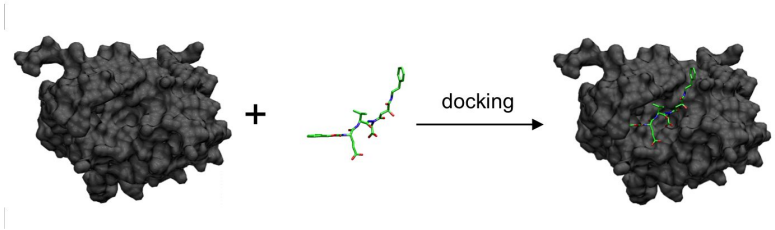
DIFFDOCK: DIFFUSION STEPS, TWISTS, AND TURNS FOR MOLECULAR DOCKING

Gabriele Corso*, **Hannes Stärk***, **Bowen Jing***, **Regina Barzilay** & **Tommi Jaakkola**
CSAIL, Massachusetts Institute of Technology

ICLR 2023
Presented by Victor Chu and Howard Yen

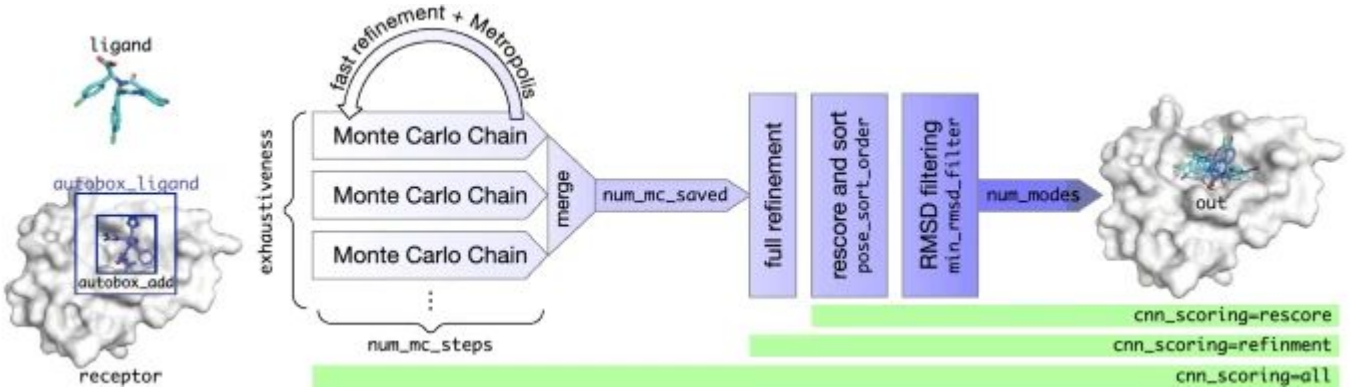
Problem Statement

- Developing a single drug costs around 1 billion dollars and takes 10 years of development and testing before potentially being FDA approved
- **Blind Docking:** identifying the correct orientation and conformation for a given ligand and protein without knowing anything about where the ligand will bind onto the protein.
- Large variety of binding mechanisms (hydrophobic, hydrogen-bonding, and π-stacking)



Search-Based Docking

- Define a scoring function (physics based or neural network) where “accurate scoring requires accurate docking”
- Stochastically modify the ligand pose to maximize score
- Performance of Search-based Method on Single Ligand-Receptor (23%):
  - Glide: > 1000 seconds
  - GNINA: ~146 seconds



Machine Learning for Blind Docking

- Attempt to remove the search process by directly predicting where the ligand protein will bind using neural network
- Very fast BUT performance has not reached traditional search methods (20%)

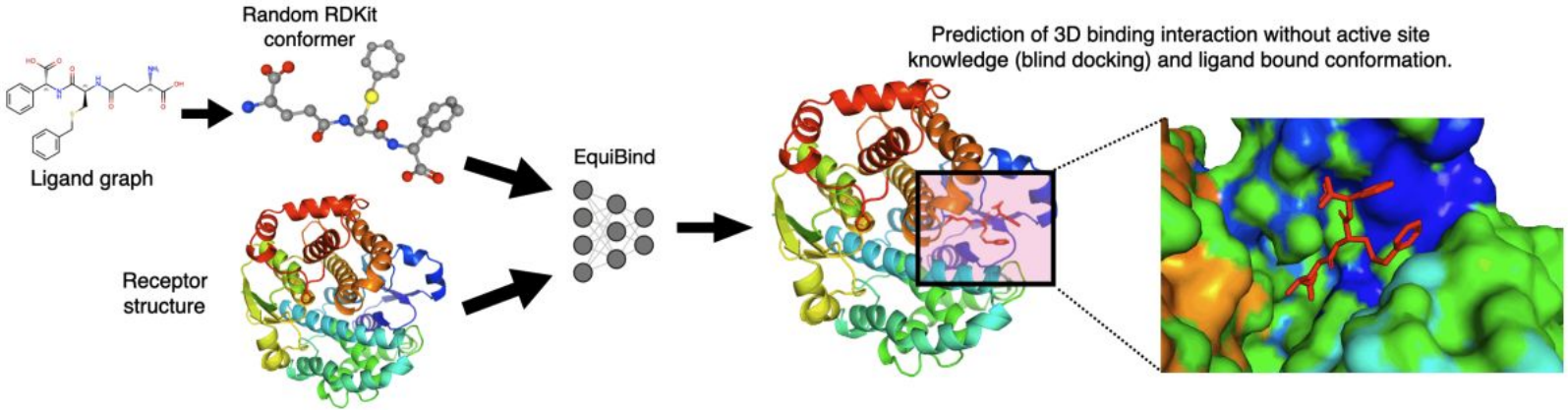
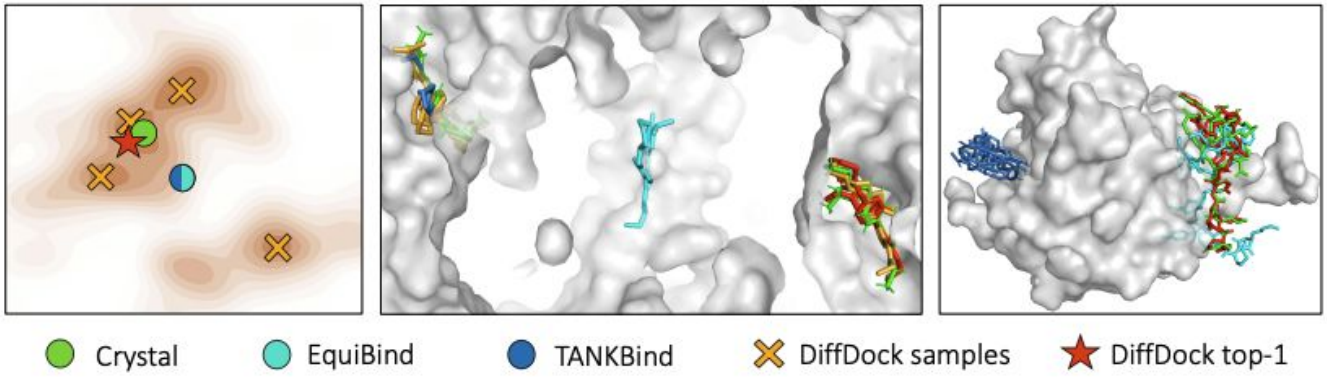


Figure 1. High-level overview of the structural drug binding problem tackled by EQUIBIND.

Stärk et al. (2022)

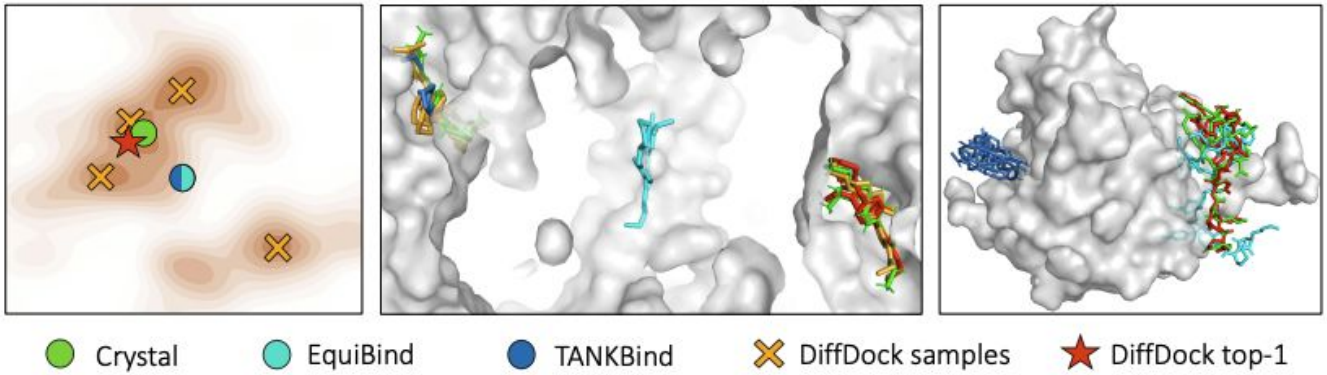
Regression Methods are not suited for Protein Docking

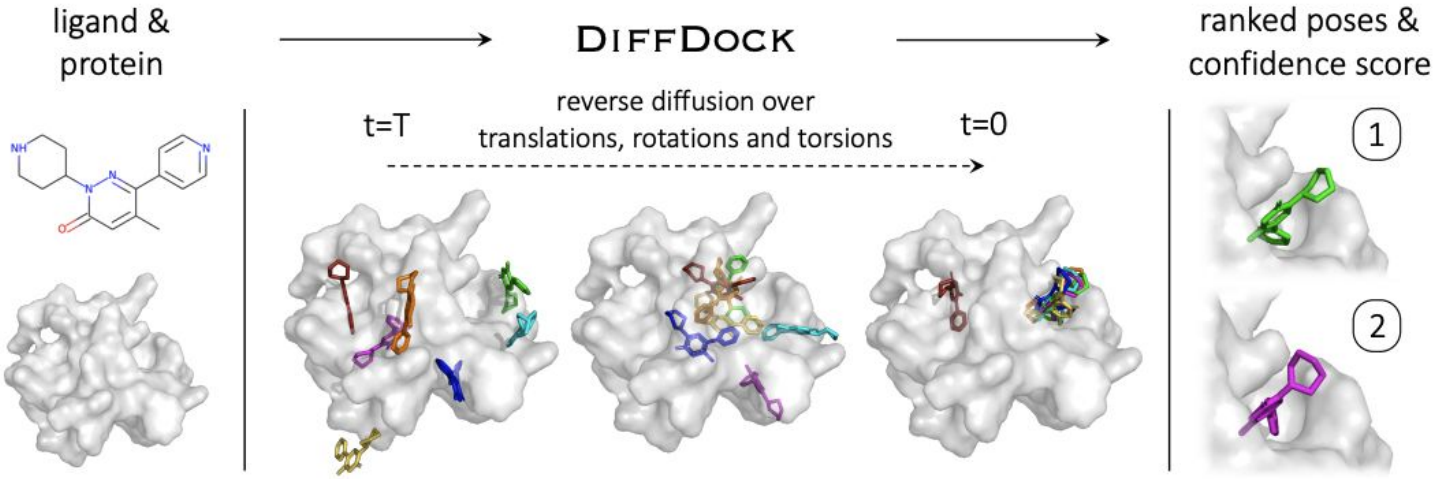
- Issues with data/problem formulation:
  - **Aleatoric Uncertainty:** ligand might bind with multiple poses to the protein
  - **Epistemic Uncertainty:** limited model is unsuitable for complexity of docking (usually results in physically unrealistic output).
- Regression-style methods select a single configuration that minimizes the expected square error -> the mean of such distributions.



Generative Models for Blind Docking

- Generative models can learn to capture the distribution unlike the alternatives
- Able to sample all/most of the significant modes




ligand &
protein
DIFFDOCK
ranked poses &
confidence score
t=T
reverse diffusion over
translations, rotations and torsions
t=0
1
2

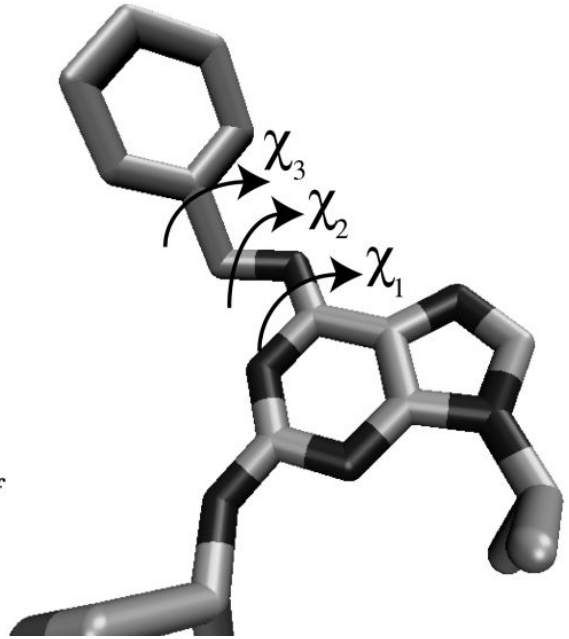
Objective

- **Traditional Docking Objective**: maximize percentage of predictions that have < eps with ground truth.
  - Not differentiable!
- **DiffDock Objective:** maximizing the likelihood of the true structure under the model's output distribution (in the limit as eps goes to 0)

Ligand Pose

- Ligand pose can be defined as atomic positions assignment in the R^{3n} dimension
- However, Ligands are relatively rigid, not completely independent atoms.

Any ligand pose consistent with a seed conformation can be reached by a combination of (1) ligand translations, (2) ligand rotations, and (3) changes to torsion angles.



Definition of Space of Ligand Poses

- x = R^{3n} ligand pose
- R = Rotation Matrix
- r = translation vector
- Theta = torsion vector (m rotatable bonds)
- c = seed pose confirmation
- M_c = Ligand Pose manifold conditioned on c

$$A((\mathbf{r}, R, \boldsymbol{\theta}), \mathbf{x}) = A_{\text{tr}}(\mathbf{r}, A_{\text{rot}}(R, A_{\text{tor}}(\boldsymbol{\theta}, \mathbf{x})))$$

$$\mathcal{M}_{\mathbf{c}} = \{A(g, \mathbf{c}) \mid g \in \mathbb{P}\}$$

Diffusion Details

- SDE: $d\mathbf{x} = \sqrt{d\sigma^2(t)/dt}\, d\mathbf{w}$ x = tr, rot, c
  motion
- Diffusion Kernels:
  - Translation Kernel (T(3)): sample and comp
  - Rotation Kernel (SO(3)): SO(3) kernel is give
  - Torsion Kernel((SO(2)^m): a wrapped normal

$$p(\omega) = \frac{1-\cos\omega}{\pi} f(\omega) \quad \text{where} \quad f(\omega) = \sum_{l=0}^{\infty} ($$

Product Space Diffusion

Product space
$\mathbb{R}^3 \times SO(3) \times \mathbb{T}^m$

↓

Forward diffusions are **independent**.

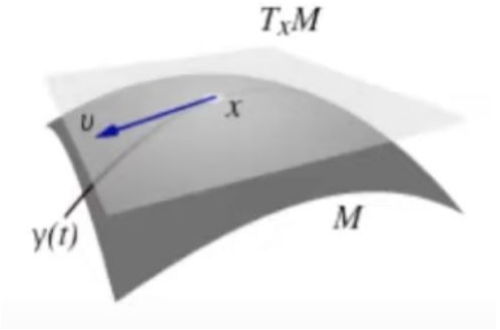
↓

Product tangent space is a **direct sum** of tangent spaces.

Space	$\mathbb{R}^3$ (position)
Tangent space	$\mathbb{R}^3$ (translation vectors)
Heat kernel	Normal
Stationary dist.	Normal
SE(3) symmetry	Equivariant

Steps

Two Neural Components of DiffDock: Score Model + Confidence Model



- Inputs:
  - x = ligand, y = protein
- Score Model s(x,y,t)
  - Predicts tangent space “score” - two SE(3)-equivariant vectors for the ligand as a whole and an SE(3)-invariant scalar at each of the m freely rotatable bonds.
  - Coarse grained: operates on α-carbon
- Confidence Model d(x,y)
  - Outputs scalar confidence value
  - Fine grained: operates on all atom

Training the Diffusion Model

- Training Data (x*, y, c):
  - x* = ground-truth ligand pose
  - y = protein structure
  - c ∈ Mx*
- Training Procedure:
  - Don’t want to train using Mx* because of poor generalization
  - Create c ∈ Mx* using RDKit and train on Mc manifold
  - Replace x* ground truth with:

$$\arg\min_{\mathbf{x}^\dagger \in \mathcal{M}_\mathbf{c}} \mathrm{RMSD}(\mathbf{x}^\star, \mathbf{x}^\dagger)$$

Score Model

Parameter	Search Space
Using all atoms for the protein graph	Yes, **No**
Using language model embeddings	**Yes**, No
Using ligand hydrogens	Yes, **No**
Using exponential moving average	**Yes**, No
Maximum number of neighbors in protein graph	10, 16, **24**, 30
Maximum neighbor distance in protein graph	5, 10, **15**, 18, 20, 30
Distance embedding method	**sinusoidal**, gaussian
Dropout	0, 0.05, **0.1**, 0.2
Learning rates	0.01, 0.008, 0.003, **0.001**, 0.0008, 0.0001
Batch size	8, **16**, 24
Non linearities	**ReLU**
Convolution layers	6
Number of scalar features	48
Number of vector features	10

Confidence Model

PARAMETER	SEARCH SPACE
USING ALL ATOMS FOR THE PROTEIN GRAPH	**YES**, NO
USING LANGUAGE MODEL EMBEDDINGS	**YES**, NO
USING LIGAND HYDROGENS	**NO**
USING EXPONENTIAL MOVING AVERAGE	**NO**
MAXIMUM NUMBER OF NEIGHBORS IN PROTEIN GRAPH	10, 16, **24**, 30
MAXIMUM NEIGHBOR DISTANCE IN PROTEIN GRAPH	5, 10, **15**, 18, 20, 30
DISTANCE EMBEDDING METHOD	**SINUSOIDAL**
DROPOUT	0, 0.05, **0.1**, 0.2
LEARNING RATES	0.03, 0.003, **0.0003**, 0.00008
BATCH SIZE	**16**
NON LINEARITIES	**RELU**
CONVOLUTION LAYERS	5
NUMBER OF SCALAR FEATURES	24
NUMBER OF VECTOR FEATURES	6

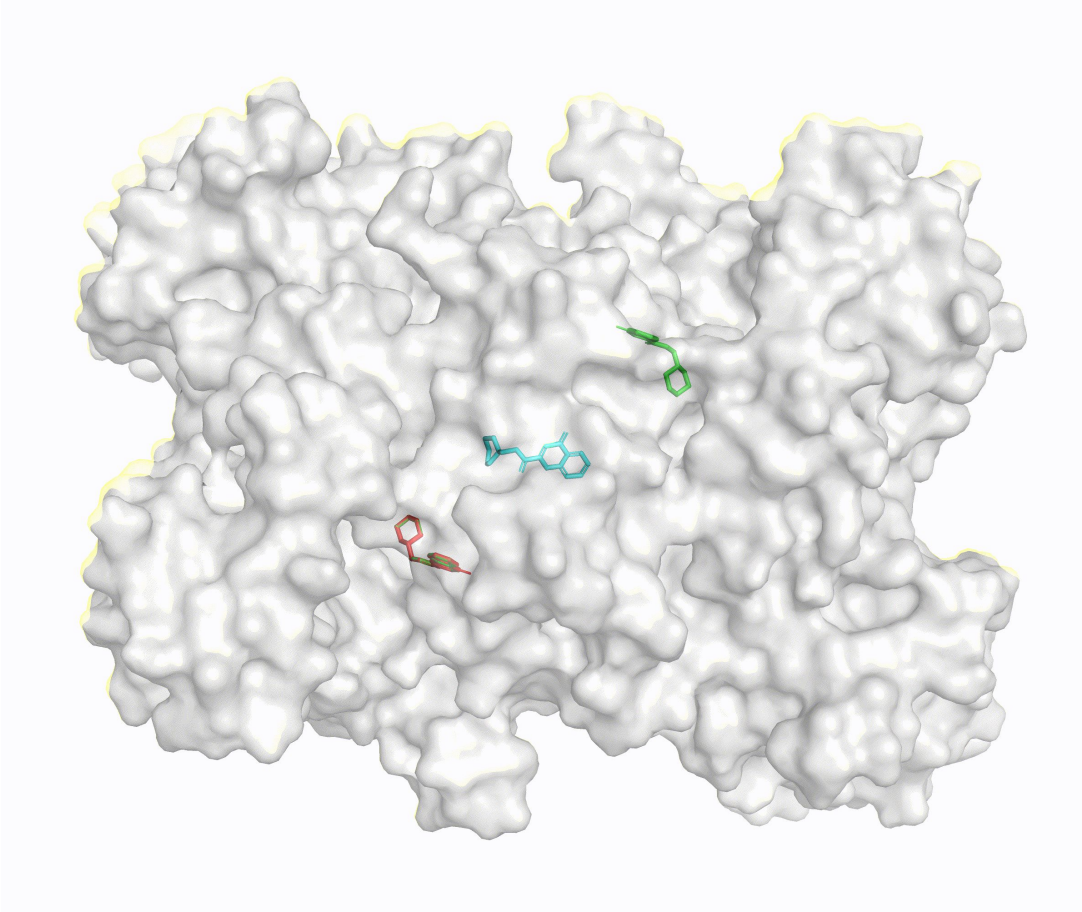
Training Confidence Model d(x,y)

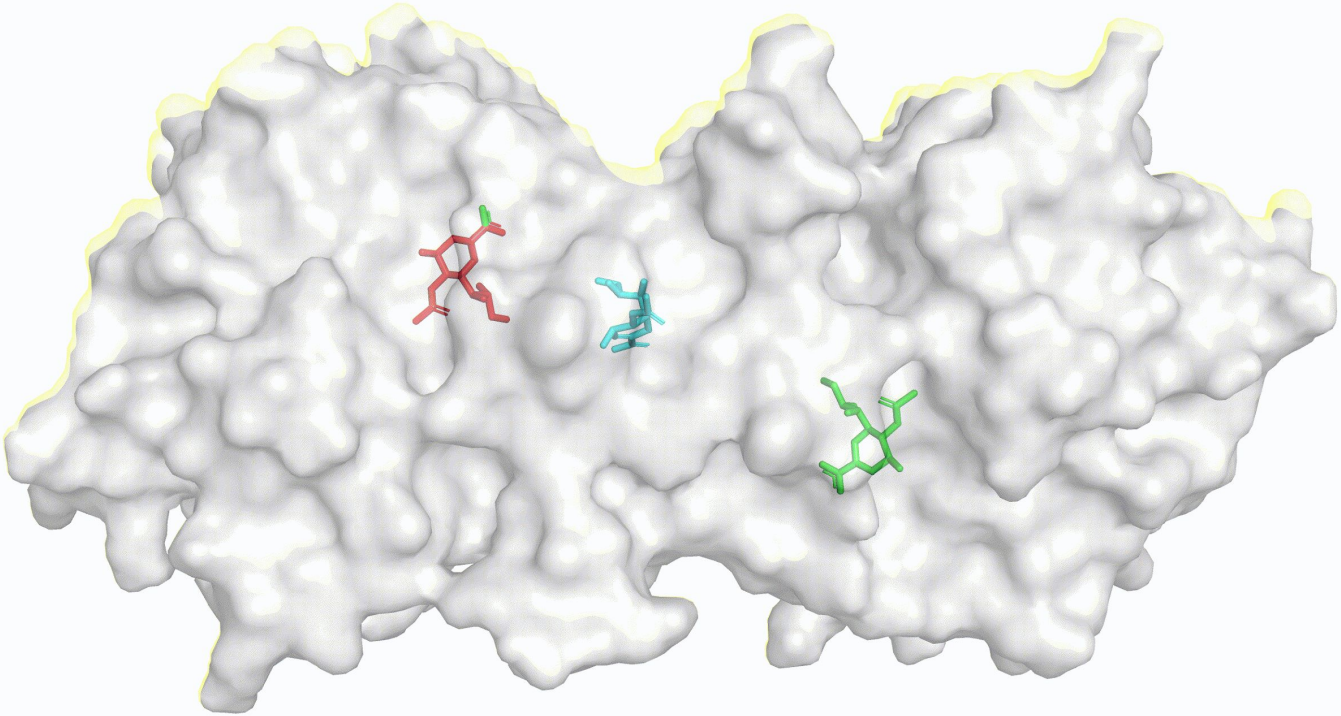
- Use trained diffusion model to generate poses for every training example.
- Generate binary labels based on which poses have RMSD below 2 Å
- Train Confidence Model (4.77 million parameters) on generated labels.

Workflow

- Generate N random poses
- Simulate Reverse Diffusion
- Rank and select top M poses

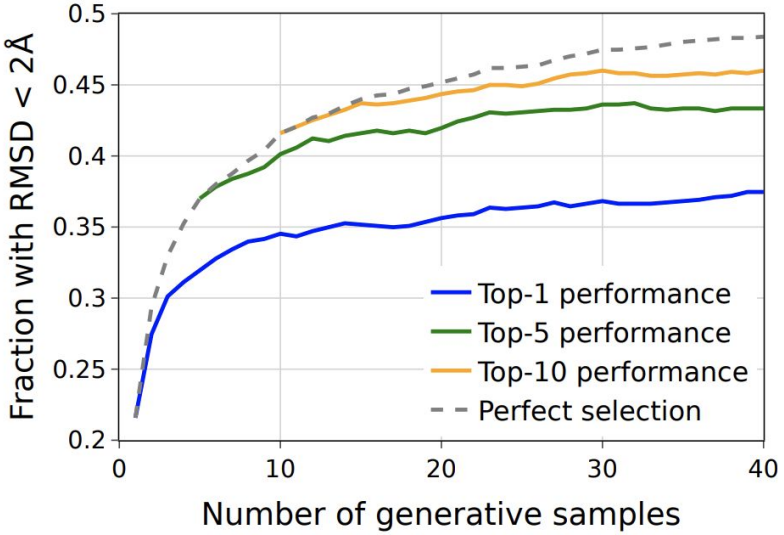
Results

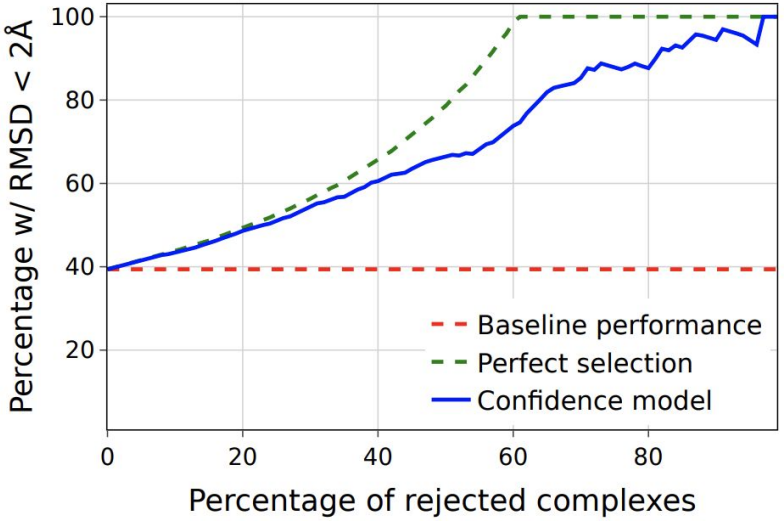




Results

Method	Holo crystal proteins Top-1 RMSD %<2	Holo crystal proteins Top-1 RMSD Med.	Holo crystal proteins Top-5 RMSD %<2	Holo crystal proteins Top-5 RMSD Med.	Apo ESMFold proteins Top-1 RMSD %<2	Apo ESMFold proteins Top-1 RMSD Med.	Apo ESMFold proteins Top-5 RMSD %<2	Apo ESMFold proteins Top-5 RMSD Med.	Average Runtime (s)
GNINA	22.9	7.7	32.9	4.5	2.0	22.3	4.0	14.22	127
SMINA	18.7	7.1	29.3	4.6	3.4	15.4	6.9	10.0	126*
GLIDE	21.8	9.3							1405*
EQUIBIND	5.5	6.2	-	-	1.7	7.1	-	-	0.04
TANKBIND	20.4	4.0	24.5	3.4	10.4	5.4	14.7	4.3	0.7/2.5
P2RANK+SMINA	20.4	6.9	33.2	4.4	4.6	10.0	10.3	7.0	126*
P2RANK+GNINA	28.8	5.5	38.3	3.4	8.6	11.2	12.8	7.2	127
EQUIBIND+SMINA	23.2	6.5	38.6	3.4	4.3	8.3	11.7	5.8	126*
EQUIBIND+GNINA	28.8	4.9	39.1	3.1	10.2	8.8	18.6	5.6	127
DIFFDOCK (10)	35.0	3.6	40.7	2.65	**21.7**	**5.0**	**31.9**	**3.3**	10
DIFFDOCK (40)	**38.2**	**3.3**	**44.7**	**2.40**	20.3	5.1	31.3	**3.3**	40


Fraction with RMSD < 2Å
0.5
0.45
0.4
0.35
0.3
0.25
0.2
0
10
20
30
40
Number of generative samples
Top-1 performance
Top-5 performance
Top-10 performance
Perfect selection


Percentage w/ RMSD < 2Å
100
80
60
40
20
0
20
40
60
80
Percentage of rejected complexes
Baseline performance
Perfect selection
Confidence model

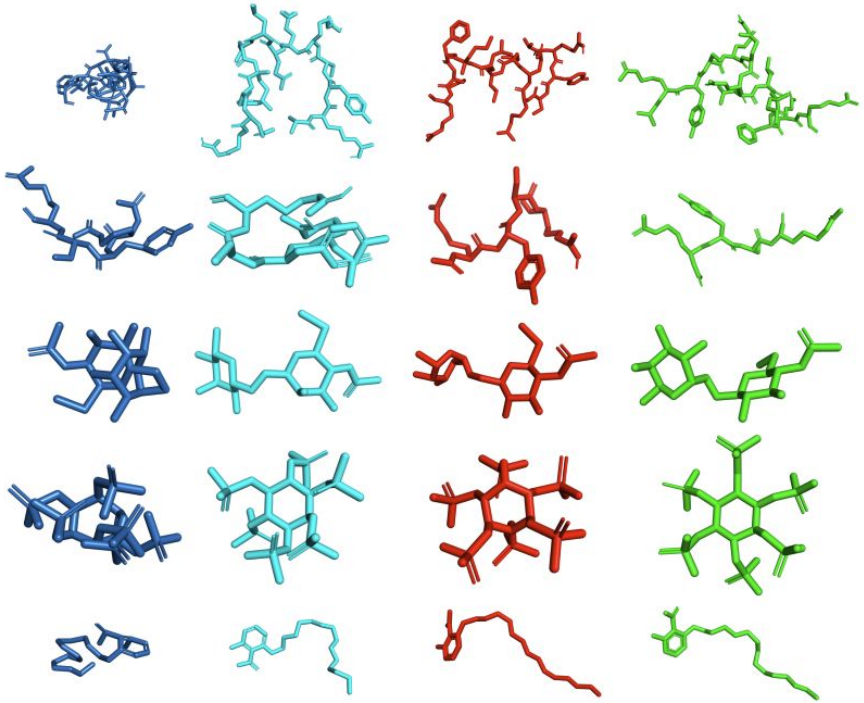


Figure 4: **Ligand self-intersections.** TANKBind (blue), EquiBind (cyan), DIFFDOCK (red), and crystal structure (green). Due to the averaging phenomenon that occurs when epistemic uncertainty is present, the regression-based deep learning models tend to produce ligands with atoms that are close together, leading to self-intersections. DIFFDOCK, as a generative model, does not suffer from this averaging phenomenon, and we never found a self-intersection in any of the investigated results of DIFFDOCK.

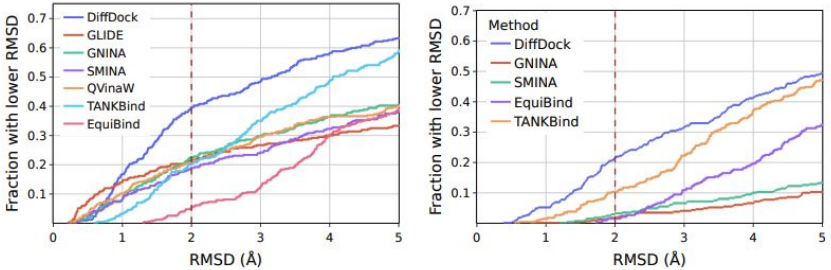


Figure 6: Cumulative density histogram of the methods' RMSD: left on holo crystal structures, right on apo ESMFold structures.

Method	Top-1 RMSD (Å) %<2	Top-1 RMSD (Å) Med.	Top-5 RMSD (Å) %<2	Top-5 RMSD (Å) Med.	Average Runtime (s)
DIFFDOCK-SMALL-NOESM (10)	26.2	4.7	32.0	3.2	7
DIFFDOCK-SMALL-NOESM (40)	28.4	3.8	37.7	2.6	28
DIFFDOCK-SMALL (10)	26.0	4.3	33.3	3.2	7
DIFFDOCK-SMALL (40)					
DIFFDOCK-NOESM (10)					
DIFFDOCK-NOESM (40)					
DIFFDOCK (10)					
DIFFDOCK (40)					

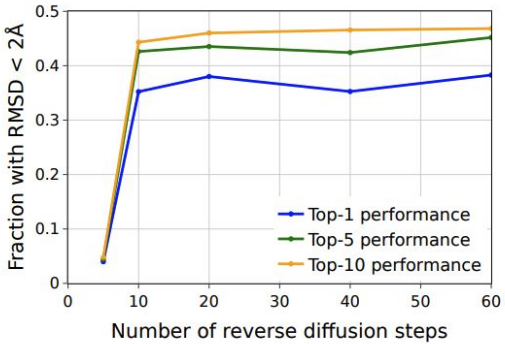
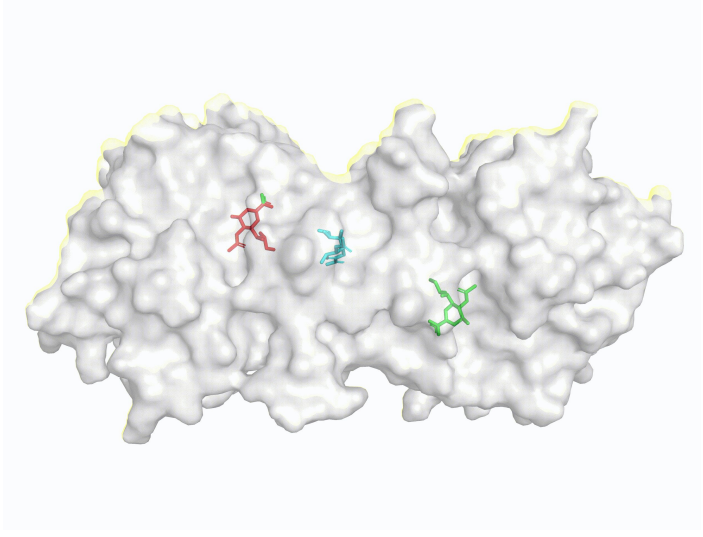


Figure 11: Ablation study on the number of reverse diffusion steps.

Takeaway

- Diffdock achieves 38% top-1 success rate (RMSD<2A) on PDB-Bind, significantly outperforming the previous state-of-the-art of traditional docking (23%)
- 3 to 12 times faster than the best search-based method (previous state of the art)!



Generalized Biomolecular Modeling and Design with RoseTTAFold All-Atom

Rohith Krishna[1,2‡], Jue Wang[1,2‡], Woody Ahern[1,2,3‡], Pascal Sturmfels[1,2,3], Preetham Venkatesh[1,2,4°], Indrek Kalvet[1,2,7°], Gyu Rie Lee[1,2,7°], Felix S Morey-Burrows[5], Ivan Anishchenko[1,2], Ian R Humphreys[1,2], Ryan McHugh[1,2,4], Dionne Vafeados[1,2], Xinting Li[1,2], George A Sutherland[5], Andrew Hitchcock[5], C Neil Hunter[5], Minkyung Baek[6], Frank DiMaio[1,2], David Baker[1,2,7*]

bioRxiv 2023

Presented by Victor Chu and Howard Yen

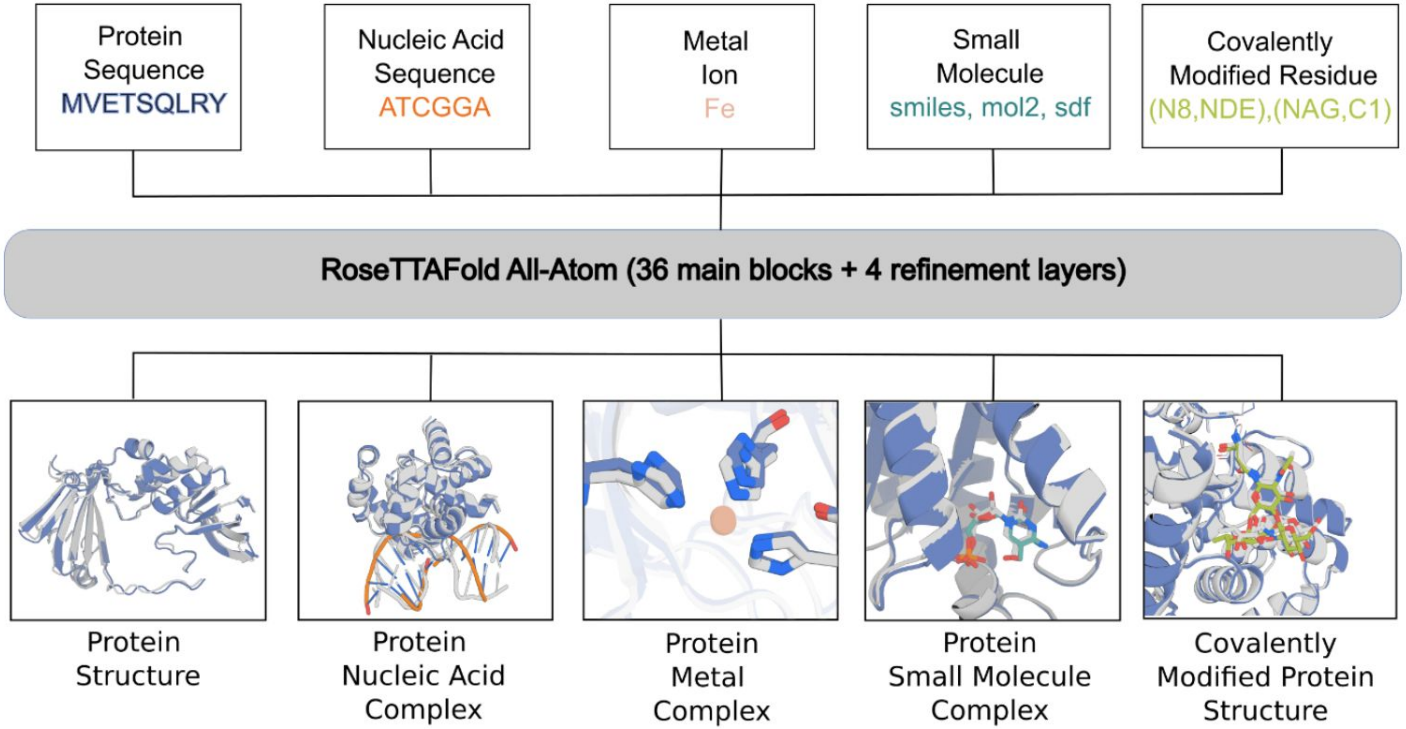
Outline

1. Introduction
2. RoseTTAFold All-Atom – dataset and training
3. Binding evaluation
4. Structure prediction
5. Experimental results
6. Takeaways

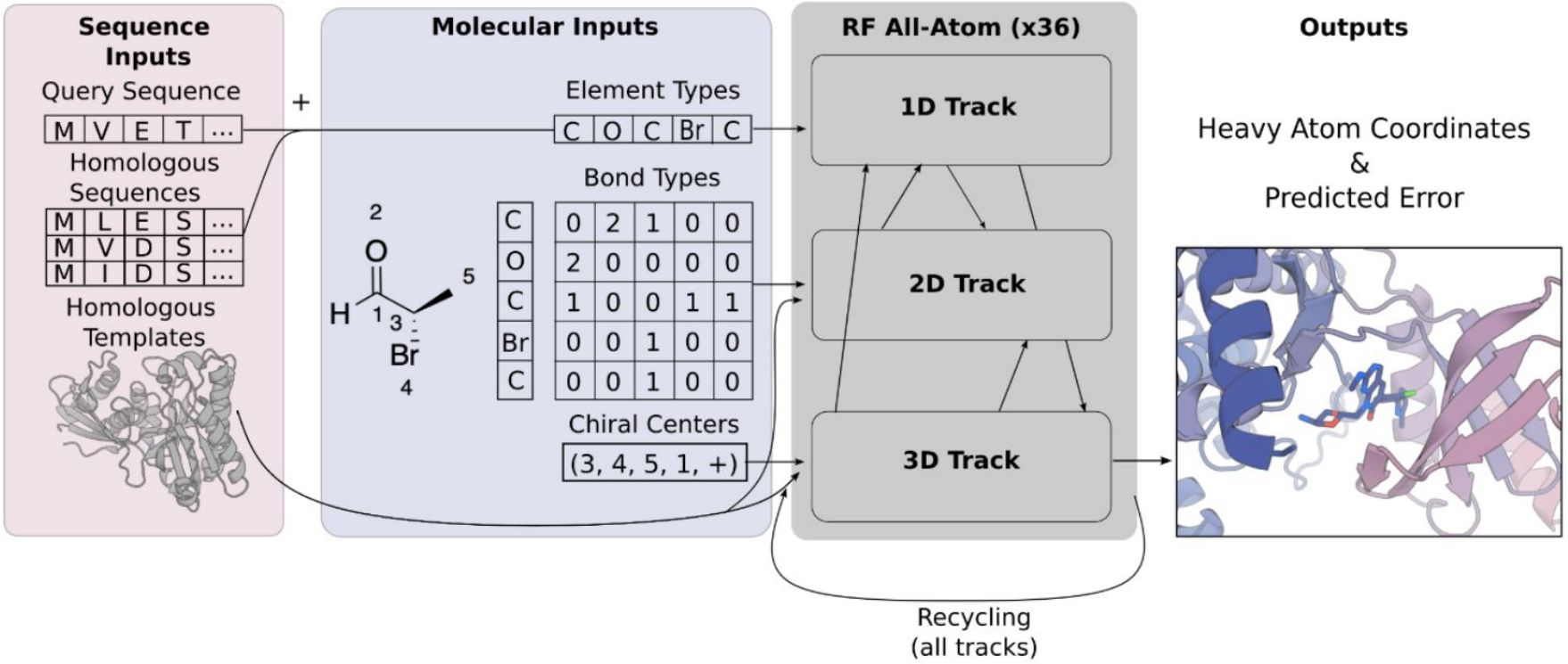
Introduction

- AlphaFold2 and RoseTTAFold allow for protein structure prediction using only the amino acid sequences
- BUT proteins are rarely alone – they interact with other proteins & molecules
- This interaction is important for applications such as drug design
- **How to model multiple molecular structures jointly?**

RoseTTAFold All-Atom



Inputs are mostly similar to RF2



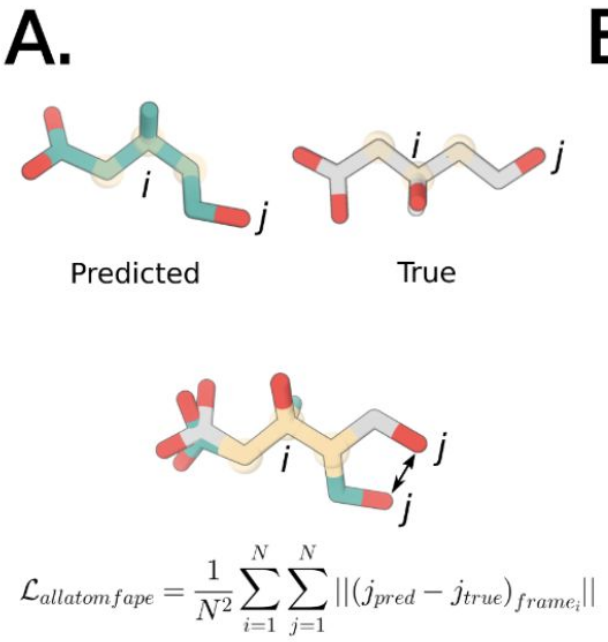
Training RFAA – Dataset

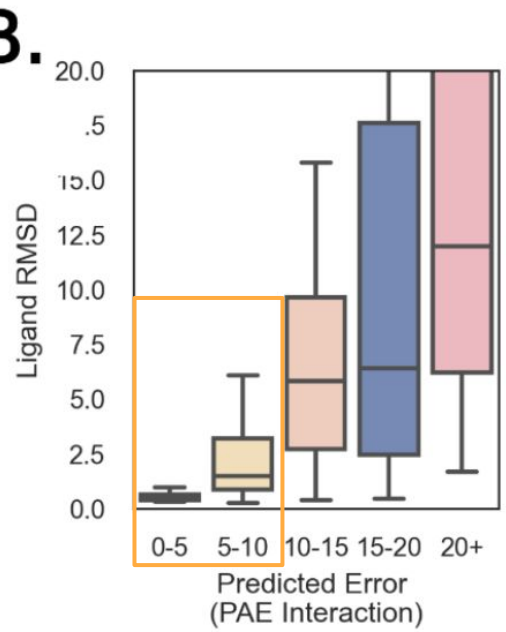
- PDB ⟶ for learning complex structures
  - protein-small molecule, protein-metal, and covalently modified protein complexes
  - 30% sequence identity clustering + some common filtering
- Cambridge Structural Database ⟶ for learning the general properties of small molecules
  - small crystal structures of organic non-polymeric molecules
- Atomize residues ⟶ for learning generic interactions
  - Randomly atomize residues in the protein
- Training on multiple modalities/tasks ⟶ better generalization!

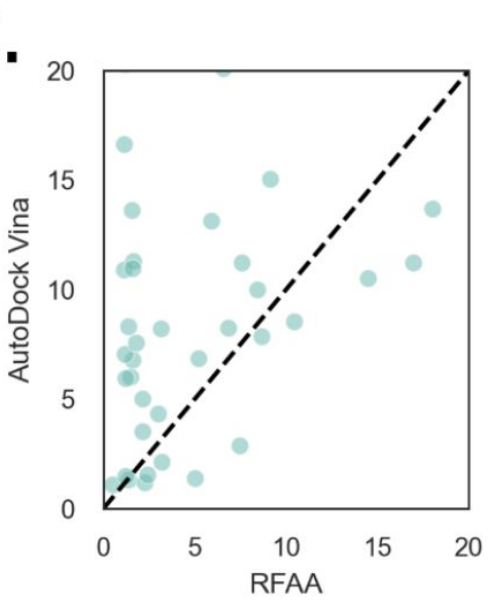
CAMEO – Continuous Automated Model Evaluation

- Commonly used for ligand docking evaluation
- Carries out predictions using structures submitted to the PDB

CAMEO Results



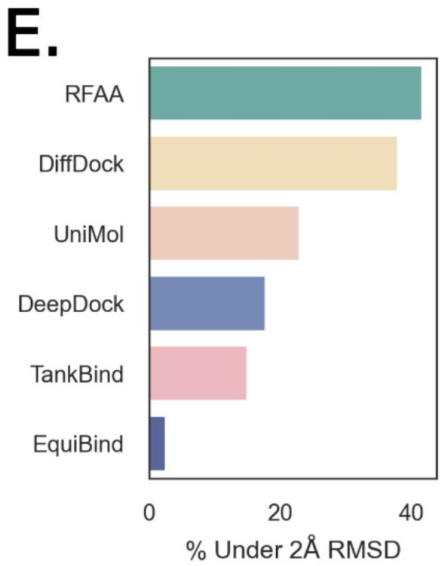


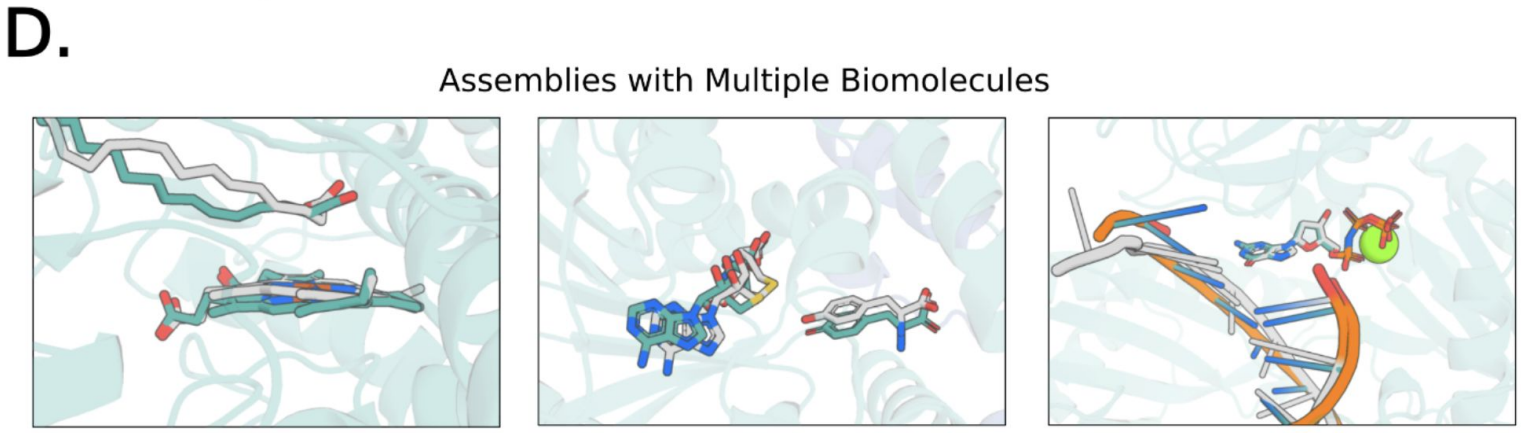


77% of the high confidence predictions have an error of < 2Å RMSD

CAMEO Results

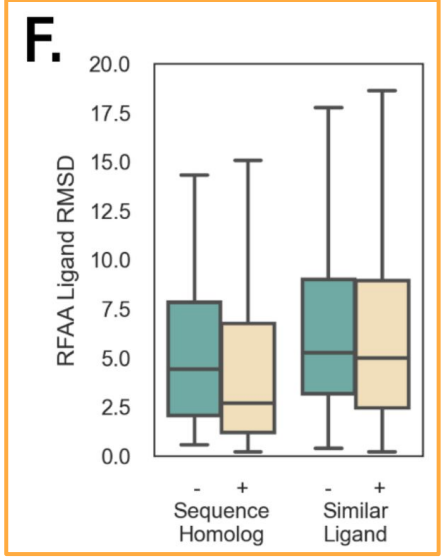
RFAA is good at assembling multiple biomolecules, much better than previous methods

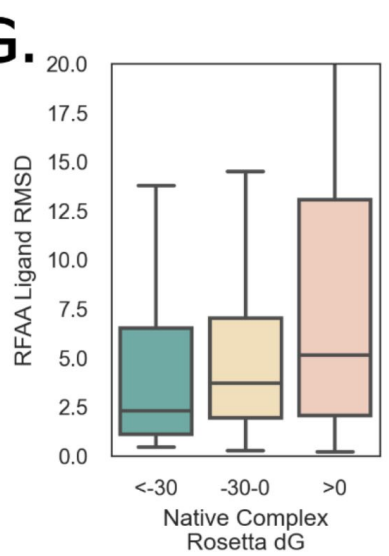


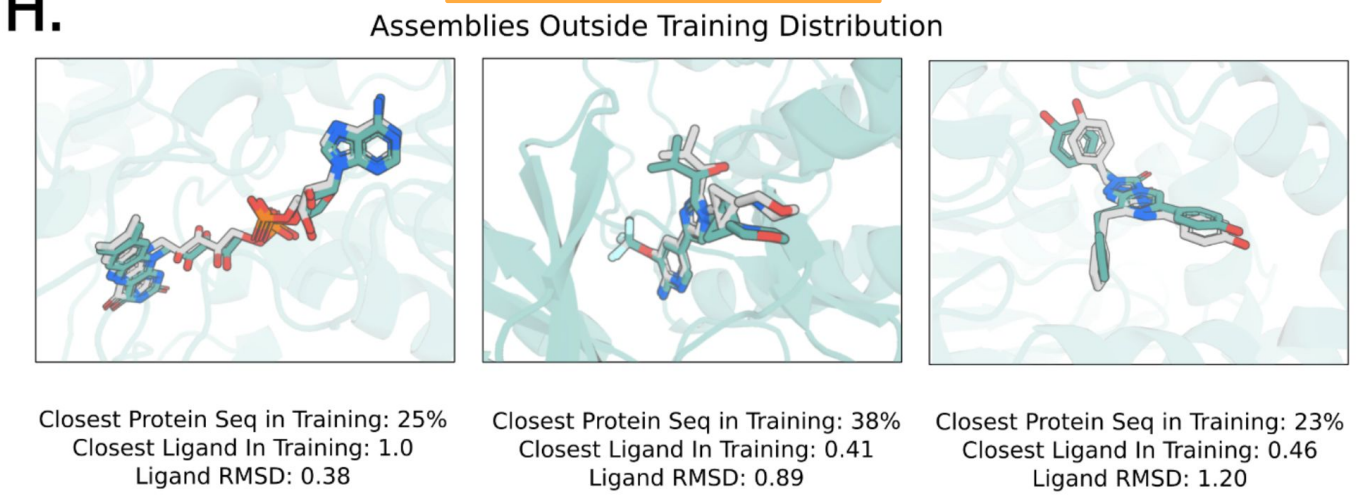


CAMEO Results

RFAA generalizes to new proteins with low similarity with the training set

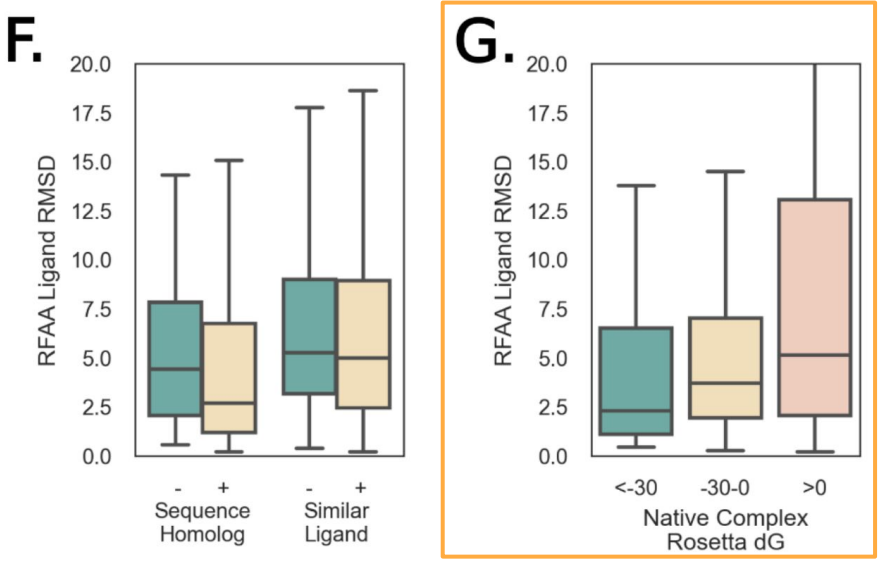


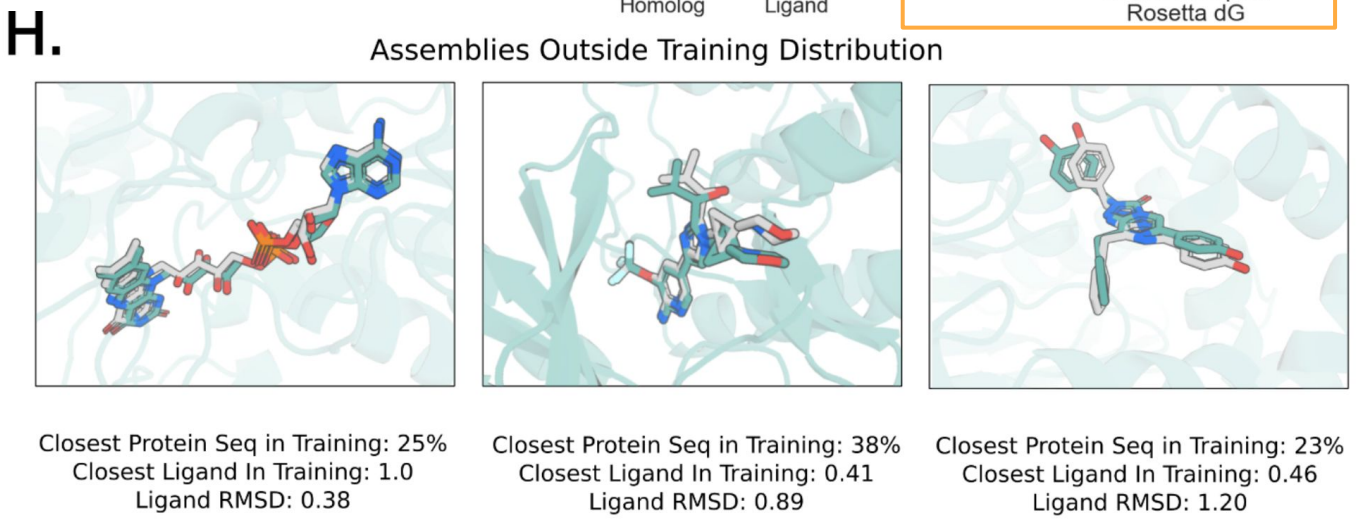




CAMEO Results

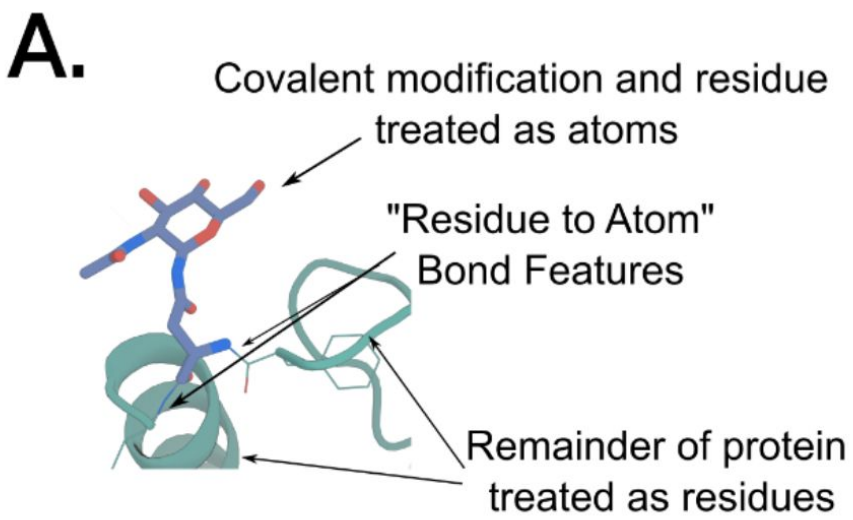
RFAA makes more accurate predictions for complexes with low Rosetta energy → correlations with physics principles



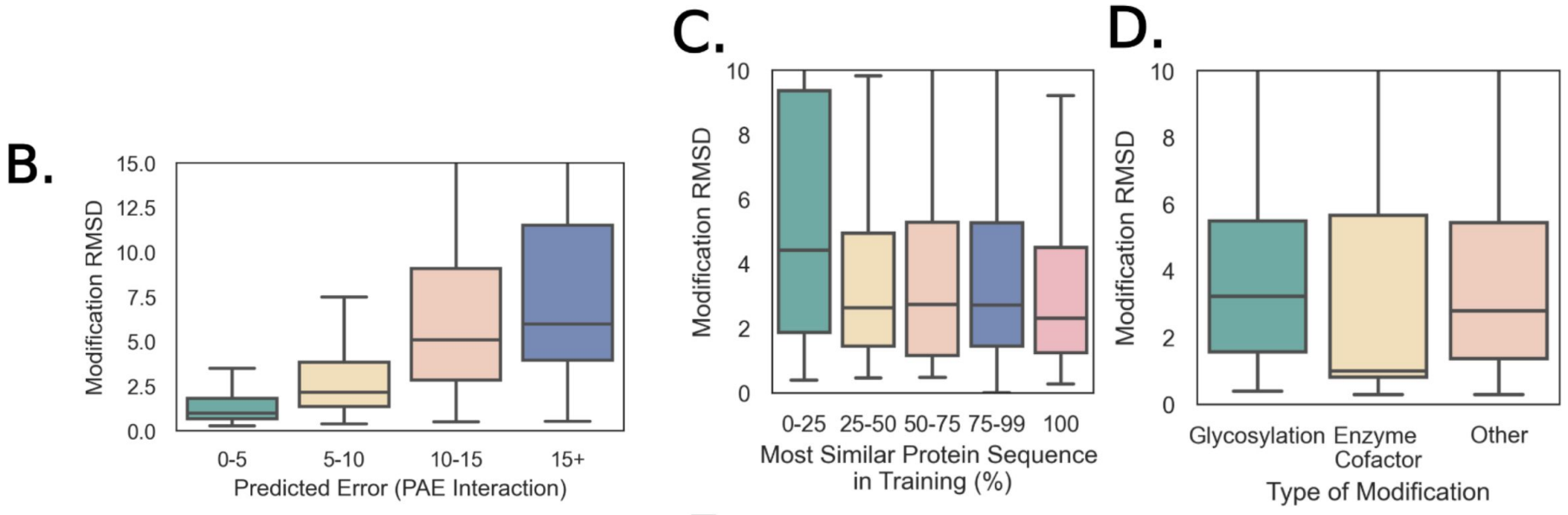


Predicting Structures of Covalent Modifications to Proteins

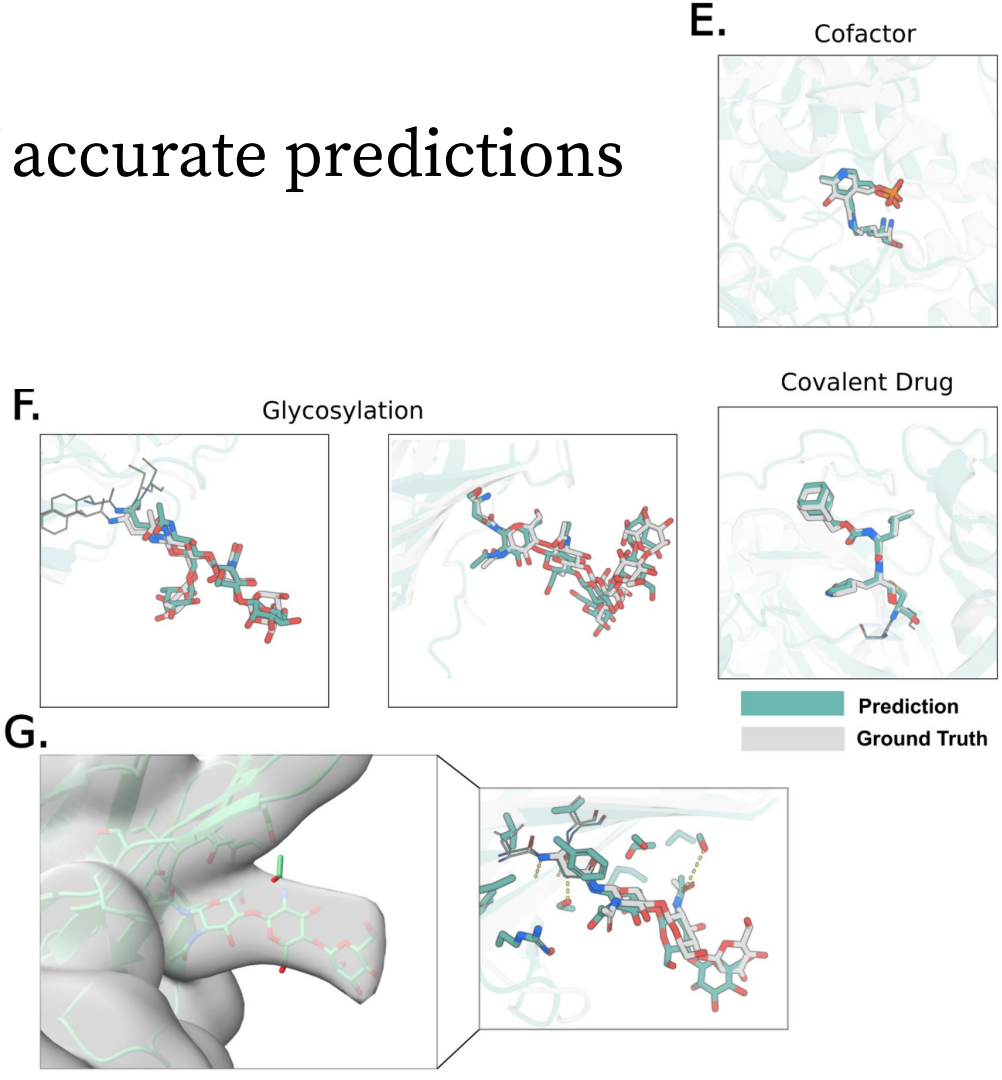
Amino acid side chains are modified with sugars, phosphates, lipids, and other molecules in many essential protein functions



Higher confidence/overlap with training → better performance

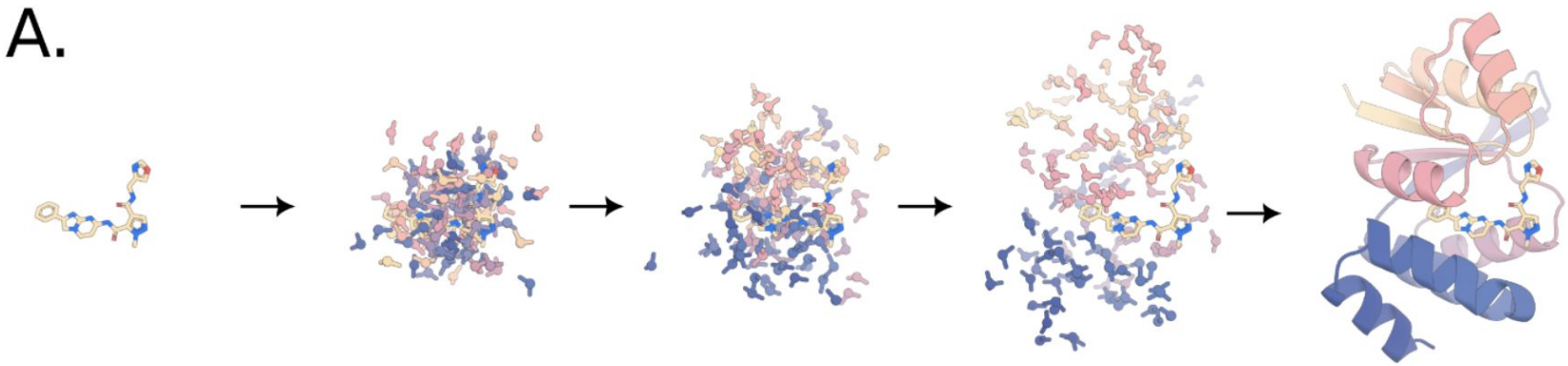


Examples of accurate predictions

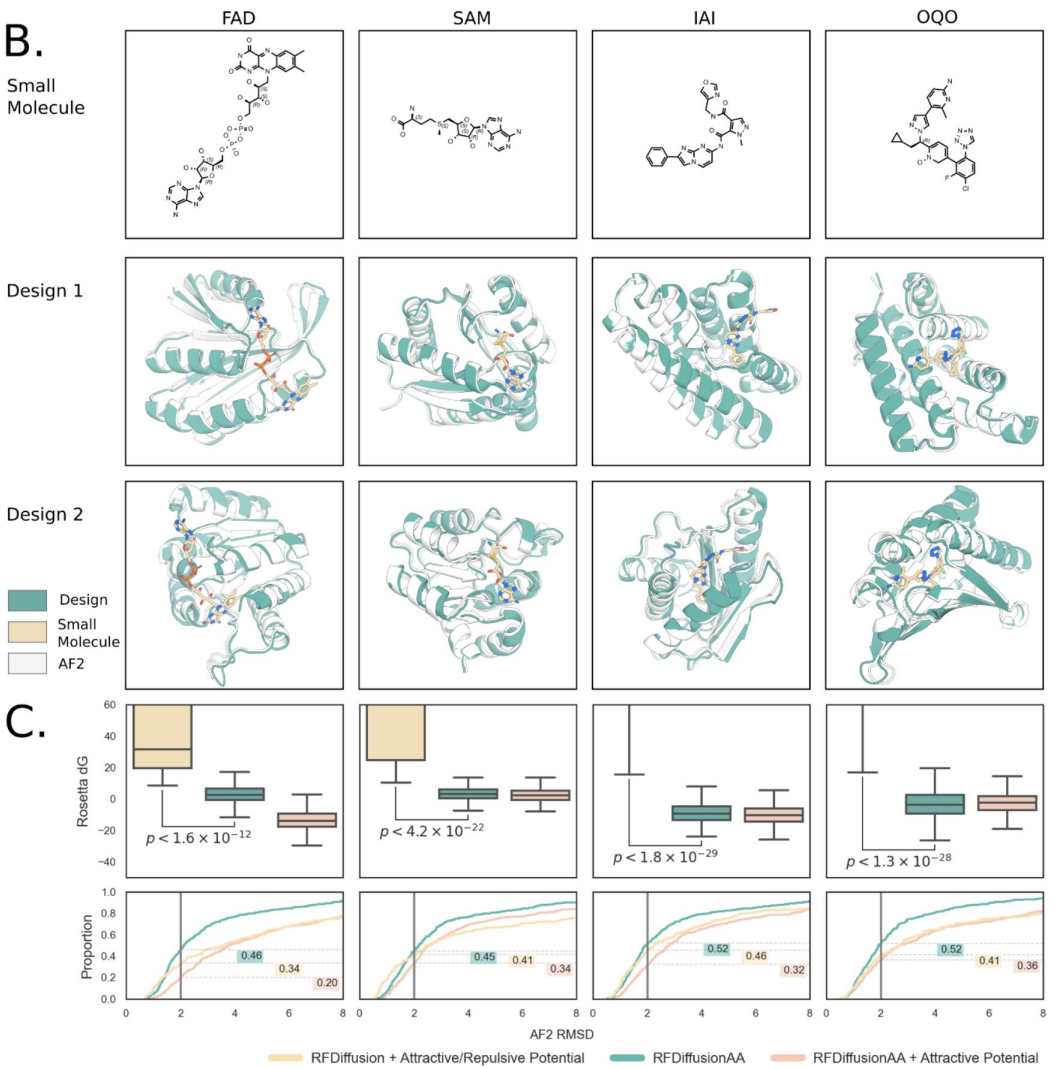


De Novo Small Molecule Binder Design

- Similar to the goal of DiffDock, we are interested in generating/designing proteins that binds small molecules
- Developed RFDiffusion All-Atom on top of RFAA

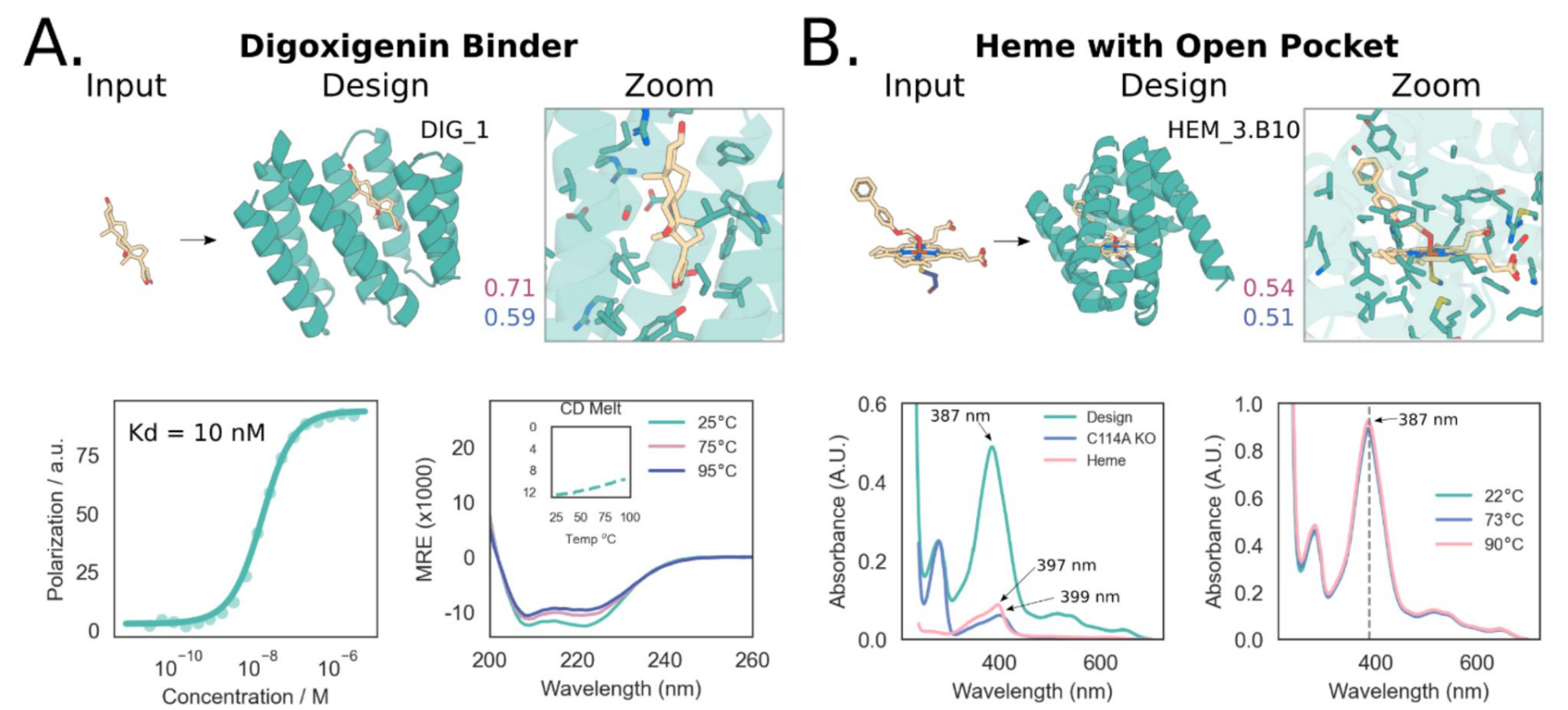


RFdiffusionAA produces protein-ligand interfaces with a lower computed ΔG than RFdiffusion



Binding Experiments

- Designed binders for three diverse small molecules
  - no protein motif, a single protein motif, a four residue protein motif
- Then measured the binding experimentally



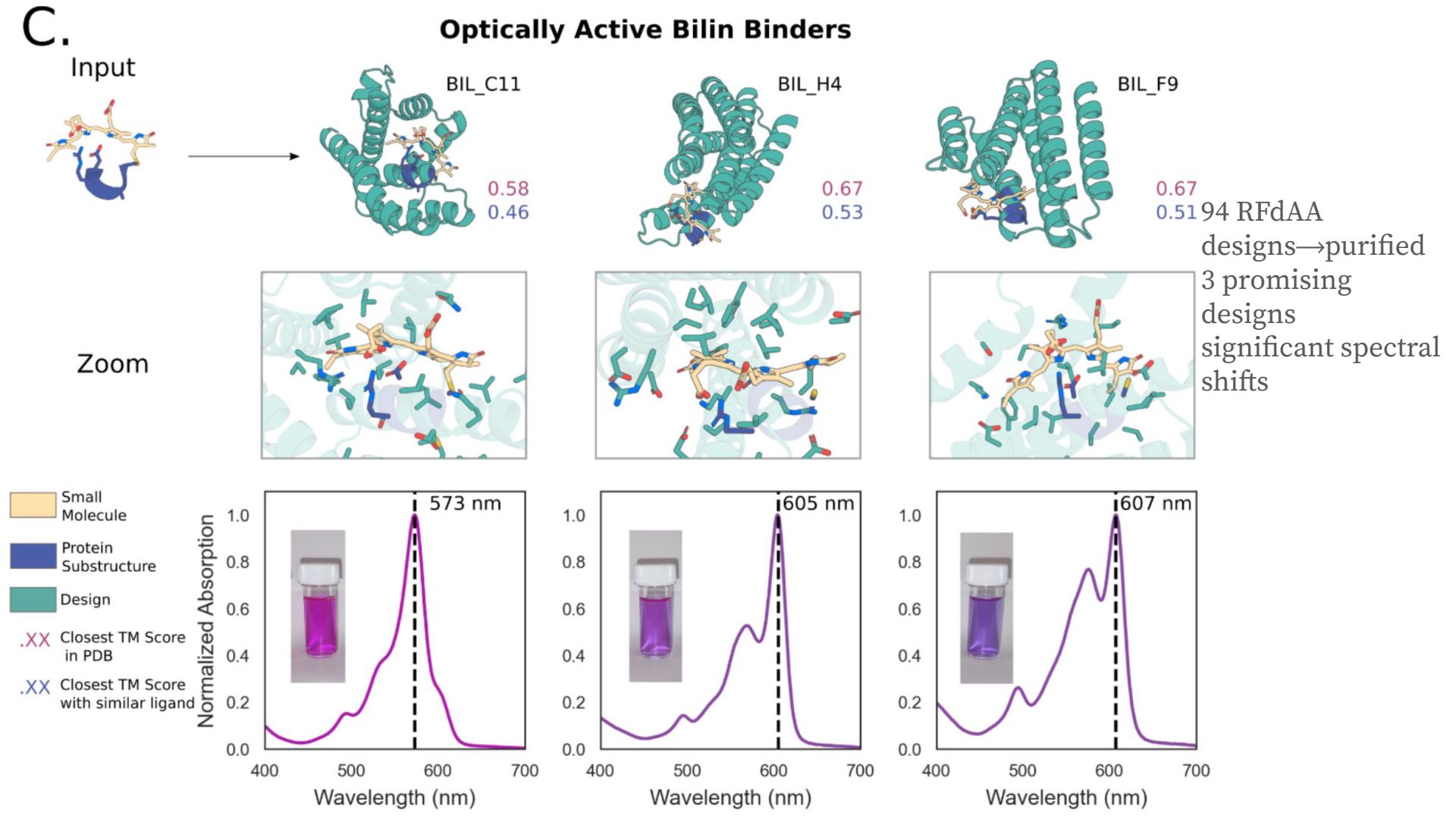
Small Molecule
Protein Substructure
Design
.XX Closest TM Score in PDB
.XX Closest TM Score with similar ligand

4,416 designs were selected
the tightest binder was stable
at temp up to 95℃

168 designs selected ⟶ 135 expressed ⟶ 96 UV/Vis spectra consistent with CYS-bound heme. 38 were monomeric and retained heme-binding through SEC

C.

Optically Active Bilin Binders



94 RFdAA designs→purified 3 promising designs significant spectral shifts

Takeaways

1. RFAA demonstrates significant ability to work on proteins with low overlap with the training set → good generalization
2. RFAA predictions are highly correlated with physical chemistry characteristics
3. RFAA can be extended to ALL-ATOMS