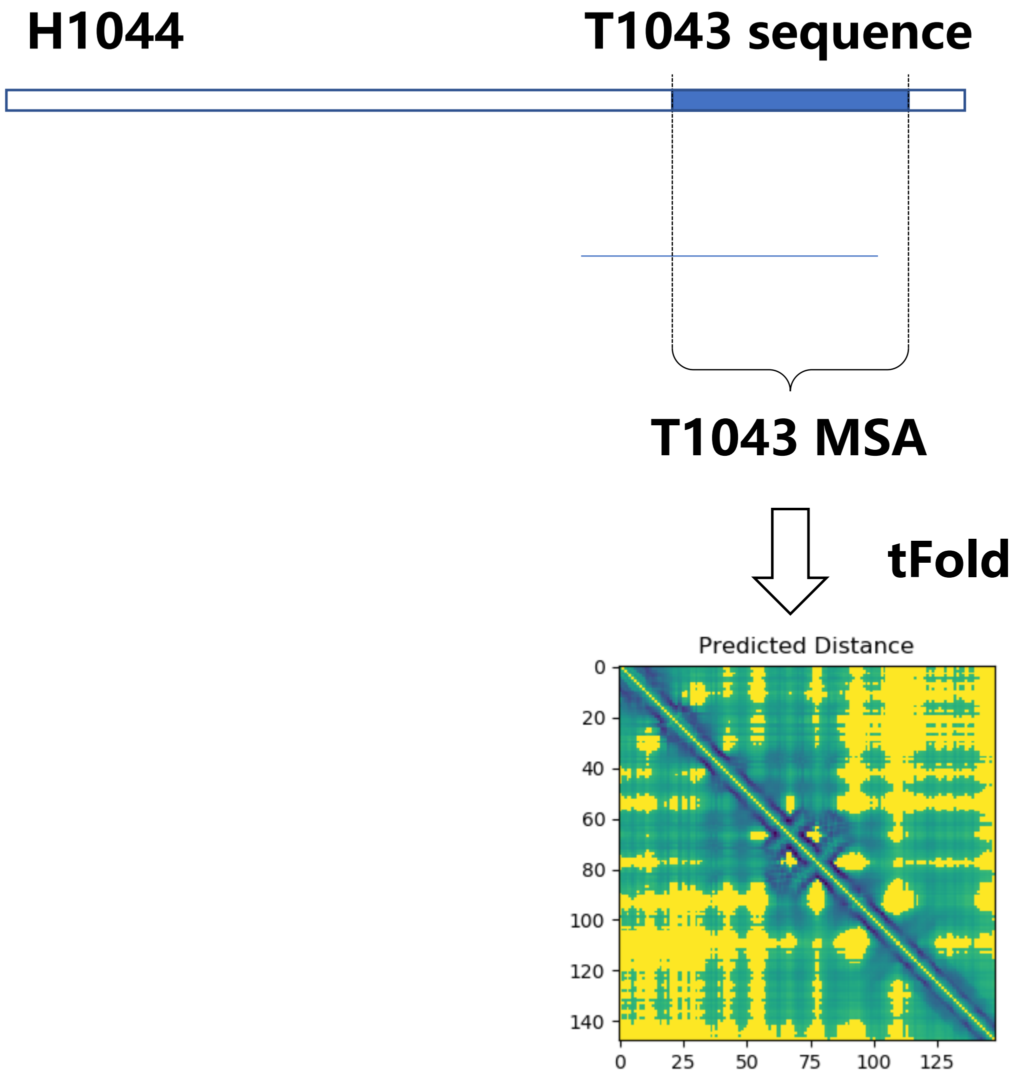
What goes right and why?

Group: --All-- **Target:** T1043-D1 **Contact Range:** long **List size:** L

Seq. Length: 148 **Depth of alignment: PSIBLAST** - 1; **HHBLITS** - 1 Similarity of models (Jaccard distance)

#	Model	No Conts.	F1	Prec	Recall
1	T1043RR368_1-D1	148	46.980	47.300	46.670
2	T1043RR009_1-D1	148	46.980	47.300	46.670
3	T1043RR488_1-D1	148	44.970	45.270	44.670
4	T1043RR183_1-D1	148	40.940	41.220	40.670
5	T1043RR351_1-D1	148	40.270	40.540	40.000
6	T1043RR238_1-D1	148	25.500	25.680	25.330
7	T1043RR157_1-D1	148	8.720	8.780	8.670
8	T1043RR326_1-D1	148	8.050	8.110	8.000

tFold series



What goes right and why?

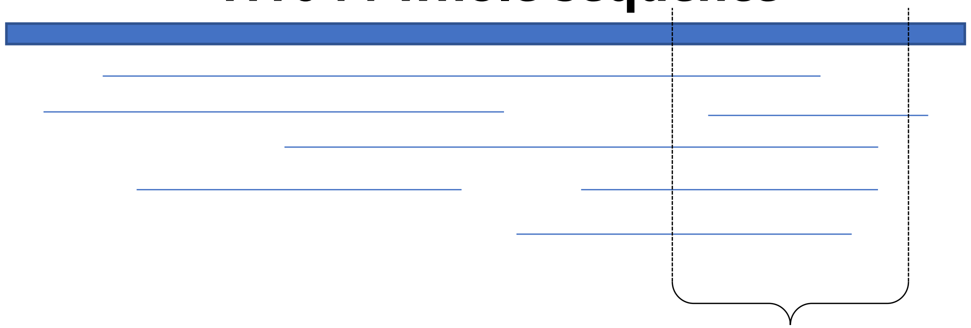
Group: --All-- **Target:** T1043-D1 **Contact Range:** long **List size:** L

Seq. Length: 148 **Depth of alignment: PSIBLAST** - 1; **HHBLITS** - 1 Similarity of models (Jaccard distance)

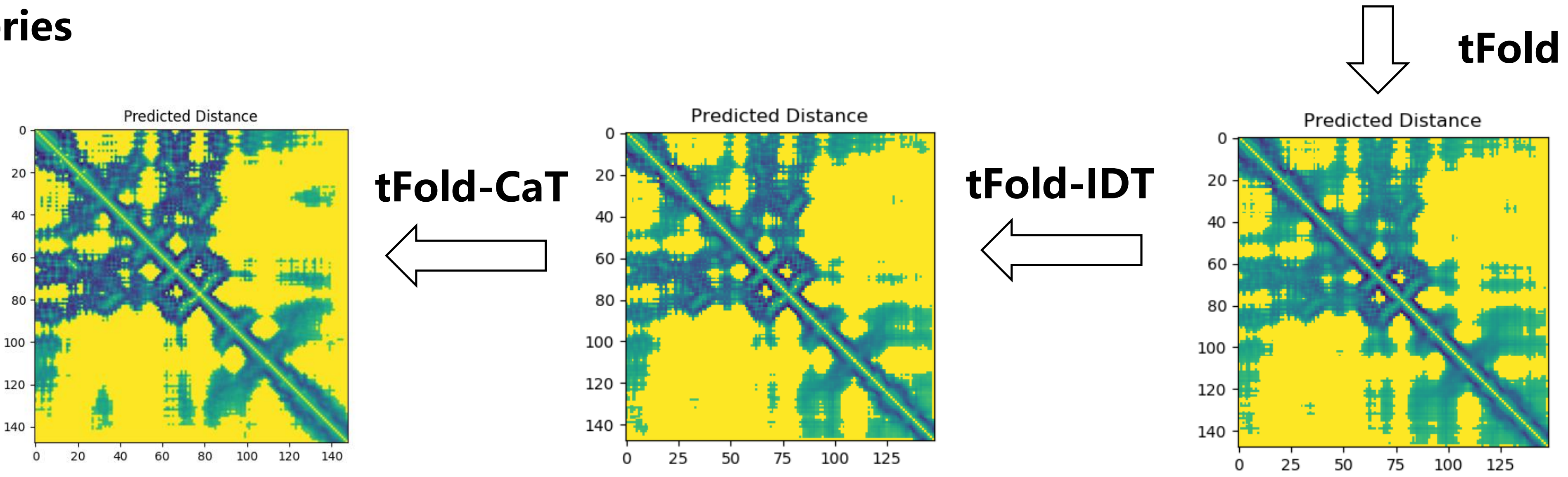
#	Model	No Conts.	F1	Prec	Recall
1	T1043RR368_1-D1	148	46.980	47.300	46.670
2	T1043RR009_1-D1	148	46.980	47.300	46.670
3	T1043RR488_1-D1	148	44.970	45.270	44.670
4	T1043RR183_1-D1	148	40.940	41.220	40.670
5	T1043RR351_1-D1	148	40.270	40.540	40.000
6	T1043RR238_1-D1	148	25.500	25.680	25.330
7	T1043RR157_1-D1	148	8.720	8.780	8.670
8	T1043RR326_1-D1	148	8.050	8.110	8.000

tFold series

H1044 whole sequence



T1043 MSA



What goes wrong and why?

We didn't use…

Modern metagenomics databases

BFD

MGnify

JGI

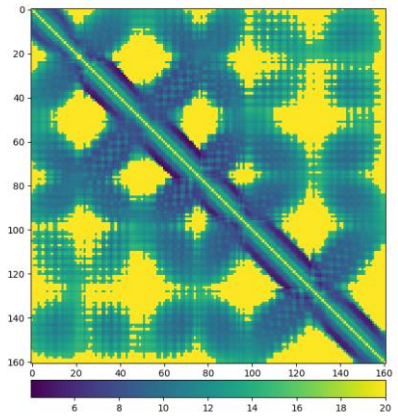
ID	TripleRes	tFold-orig	tFold-BFD	diff
T1027-D1	0.4343	0.3939	0.5152	0.1213
T1029-D1	0.0560	0.0400	0.0320	(0.0080)
T1031-D1	0.3789	0.0632	0.2526	0.1894
T1033-D1	0.1200	0.1400	0.1600	0.0200
T1037-D1	0.5767	0.4455	0.5347	0.0892
T1038-D1	0.2719	0.3070	0.3333	0.0263
T1039-D1	0.4596	0.0994	0.6149	0.5155
T1040-D1	0.4000	0.0769	0.2000	0.1231
T1041-D1	0.6901	0.6322	0.7107	0.0785
T1042-D1	0.4891	0.3225	0.5290	0.2065
T1043-D1	0.0473	0.2568	0.2568	0.0000
T1047s1-D1	0.4834	0.6256	0.6445	0.0189
T1049-D1	0.7388	0.8284	0.8284	0.0000
T1061-D2	0.7232	0.5867	0.7048	0.1181
T1064-D1	0.0652	0.0326	0.0543	0.0217
T1074-D1	0.3864	0.2879	0.5758	0.2879
T1090-D1	0.6138	0.7831	0.7696	(0.0135)
T1093-D1	0.0284	0.1631	0.2553	0.0922
T1093-D3	0.0566	0.6321	0.5566	(0.0755)
T1094-D2	0.6473	0.4541	0.5411	0.0870
T1096-D1	0.5569	0.3412	0.3765	0.0353
T1096-D2	0.4971	0.2164	0.4035	0.1871
Average	0. 3964	0.3513	**0.4477**	0.0964

			No. Domains		F1		Prec	
#	Gr.#	Gr. Name	No Submit.	No Total	Submit.	Total	Submit.	Total
1.	368	tFold-CaT_human	22	22	41.158	41.158	41.783	41.783
2.	488	tFold-IDT_human	22	22	39.374	39.374	40.504	40.504
3.	010	TripletRes	22	22	39.282	39.282	39.641	39.641
4.	125	PreferredFold	22	22	38.696	38.696	39.440	39.440
5.	024	DeepPotential	22	22	38.286	38.286	38.586	38.586
6.	009	tFold_human	22	22	36.821	36.821	38.056	38.056
7.	183	tFold-CaT	22	22	35.465	35.465	37.107	37.107
8.	351	tFold-IDT	22	22	34.774	34.774	36.516	36.516
9.	238	tFold	22	22	33.548	33.548	35.130	35.130

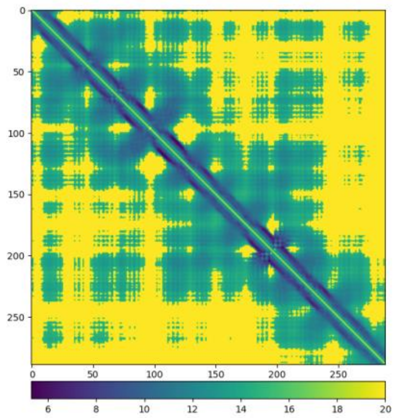
tFold-orig only uses metaclust50 (year 2018) as the metagenomics databases.
tFold-BFD adds BFD (year 2019) as the additional metagenomics databases.

What goes wrong and why?

T1039 (orig)

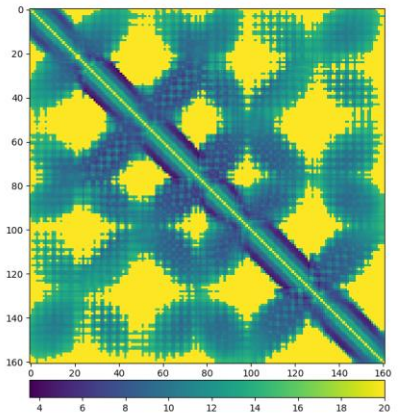


T1042 (orig)

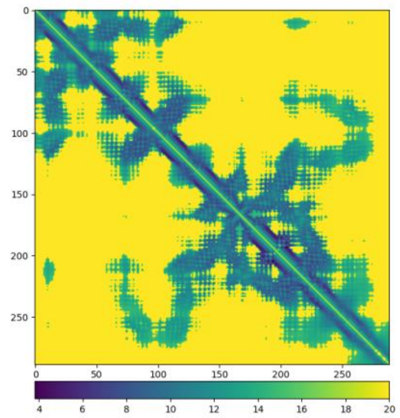


ID	tFold-orig	tFold-BFD	diff
T1027-D1	0.3939	0.5152	0.1213
T1029-D1	0.0400	0.0320	(0.0080)
T1031-D1	0.0632	0.2526	0.1894
T1033-D1	0.1400	0.1600	0.0200
T1037-D1	0.4455	0.5347	0.0892
T1038-D1	0.3070	0.3333	0.0263
T1039-D1	0.0994	0.6149	0.5155
T1040-D1	0.0769	0.2000	0.1231
T1041-D1	0.6322	0.7107	0.0785
T1042-D1	0.3225	0.5290	0.2065
T1043-D1	0.2568	0.2568	0.0000
T1047s1-D1	0.6256	0.6445	0.0189
T1049-D1	0.8284	0.8284	0.0000
T1061-D2	0.5867	0.7048	0.1181
T1064-D1	0.0326	0.0543	0.0217
T1074-D1	0.2879	0.5758	0.2879
T1090-D1	0.7831	0.7696	(0.0135)
T1093-D1	0.1631	0.2553	0.0922
T1093-D3	0.6321	0.5566	(0.0755)
T1094-D2	0.4541	0.5411	0.0870
T1096-D1	0.3412	0.3765	0.0353
T1096-D2	0.2164	0.4035	0.1871
Average	0.3513	0.4477	0.0964

T1039 (BFD)



T1042 (BFD)



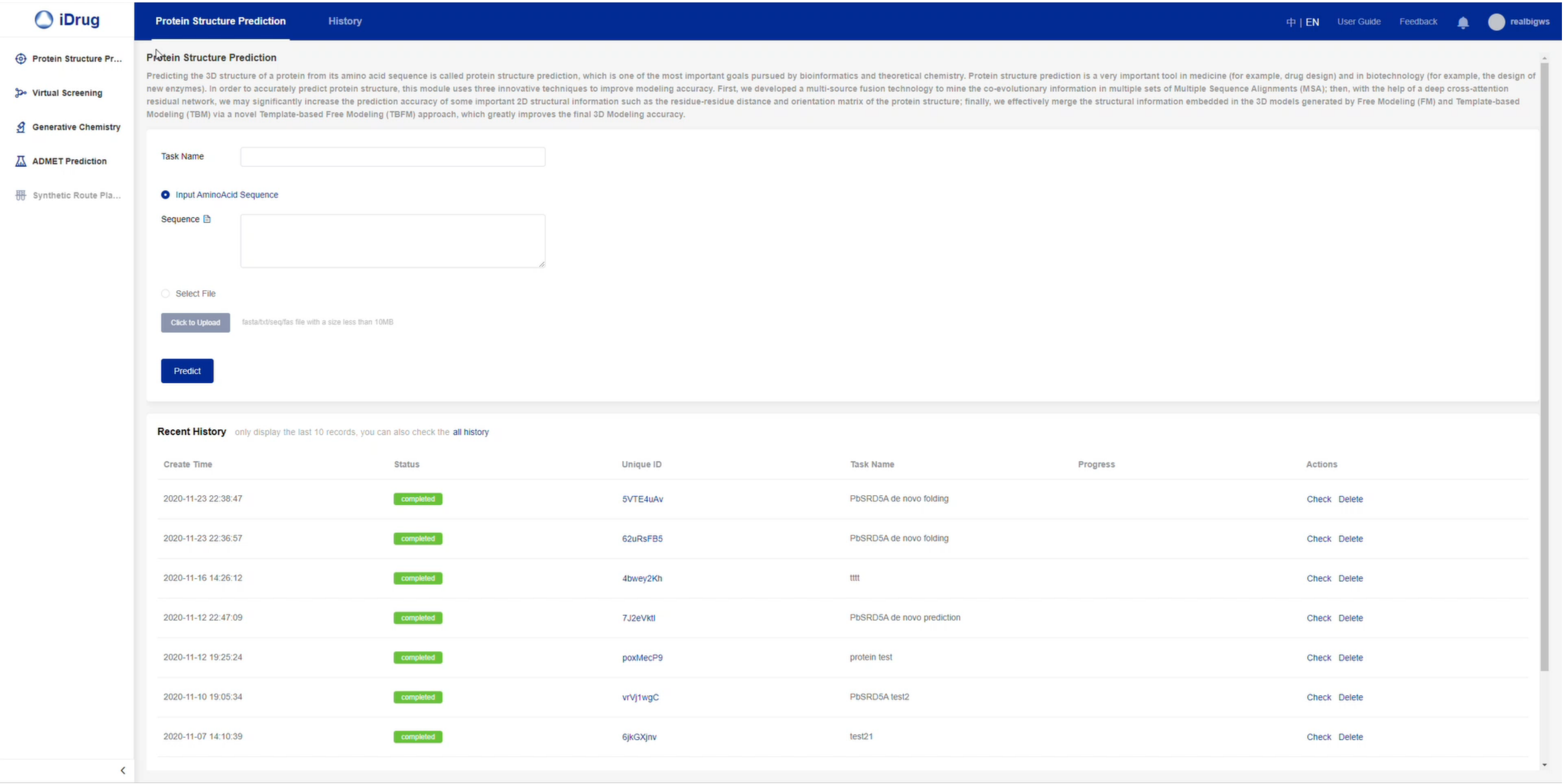
tFold-orig only uses metaclust50 (year 2018) as the metagenomics databases.
tFold-BFD add BFD (year 2019) as the additional metagenomics databases.

Take home messages

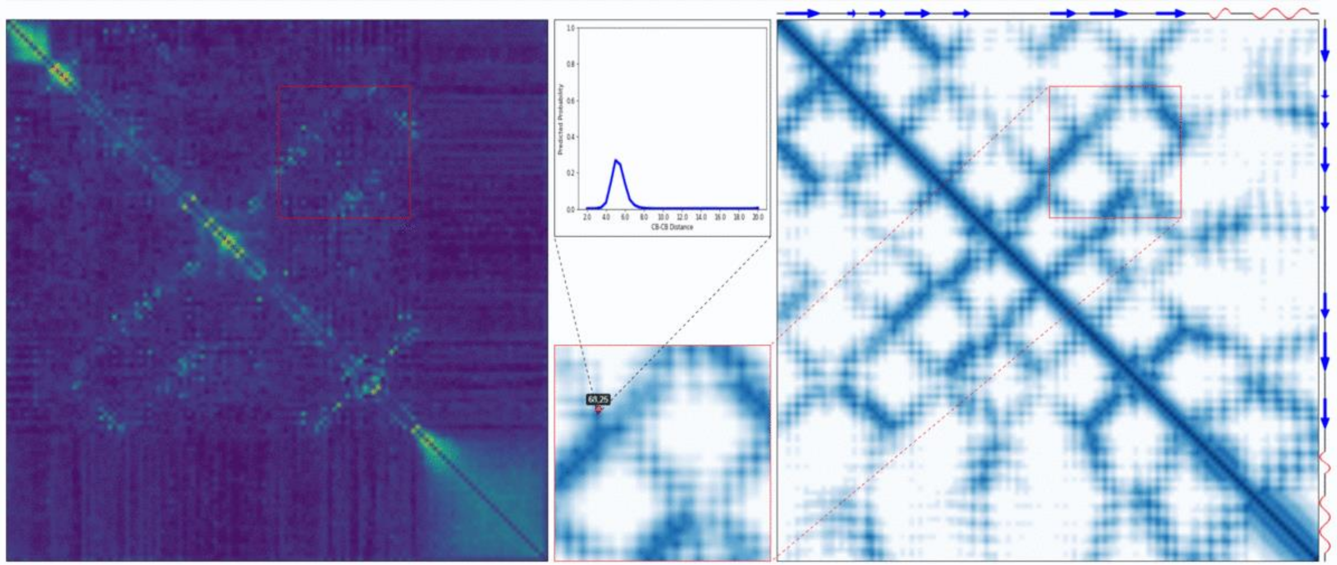
- Distance/orientation matters.
- Deeper and attention-based network works.
- Sufficient data usage will increase the robustness:
  a. More input MSAs
  b. More input decoys

Accurate De Novo Protein Structure Prediction by tFold Server

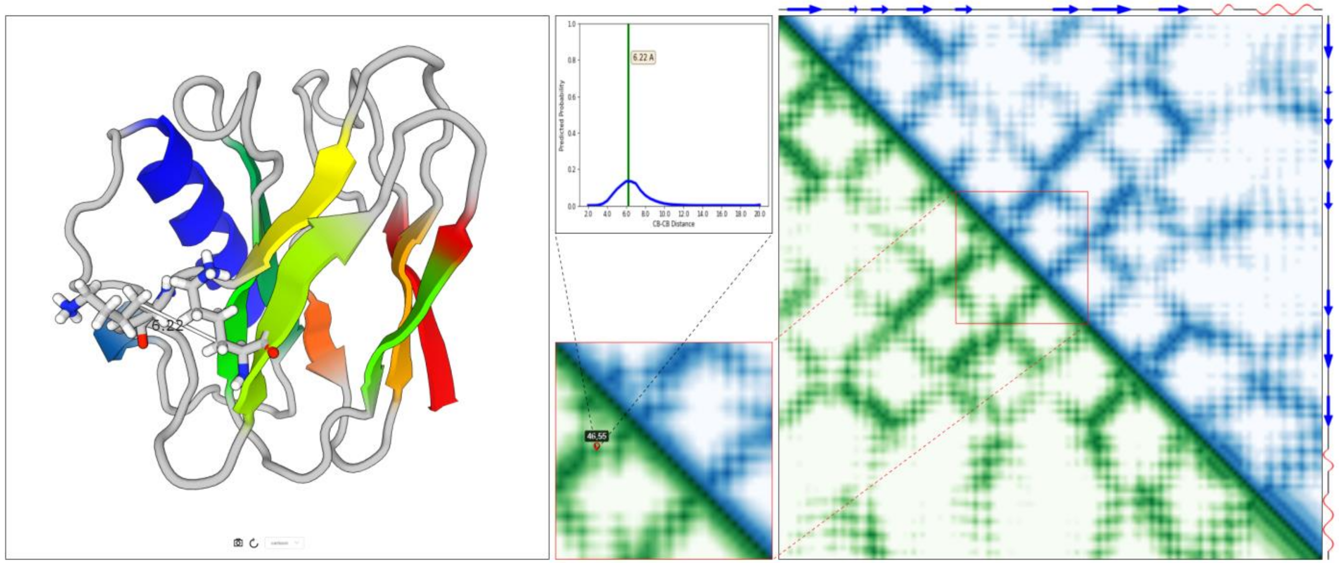
https://drug.ai.tencent.com/console/en/tfold



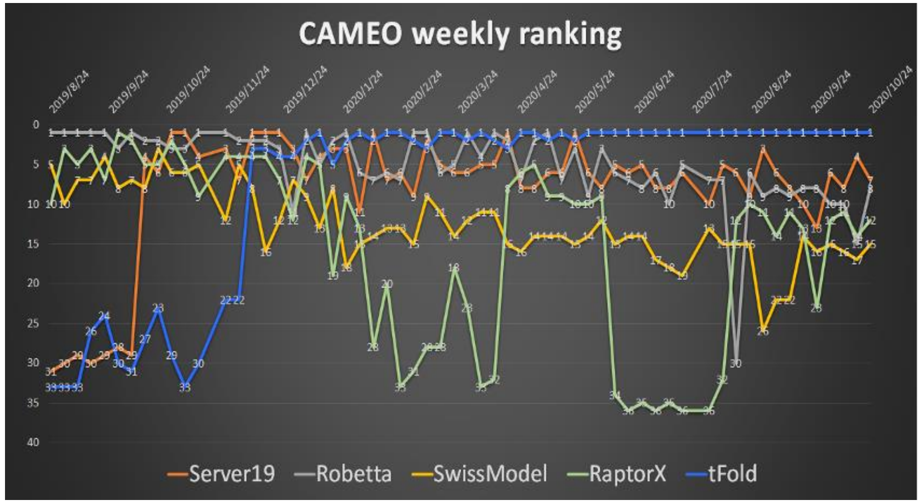
From MRF to Distance



From Distance to 3D model



Excellent performance of tFold Server on CAMEO



Performance on hard targets

1-month - [2020-10-23 - 2020-11-14] - "All targets" dataset

Server Name (Common Subset - Start Comparison)	Avg. response time (hh:mm:ss)	Targets #Submitted	#Modeled	#Submitted Oligo	#Modeled Oligo	Average lDDT All	Modeled
tFold	67:43:39	12	12	25	0	47.1	47.1
BestSingleStructuralTemplate	02:34:29	12	11	25	0	42.8	46.7
Robetta	34:18:36	12	12	25	15	38.2	38.2
IntFOLD5-TS	43:25:43	12	11	25	0	35.5	38.7
IntFOLD3-TS	34:42:34	12	12	25	0	33.4	33.4
SPARKS-X	02:15:58	12	12	25	0	31.1	31.1
RaptorX	16:17:44	12	10	25	0	29.5	35.4
IntFOLD4-TS	54:52:41	12	9	25	0	26.5	35.3
SWISS-MODEL	00:03:59	12	10	25	14	21.2	25.5
PRIMO	00:27:04	12	8	25	0	17.8	26.7

6-months - [2020-05-22 - 2020-11-14] - "All targets" dataset

Server Name (Common Subset - Start Comparison)	Avg. response time (hh:mm:ss)	Targets #Submitted	#Modeled	#Submitted Oligo	#Modeled Oligo	Average lDDT All	Modeled	Avera CAD- All
tFold	63:18:31	84	84	133	0	57.3	57.3	61.7
BestSingleStructuralTemplate	03:49:03	84	78	133	0	47.1	50.7	51.5
Robetta	31:26:44	84	82	133	79	45.7	46.9	53.9
IntFOLD5-TS	44:15:28	84	79	133	0	40.4	42.9	48.2
IntFOLD3-TS	32:44:48	84	84	133	0	38.5	38.5	48.8
SWISS-MODEL	00:04:28	84	79	133	80	29.0	30.9	33.6
IntFOLD4-TS	62:37:38	84	57	133	0	28.4	41.9	33.9
RaptorX	18:24:48	84	54	133	0	27.0	42.0	33.6
SPARKS-X	05:03:24	84	69	133	0	26.9	32.7	37.8
PRIMO_HHS_CL	00:49:04	84	73	133	0	21.5	24.7	28.2

3-months - [2020-08-21 - 2020-11-14] - "All targets" dataset

Server Name (Common Subset - Start Comparison)	Avg. response time (hh:mm:ss)	Targets #Submitted	#Modeled	#Submitted Oligo	#Modeled Oligo	Average lDDT All	Modeled
tFold	59:16:17	42	42	76	0	55.5	55.5
BestSingleStructuralTemplate	03:10:03	42	38	76	0	46.9	51.8
Robetta	29:58:37	42	42	76	47	46.6	46.6
IntFOLD5-TS	39:26:14	42	41	76	0	41.4	42.4
IntFOLD3-TS	32:26:40	42	42	76	0	37.8	37.8
RaptorX	14:58:13	42	38	76	0	37.3	41.2
IntFOLD4-TS	56:41:27	42	33	76	0	33.0	41.9
SWISS-MODEL	00:04:43	42	39	76	45	27.1	29.2
SPARKS-X	05:40:16	42	27	76	0	21.4	33.4
PRIMO_HHS_CL	00:51:16	42	35	76	0	21.3	25.5


Junhong Huang


Tao Shen


Junzhou Huang


Wei Liu


Jiaxiang Wu


Jianguo Pei


Ningqiao Huang


Haidong Lan

Thank you


Liangzhen Zheng


Yuyi Liu


Zhenlei Xu

Tencent AI Lab


Sheng Wang