Manifold Learning and Artificial Intelligence

Lecture 16

Diffusion Model for Molecular Docking

Momiao Xiong, University of Texas School of Public Health

- Time: 10:00 pm, US East Time, 05/27/2023
- 10:00 am, Beijing Time. 05/28/2023

Github Address: https://ai2healthcare.github.io/

Procedure for Joining the Meeting

Join from PC, Mac, Linux, iOS or Android: Click Here to Join
https://uwmadison.zoom.us/w/93316139423?tk=wfbmsTfN2fgERto_HI1WKtBzh94d3HO02XVRDCexqd8.DQMAAAAVuhNJnxZ2dGVCbllIYlR3ZWJBNHl5LXlYMjBnAAAAAAAAAAAAAAAAAAAAAAAAAAAA&pwd=Q0NVWFYvRFg5RmxCNkwxMmYrbW41dz09#success
Passcode: 416262
Note: This link should not be shared with others; it is unique to you.
Add to Calendar Add to Google Calendar Add to Yahoo Calendar

Or One tap mobile :
US: +12532158782,,93316139423# or +13017158592,,93316139423#
Or Telephone:
Dial(for higher quality, dial a number based on your current location):
US: +1 253 215 8782 or +1 301 715 8592 or +1 305 224 1968 or +1 309 205 3325 or +1 312 626 6799 or +1 346 248 7799 or +1 360 209 5623 or +1 386 347 5053 or +1 507 473 4847 or +1 564 217 2000 or +1 646 931 3860 or +1 669 444 9171 or +1 669 900 6833 or +1 689 278 1000 or +1 719 359 4580 or +1 929 205 6099 or +1 253 205 0468
Meeting ID: 933 1613 9423
Passcode: 416262
International numbers available: https://uwmadison.zoom.us/u/aguj5kFFd

Or an H.323/SIP room system:
H.323: 162.255.37.11 (US West) or 162.255.36.11 (US East)
Meeting ID: 933 1613 9423
Passcode: 416262

SIP: 93316139423@zoomcrc.com
Passcode: 416262

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

Gabriele Corso et al. 2023, MIT

all code is available at https://github.com/gcorso/DiffDock.

Riemannian Score-Based Generative Modelling

Valentin De Bortoli et al. 2022

Concepts

- **Molecular docking:**

  binding structure of a small molecule ligand to a protein

  We instead frame molecular docking as a generative modeling problem and develop DIFFDOCK, a diffusion generative model over the non-Euclidean manifold of ligand poses.

  The biological functions of proteins can be **modulated by small molecule ligands (such as drugs) binding to them**

- **Traditional approaches for docking**

  scoring-functions that estimate the correctness of a proposed structure or pose,

  optimization algorithm that searches for the global maximum of the scoring function

- **frame molecular docking as a generative modeling problem—** given a ligand and target protein structure, we learn a distribution over ligand poses

DIFFDOCK, a diffusion generative model (DGM) over the space of ligand poses for molecular docking.

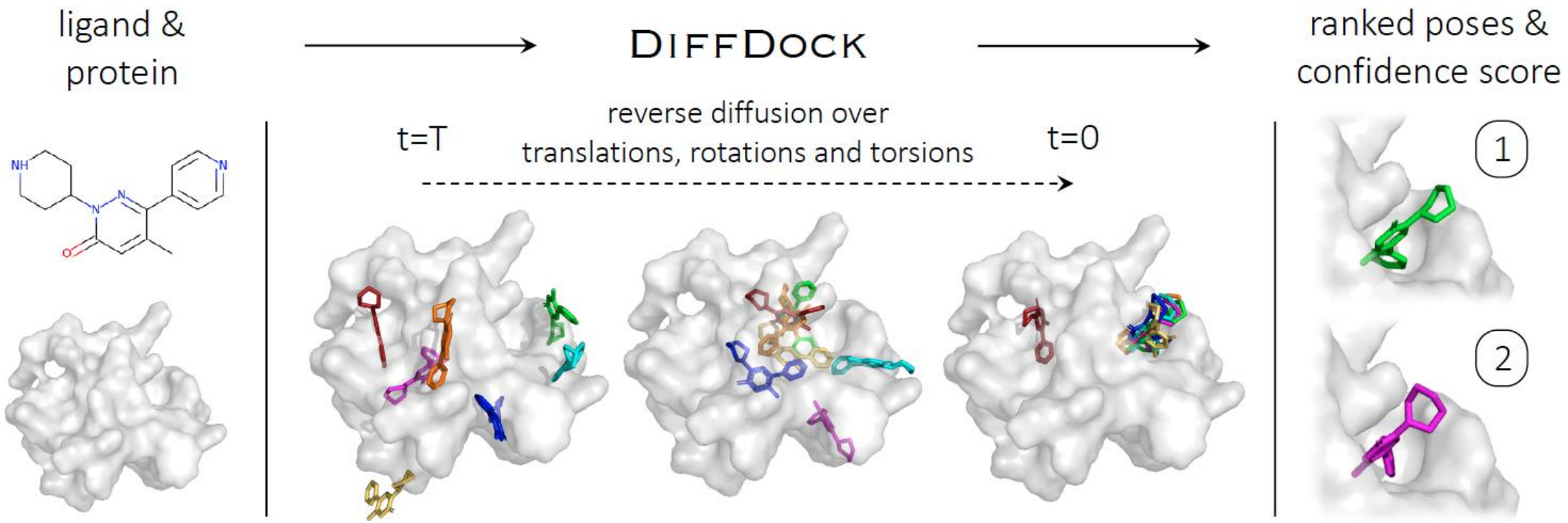


Figure 1: Overview of DIFFDOCK. Left: The model takes as input the separate ligand and protein structures. Center: Randomly sampled initial poses are denoised via a reverse diffusion over translational, rotational, and torsional degrees of freedom. Right:. The sampled poses are ranked by the confidence model to produce a final prediction and confidence score.

Diffusion generative models (DGMs) In Euclidean Space.

- **The initial distribution $P_0(x)$**
- **Forward diffusion process:**

$$dx = f(x,t)dt + g(t)dw$$

- **Reverse diffusion process:**

$$dx = [f(x,t) - g^2(t)\nabla_x \log P_t(x)]dt + g(t)dw$$

- **Assumption:**

$$f(x,t) = 0$$

DOCKING AS GENERATIVE MODELING

- **Molecular docking objective**

Concretely, a prediction is considered acceptable when the distance between the structures (measured in terms of ligand RMSD) is below some small tolerance on the order of the length scale of atomic interactions (a few A°). Consequently, the standard evaluation metric used in the field has been the percentage of predictions with a ligand RMSD (to the crystal ligand pose) below some value .

Thus, we view **molecular docking as the problem of learning a distribution over ligand poses conditioned on the protein structure** and **develop a diffusion generative model over this space**

- **Confidence model**

With a trained diffusion model, it is possible to sample an arbitrary number of ligand poses from the posterior distribution according to the model. However, researchers are **often interested in seeing only one or a small number of predicted poses and an associated confidence measure for downstream analysis**. Thus, we train a confidence model over the poses sampled by the diffusion model and rank them based on its confidence that they are within the error tolerance.

The top-ranked ligand pose and the associated confidence are then taken as DIFFDOCK's top-1 prediction and confidence score.

Problem with regression-based methods.

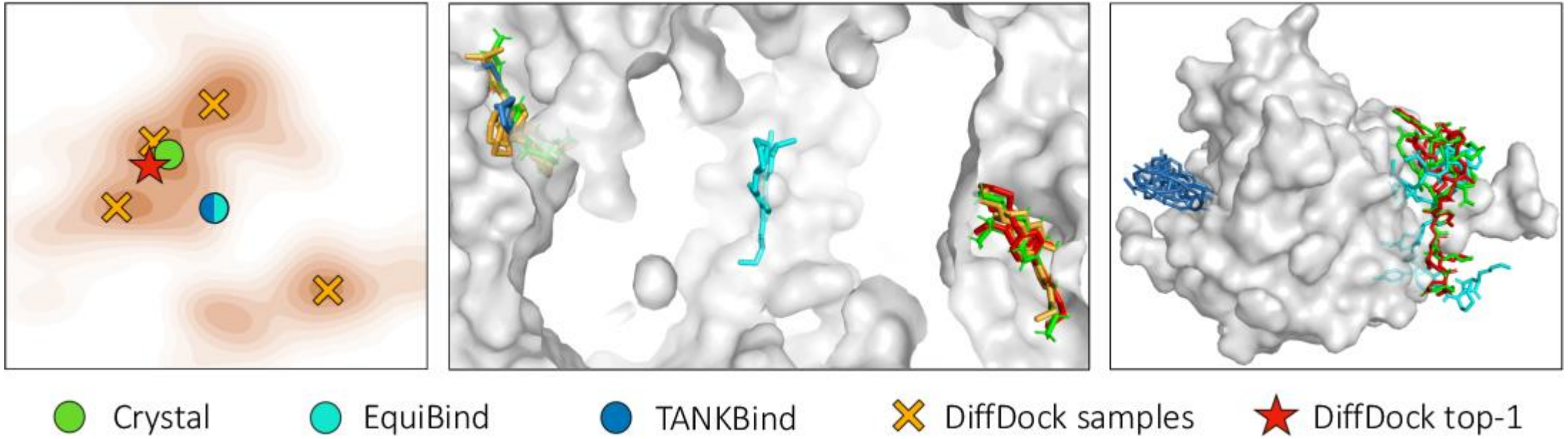
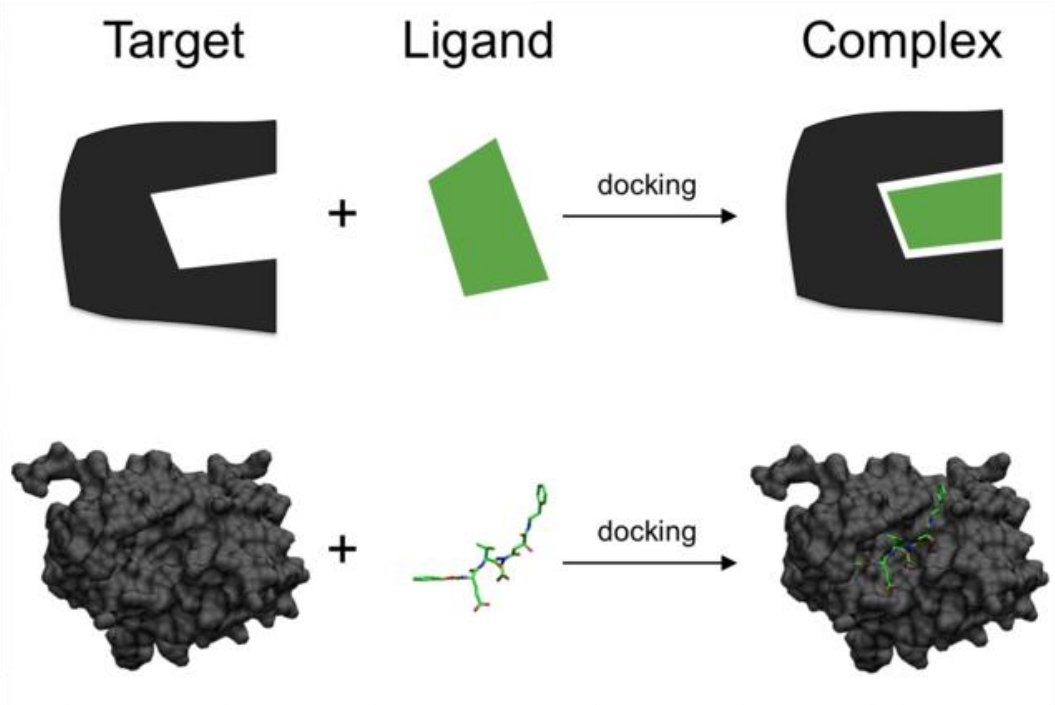
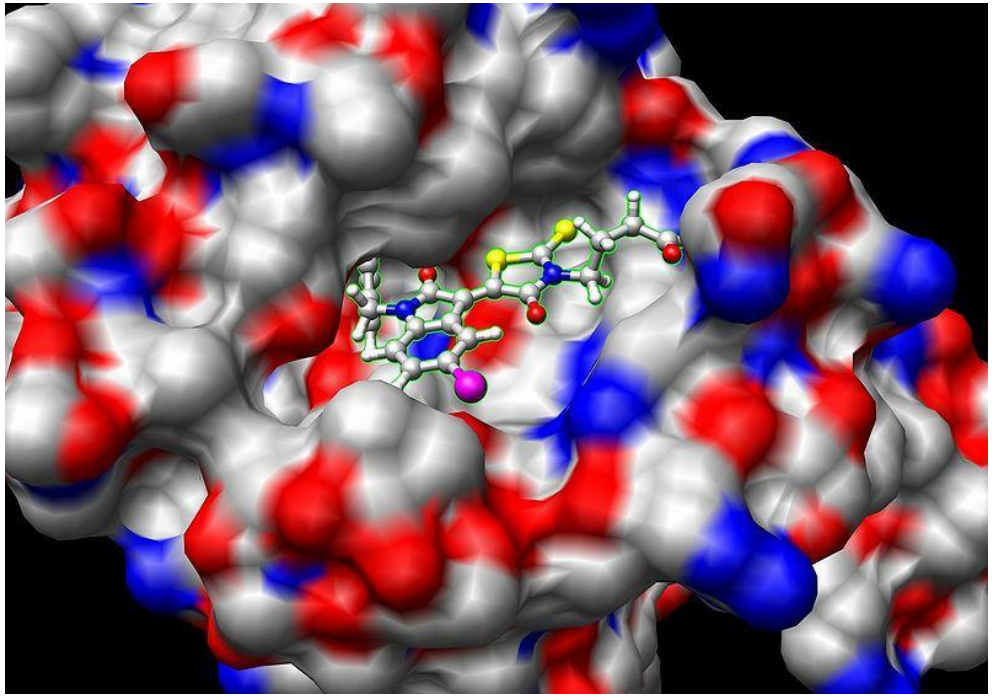


Figure 2: “DIFFDOCK top-1” refers to the sample with the highest confidence. “DIFFDOCK samples” to the other diffusion model samples. Left: Visual diagram of the advantage of generative models over regression models. **Given uncertainty in the correct pose (represented by the orange distribution),** regression models tend to predict the mean of the distribution, which may lie in a region of low density. Center: when there is a global symmetry in the protein (aleatoric uncertainty), EquiBind places the molecule in the center while DIFFDOCK is able to sample all the true poses. Right: even in the absence of strong aleatoric uncertainty, the epistemic uncertainty causes EquiBind’s prediction to have steric clashes and TANKBind’s to have many self-intersections.

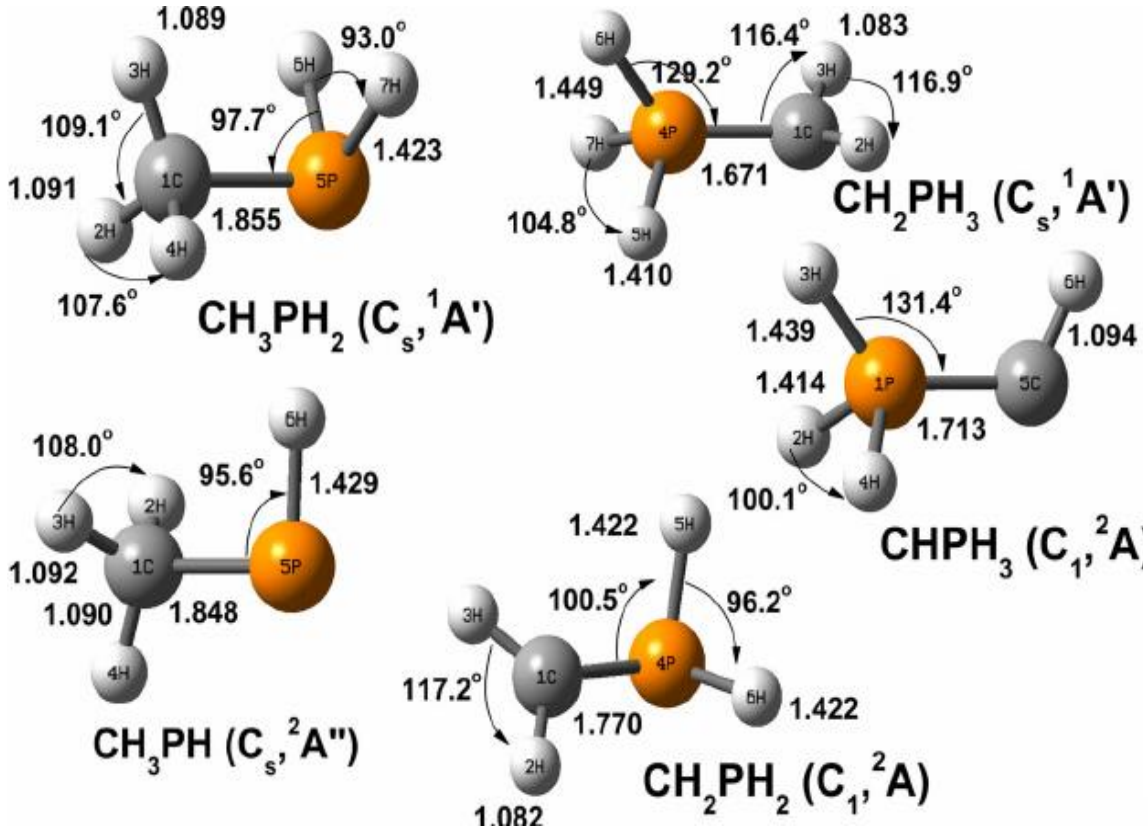
METHOD

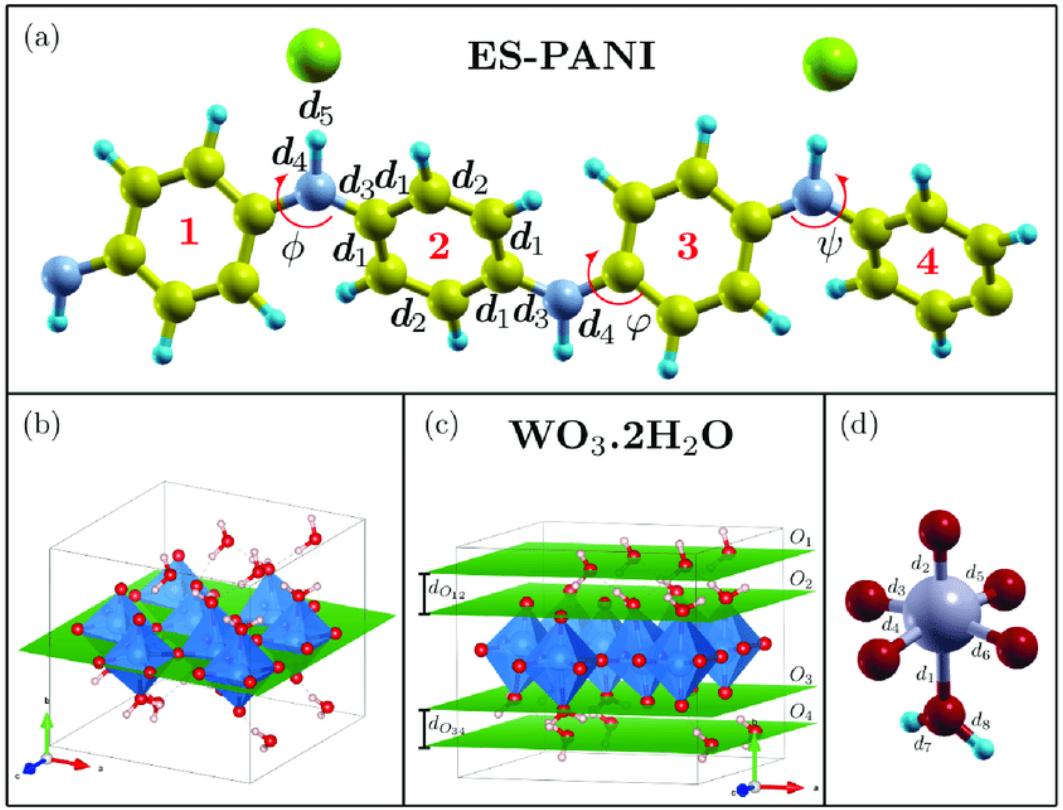
- **Overview**



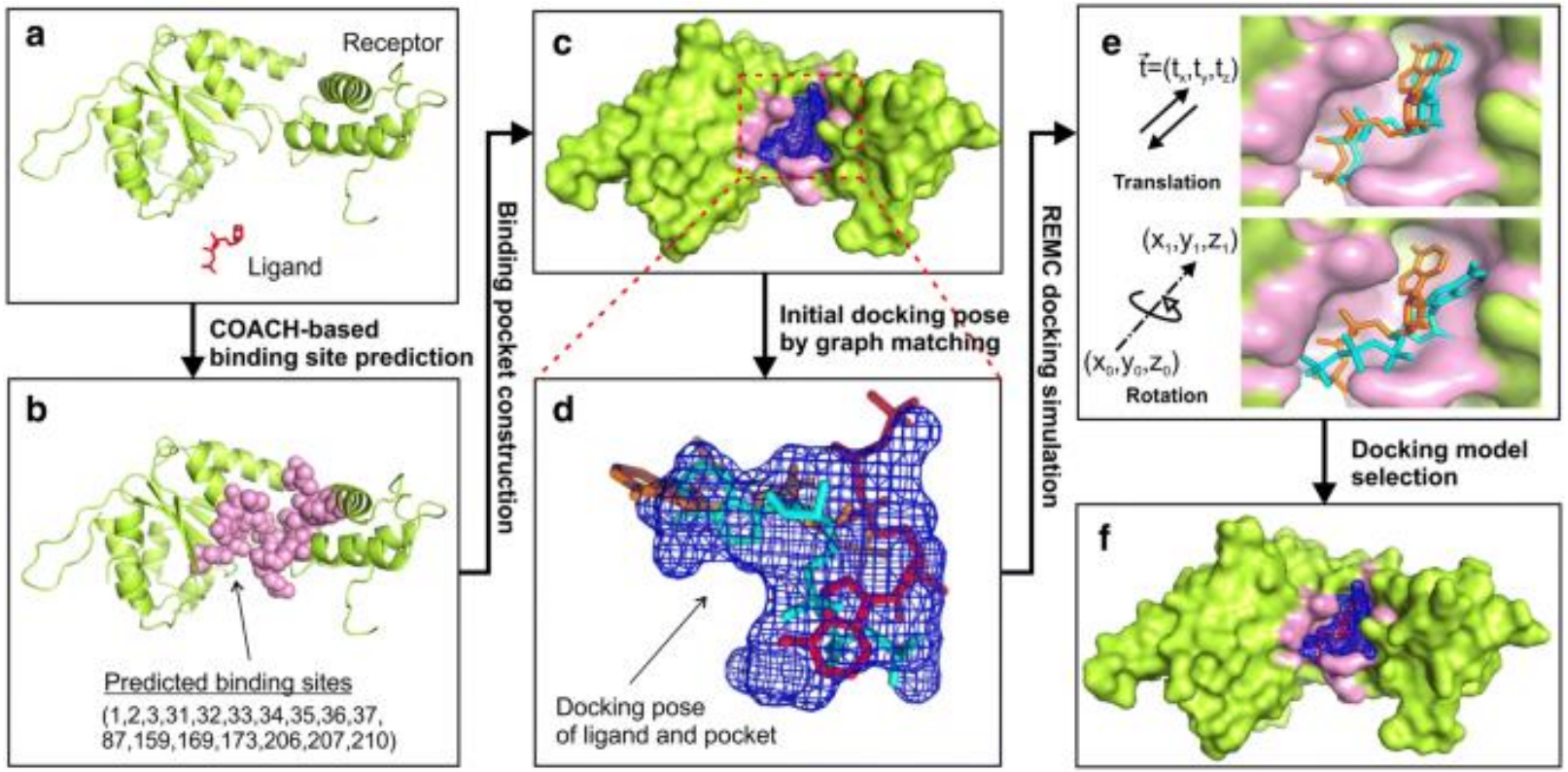


A ligand pose x is an assignment of atomic positions in $R^3, x \in R^{3n}, n.:$ number of atoms.

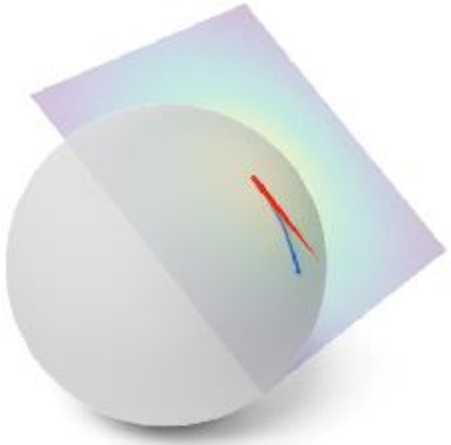




A ligand pose x is an assignment of atomic positions in $R^3, x \in R^{3n}, n.:$ number of atoms. bond lengths, angles, and small rings in the ligand are essentially rigid, such that the ligand flexibility lies almost entirely in the torsion angles at rotatable bonds. Traditional docking methods, as well as most ML ones, take as input a seed conformation $c \in R^{3n}$ of the ligand in solation and change only the relative position and the torsion degrees of freedom in the final bound conformation.

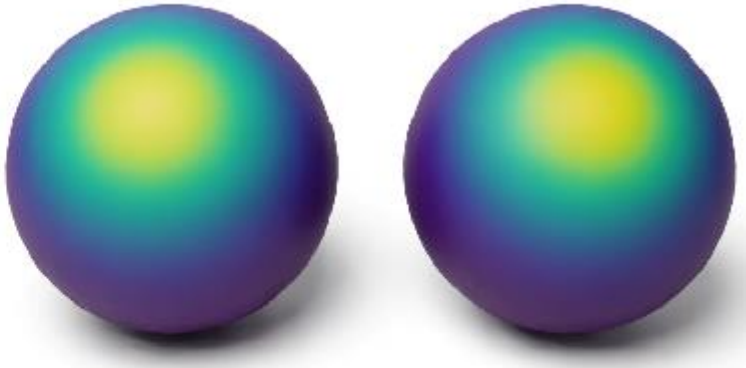


The space of ligand poses consistent with c is, therefore, an (m + 6)-dimensional submanifold $M_c \subset R^{3n}$, where **m is the number of rotatable bonds**, and the six additional degrees of freedom come from rototranslations relative to the fixed protein. We follow this paradigm of taking as input a seed conformation c, and formulate molecular docking as learning a probability distribution $P_c(x|y)$ over the manifold M_c, conditioned on a protein structure y.



a) A single step of a Geodesic Random Walk.

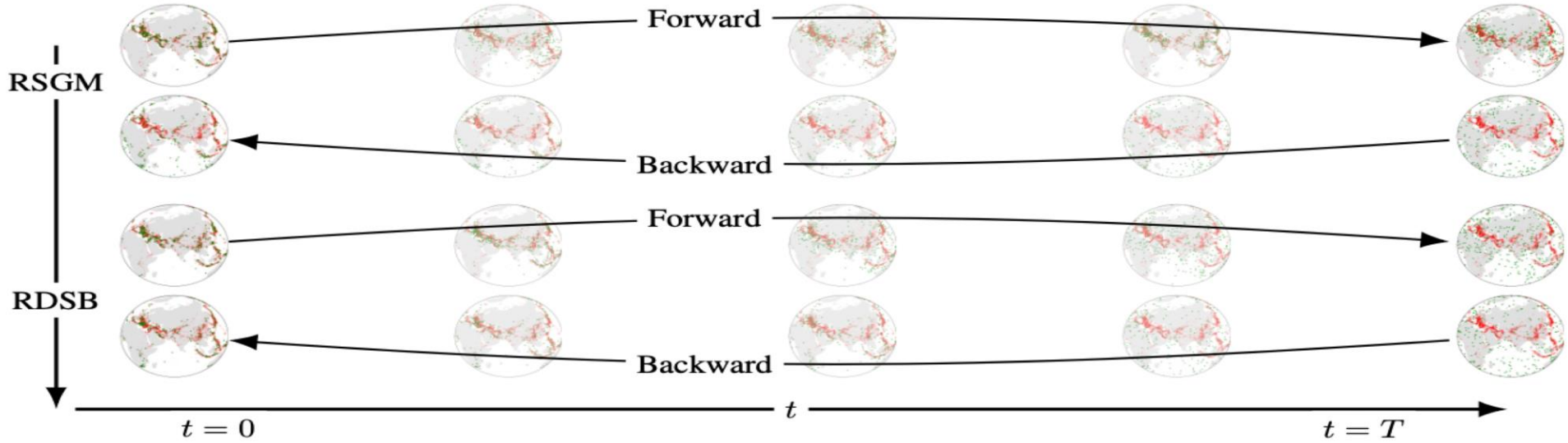
(b) Many steps yield an approximate trajectory.

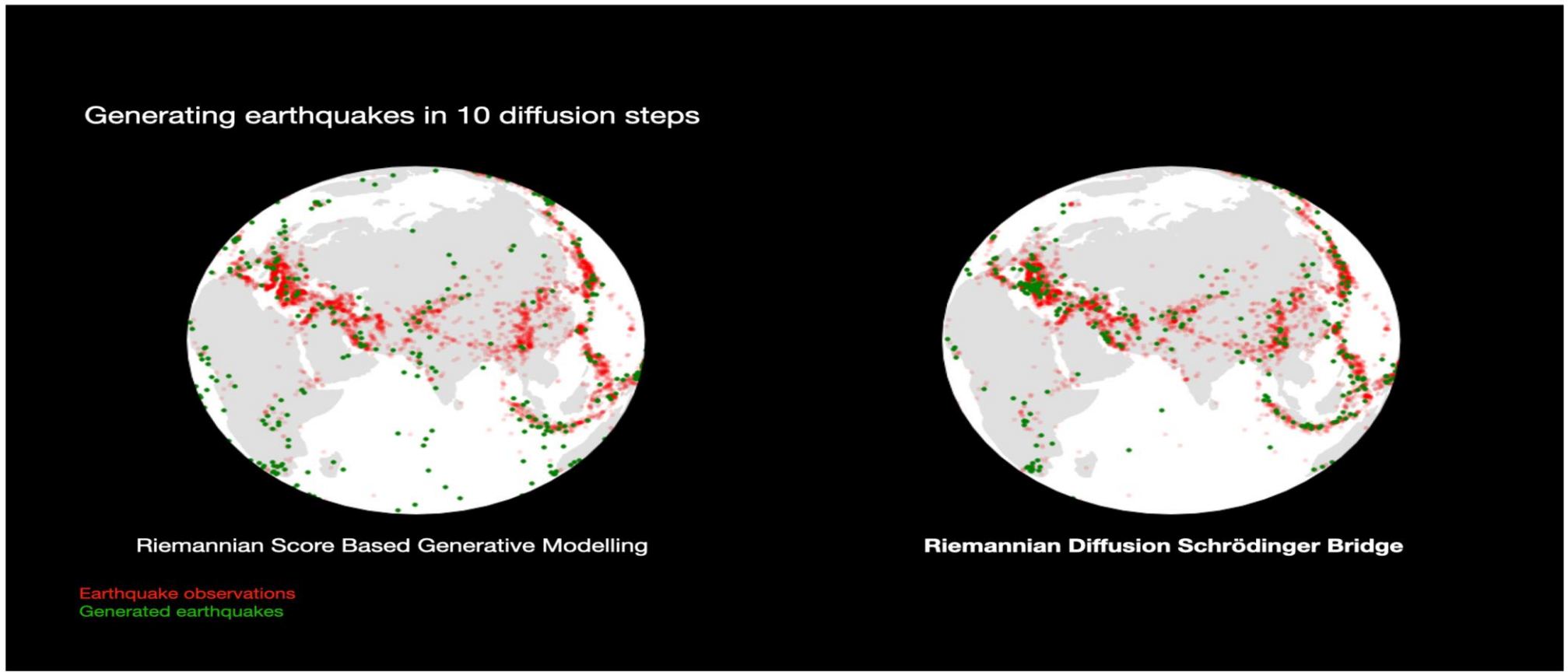


(c) Gaussian Random Walk [Left] and the Brownian motion density [Right] agree well for small time steps.

- **Forward SDE**

$$dx = \sqrt{\frac{d\sigma^2(t)}{dt}}\, dW, \qquad \sigma^2 = \sigma_{tr}^2, \sigma_{rot}^2, \sigma_{tor}^2$$


Forward
RSGM
Backward
Forward
RDSB
Backward
t
$t = 0$
$t = T$


Generating earthquakes in 10 diffusion steps
Riemannian Score Based Generative Modelling
Riemannian Diffusion Schrödinger Bridge
Earthquake observations
Generated earthquakes

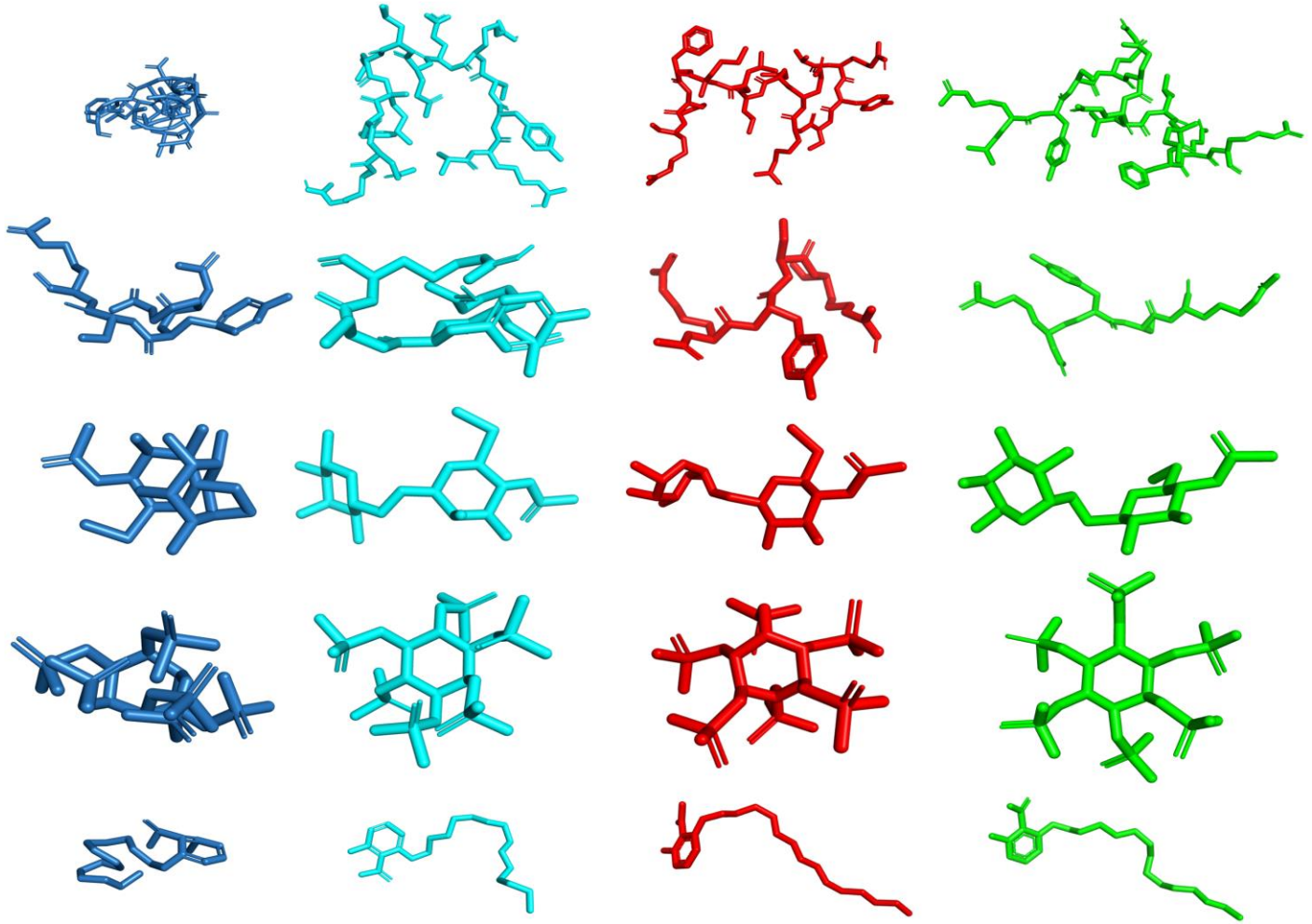
- Any ligand pose consistent with a seed conformation can be reached by a combination of (1) ligand translations, (2) ligand rotations, and (3) changes to torsion angles.

This can be viewed as an informal definition of the manifold M_c

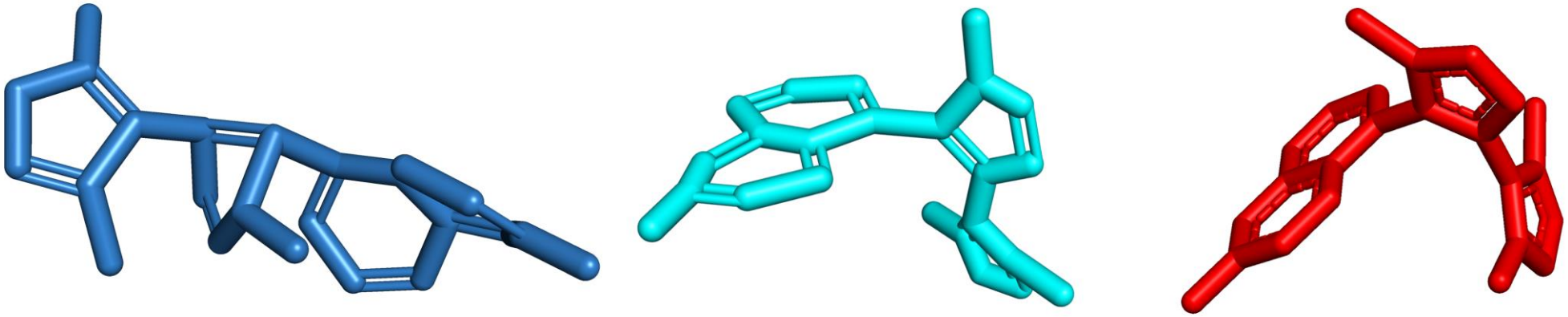
MODEL ARCHITECTURE

We construct the score model s(x; y; t) and the confidence model d(x; y) to take as input the current ligand pose x and protein structure y in 3D space.

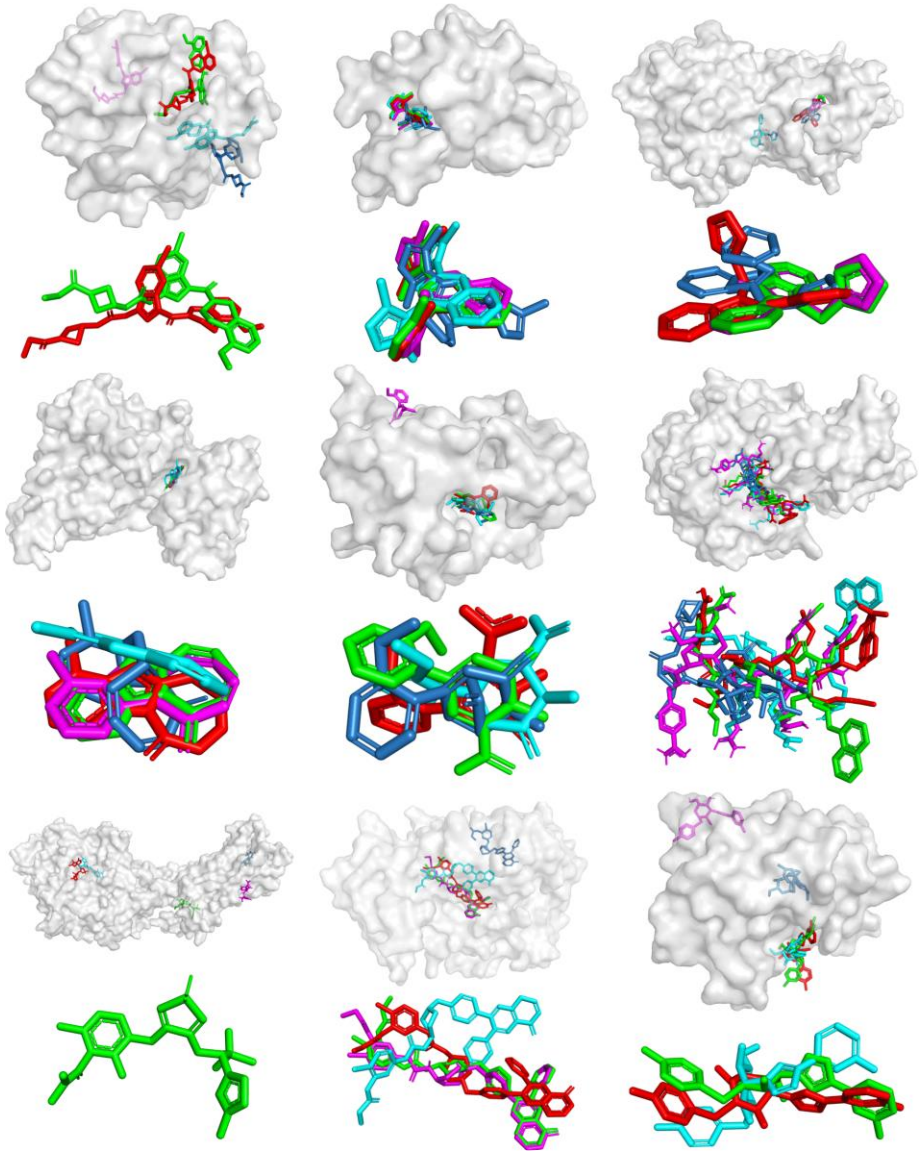
The output of the confidence model is a single scalar, as ligand pose distributions, are defined relative to the protein structure, which can have arbitrary location and orientation. On the other hand, the output of the score model must be in the tangent space.



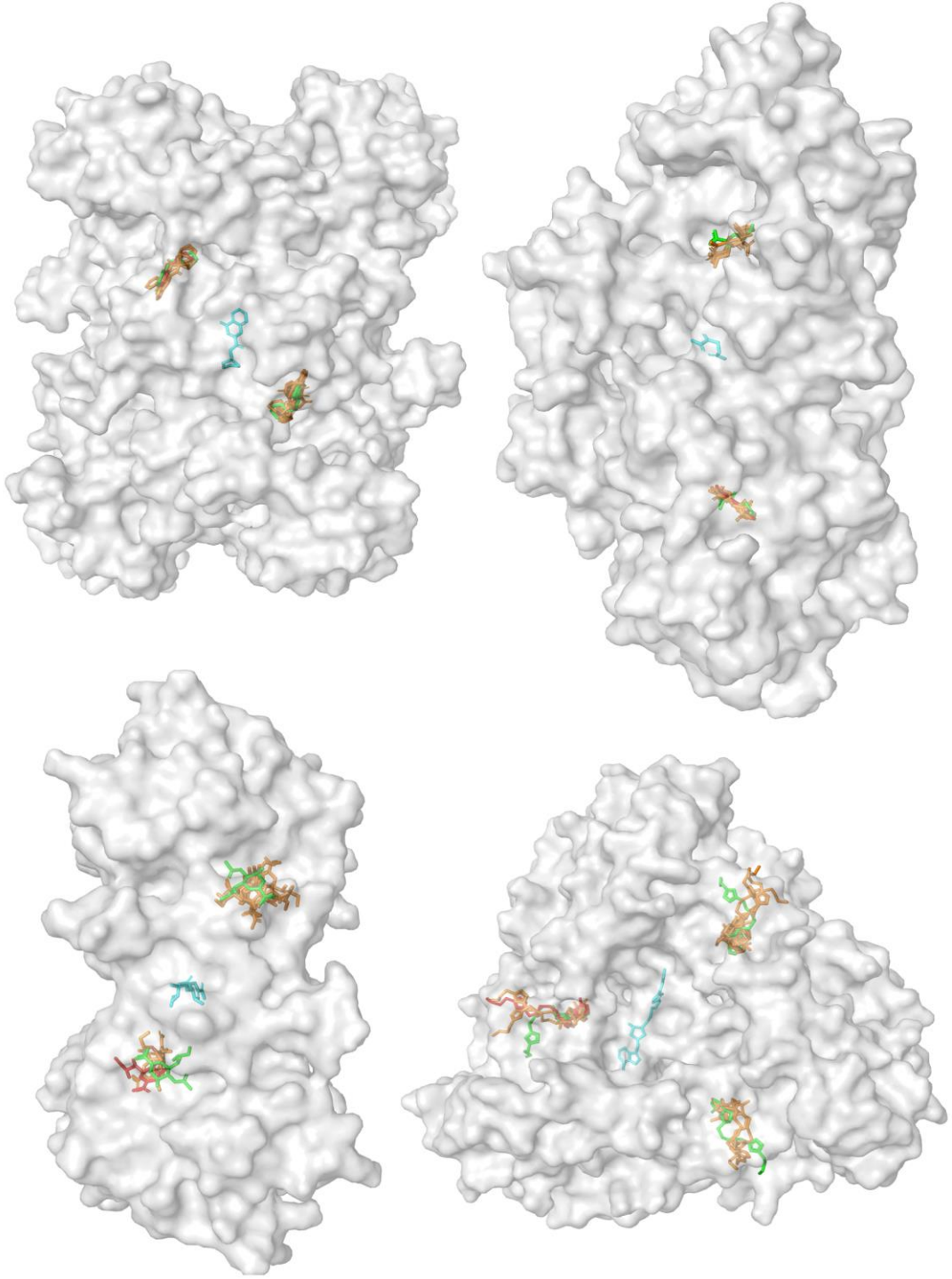
Ligand self-intersections. TANKBind (blue), EquiBind (cyan), DIFFDOCK (red), and crystal structure (green). Due to the averaging phenomenon that occurs when epistemic uncertainty is present, the regression-based deep learning models tend to produce ligands with atoms that are close together, leading 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.



Chemically plausible local structures. TANKBind (blue), EquiBind (cyan), and DIFFDOCK (red) structures for complex 6g2f. EquiBind (without their correction step) produces very unrealistic local structures and TANKBind, e.g., produces non-planar aromatic rings. DIFFDOCK's local structures are the realistic local structures of RDKit

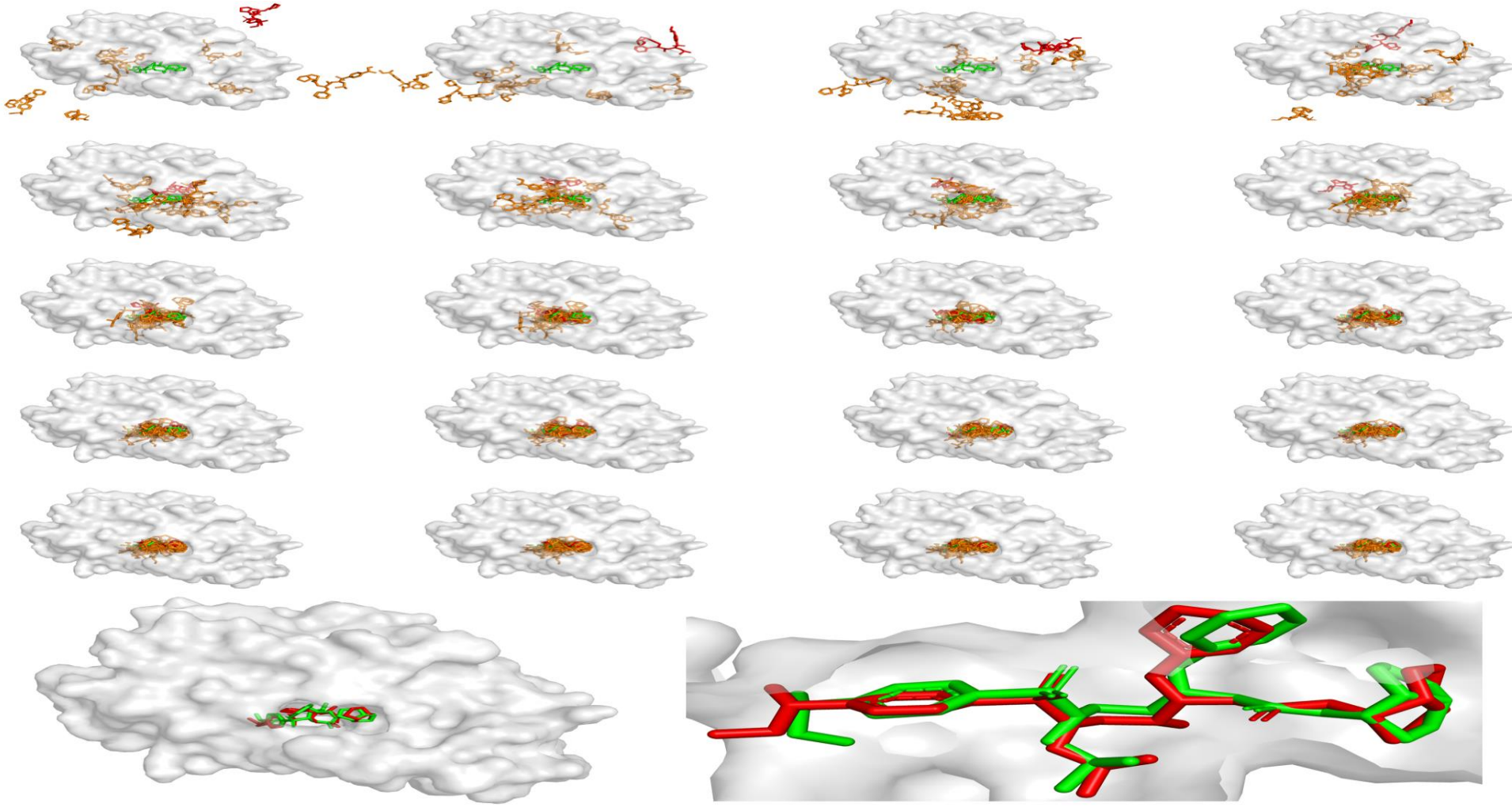


Randomly picked examples. The predictions of TANKBind (blue), EquiBind (cyan), GNINA (magenta), DIFFDOCK (red), and crystal structure (green). Shown are the predictions once with the protein and without it below. The complexes were chosen with a random number generator from the test set. TANKBind often produces self intersections (examples at the top-right; middle middle; middle-right; bottom-right). DIFFDOCK and GNINA sometimes almost perfectly predict the bound structure (e.g., top-middle).



Symmetric complexes and multiple modes. **EquiBind (cyan)**, **DIFFDOCK highest confidence sample (red), all other DIFFDOCK samples (orange), and the crystal structure (green).**

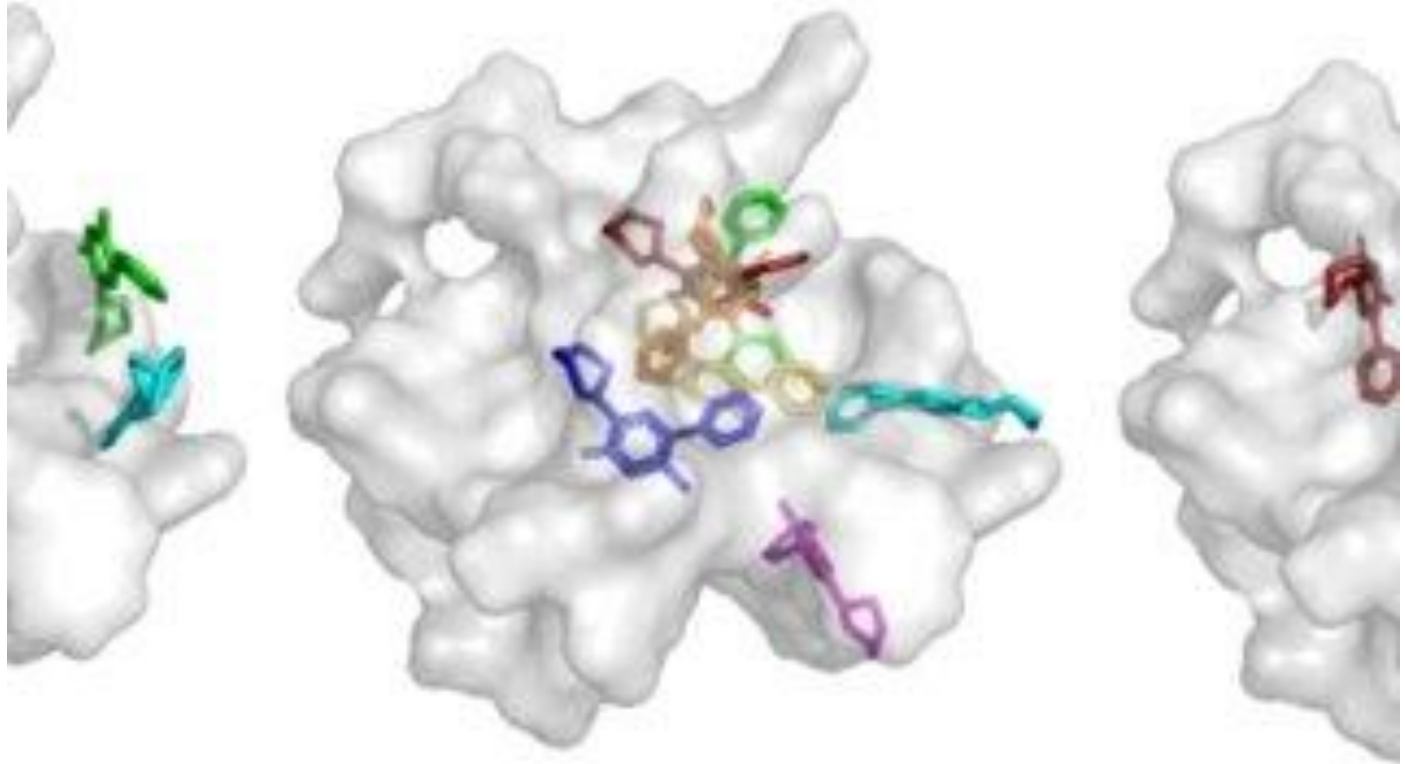
We see that, since it is a generative model, DIFFDOCK is able to produce multiple correct modes and to sample around them. Meanwhile, as a regression-based model, EquiBind is only able to predict a structure at the mean of the modes. The complexes are unseen during training. The PDB IDs in reading order: 6agt, 6gdy, 6ckl, 6dz3.

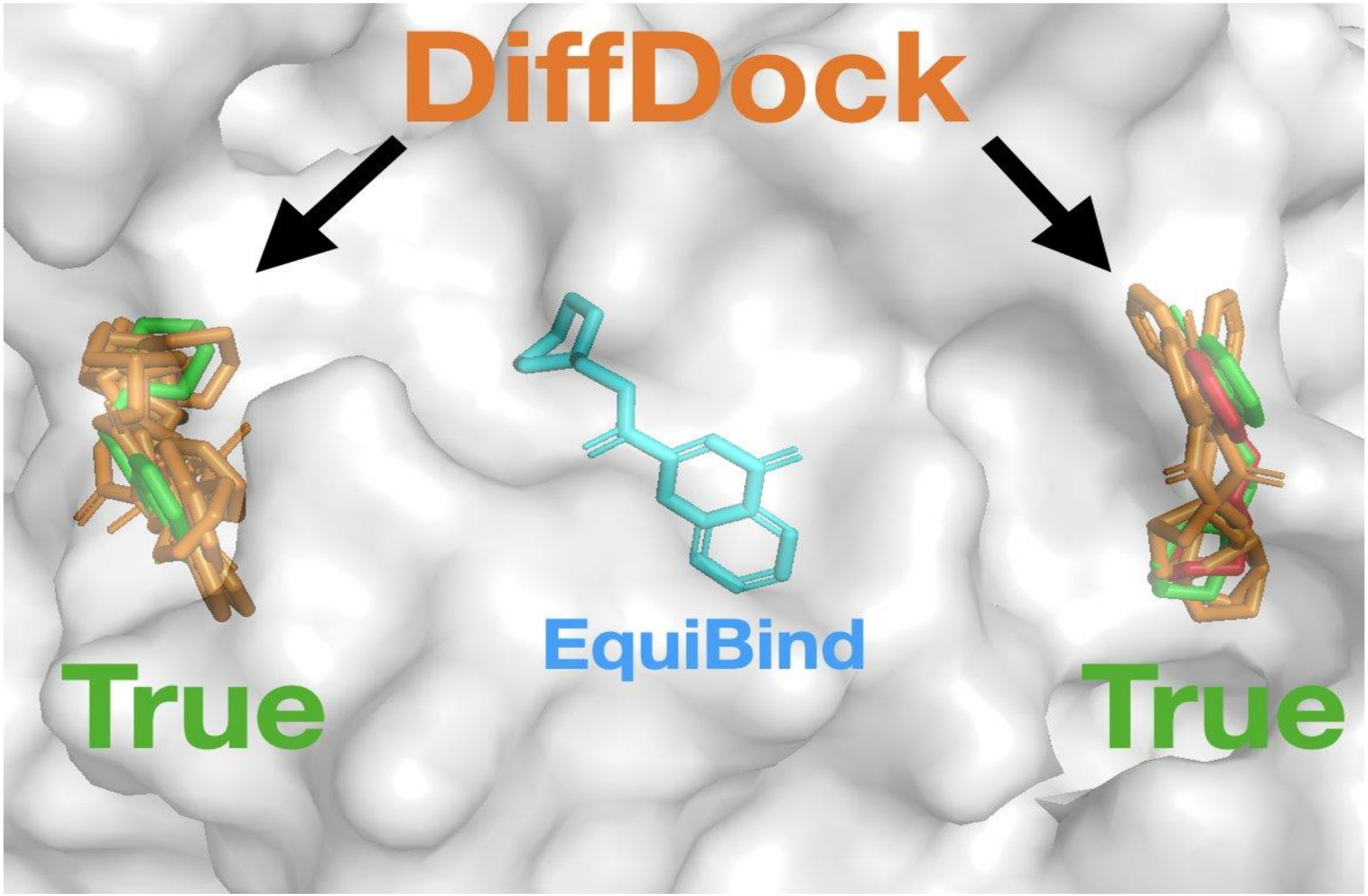


Reverse Diffusion. Reverse diffusion of a randomly picked complex from the test set. Shown are DIFFDOCK highest confidence sample (red), all other DIFFDOCK samples (orange), and the crystal structure (green). Shown are the 20 steps of the reverse diffusion process (in reading order) of DIFFDOCK for the complex 6oxx. Videos of the reverse diffusion are available

DIFFDOCK

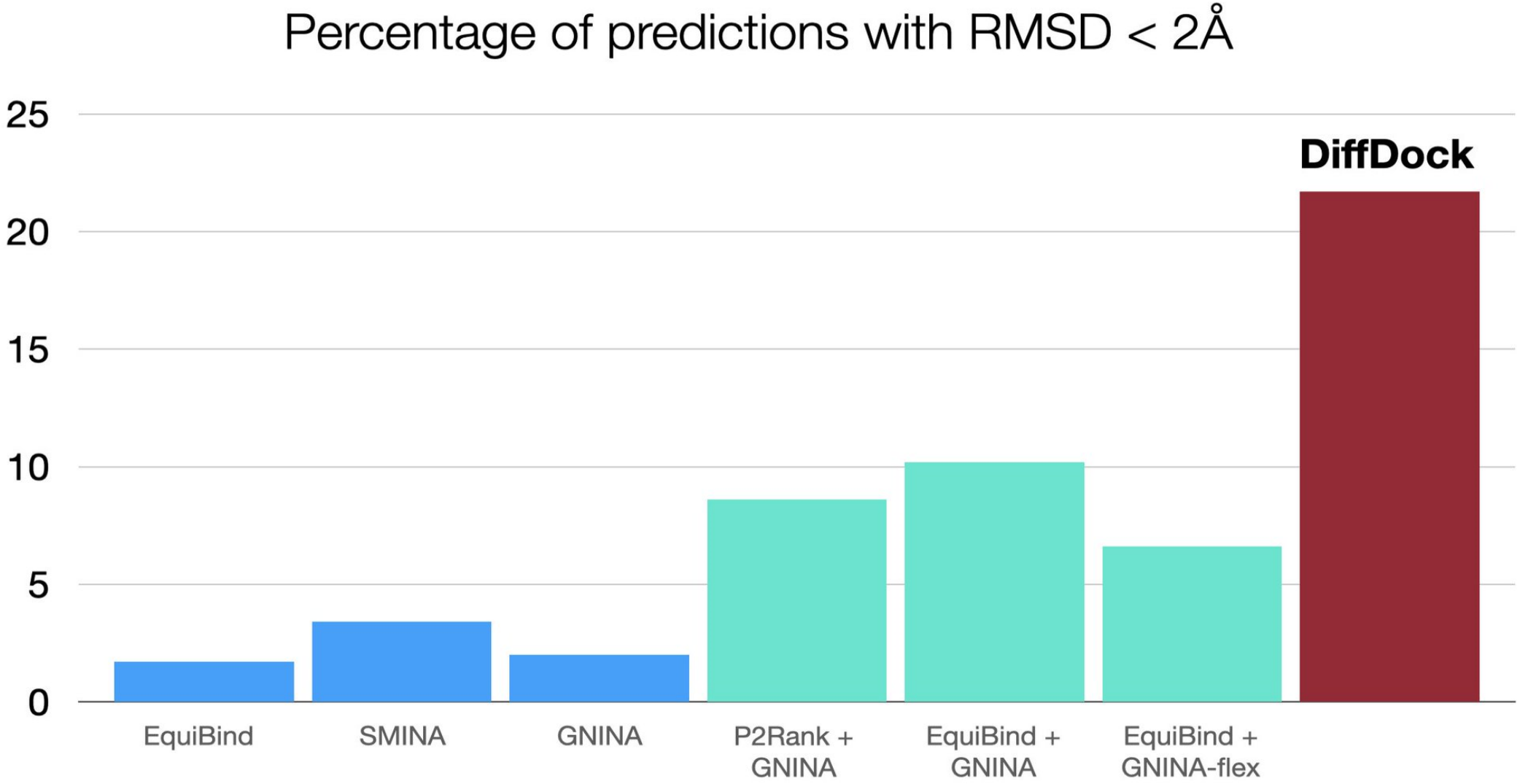
reverse diffusion over
translations, rotations and torsions

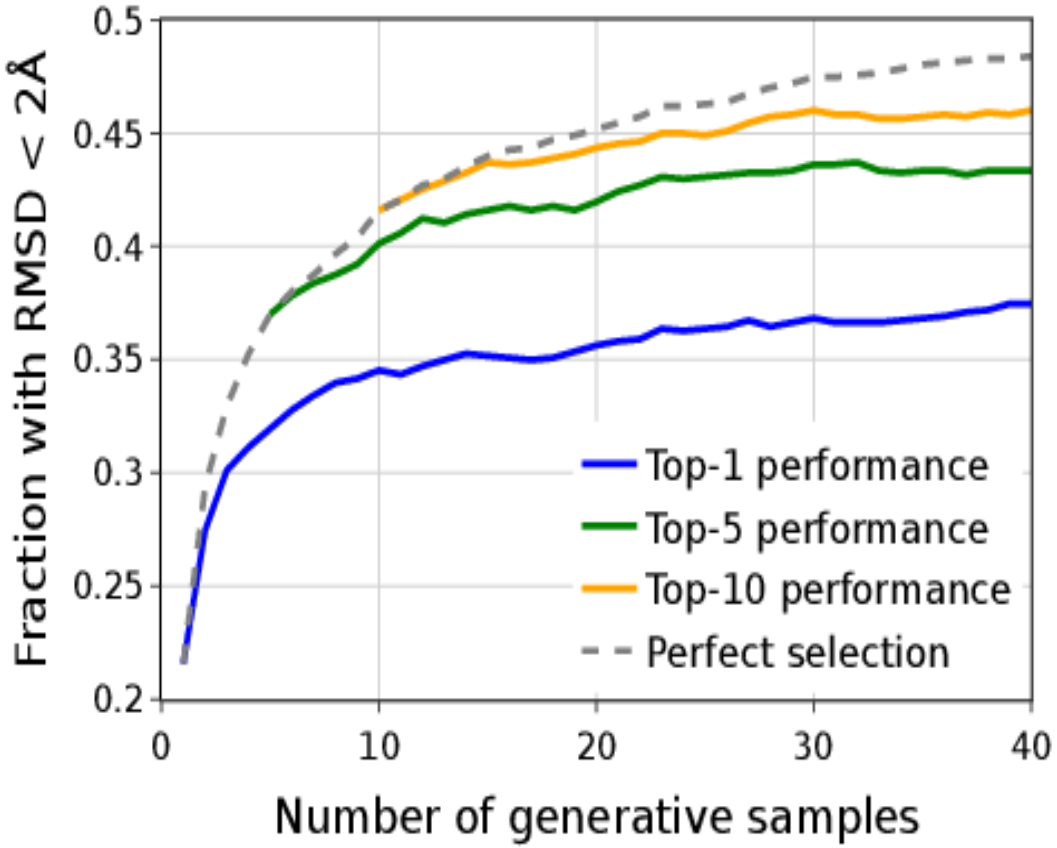


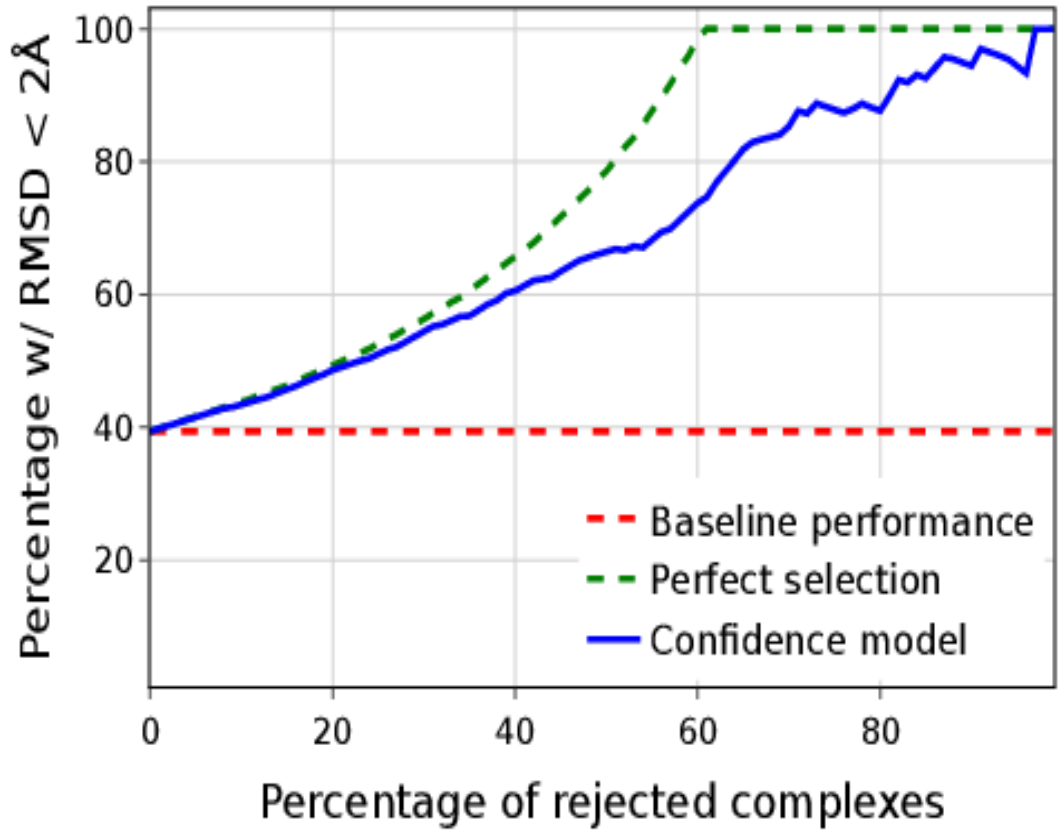

DiffDock
True
EquiBind
True

Method	Top-1 RMSD (Å) %<2	Top-1 RMSD (Å) Med.	Top-5 RMSD (Å) %<2	Top-5 RMSD (Å) Med.	Average Runtime (s)
QVINAW	20.9	7.7			49*
GNINA	22.9	7.7	32.9	4.5	127
SMINA	18.7	7.1	29.3	4.6	126*
GLIDE	21.8	9.3			1405*
EQUIBIND	5.5	6.2	-	-	**0.04**
TANKBIND	20.4	4.0	24.5	3.4	0.7/2.5
P2RANK+SMINA	20.4	6.9	33.2	4.4	126*
P2RANK+GNINA	28.8	5.5	38.3	3.4	127
EQUIBIND+SMINA	23.2	6.5	38.6	3.4	126*
EQUIBIND+GNINA	28.8	4.9	39.1	3.1	127
DIFFDOCK (10)	35.0±1.4	3.56±0.05	40.7±1.6	2.65±0.10	10
DIFFDOCK (40)	**38.2**±**1.0**	**3.30**±**0.11**	**44.7**±**1.7**	**2.40**±**0.12**	40

Performance with ESMFold structures




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