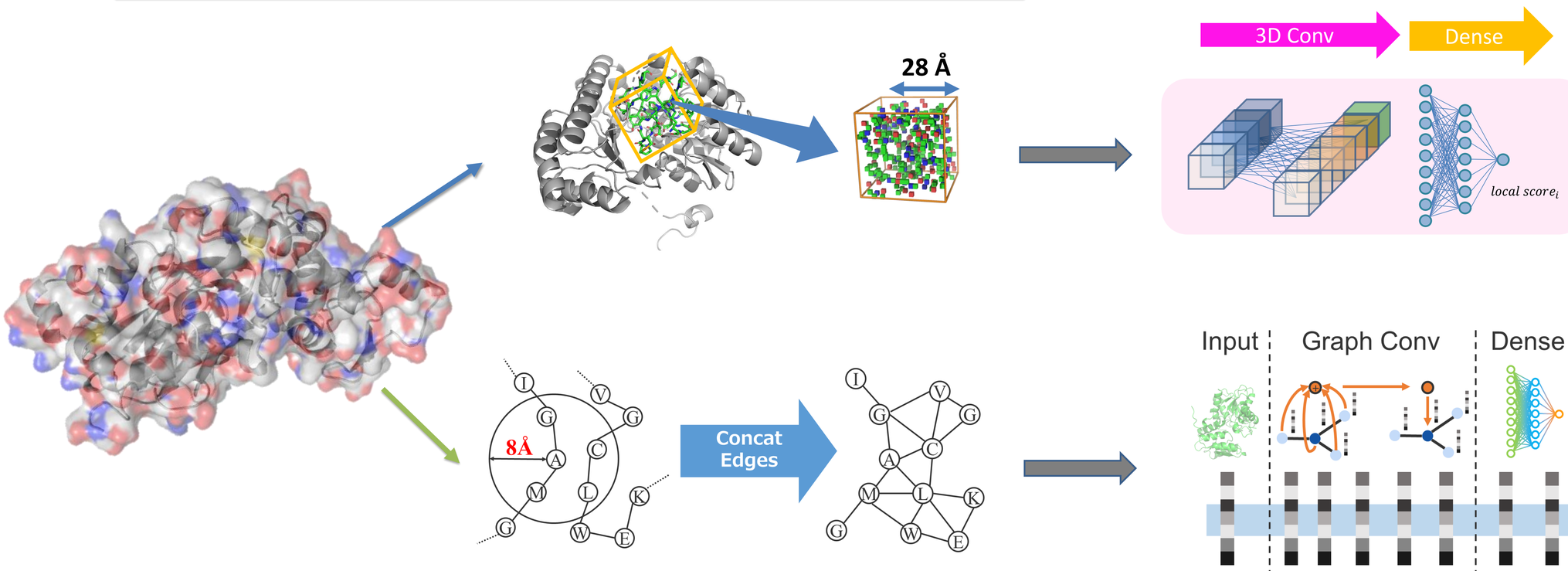
深層学習による生命科学3次元データの解析

情報理工学院 情報工学系 石田研究室

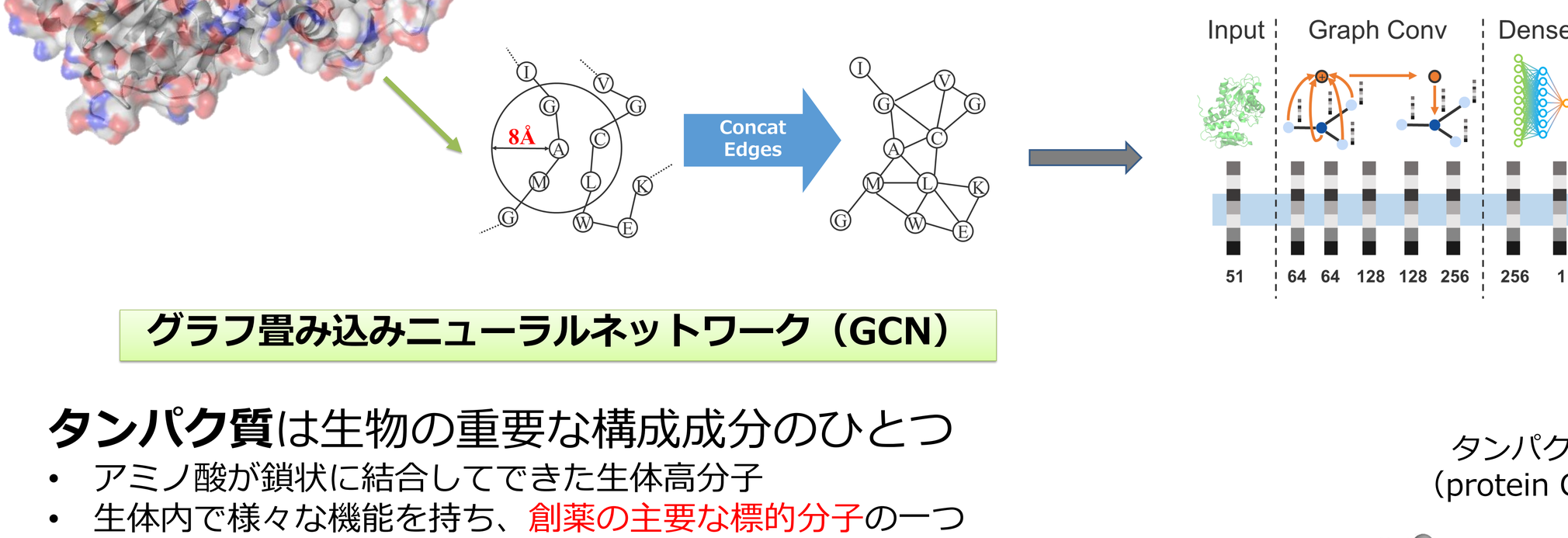
Tokyo Tech

複雑な階層構造を持つ生体高分子であるタンパク質の立体構造情報をディープニューラルネットワーク上で効率的に解析

3次元畳み込みニューラルネットワーク（3D-CNN）

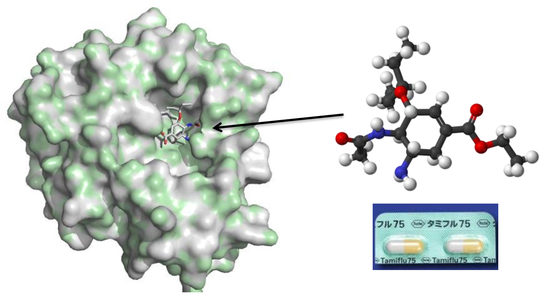


グラフ畳み込みニューラルネットワーク（GCN）



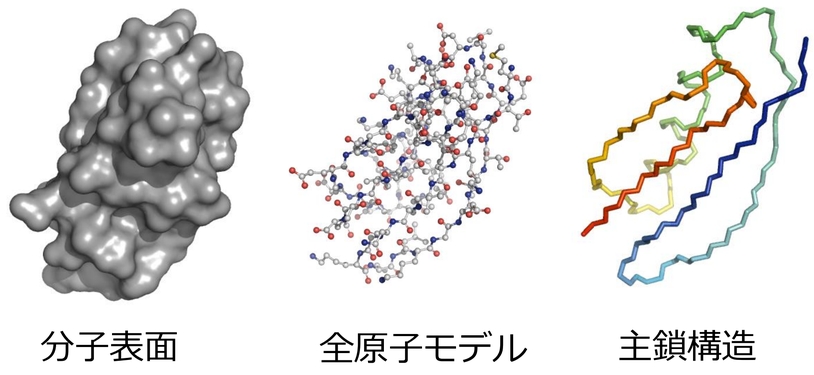
タンパク質は生物の重要な構成成分のひとつ

- アミノ酸が鎖状に結合してできた生体高分子
- 生体内で様々な機能を持ち、創薬の主要な標的分子の一つ
- その立体（3次元）構造は機能を理解する上で重要な情報



インフルエンザ薬タミフルは標的タンパク質であるインフルエンザウイルスのneuraminidaseの3次元構造を利用して最適化することで開発
[Kim, *et al.*, 1997]

タンパク質立体構造の例
(protein G, PDB ID: 1pgp)

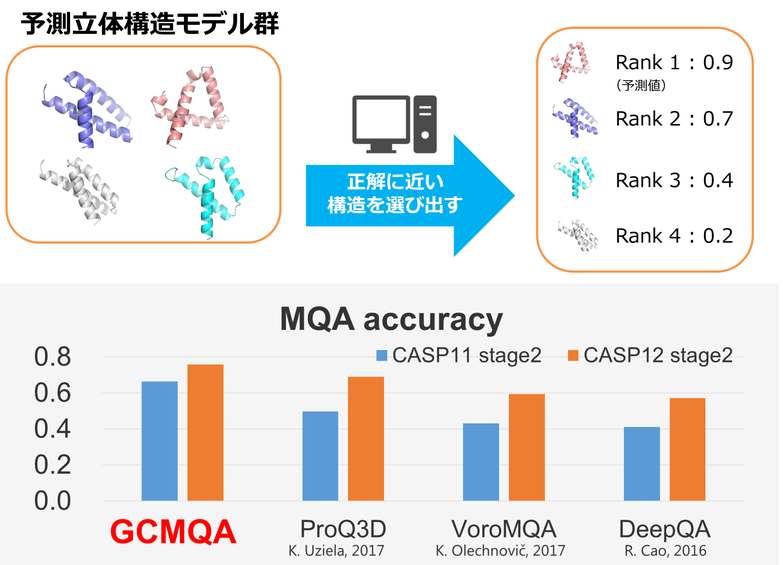


分子表面　全原子モデル　主鎖構造

応用例

タンパク質構造モデルの評価

- タンパク質立体構造予測ではどんな対象でも最良の結果を出力する手法が存在しないため、予測モデルの評価が重要
- GCNを用いて構造モデルを評価することで世界最高の予測精度を達成

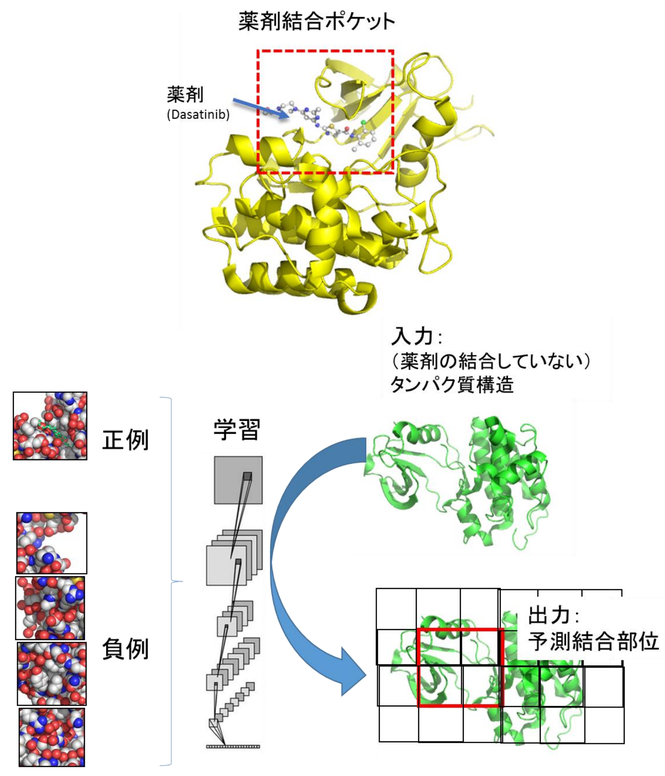


最先端の他手法に比べ予測精度が大きく向上

1. Sato and Ishida, PLOS ONE, 2019
2. Sato and Ishida, Bioinformatics, in revision

薬剤化合物結合部位の予測

- タンパク質の薬剤が結合可能なポケット部位を3D-CNNにより予測



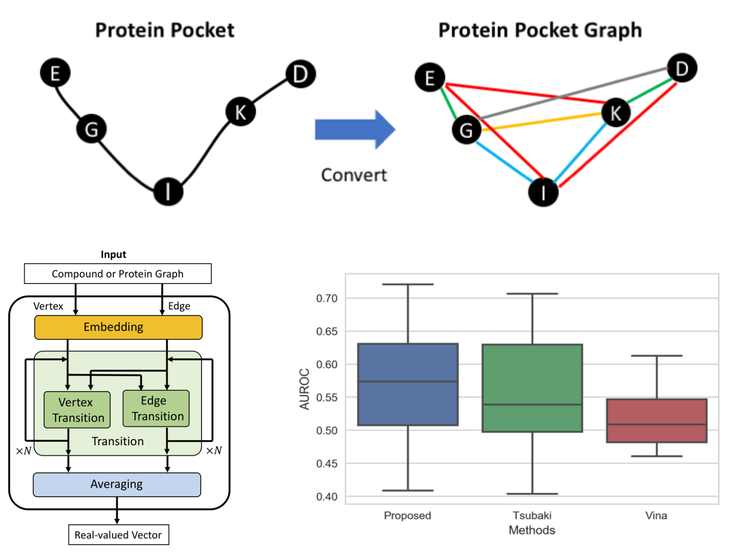
3次元畳み込みニューラルネットワーク

1. Kagami and Ishida, IPSJ SIGBIO, 2016

タンパク質構造情報を用いた化合物結合の予測

ある化合物が標的タンパク質に結合するかどうかをGCNにより予測

- タンパク質の結合ポケット構造をグラフとして扱うことで予測精度を向上



アミノ酸配列によってタンパク質を特徴化した手法に比べ、予測精度が向上

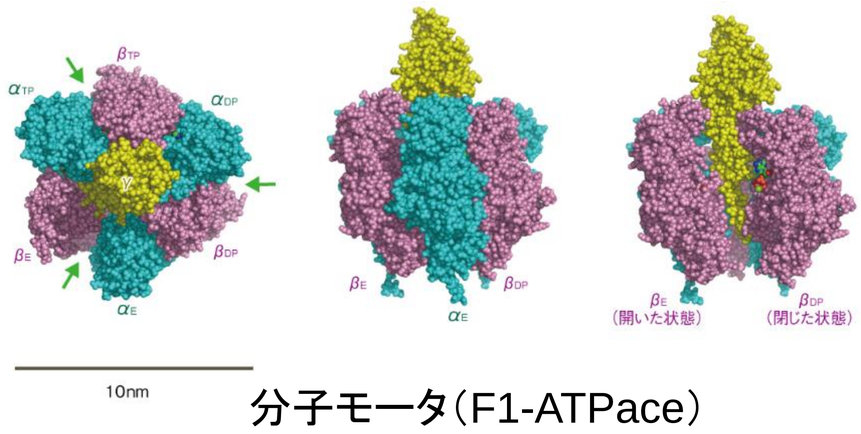
1. Tanebe and Ishida, ICIC2019, 2019
2. Tanebe and Ishida, BMC Bioinformatics, in revision

タンパク質構造予測技術の開発

情報理工学院 情報工学系 石田研究室

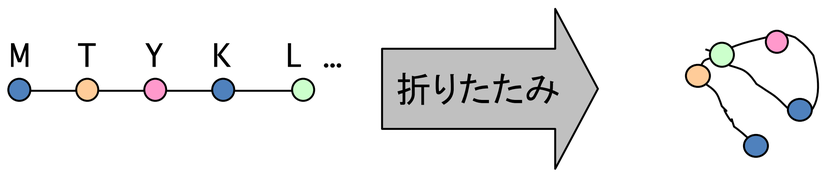
Tokyo Tech

- タンパク質構造は生命現象の理解に重要



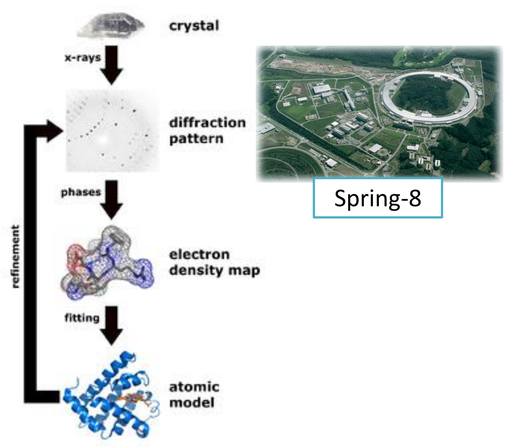
分子モータ(F1-ATPace)

- タンパク質の折りたたみ(Folding)
  - タンパク質は20種類のアミノ酸からなる鎖状の分子
  - タンパク質は一定の構造に自発的に折りたたむ事で機能

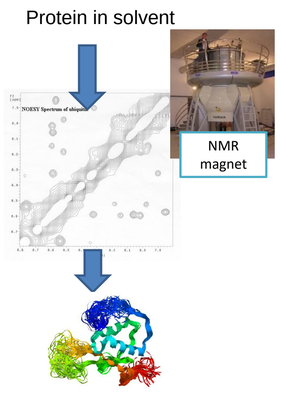


- タンパク質構造の決定は高コスト

X線結晶構造解析



核磁気共鳴法(NMR)



知識ベースの方法

- ホモロジーモデリング

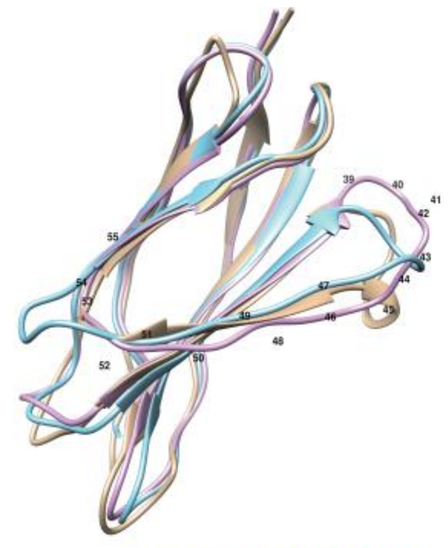
  アミノ酸配列の似たタンパク質は似た構造を持つ

  既知構造を使って未知のタンパク質構造をモデリング

Hemoglobin (human)

Hemoglobin (chicken)

```
Human   : VLSPADKTNVKAAWGKVGAHAGEYGAEALERMFLSFPTTKTYFPHFDLSHGSAQVKGHGKKV・・・・・
          +L+  DK  ++ AW K  +H  E+GAEAL RMF ++P TKTYFPHFDLS GS QV+GHGKKV
Chicken: MLTAEDKKLIQQAWEKAASHQEEFGAEALTRMFTTYPQTKTYFPHFDLSPGSDQVRGHGKKV・・・・・
```

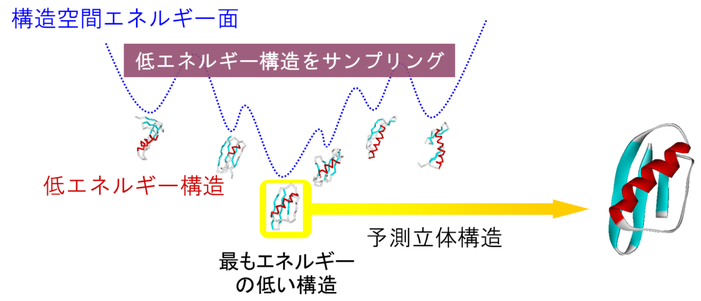



```
Template  ISTEEAAPDGPPMDVTLQPVTSQSIQVTWK 30
HHsearch  -------DLGAPQNPNAKAAGSRKIHFNWL 23
Structural-------DL-GAPQNPNAKAAGSRKIHFNWL 23
Proposed  -------DLGAPQNPNAKAAGSRKIHFNWL 23

Template  APKKELQNGVIRGYQIGYRENSPGSNGQYS 60
HHsearch  PPS-----GKPMGYRVKYWIQGDSESEA--  46
Structural PP-S--G--KPMGYRVKYWIQGD-SESEAH 47
Proposed  PPS-----GKPMGYRVKYWIQGD-SESEAH 47

Template  IVEMKATGDSEVYTLDNLKKFAQYGVVVQA 90
HHsearch  ---HLLDSKVPSVELTNLYPYCDYEMKVCA 73
Structural LLDSKV----PSVELTNLYPYCDYEMKVCA 73
Proposed  LLDSKV----PSVELTNLYPYCDYEMKVCA 73

Template  FNRAGTGPSSSEINATTLE 109
HHsearch  YGAQGEGPYSSLVSCRTHQ 92
Structural YGAQGEGPYSSLVSCRTHQ 92
Proposed  YGAQGEGPYSSLVSCRTHQ 92
```

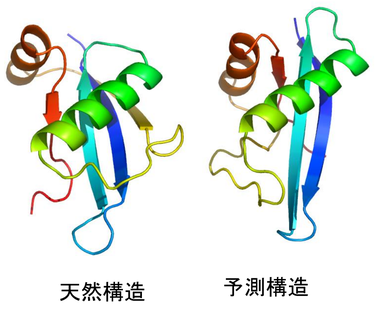
機械学習を用いてホモロジーモデリングの
配列アラインメントを改良 [Makigaki+, 2019]

構造最適化による方法

- Anfinsenの仮説に基づいて、自由エネルギー最小となる構造を探索



- 機械学習を用いて計算量の少なく正確なポテンシャル関数を開発 [Ishida+, 2006]

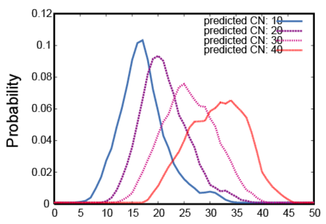


天然構造

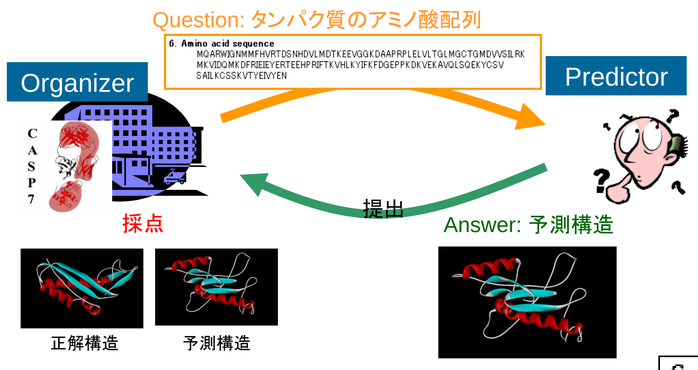
予測構造

GDT_TS: 49.7
RMSD: 7.8 Å

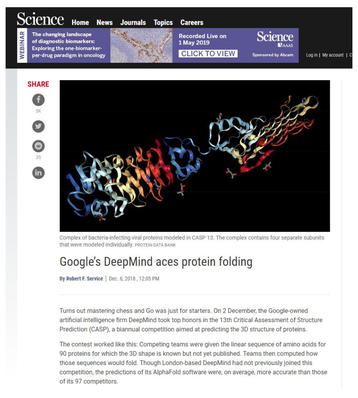
$$E^{CNpred} = -\sum_i \log P(CN_i | CN_i^{pred})$$



CASP – 構造予測の国際コンテスト



CASP13(2018年)にはGoogle(DeepMind)も参加



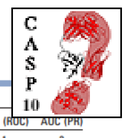


Table II
Performance of the Groups Participating in the DR Prediction Category and the *Naïve* Method

ID	Group name	Targ	TP	FP	TN	FN	prec	Acc	MCC	AUC (ROC)	AUC (PR)	Ranks prec	Acc	MCC	AUC (ROC)	AUC (PR)
369	Prdos-CNF	94	657	287	22401	845	0.696	0.712	**0.529**	**0.907**	0.581	2	18	2	1	2
170	DISOPRED3	94	607	201	22487	895	0.751	0.698	**0.531**	**0.897**	0.603	1	22	1	2	1
478	biomine_dr_mixed	94	628	368	22320	874	0.631	0.701	**0.488**	**0.890**	0.526	4	20	3	3	3
288	biomine_dr_pdb_c	94	579	290	22398	923	0.666	0.686	**0.483**	**0.886**	0.526	3	25	4	4	3
340	metaprdos2	88	918	2228	18603	467	0.292	0.778	**0.385**	**0.879**	0.496	15	2	10	5	7
216	POODLE	94	980	2064	20624	522	0.322	0.781	**0.409**	**0.875**	0.416	12	1	6	6	15
222	MULTICOM-construct	94	940	1972	20716	562	0.323	0.769	**0.400**	**0.873**	0.502	11	4	7	7	5
180	Yang test	94	828	1702	20986	674	0.327	0.738	**0.376**	**0.872**	0.483	10	10	11	8	8
413	ZHOU-SPARKS-X	94	994	3065	19623	508	0.245	0.763	**0.340**	**0.870**	0.475	21	5	18	9	10
129	CASPITAv2	93	863	1610	20766	639	0.349	0.7								
424	MULTICOM-novel	94	944	2630	20058	558	0.264	0.7								
380	Espritz	94	916	2938	19750	586	0.238	0.740	**0.317**	**0.855**	0.465	23	9	20	12	11
327	Espritzv2	94	780	1639	21049	722	0.322	0.724	**0.360**	**0.852**	0.460	13	16	13	13	12
125	MULTICOM-refine	94	1029	3059	19629	473	0.252	0.775	**0.354**	**0.846**	0.459	20	3	14	14	13
003																
084																

[Monastyrskyy, *et al.*, 2014]

CASP10天然変性タンパク質領域予測部門で世界1位を達成