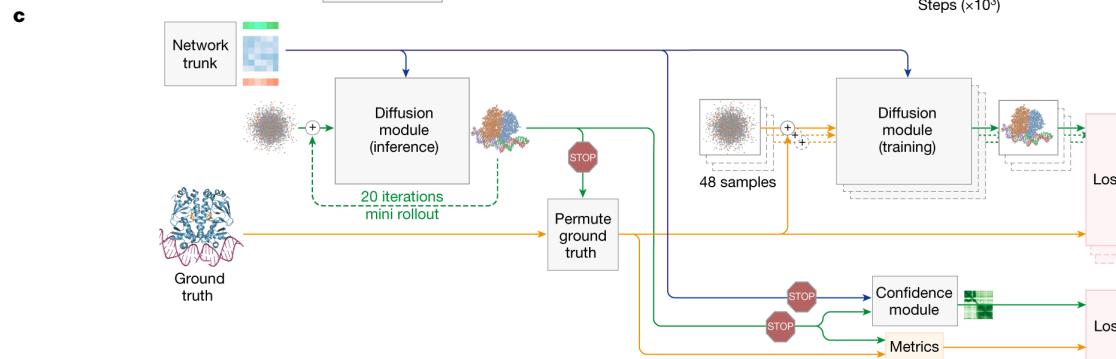
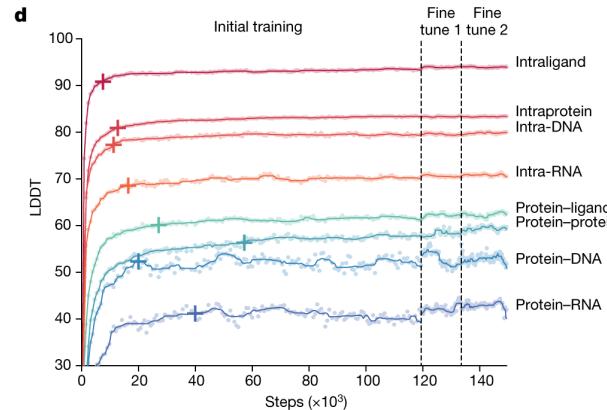
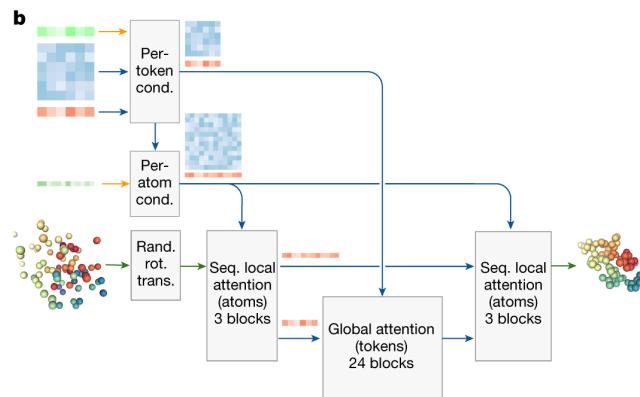
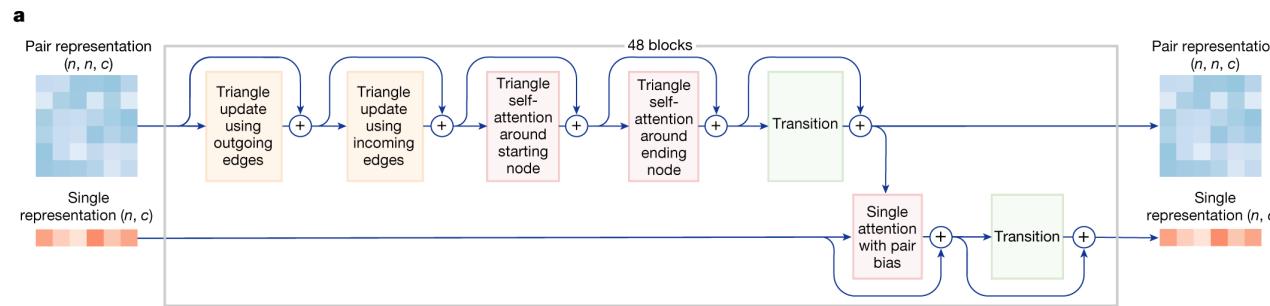


AlphaFold3による生体分子相互作用の構造予測

Accurate structure prediction of biomolecular interactions with AlphaFold 3



根本班のテーマ: 今年のノーベル賞

物理学賞

“for foundational discoveries and inventions that enable machine learning with artificial neural networks”

人工ニューラルネットワークを使った
機械学習の基盤となる発見と発明



John Hopfield (91)
Princeton University
1/2



Geoffrey Hinton (77)
University of Toronto
1/2

化学賞

“for computational protein design”
“for protein structure prediction”

コンピュータによるタンパク質の設計/
タンパク質の構造予測



David Baker (62)
University of Washington
1/2



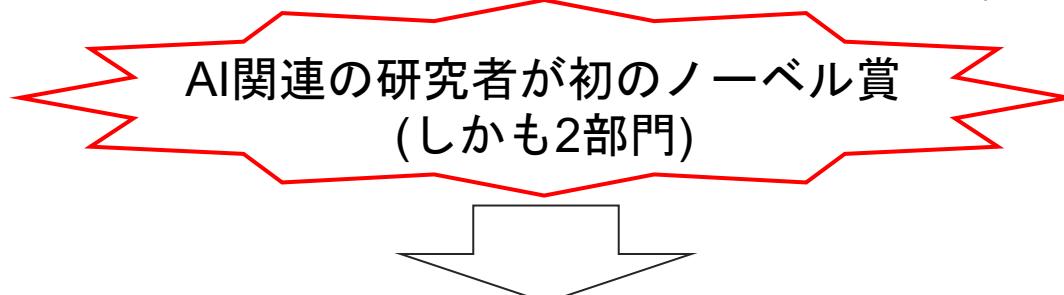
Demis Hassabis (48)
Google DeepMind
1/4



John Jumper (39)
Google DeepMind
1/4

根本班のテーマ

ノーベル物理学賞・化学賞受賞者の最新研究



論文の概要

背景

タンパク質単体についてはアミノ酸配列から高い精度で立体構造予測が可能となったが、多くのタンパク質は生体内では核酸・低分子などと複合体を形成している

目的

タンパク質と核酸・低分子等の複合体の立体構造を予測する

方法

複合体構造予測モデルAlphaFold3を開発した。
AlphaFold2をベースとし、拡散モデルによる生成に変更

結果

タンパク質構造の予測精度が向上した。
既存モデルを上回る複合体構造予測性能を示した。

著者紹介

Article | [Open access](#) | Published: 08 May 2024

Accurate structure prediction of biomolecular interactions with AlphaFold3

Josh Abramson, Jonas Adler, Jack Dunger, Richard Evans, Tim Green, Alexander Pritzel, Olaf Ronneberger, Lindsay Willmore, Andrew J. Ballard, Joshua Bambrick, Sebastian W. Bodenstein, David A. Evans, Chia-Chun Hung, Michael O'Neill, David Reiman, Kathryn Tunyasuvunakool, Zachary Wu, Akvilė Žemgulytė, Eirini Arvaniti, Charles Beattie, Ottavia Bertolli, Alex Bridgland, Alexey Cherepanov, Miles Congreve, Alexander I. Cowen-Rivers, Andrew Cowie, Michael Figurnov, Fabian B. Fuchs, Hannah Gladman, Rishabh Jain, Yousuf A. Khan, Caroline M. R. Low, Kuba Perlin, Anna Potapenko, Pascal Savy, Sukhdeep Singh, Adrian Stecula, Ashok Thillaisundaram, Catherine Tong, Sergei Yakneen, Ellen D. Zhong, Michał Zieliński, Augustin Žídek, Victor Bapst, Pushmeet Kohli, Max Jaderberg  & Demis Hassabis  & John M. Jumper  [Show fewer authors](#)



Nature 2024.5.8 掲載

Corresponding author (一部)



Demis Hassabis
Google DeepMind

人工知能の開発を行うDeepMind社の創業者

経歴・業績

1976年生まれ(48歳)

~2009 人工知能の開発のため神経科学を研究

(2010) DeepMind社を設立

(2020) タンパク質構造予測モデルAlphaFold2を発表 (Last author)

(2024) ノーベル化学賞受賞 “for protein structure prediction”



John Jumper
Google DeepMind

DeepMind社に所属

経歴・業績

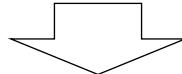
1985年生まれ

(2020) タンパク質構造予測モデルAlphaFold2を発表 (First author)

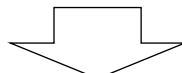
(2024) ノーベル化学賞受賞 “for protein structure prediction”

背景: タンパク質の構造予測

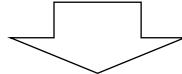
- ✓ 現存するほとんどのタンパク質のうち, ほとんどは立体構造が不明
アミノ酸配列だけが分かっているタンパク質が多い。
- ✓ 1961年 Anfinsenが, 構造を分解したリボヌクレアーゼが自発的に元の構造に戻ることを発見
その後, 他のタンパク質についてもアミノ酸配列から構造が決まることが明らかになる



アミノ酸配列からタンパク質構造を予測する研究が長年行われてきた
未公開の構造を予測する大会CASPが隔年で開かれている。



- ✓ 徐々に精度が向上していたが, 2020年のCASP14で
AlphaFold2が飛躍的に高い精度を達成したことで話題となった。



- ✓ 一方, 生体内のタンパク質の多くは核酸・低分子などと複合体を形成している
これら複合体構造の予測が次の課題となった。

背景: AlphaFold2 (AF2)

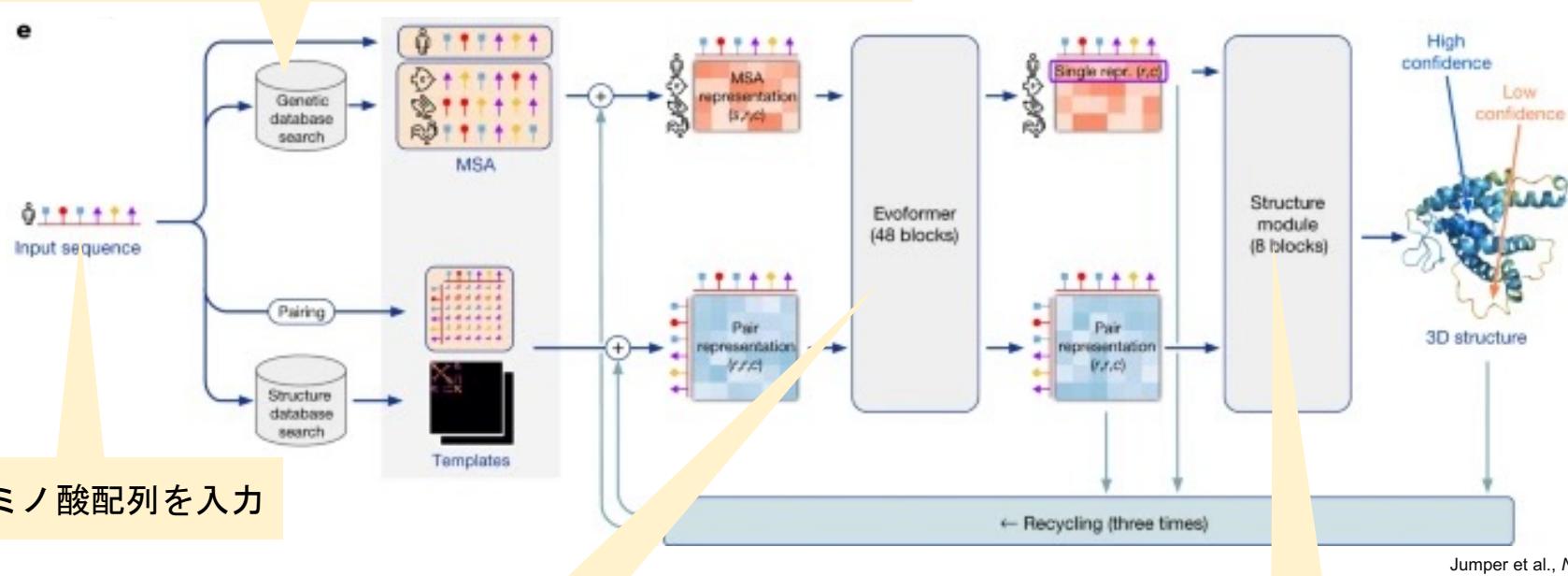
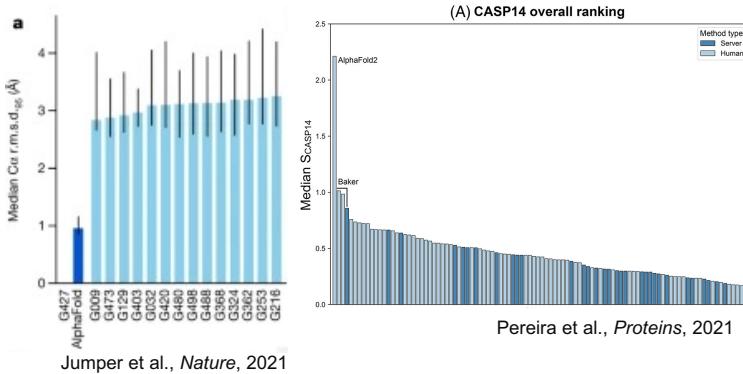
概要

2020年にDeepMindが発表した、タンパク質のアミノ酸配列から立体構造を予測するモデル

それ以前のAlphaFold(2018)とは構造が全く異なる

モデル構造

他生物種のゲノムから類似配列(MSA)を検索して入力に追加
→ 進化的に保存された、構造上重要な領域を考慮



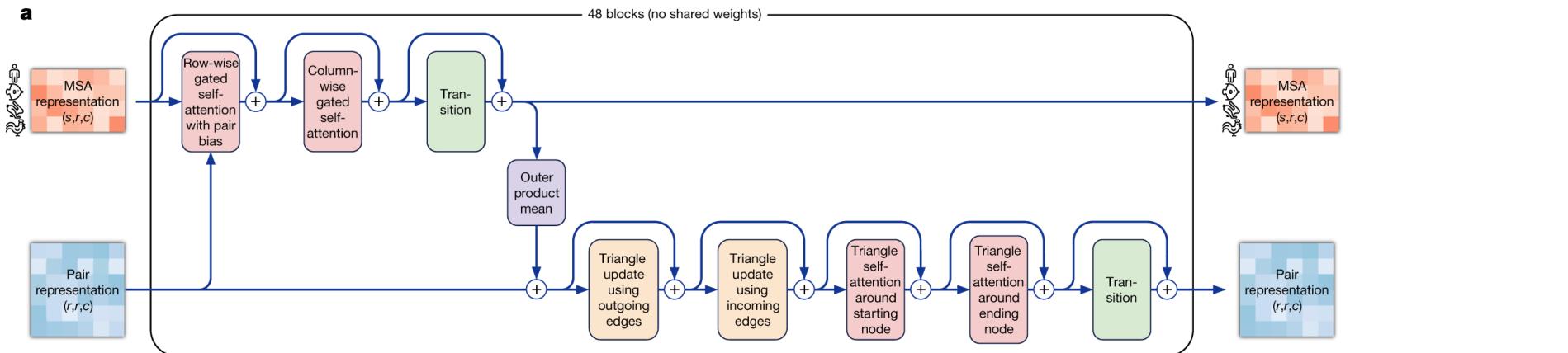
アミノ酸配列を入力

① MSA表現と残基ペア表現を処理する
Evoformer

② Evoformerの表現から残基の座標・姿勢を予測する
Structure module

背景: AF2の詳細

Evoformer



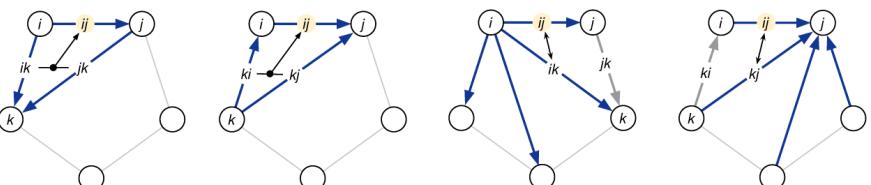
MSA表現: MSA方向と残基数方向のattentionにより更新

Triangle multiplicative update using 'outgoing' edges

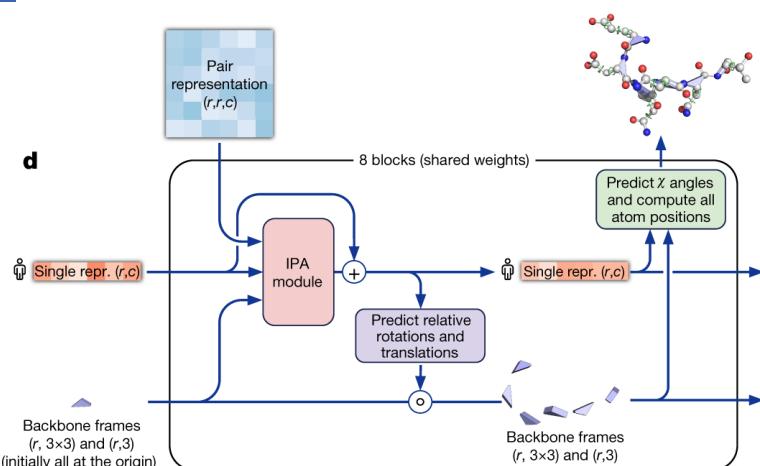
Triangle multiplicative update using 'incoming' edges

Triangle self-attention around starting node

Triangle self-attention around ending node



Structure module



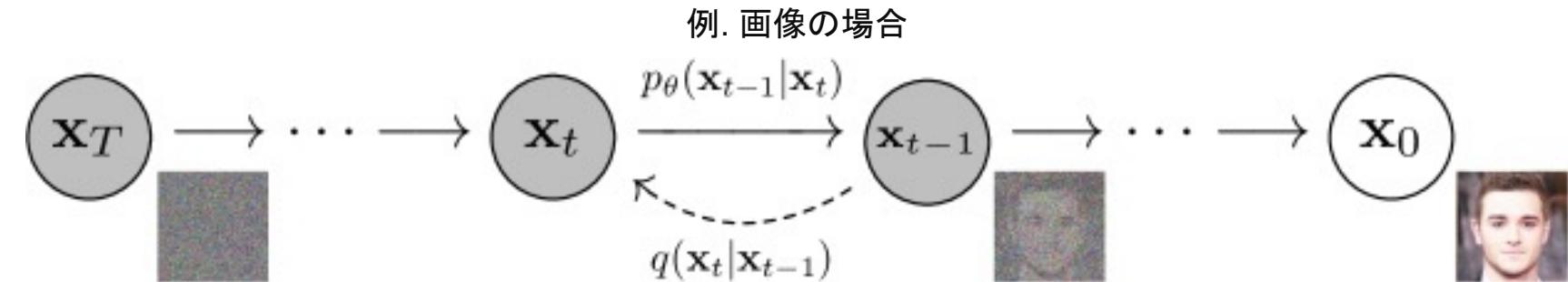
MSA表現の元の配列に対応する部分とペア表現から各残基の位置と姿勢を直接出力

背景: 拡散モデル

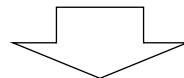
概要

学習データをもとに、それと類似したデータを生成する深層学習フレームワーク

学習方法



1. 学習データについて、徐々にノイズを加えたデータを作成
 2. 各段階についてノイズを除去(denoising)、つまり1つ前の状態を予測するようモデルを学習
 3. ランダムなノイズのみのデータを徐々にノイズ除去することで、データを生成
- ✓ データについての情報をノイズ除去時に与えることで、条件付き生成が可能になる。



AF3では、アミノ酸配列等から抽出した特徴量を条件として与えることで、実質的に予測モデルとして使っている。

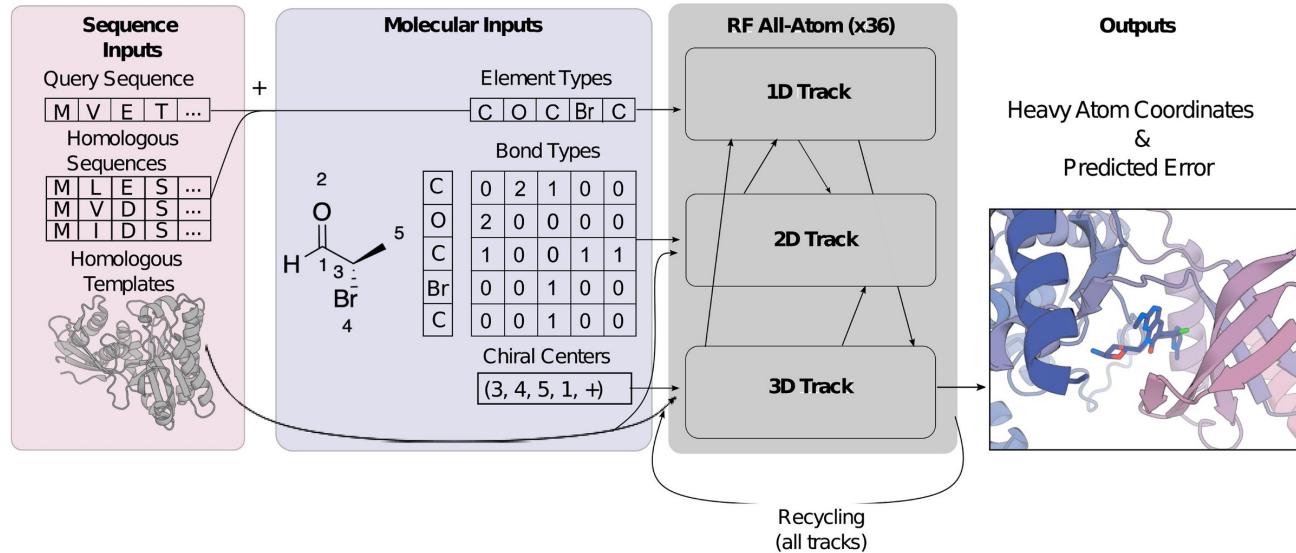
背景: RoseTTAFold All-Atom (RFAA)

概要

DeepMind社と同じくタンパク質構造予測の研究を行っているBakerらのグループが, AF3の1ヶ月前に発表したモデル

Bakerは, 構造予測ではなくそのモデルをタンパク質生成(RFdiffusion)に応用したことでノーベル賞を受賞
AF3と同じく, タンパク質と低分子等との複合体構造の予測を目指している → 本論文の比較対象

モデル構造



- ✓ 低分子の原子情報をMSA表現・ペア表現につなげて入力とし, まとめて処理する

結果

- ✓ タンパク質構造の予測精度を保つつつ, 低分子等との複合体構造予測に成功した。
- ✓ 低分子等に結合するタンパク質の設計も行った。

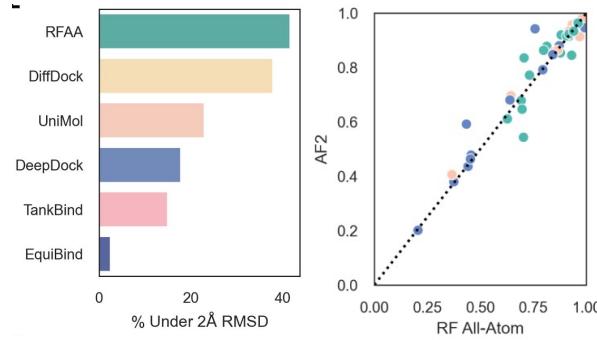


Fig. 1,2 提案するモデル構造と学習手法

Fig. 1,3 様々な複合体に対する予測精度

Fig. 4 予測の信頼性評価

Fig. 5 このモデルの課題

Fig. 1,2 提案するモデル構造と学習手法

Fig. 1,3 様々な複合体に対する予測精度

Fig. 4 予測の信頼性評価

Fig. 5 AF3のモデルの課題

Fig. 1,2 提案するモデルと学習手法

全体構造

- ✓ AF2と同じく、大きなbackboneと予測モジュールからなる。
- ✓ AF2のablation studyを元に、余分な処理を省略し、ペア表現を中心とするモデルにする。

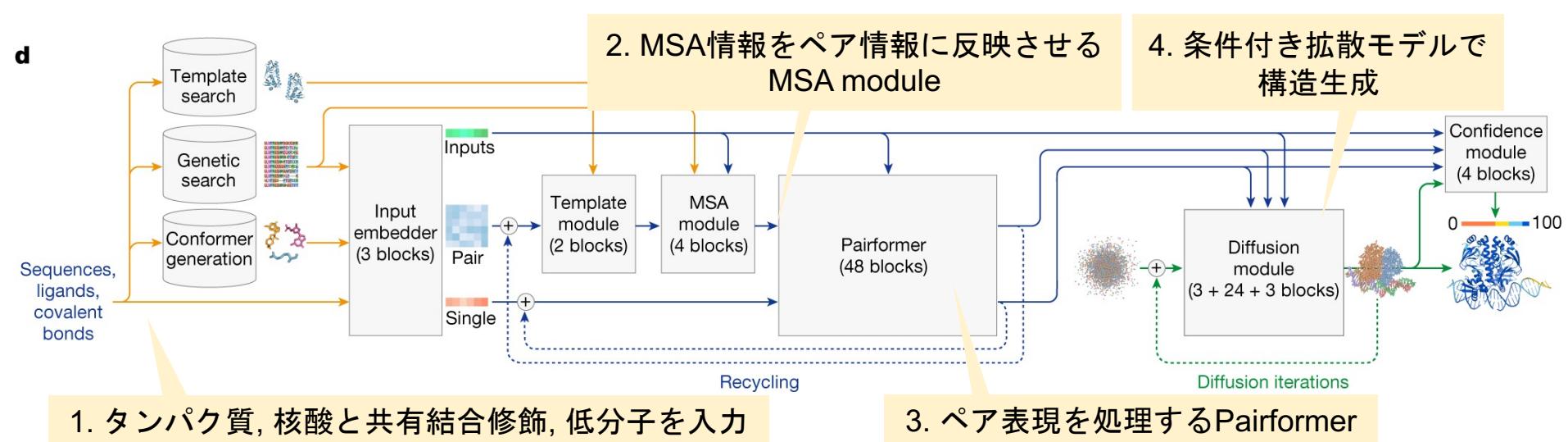
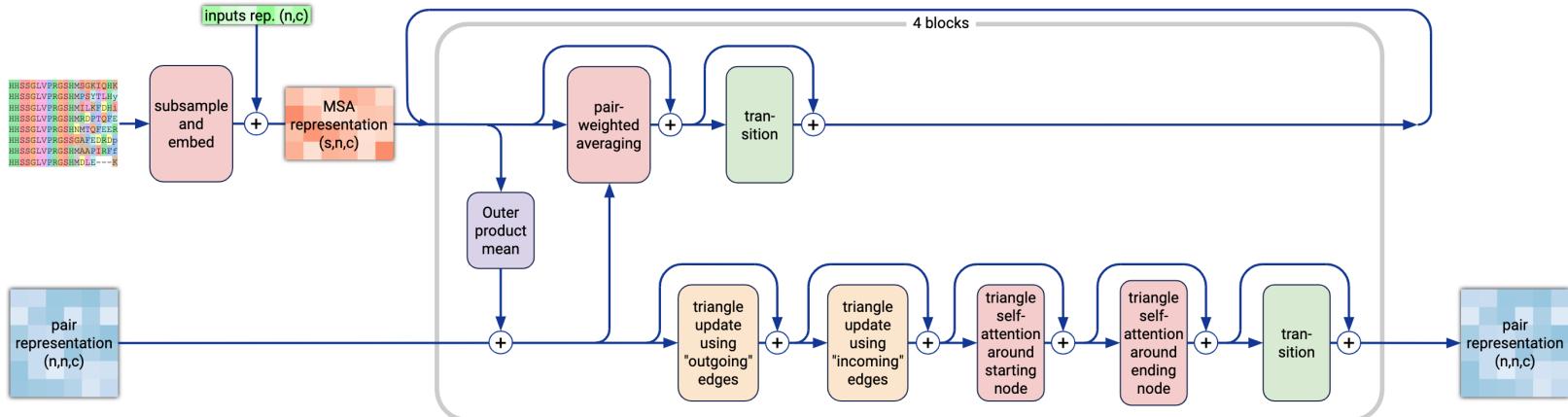


Fig. 1,2 提案するモデルと学習手法

2. MSA module

- ✓ AF2のEvoformerをベースに, MSAの処理を簡略化
 - ✓ MSA間の情報伝達をペア表現のみを使ったattentionに限定することで, ペア表現に情報を集約



参考: AF2のEvoFormer

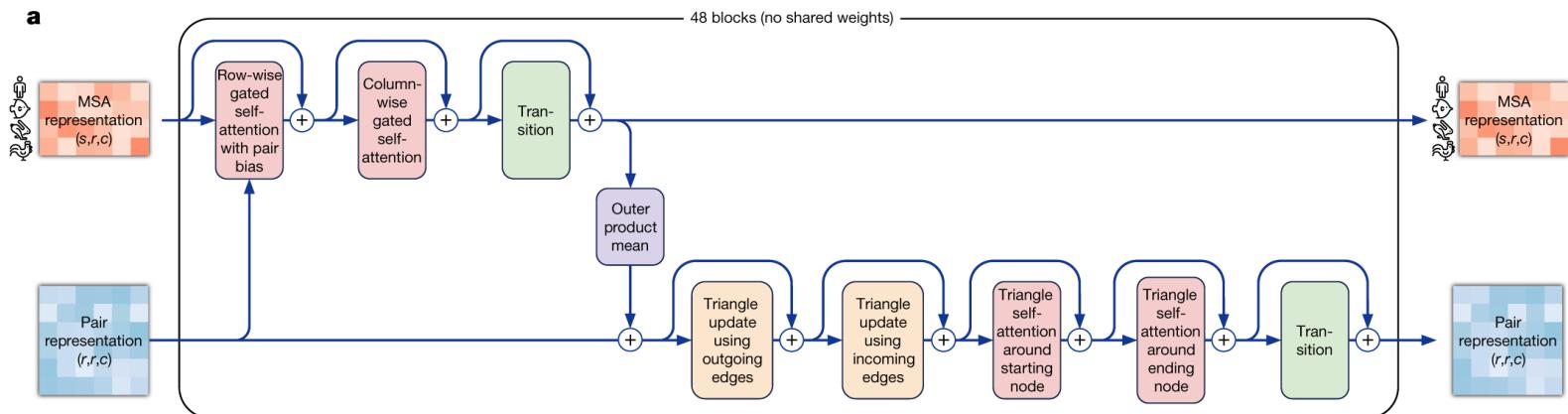
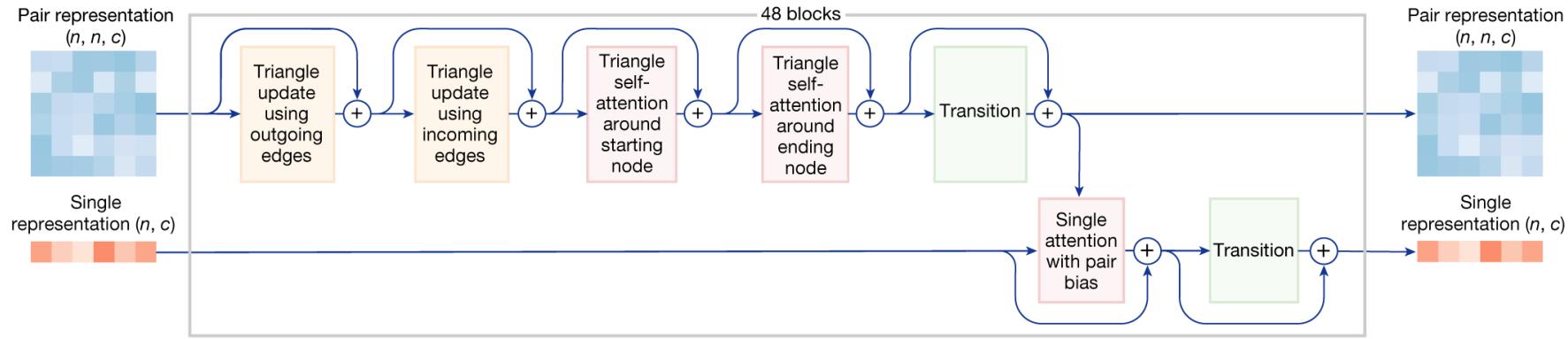


Fig. 1,2 提案するモデルと学習手法

3. Pairformer (メインブロック)

- ✓ Evoformerをベースに、MSA表現を省略し元の配列表現のみを処理する。



<参考: AF2のEvoFormer>

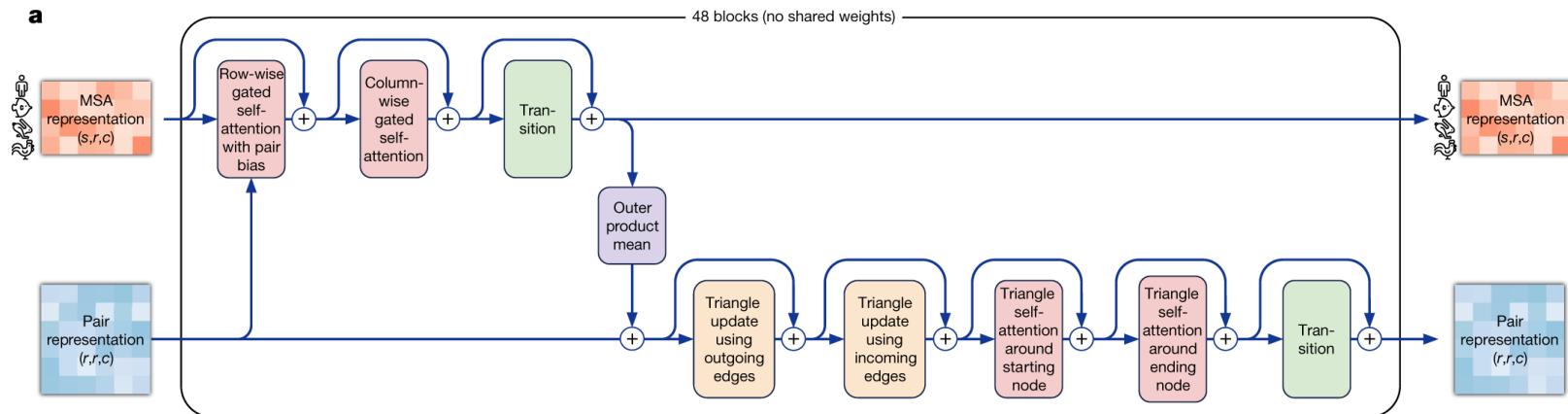
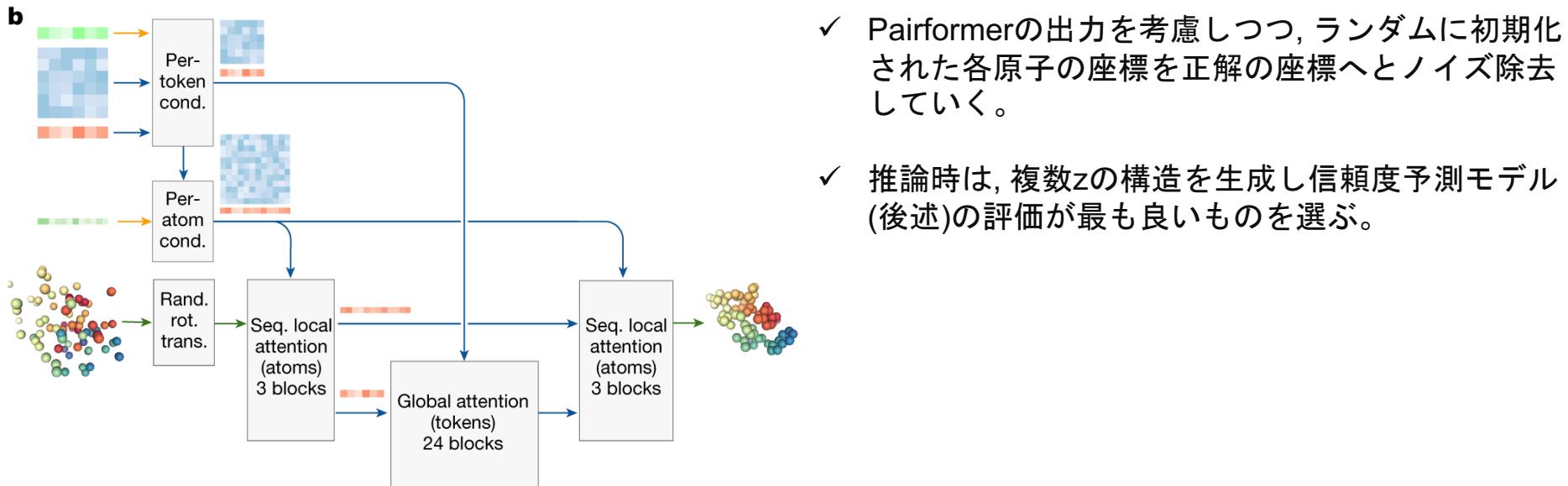


Fig. 1,2 提案するモデルと学習手法

Diffusion module



評価指標

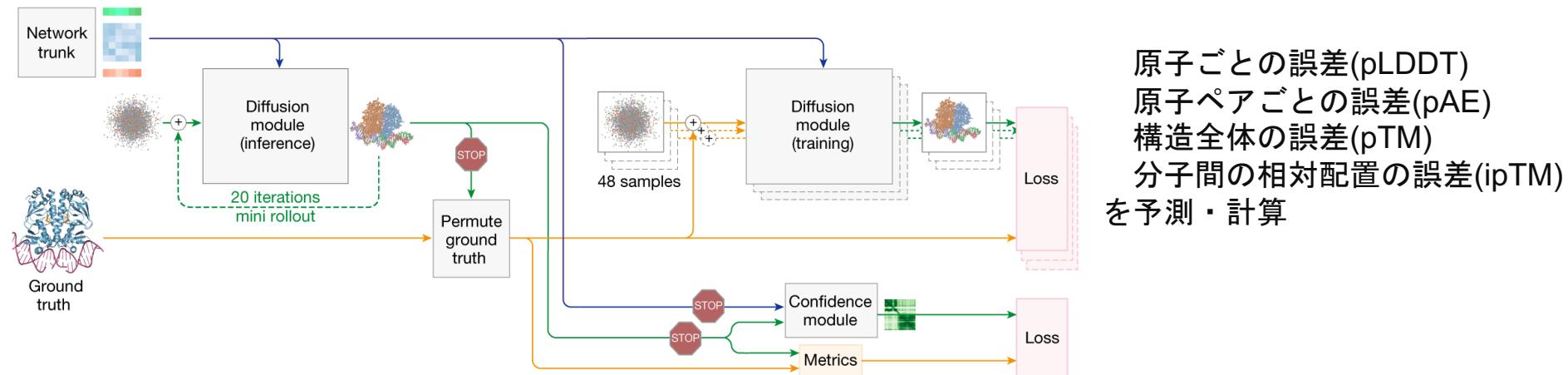
- ✓ 概ね先行研究やコンペなどで使われる指標を用いている。
- ✓ 全て高い方が予測精度が高いことを表す。

対象	名前	説明
タンパク質(単体)	LDDT	一定距離以内の原子ペアのうち、その距離の誤差が一定以下だったものの割合
核酸(単体)		AF2などで使われている。
タンパク質複合体	DockQ	複合体構造予測のコンペ(CAPRI)の指標を定量化したもの。
タンパク質-核酸複合体	iLDDT	LDDTを、異なる鎖間の原子ペアに限定したもの。
タンパク質-低分子複合体	% RMSD < 2 Å	タンパク質部分を揃えたときのリガンドの位置の二乗平均平方根偏差が2 Åだったものの割合。Dockingの評価によく使われる。
タンパク質/核酸の共有結合修飾		

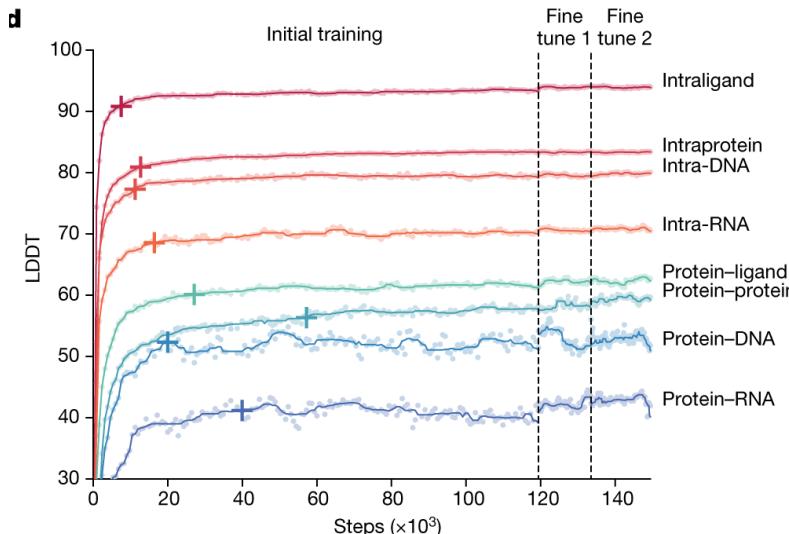
Fig. 1,2 提案するモデルと学習手法

信頼度の予測

- ✓ モデルの予測がどれくらい信頼できるかも予測するため, 学習中に荒いstepで構造を生成し誤差を予測した。



学習方法



- ✓ Protein Data Bank(PDB, 実験的に決定されたタンパク質の構造データベース)のデータをtime-splitして学習・評価に使用
- ✓ 最初の学習が終了後, より長い配列でfine tuning
- ✓ 局所的な構造は早期に学習してoverfittingを始めるが, 相対的な位置の学習には時間がかかる
- ✓ 相互作用面を多くとるようサンプリング比率を修正
複数の評価指標の組み合わせで最も良い時点のモデルを選定

Fig. 1,2 提案するモデル構造と学習手法

Fig. 1,3 様々な複合体に対する予測精度

Fig. 4 予測の信頼性評価

Fig. 5 AF3のモデルの課題

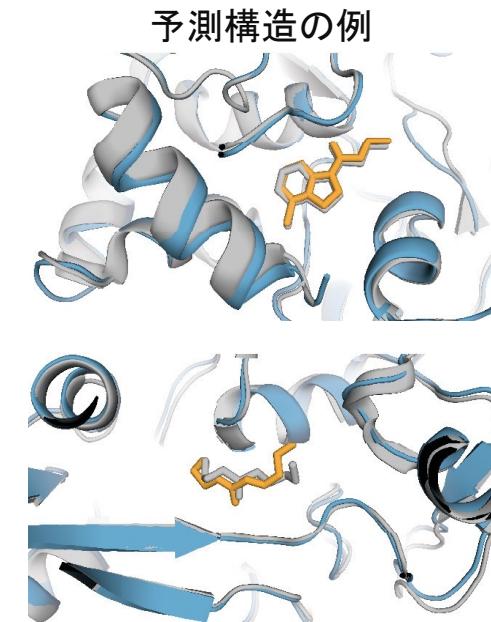
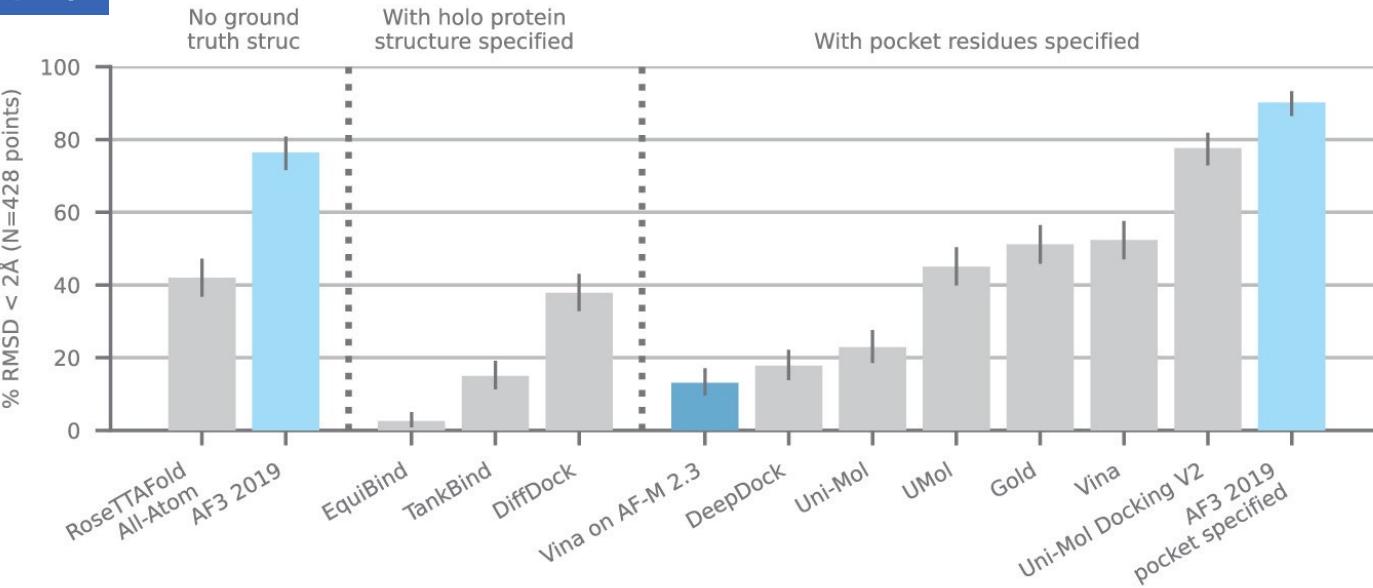
Fig. 1,3 様々な複合体に対する予測精度

タンパク質-低分子複合体

タンパク質 - 低分子複合体の構造予測を、ドッキング手法として既存手法と比較

複合体を生成した後、タンパク質部分を正解のタンパク質に揃えた上でリガンド部分を正解と比較

結果



RFAA(AF3と同じくタンパク質構造を使わないbaseline)を上回る予測精度を示した。

タンパク質構造や、それに加えてpocketのある残基の情報も使う手法もほとんど上回った。

AF3でpocketのある残基情報を使うモデルを学習した場合、全ての手法を上回った。

Fig. 1,3 様々な複合体に対する予測精度

核酸とタンパク質-核酸複合体

✓ 3つのデータセットで評価を行った。

1. PDBテストデータのタンパク質と核酸の複合体

2. CASP15のRNA部門の利用可能なデータ

3. PDBテストデータの核酸のみの構造

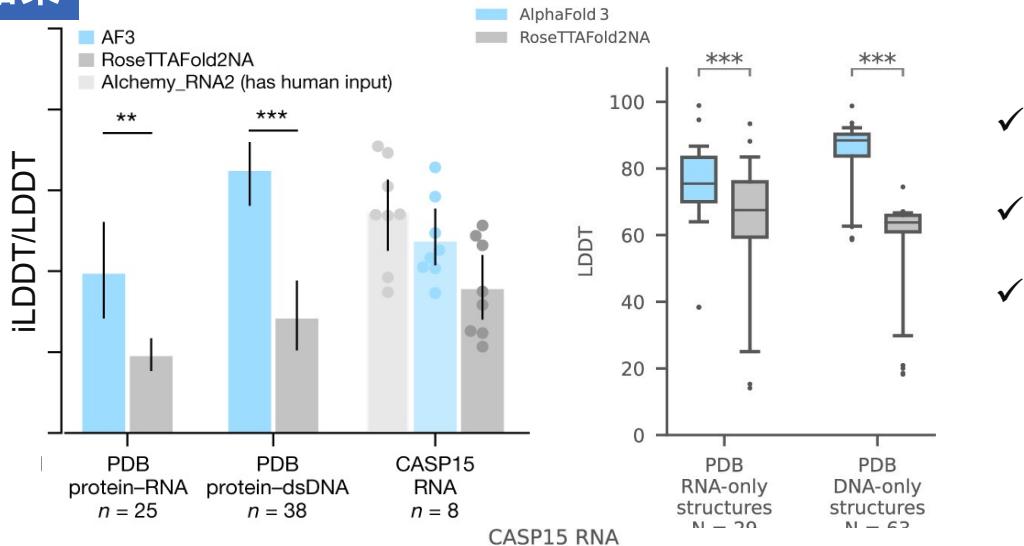
✓ 比較対象のモデル:

1. RoseTTAFold2NA: RFAAの前身で、核酸複合体予測に特化

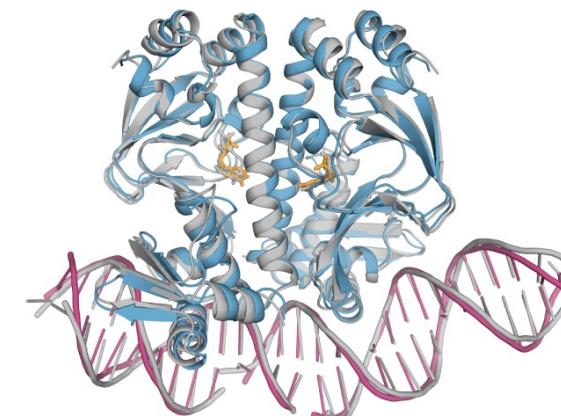
2. Alchemy_RNA: CASP15 RNA部門のAIベースの最善モデル

3. Alchemy_RNA2: CASP15 RNA部門の最善モデル
(構造の決定に人間が関与している)

結果



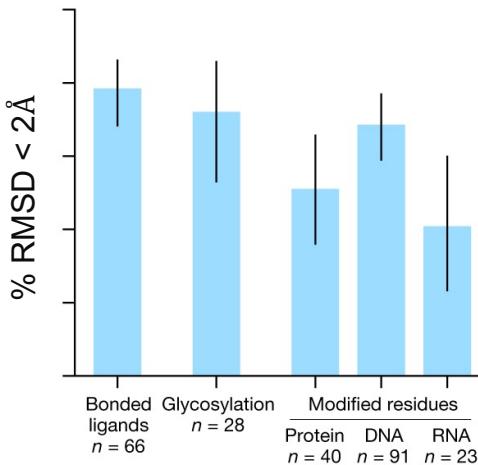
- ✓ 複合体予測ではRoseTTAFold2NAを上回った
- ✓ 単量体予測ではAlchemy_RNA2に次ぐ精度
- ✓ 予測構造の例



バクテリアCtrタンパク
核酸(DNA)と低分子(cGMP)を含む

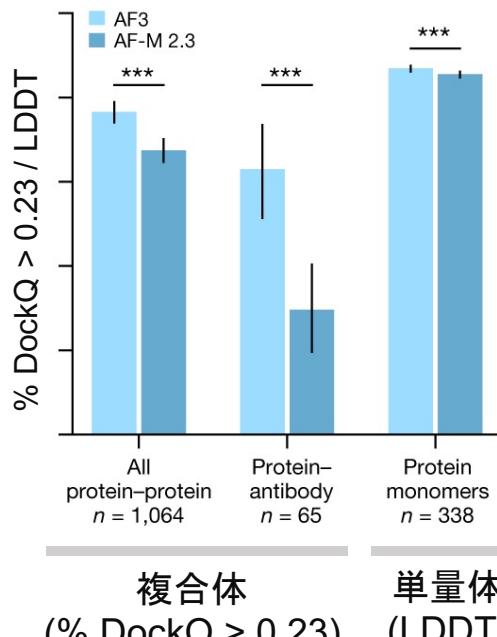
Fig. 1,3 様々な複合体に対する予測精度

共有結合修飾



- ✓ PDBのテストデータのうち、リガンドや糖鎖が共有結合している構造についても正確に予測できた。

タンパク質の単量体と複合体



- ✓ AF2を複合体予測に応用したAlphaFold-Multimerと比較
- ✓ 全体的に精度が向上し、特にタンパク質と抗体の複合体構造について精度が大きく向上した。

Fig. 1,2 提案するモデル構造と学習手法

Fig. 1,3 様々な複合体に対する予測精度

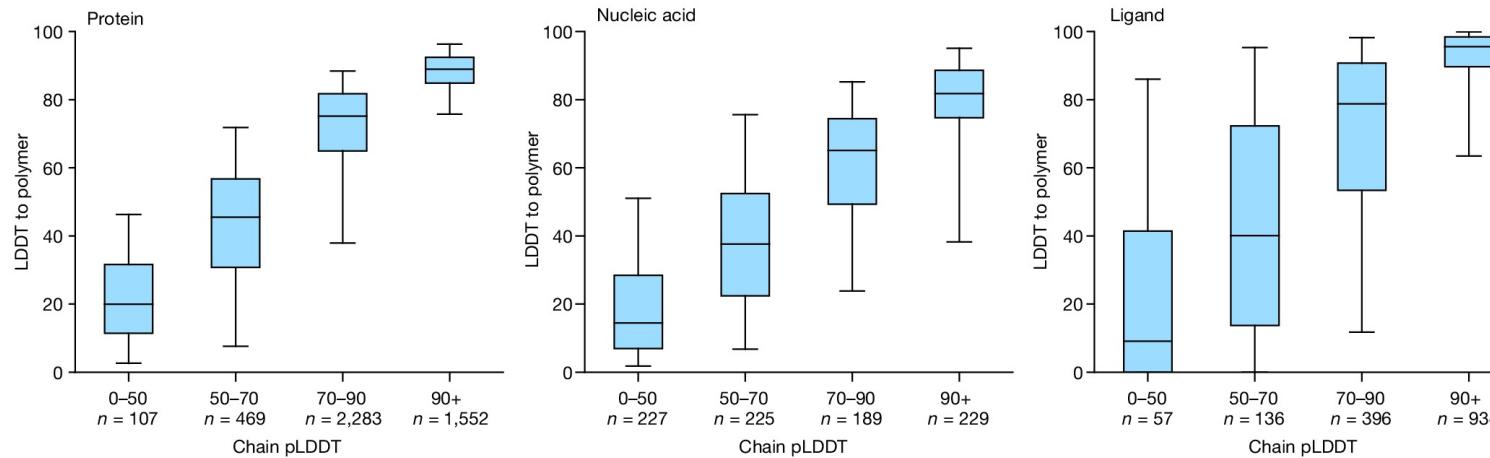
Fig. 4 予測の信頼性評価

Fig. 5 AF3のモデルの課題

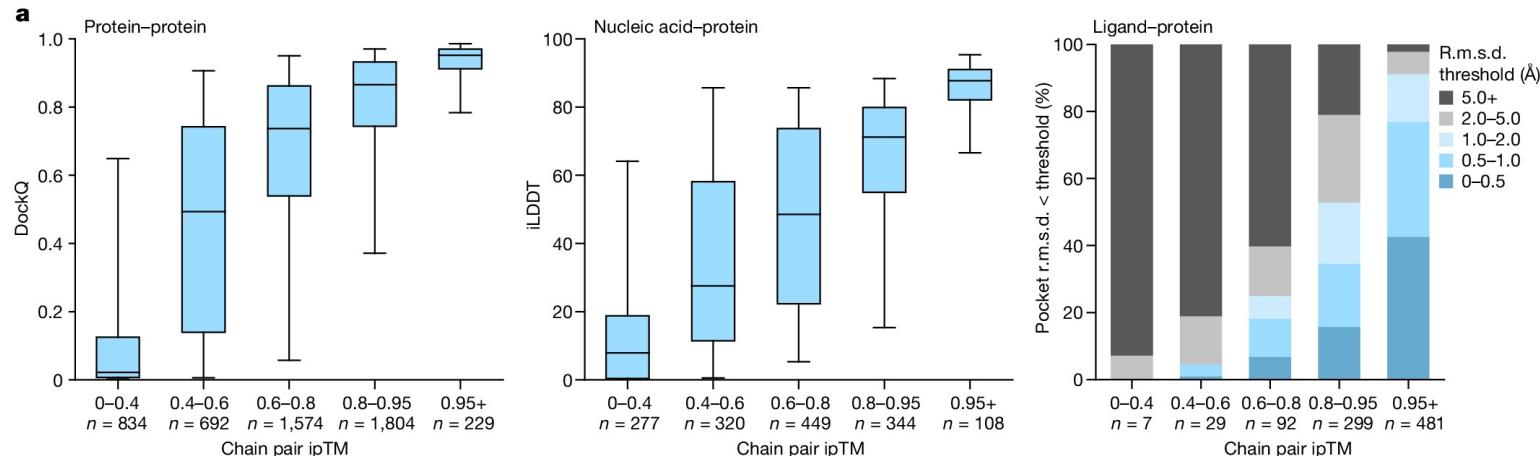
Fig. 4 予測の信頼性評価

各種生体分子とその複合体について、信頼性の予測値と実際の誤差が相関していた

- ✓ タンパク質・核酸自体の構造予測とリガンドの位置の予測



- ✓ タンパク質間 / タンパク質と核酸・リガンドの相互作用の予測

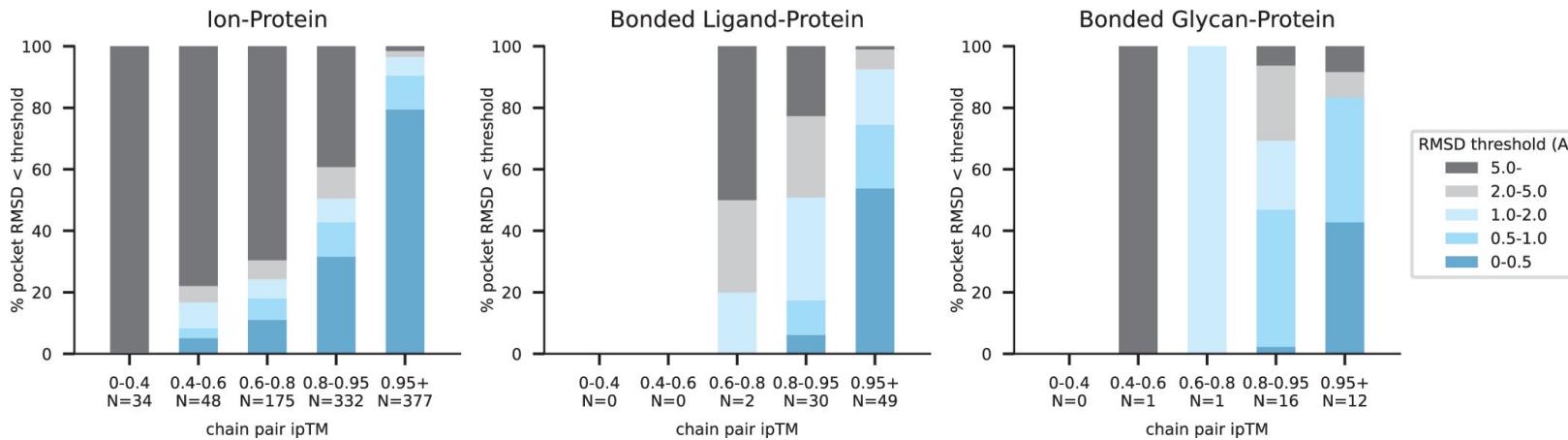


AF2等との比較を行っていないなど、この辺りは自信がないのかもしれない。

Fig. 4 予測の信頼性評価

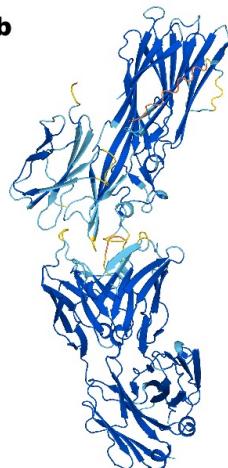
各種生体分子とその複合体について、信頼性の予測値と実際の誤差が相関していた

✓ イオン・共有結合修飾の予測

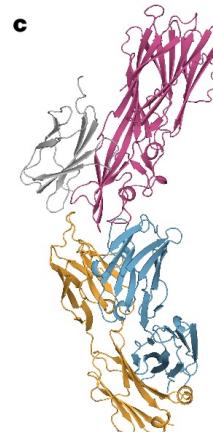


信頼性予測の具体例 (PDB7T82: ロイコシジン)

✓ 残基ごとの信頼性予測



✓ サブユニット間の境界面の信頼性が高いほど、実際の精度(DockQ score)も高かった。



		Interface DockQ score			
		A	C	D	F
A	A	0.003	0.003	0.721	
	C	0.003		0.740	
D	A	0.003		0.740	
	C				0.721

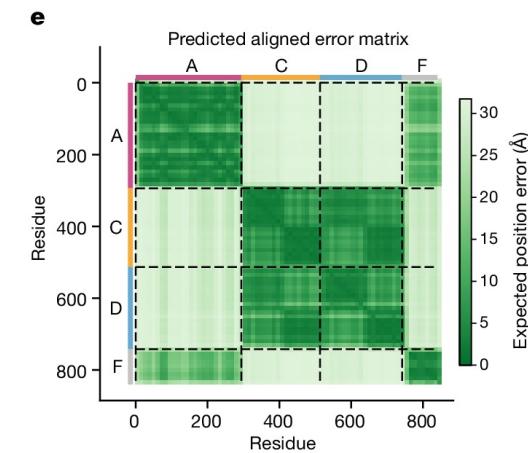


Fig. 1,2 提案するモデル構造と学習手法

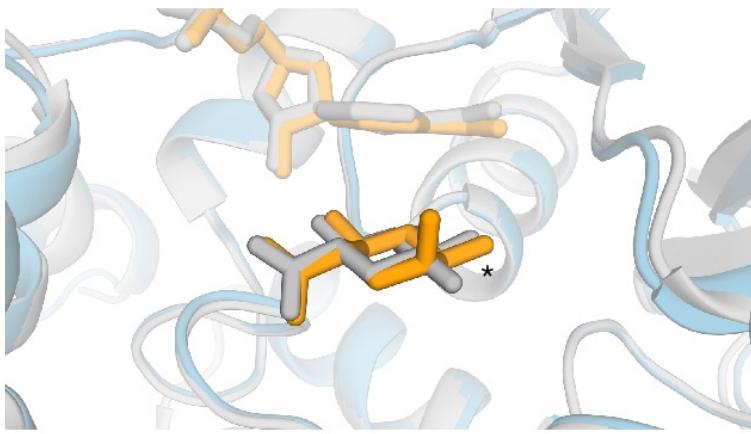
Fig. 1,3 様々な複合体に対する予測精度

Fig. 4 予測の信頼性評価

Fig. 5 AF3の課題

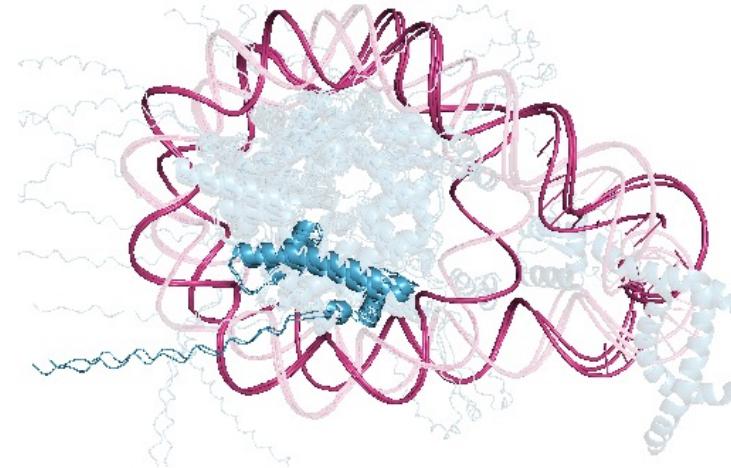
Fig. 5 AF3の課題

1. 誤った立体化学構造を出力しうる



- ✓ 例: 左図では正解の β -D-グルクロン酸に対して, α -D-グルクロン酸が出力されている。
- ✓ 出力の順位付けでは, このような立体化学構造の誤りに対するペナルティ項を加えた。
- ✓ PoseBustersデータセットでは, 全体の4.4%に立体化学構造の誤りが見られた。

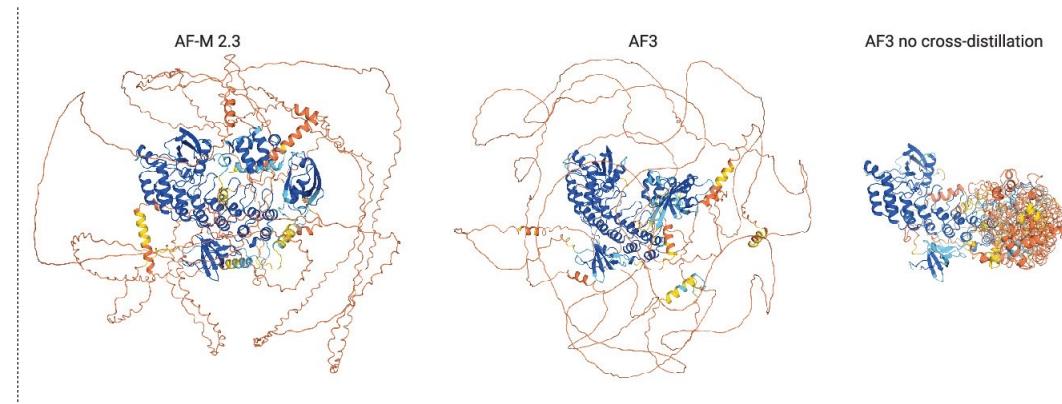
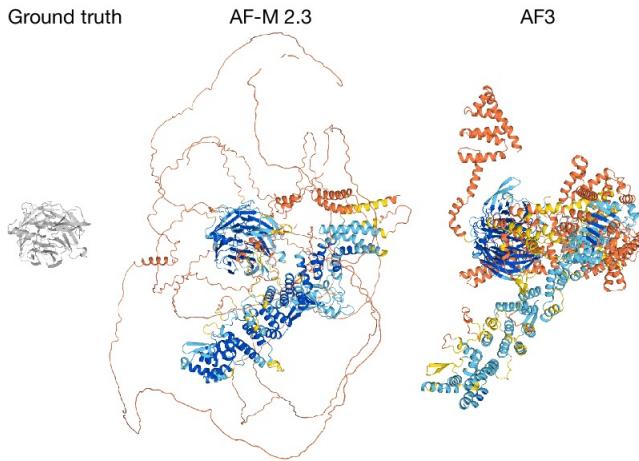
2. 複数原子が衝突した構造を出力しうる



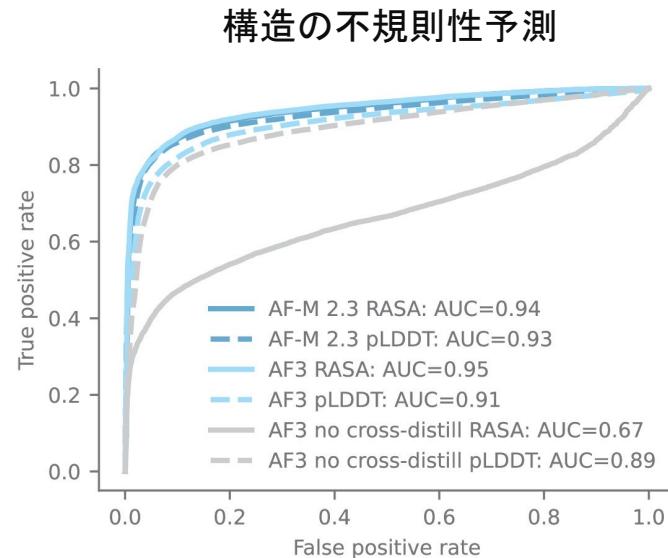
- ✓ 最悪の場合, 左のようにホモマーの同じ鎖が全く同じ位置に重なることがある。
- ✓ その他の場合, 長い配列の構造で衝突が起きやすかった。

Fig. 5 AF3の課題

3. 不規則な領域にも構造を生成することがある



- ✓ タンパク質には、細胞中で特定の構造を取りらない天然変性領域(IDR)が存在する。
- ✓ AlphaFold-Multimerではそのような領域はリボン状に予測されるが、AF3はノイズ除去モデルのためそのような領域にも構造を生成する場合がある（信頼度は低くなる）



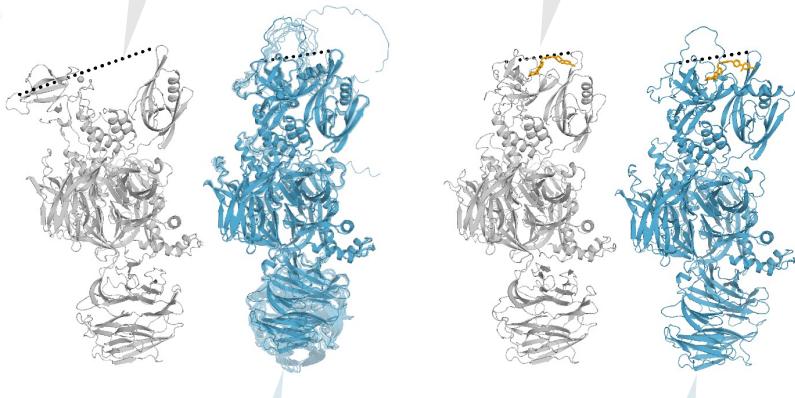
- ✓ AF-Multimerの予測結果を学習データに加えることで、ある程度改善する。

Fig. 5 AF3の課題

4. 複数のコンフォメーションを予測できない場合がある

- ✓ 例: E3ユビキチンリガーゼ(左図)

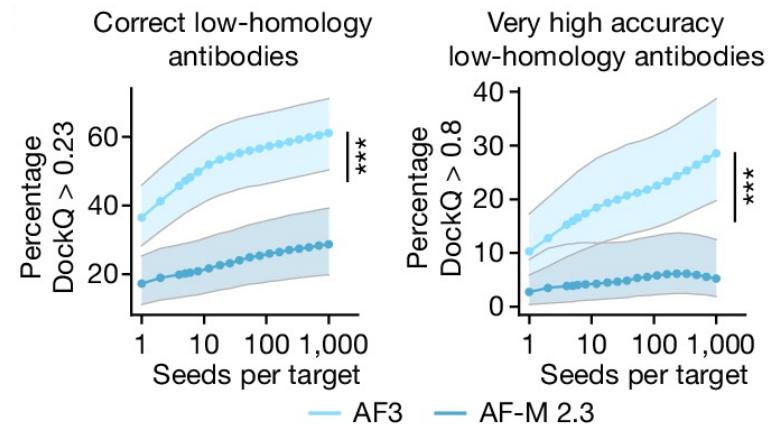
本来(灰色): 通常は開いた状態、リガンドと結合すると閉じた状態になる



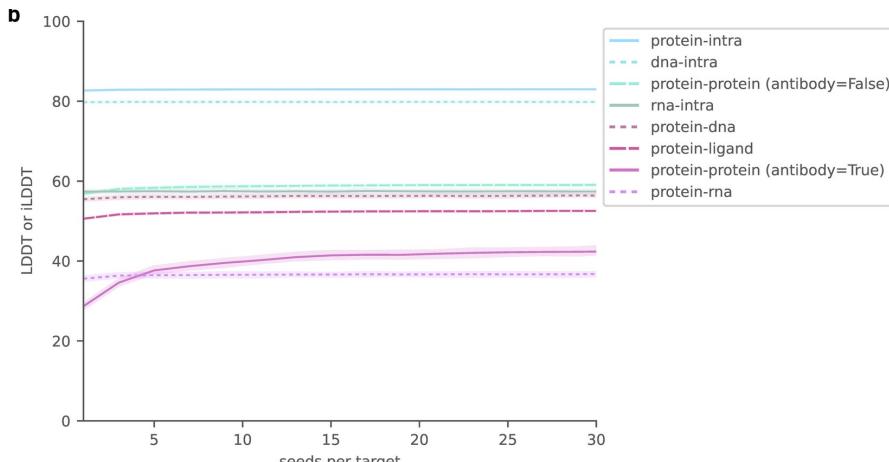
予測(水色): 常に閉じた状態で予測される

- ✓ このように、複数の状態を取りうるものについて全て予測できない場合がある

5. 抗原-抗体複合体の予測には、多くの出力から最適なものを選ぶ必要がある



- ✓ 候補分子を多くサンプリングするほど、その中で最善のものを選んだときの精度が向上する



- ✓ その他の分子種ではサンプリング数の影響は少ない

著者らのDiscussion

- ✓ AlphaFold3により、1つの深層学習モデルで幅広い種類の分子をモデリングできることを示した。
- ✓ MSAのような進化的情報がない分子についても、予測を行うことができた。
- ✓ 本モデルにより、これまでに得られたデータの利用可能性を拡大した。
- ✓ 実験的な構造解析技術の進展によりさらなるデータが得られることで、モデルの一般化能がさらに拡大することが期待される。

展望・所感

- ✓ 本モデルの新規性:

手法: 予測への拡散モデルの導入

MSA表現よりもペア表現を重視する予測モデル

結果: 幅広い入力に対し、特化したモデルをも上回る精度を示した。

タンパク質の構造予測自体も精度が向上

- ✓ モデルを残基/原子単位の拡散モデルにすることで、幅広い入力に対応するというのは、今後さらなる一般化・適用範囲の拡大が可能になると思った。

- ✓ これまでタンパク質構造予測に重要とされてきたMSAが、実はあまり重視する必要がないと明らかにしたことも興味深かった。

- ✓ 拡散モデルなので、推論には時間がかかる。

SIにのみ推論時間の言及があった(右表)。

抗体-抗原複合体など、多くのサンプリングにはさらに時間がかかる。

少なくともスクリーニング等大規模な計算には向かないと思われる。

A100 GPU 16個での推論時間

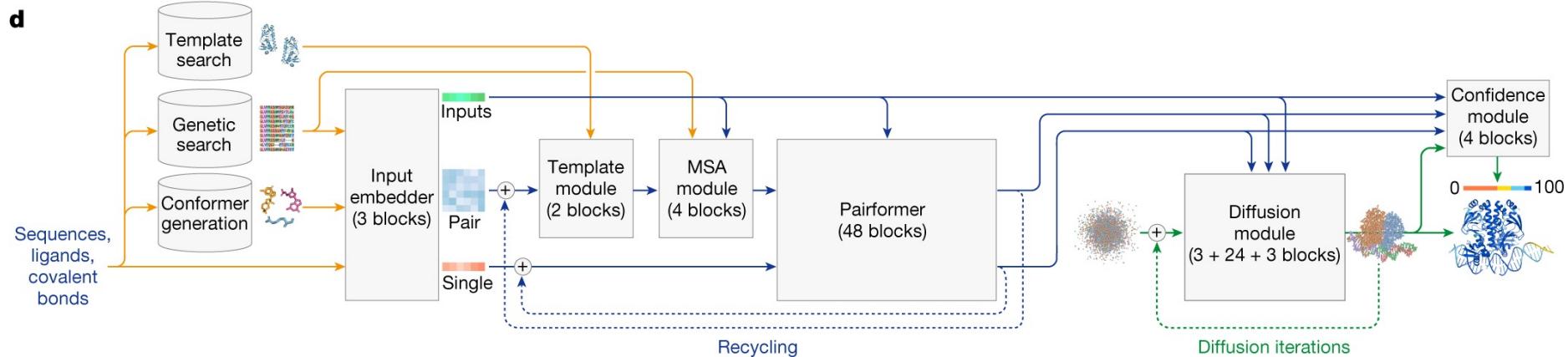
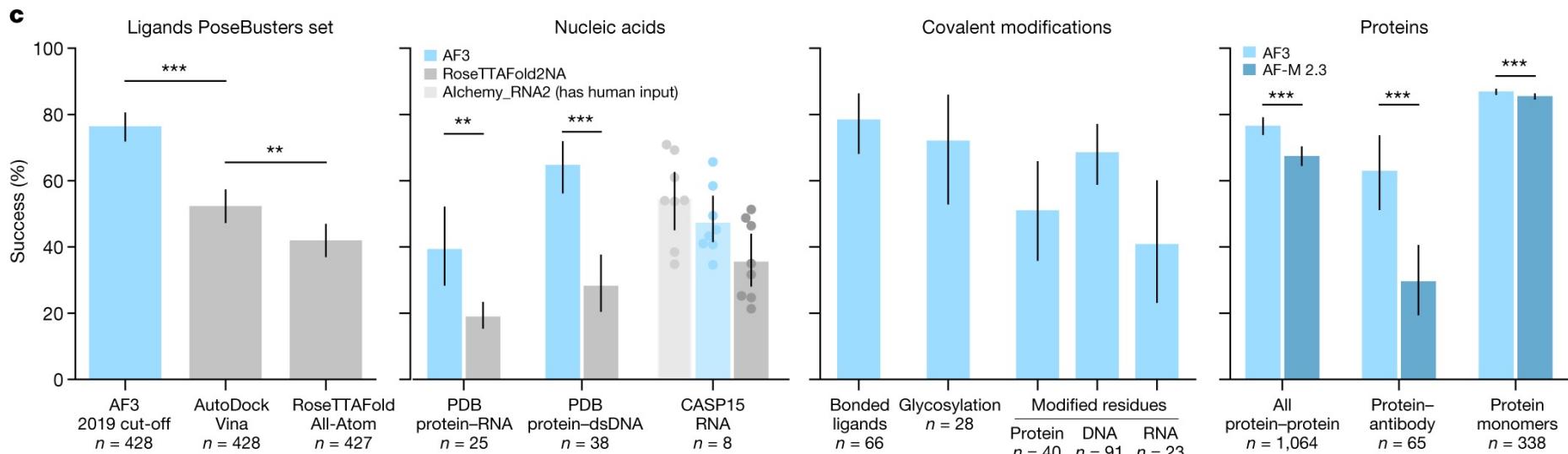
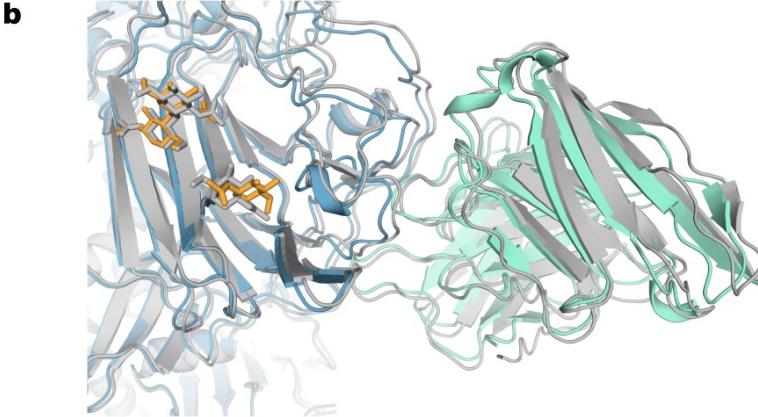
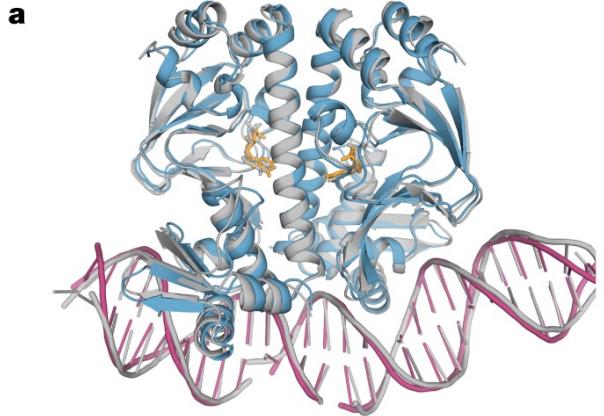
Number of tokens	Inference time (seconds)
1024	22
2048	71
3072	126
4096	228
5120	347

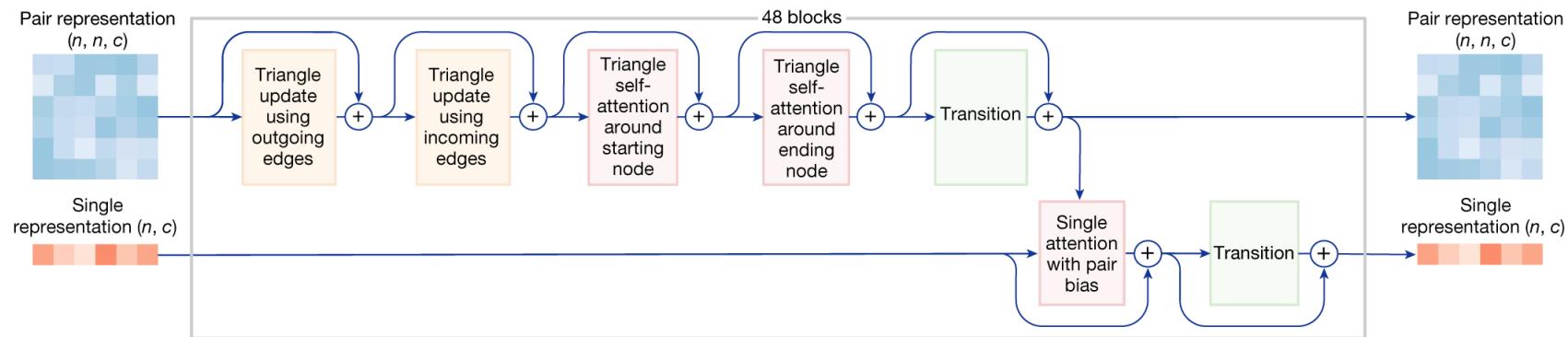
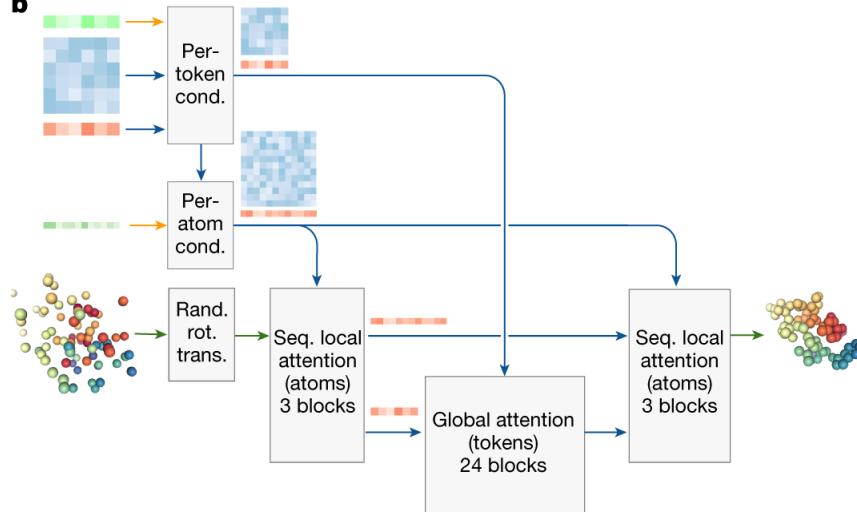
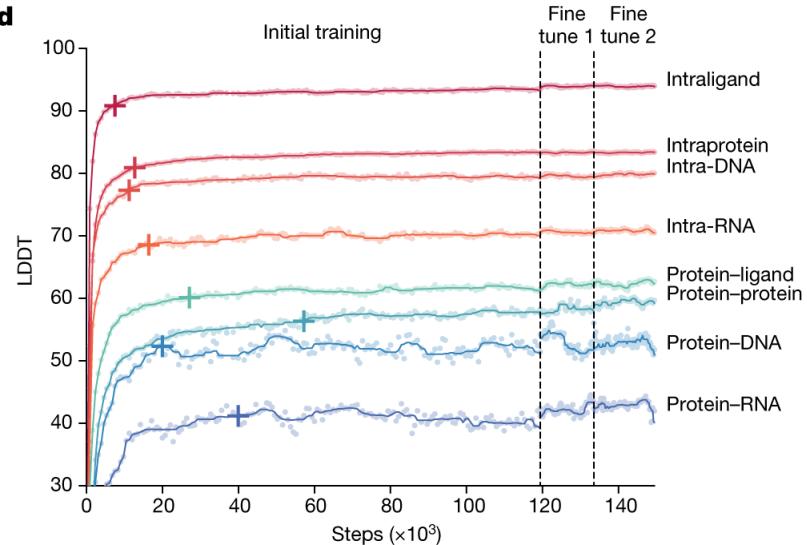
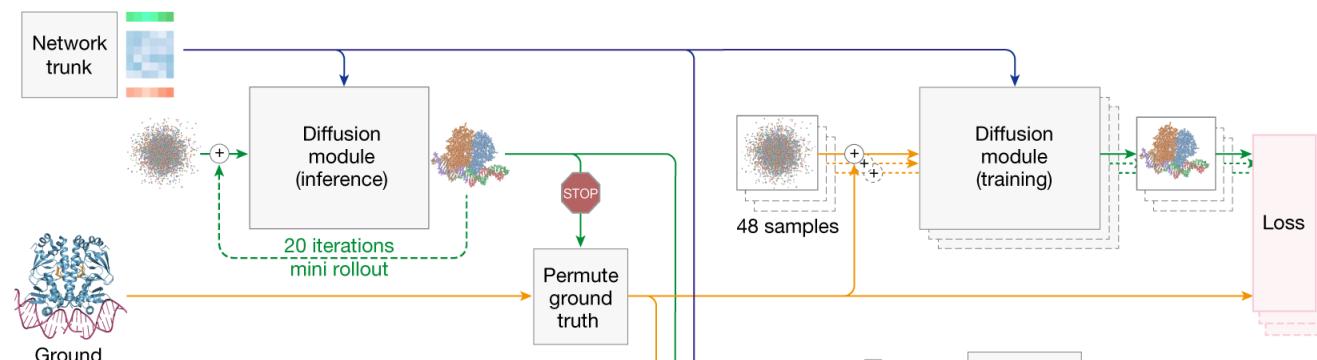
- ✓ validationの結果を見ながらモデルを改善しているようなのが気になる。

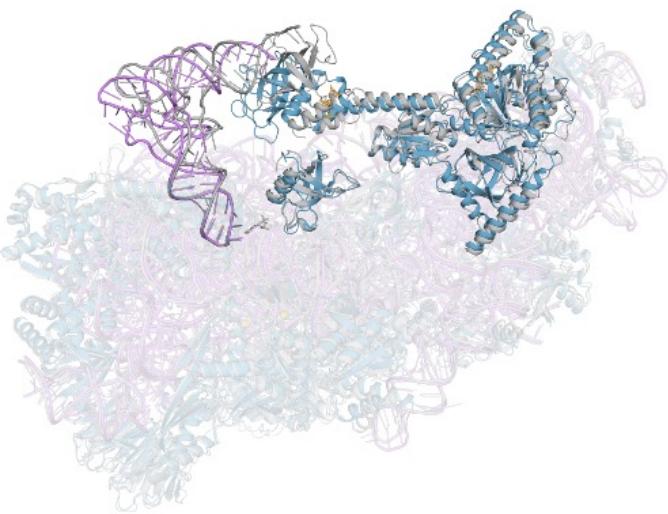
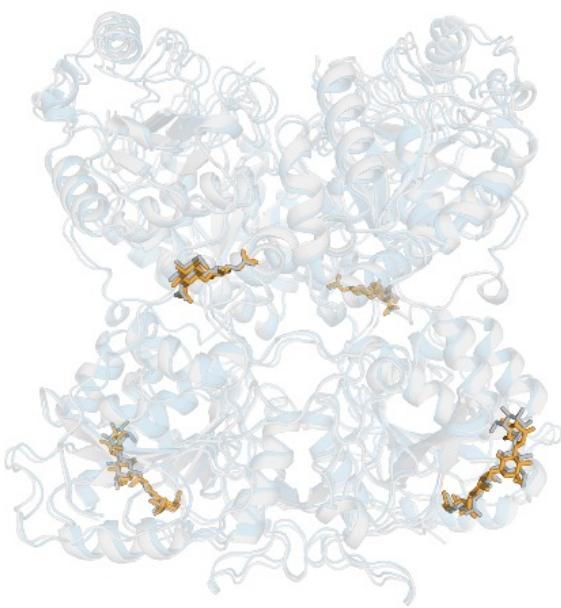
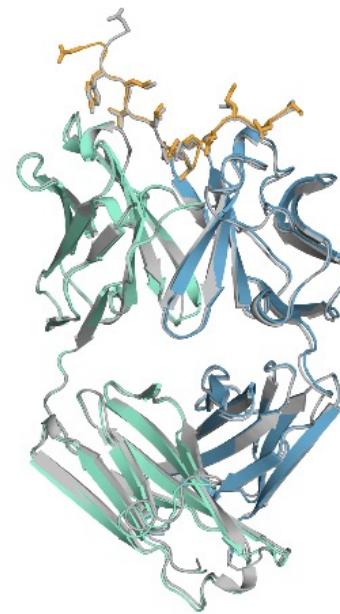
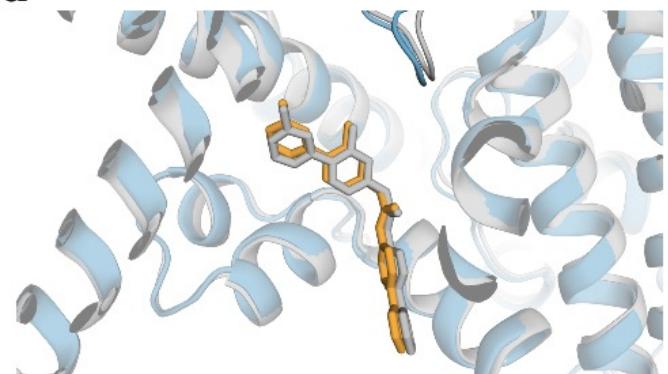
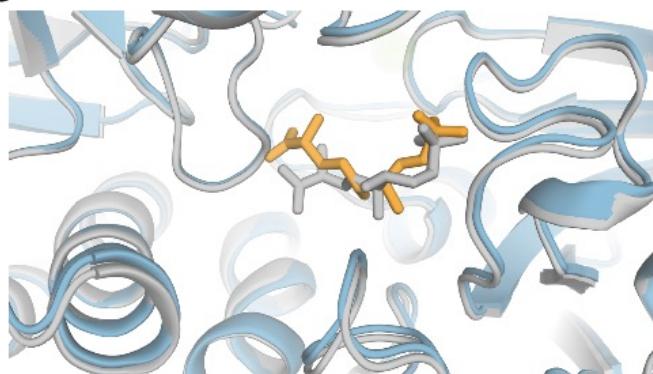
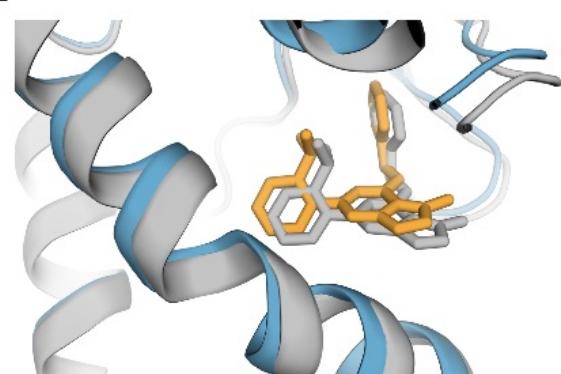
CASP等の大会で比較した方がよいかもしれない。

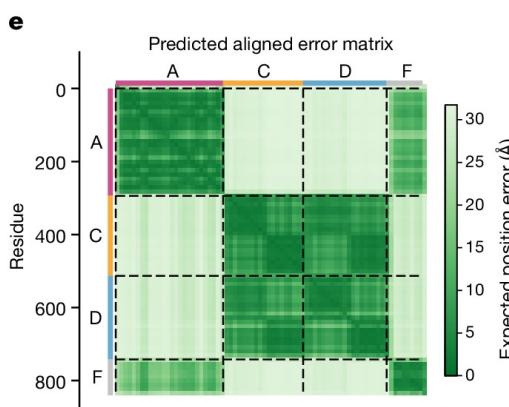
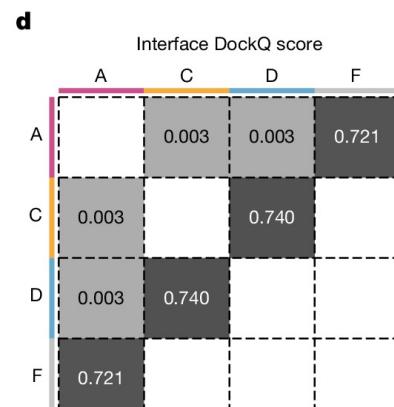
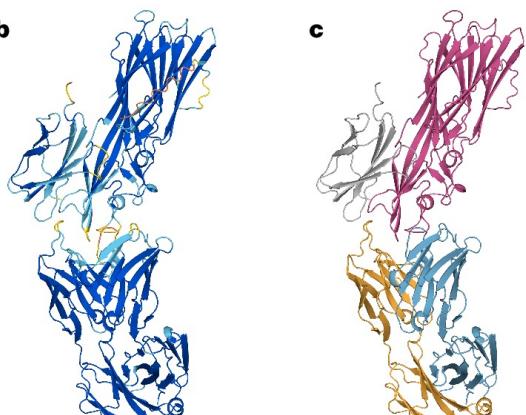
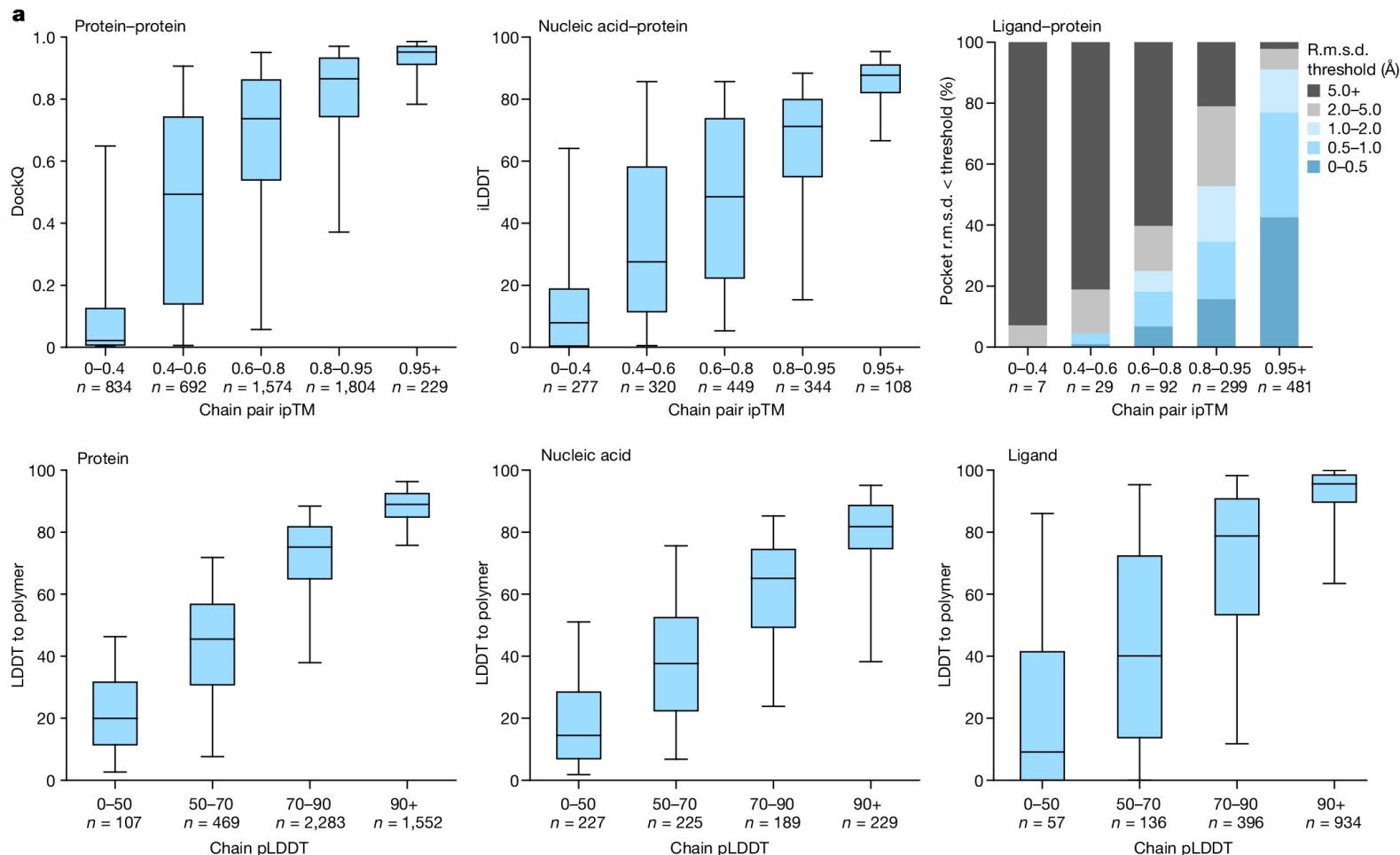
Thank you for listening

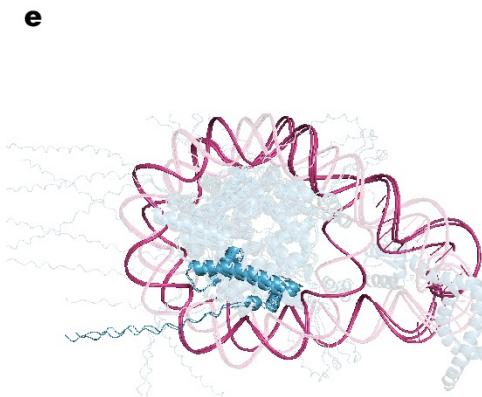
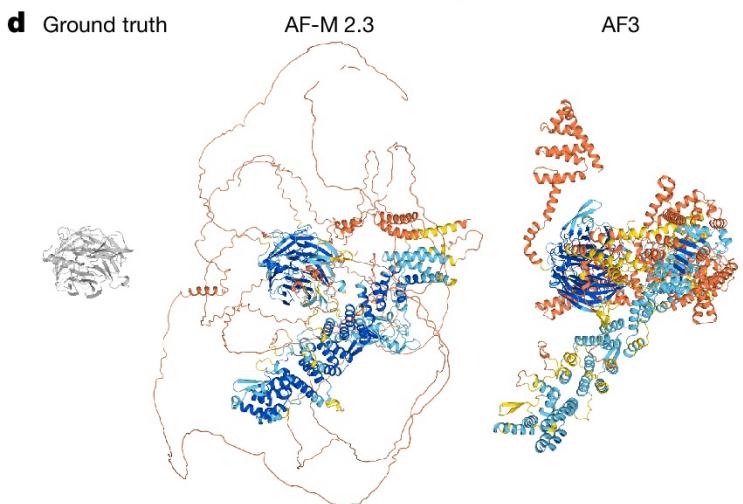
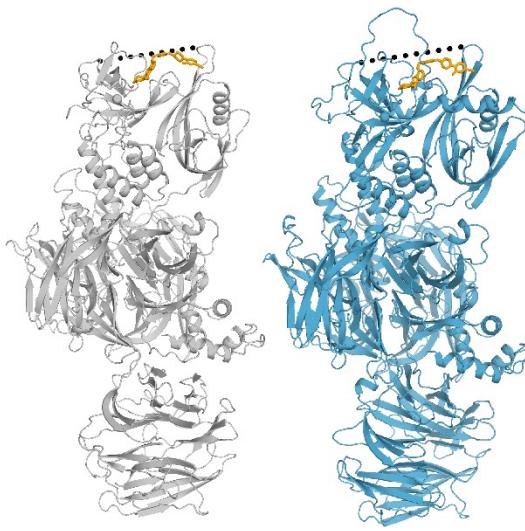
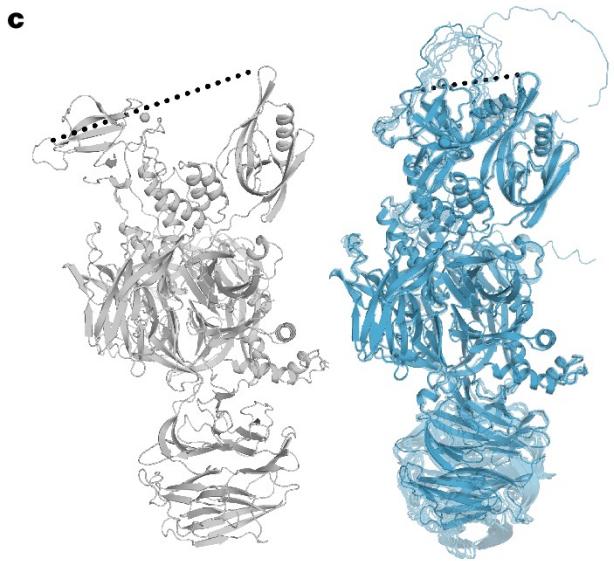
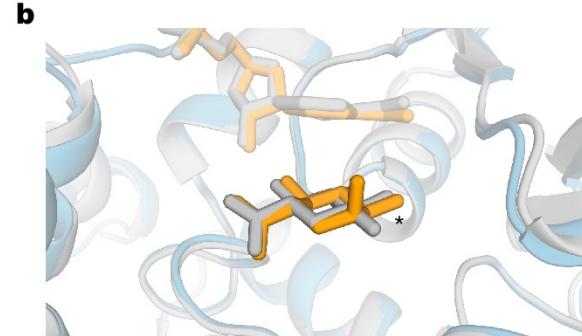
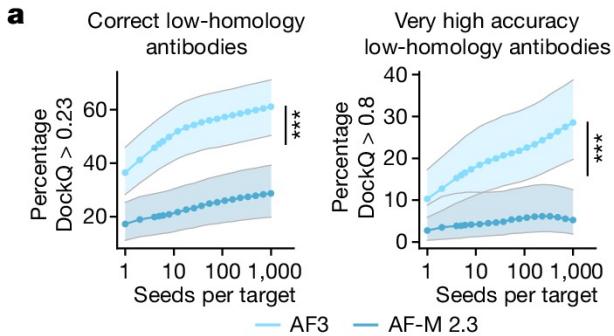
Figures

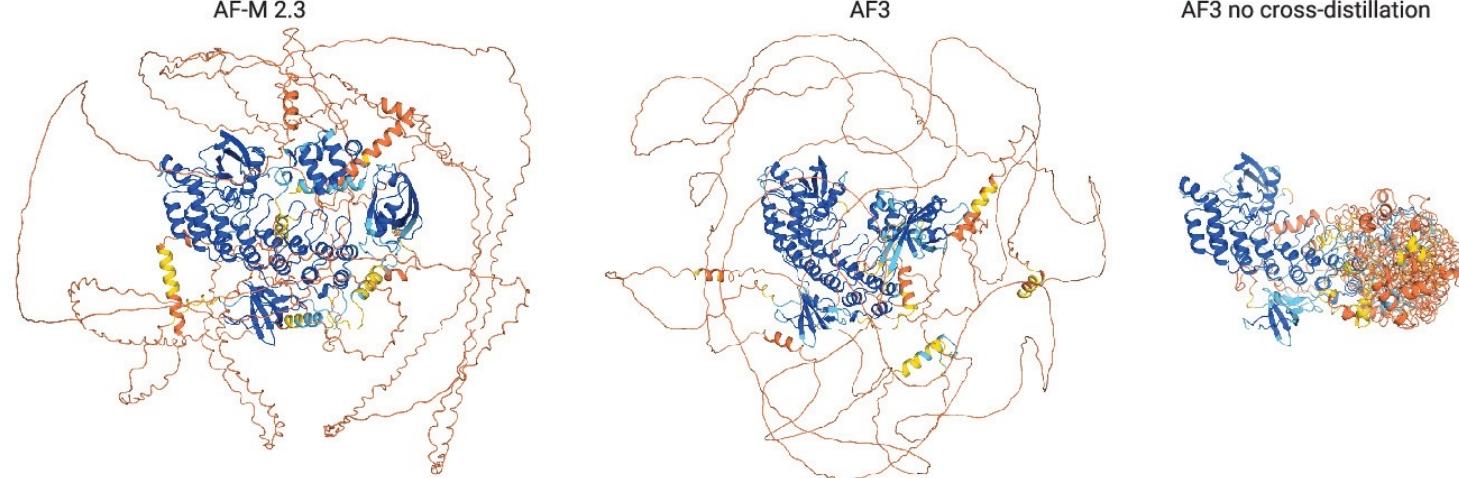
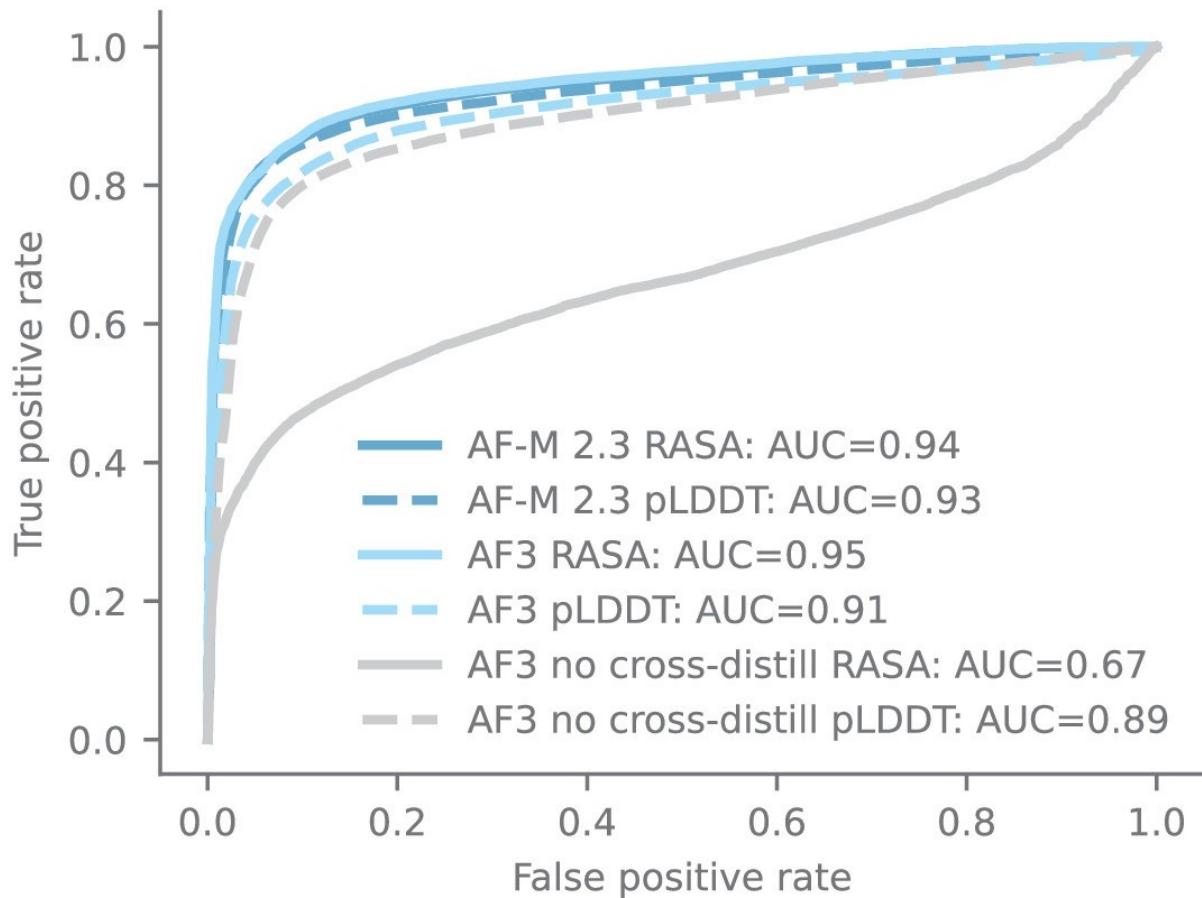


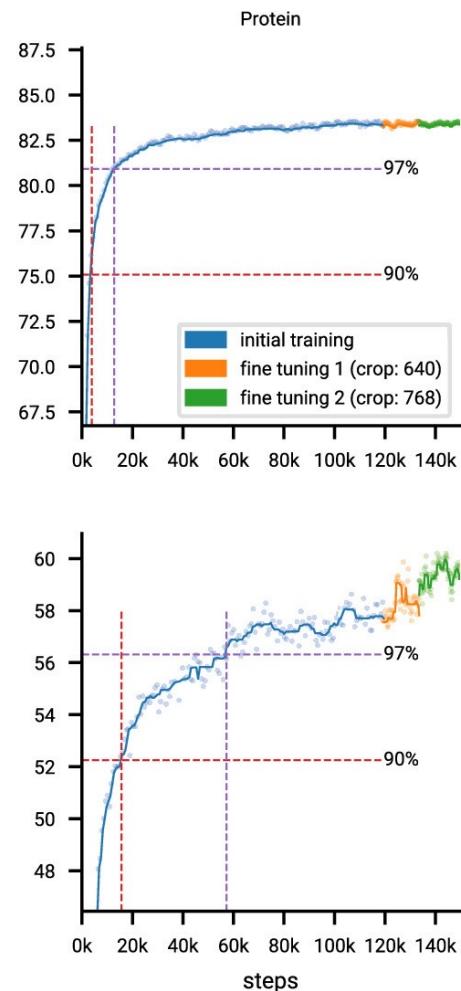
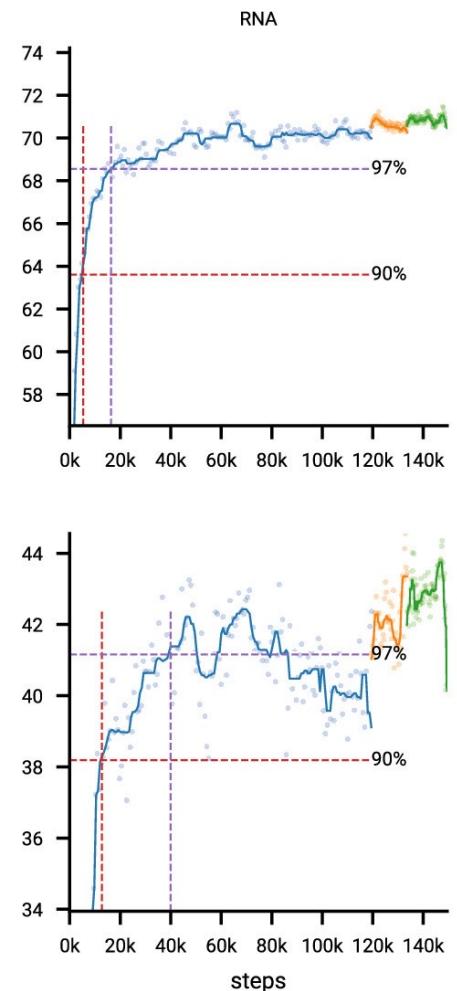
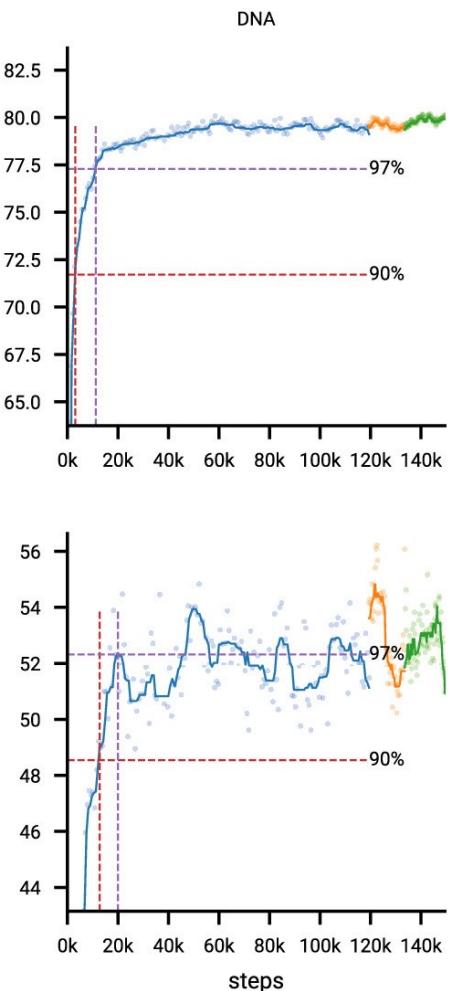
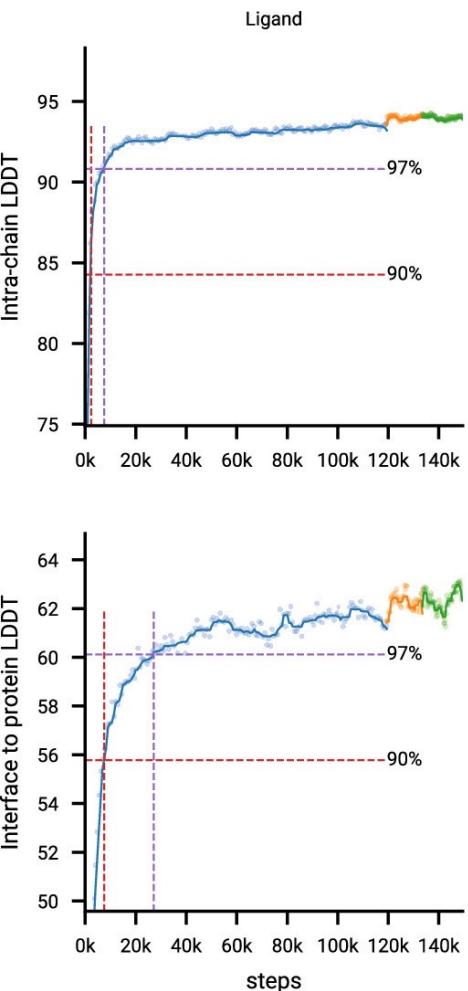
a**b****d****c**

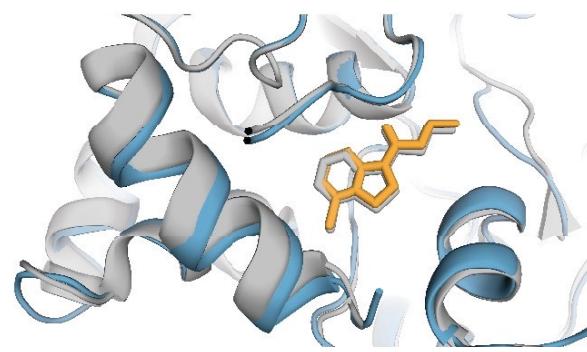
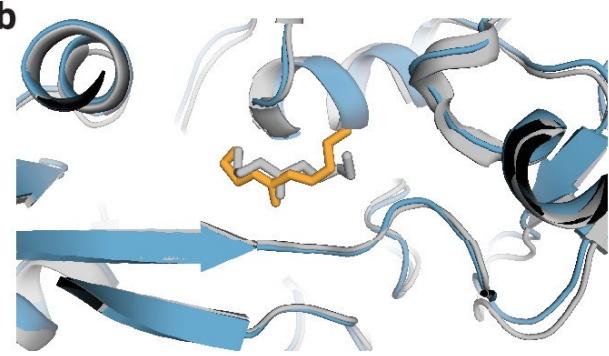
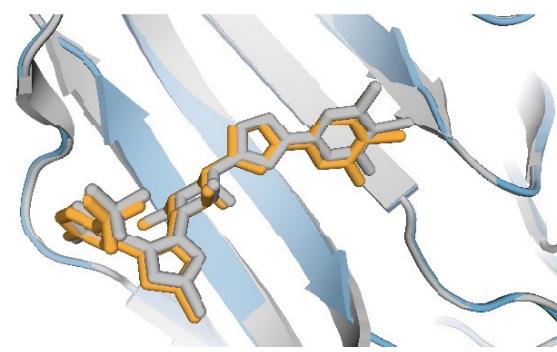
a**b****c****d****e****f**

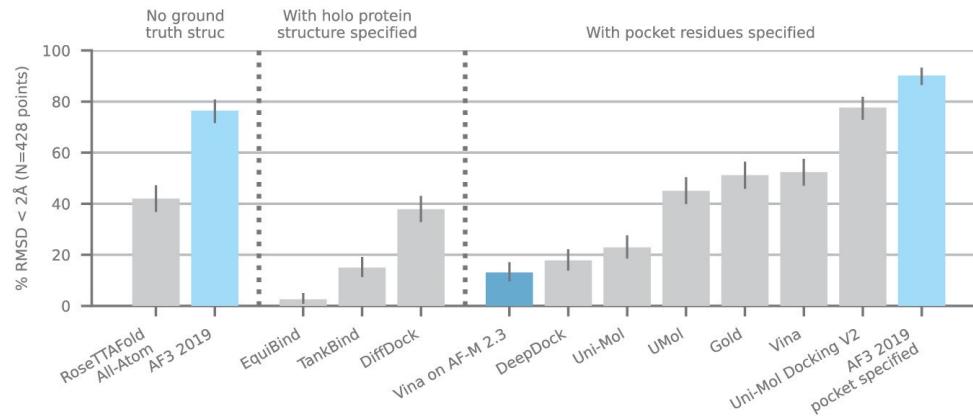
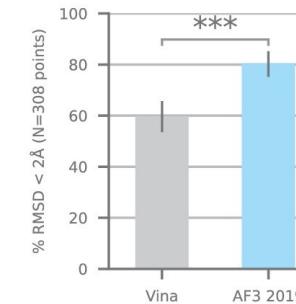
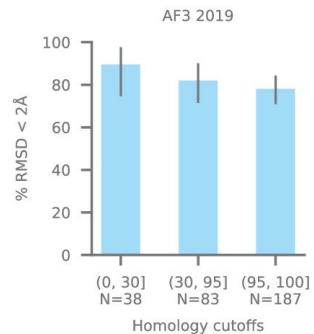
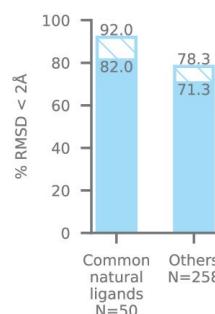
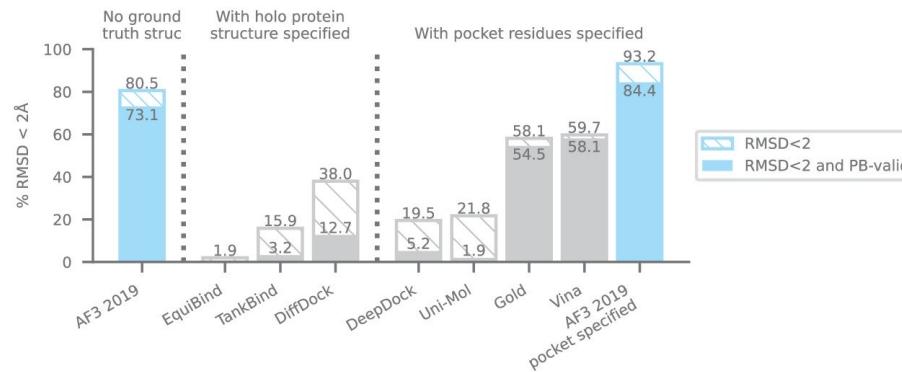
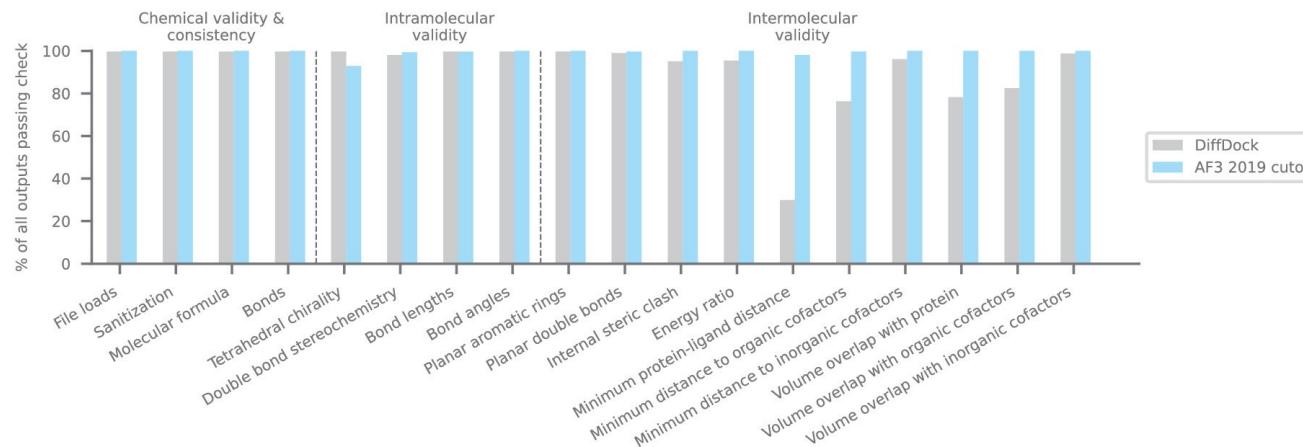


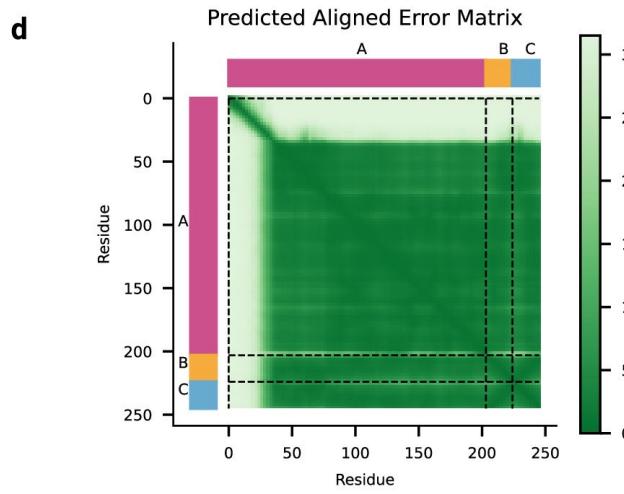
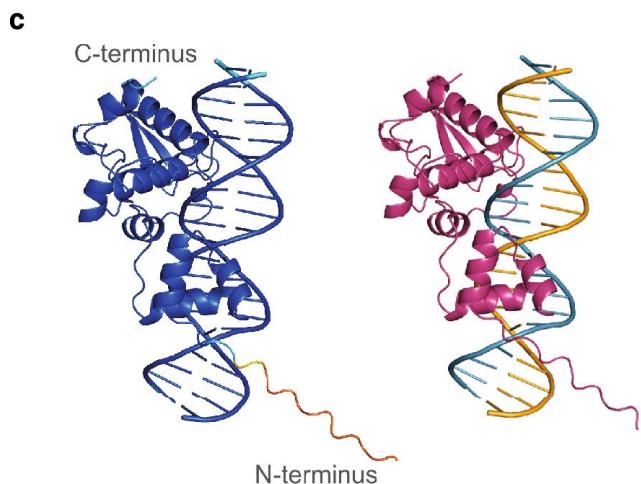
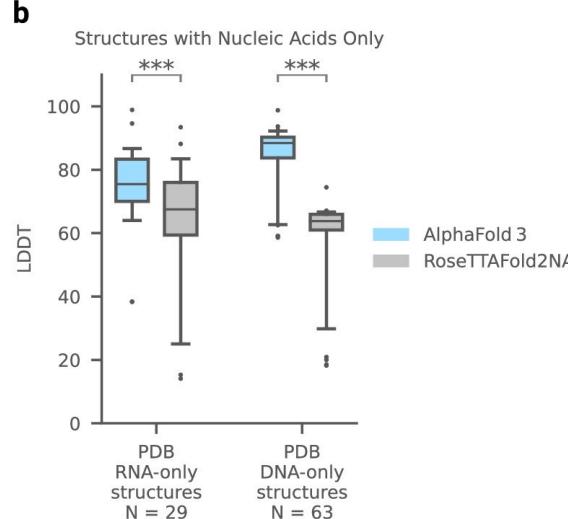
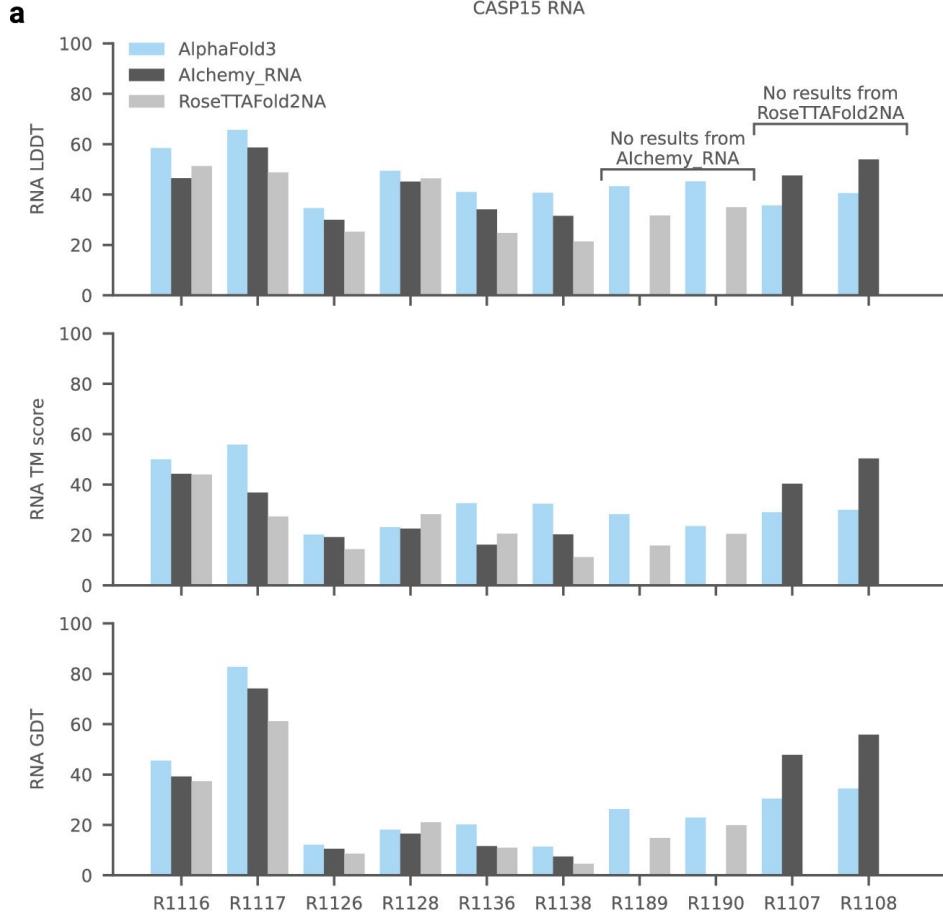


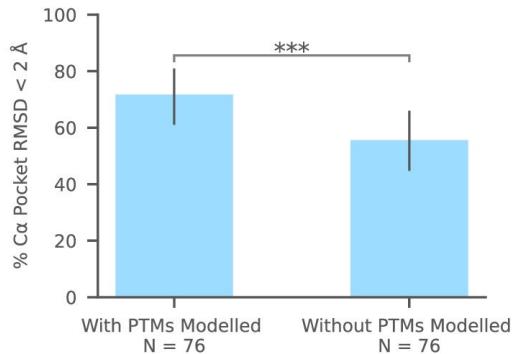
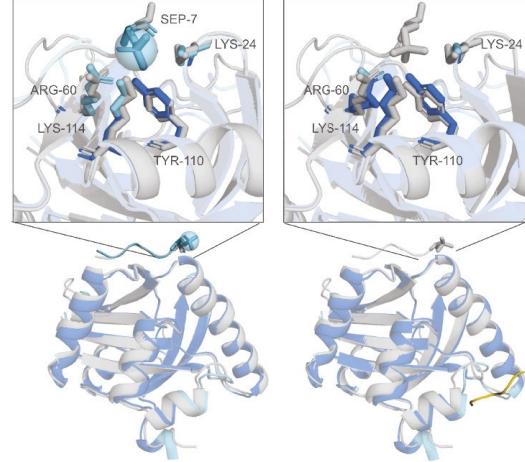
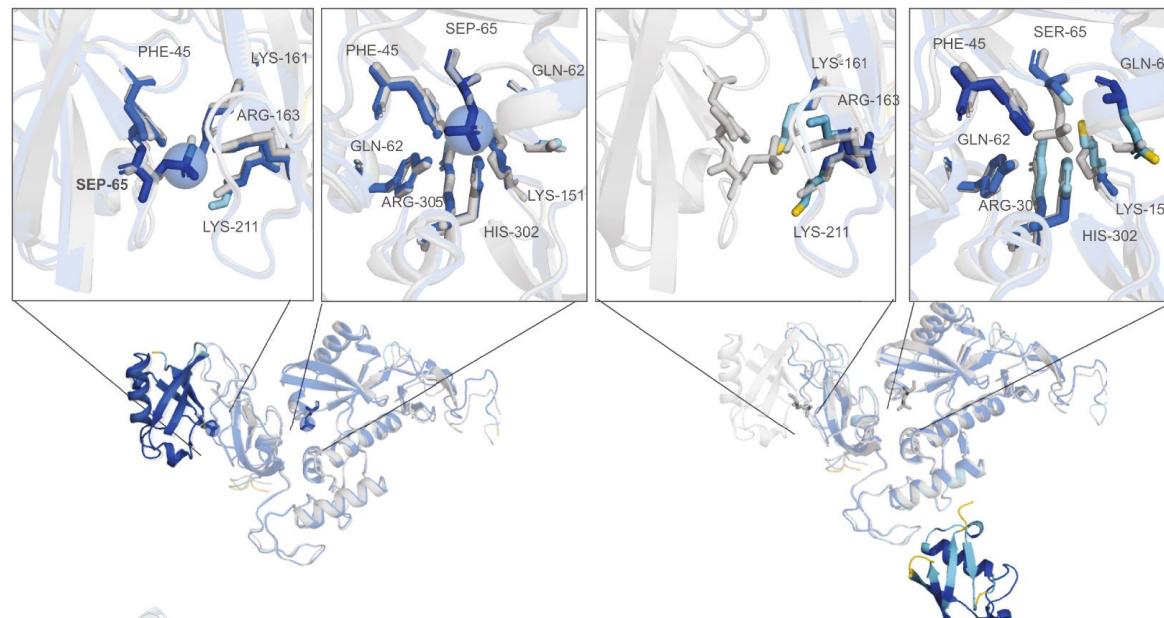
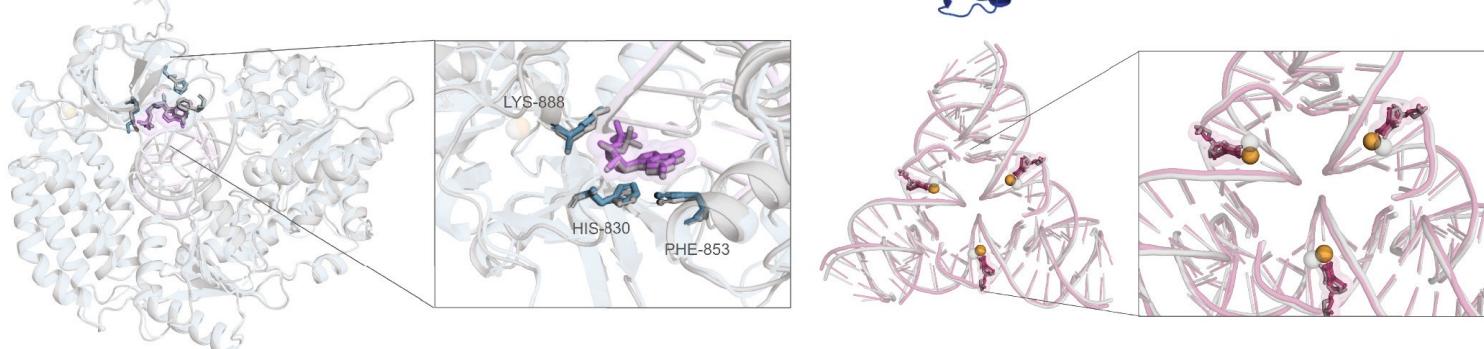
a**b**

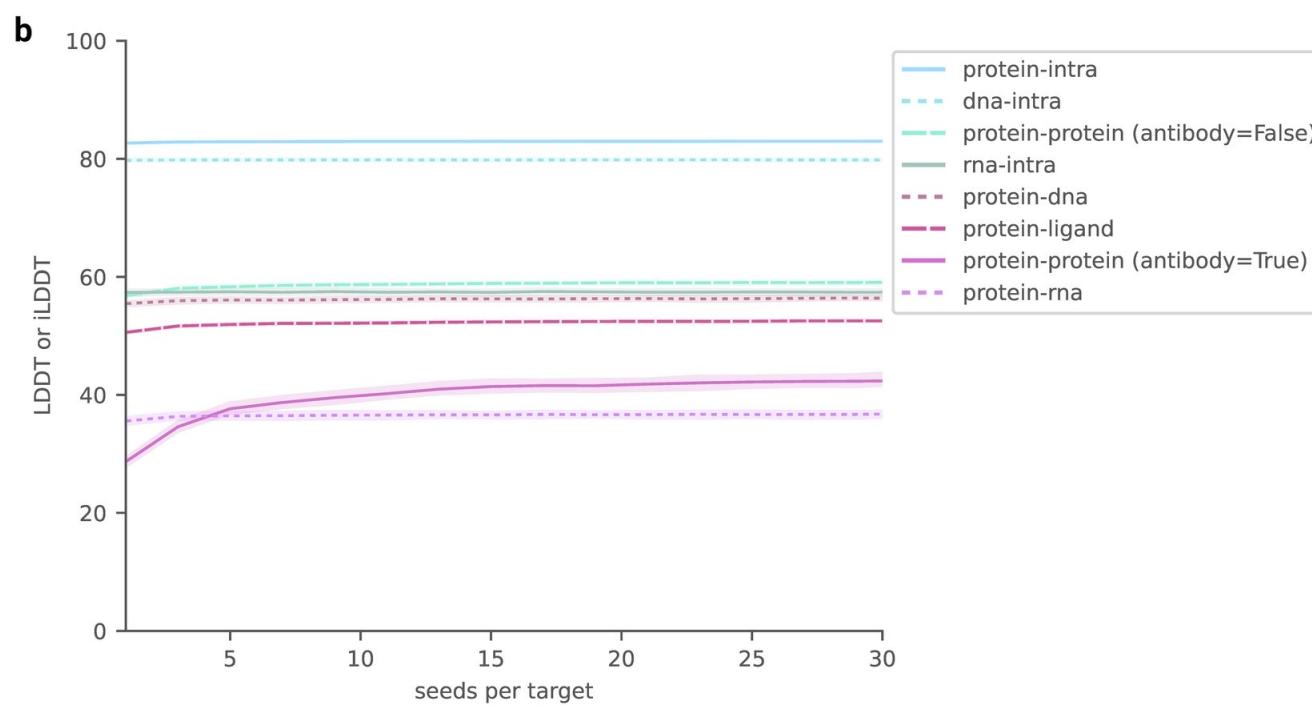
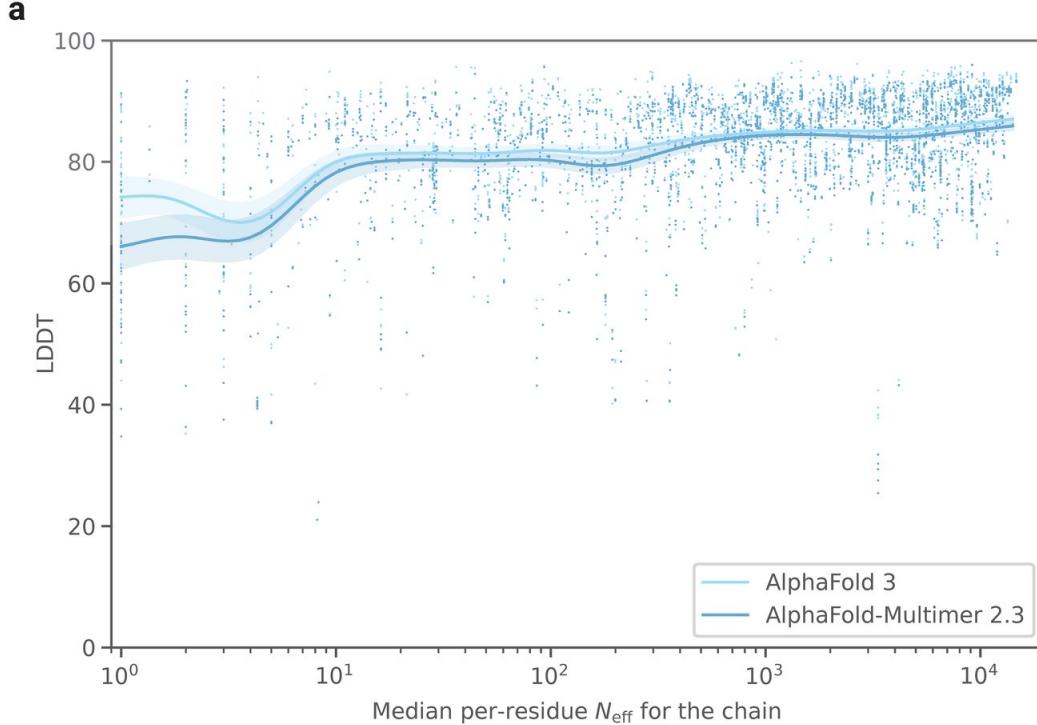


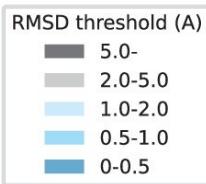
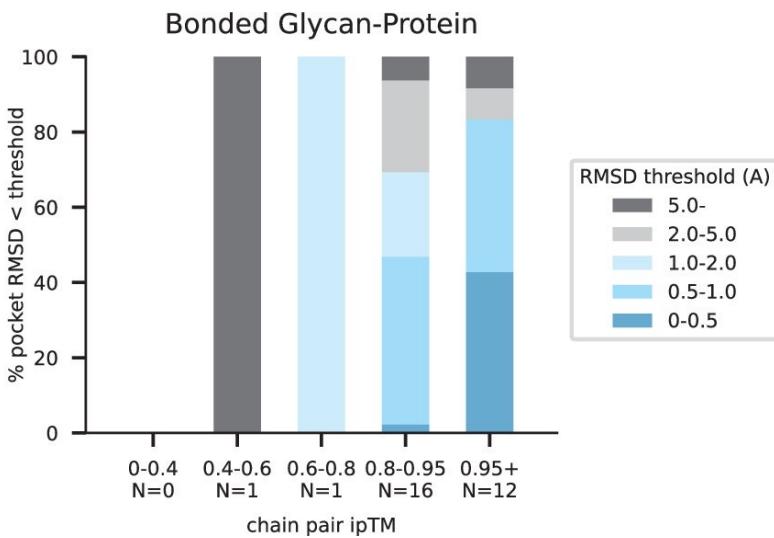
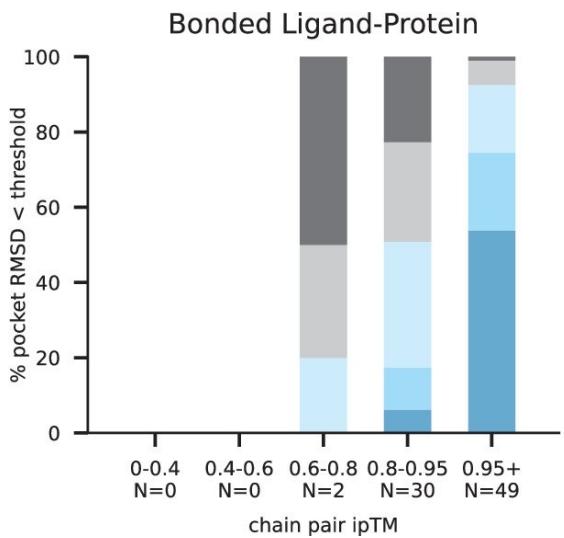
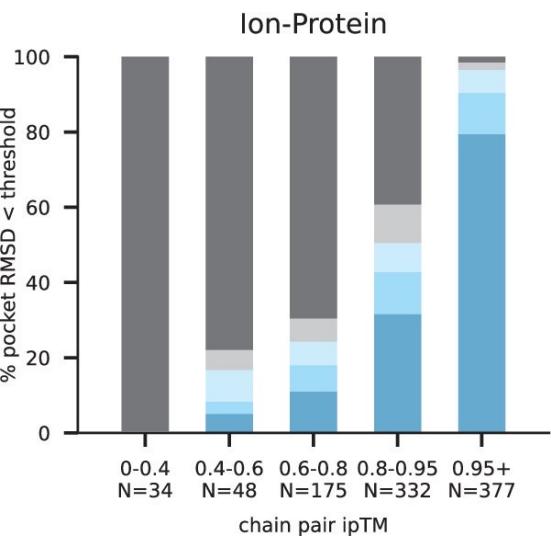
a**b****c**

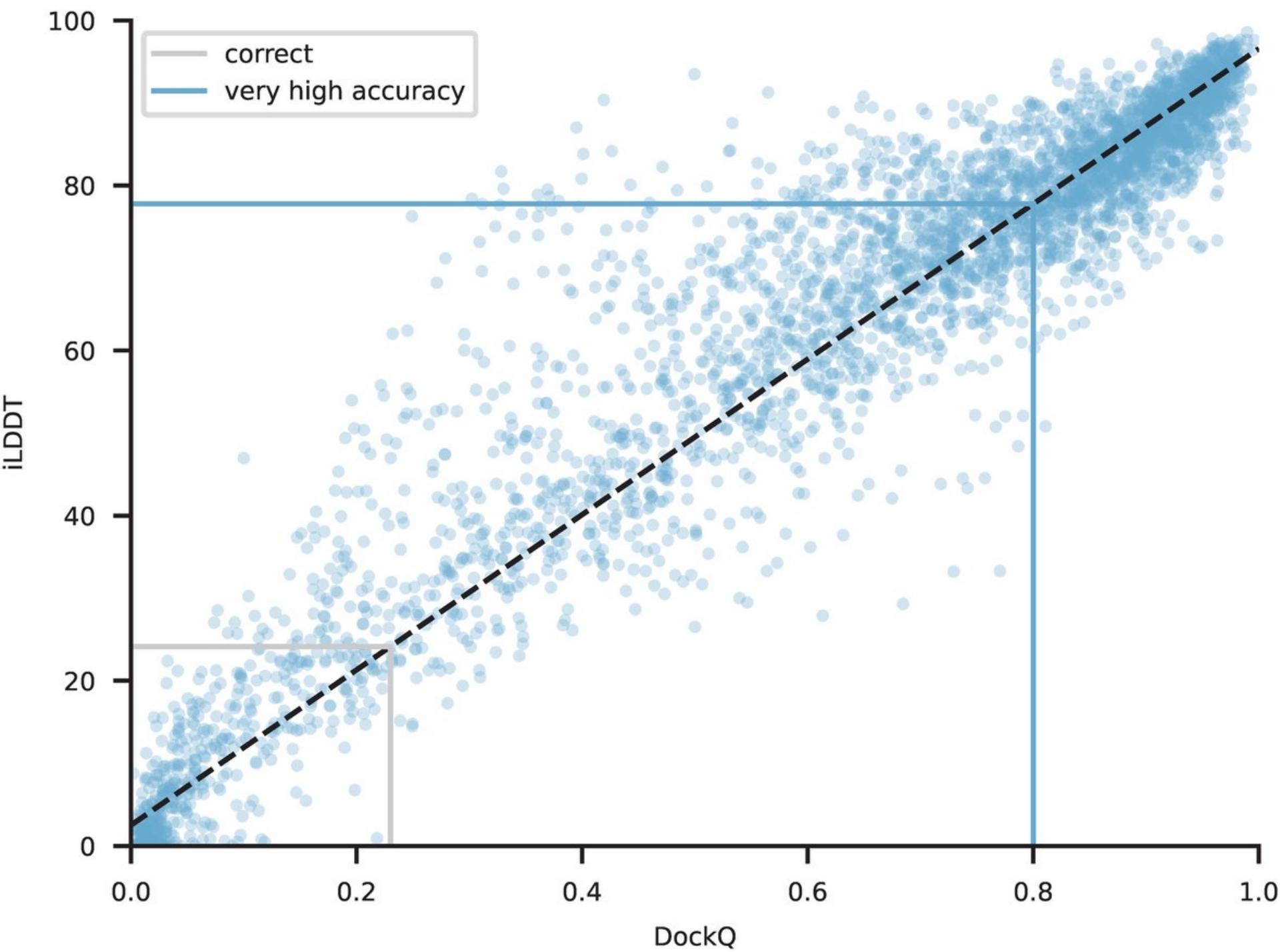
a PoseBusters Version 1**b PoseBusters Version 2****c PoseBusters Version 2****d PoseBusters Version 2****e PoseBusters Version 2****f PoseBusters Version 2**



a Common Phosphorylated Residues**b****c****d**







Task	Dataset	Metric	Notes	Method	N	Mean	95% CI
Ligands	PoseBusters V1	% RMSD < 2 Å	–	RoseTTAFold All-Atom AF3 (2019 cutoff)	427	42.0	37.2 – 46.8
					428	76.4	72.1 – 80.3
				Holo protein struct. given	428	2.6	1.3 – 4.6
			Pocket residues specified	EquiBind	428	15.0	11.7 – 18.7
				TankBind	428	37.9	33.2 – 42.6
				DiffDock	428	13.1	10.0 – 16.7
				Vina on AF-M 2.3	428	17.8	14.3 – 21.7
				DeepDock	428	22.9	19.0 – 27.2
				Uni-Mol	428	45.0	40.3 – 49.9
				Gold	428	51.2	46.3 – 56.0
			–	Vina	428	52.3	47.5 – 57.2
				Uni-Mol Docking V2	428	77.6	73.3 – 81.4
				AF3 (2019 cutoff) pocket specified	428	90.2	87.0 – 92.8
Ligands	PoseBusters V2	% RMSD < 2 Å	–	AF3 (2019 cutoff)	308	80.5	75.6 – 84.8
				Holo protein struct. given	308	1.9	0.7 – 4.2
				EquiBind	308	15.9	12.0 – 20.5
			Pocket residues specified	TankBind	308	38.0	32.5 – 43.7
				DiffDock	308	15.3	11.4 – 19.8
				Vina on AF-M 2.3	308	19.5	15.2 – 24.4
				DeepDock	308	21.8	17.3 – 26.8
				Uni-Mol	308	58.1	52.4 – 63.7
			–	Gold	308	59.7	54.0 – 65.3
				Vina	308	93.2	89.8 – 95.7
Nucleic Acids	Protein-RNA	iLDDT	–	RoseTTAFold2NA	25	19.0	15.6 – 23.2
				AF3	25	39.4	28.5 – 51.9
	Protein-dsDNA	iLDDT	–	RoseTTAFold2NA	38	28.3	20.7 – 37.5
				AF3	38	64.8	56.4 – 71.7
	CASP 15 RNA	RNA LDDT	–	RoseTTAFold2NA	8	35.5	28.3 – 43.8
				AF3	8	47.3	41.7 – 55.2
				Alchemy_RNA2 (has human input)	8	54.5	45.3 – 62.4
				RNApolis (has human input)	8	50.5	45.2 – 55.8
				Chen (has human input)	8	49.8	40.7 – 58.5
				Kiharalab	8	40.9	35.1 – 54.3
				UltraFold	8	37.8	32.5 – 45.0
Covalent Mod.	Bonded ligands	% RMSD < 2 Å	–	AF3	66	78.5	68.3 – 86.2
	Glycosylation	% RMSD < 2 Å		AF3	28	72.1	53.1 – 85.7
		high-quality, single-residue	AF3	167	46.0	40.0 – 52.1	
			AF3	131	42.4	35.4 – 49.3	
	Modified residues	% RMSD < 2 Å	–	AF3	154	59.9	52.4 – 67.0
	Modified protein residues	% RMSD < 2 Å		AF3	40	51.0	36.0 – 65.6
	Modified DNA residues	% RMSD < 2 Å		AF3	91	68.6	59.0 – 76.9
	Modified RNA residues	% RMSD < 2 Å	–	AF3	23	40.9	23.4 – 59.9
Proteins	All Protein-Protein	% dockq > 0.23	–	AF-M 2.3	1064	67.5	64.7 – 70.1
				AF3	1064	76.6	74.0 – 78.9
	Protein-Antibody	% dockq > 0.23	–	AF-M 2.3	65	29.6	19.6 – 40.4
				AF3	65	62.9	51.4 – 73.5
	Monomers	LDDT	–	AF-M 2.3	338	85.5	84.7 – 86.1
				AF3	338	86.9	86.2 – 87.6