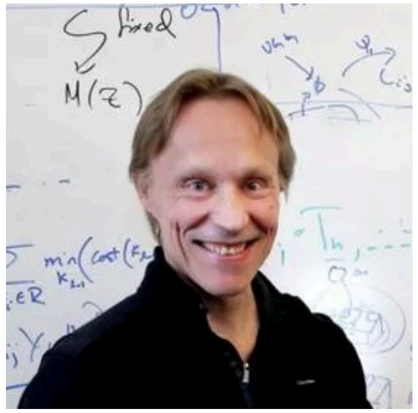
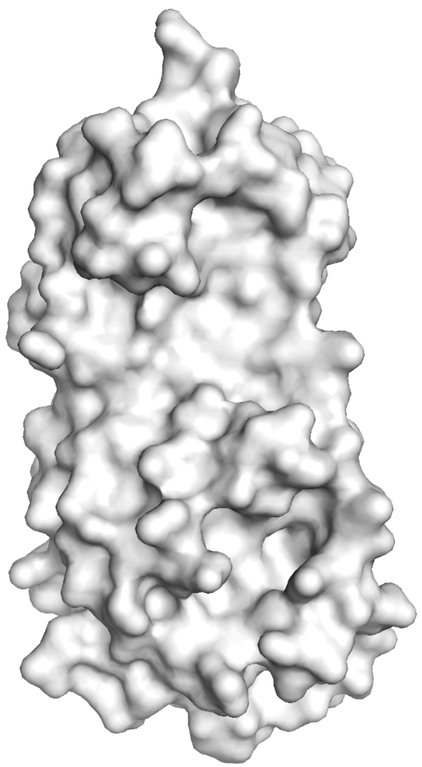
DiffDock

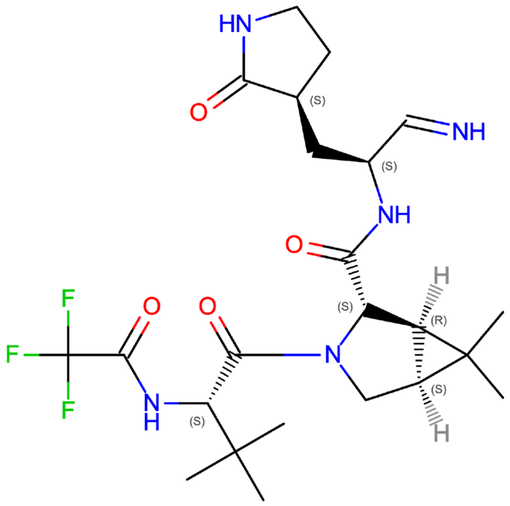
Equivariant Diffusion Models for Molecular Docking

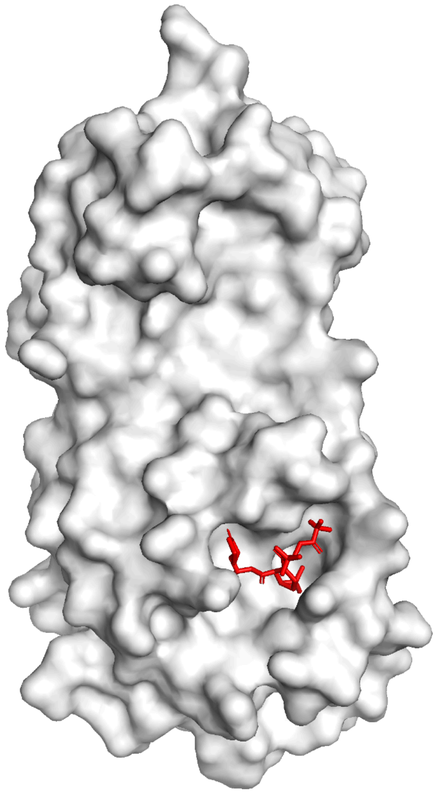
Gabriele Corso* Hannes Stärk* Bowen Jing* Regina Barzilay Tommi Jaakkola



Blind Protein-Ligand Docking







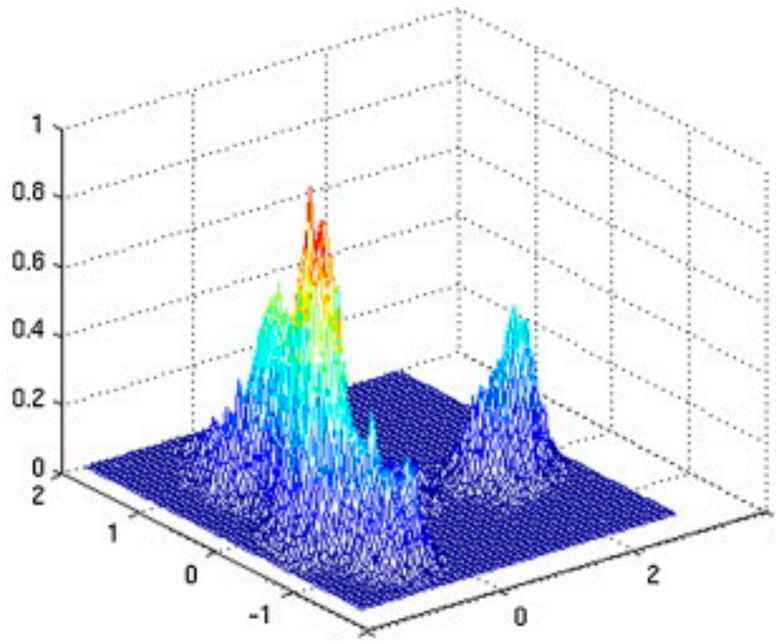
Input: protein structure + molecule

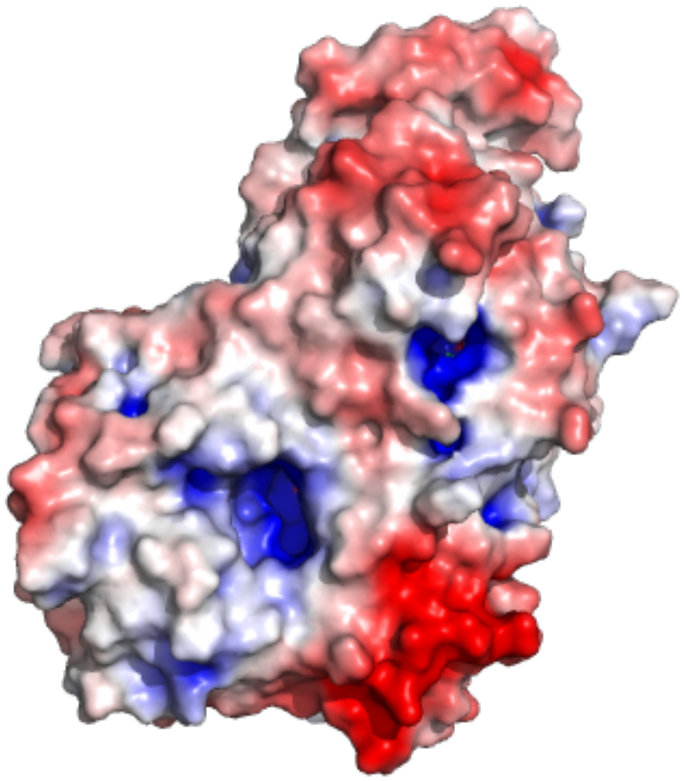
Output: bound structure

Traditional search based methods

Problems of search algorithm + scoring function approach

- fail to grasp with the vast search space of blind docking

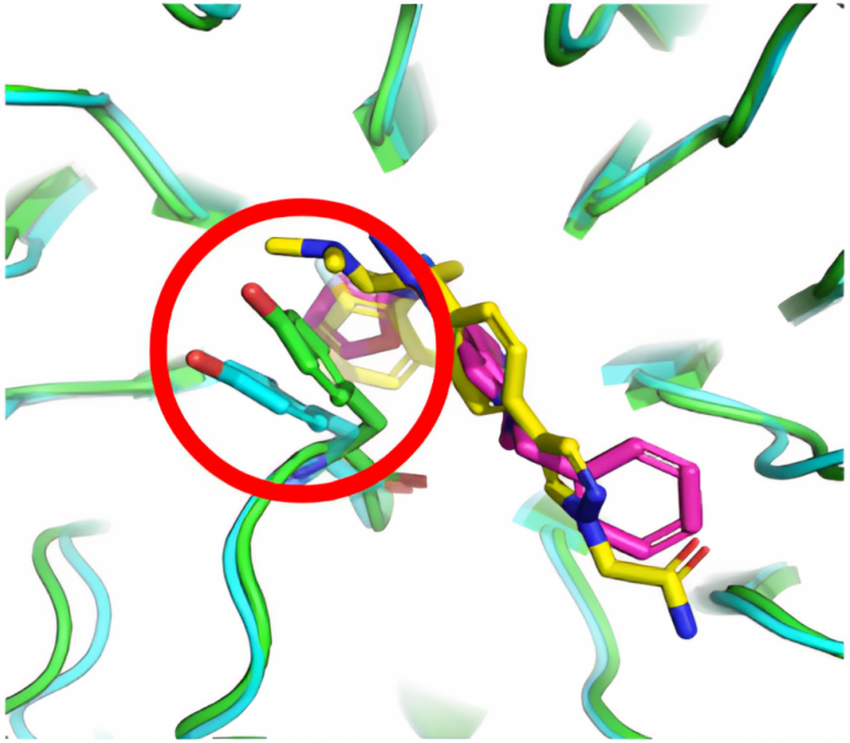




Traditional search based methods

Problems of search algorithm + scoring function approach

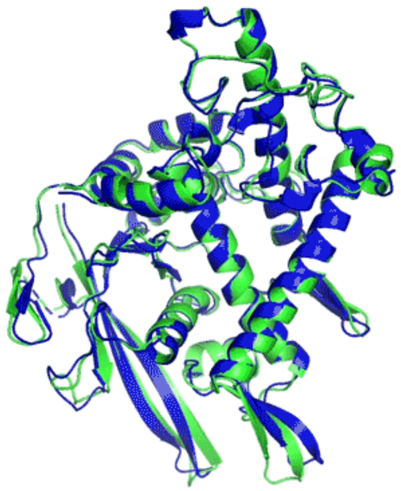
- fail to grasp with the vast search space of blind docking
- struggle with, e.g., side chain changes from unbound to bound protein structures



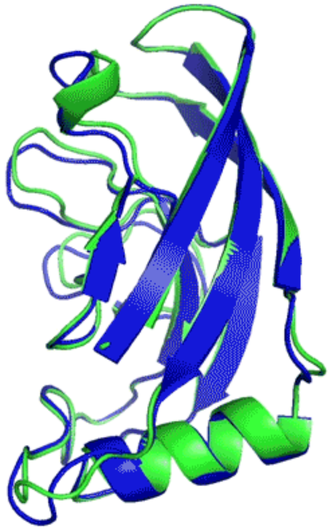
Traditional search based methods

Problems of search algorithm + scoring function approach

- fail to grasp with the vast search space of blind docking
- struggle with, e.g., side chain changes from unbound to bound protein structures
- unable to dock to imperfect computationally generated protein structures



T1037 / 6vr4
90.7 GDT
(RNA polymerase domain)



T1049 / 6y4f
93.3 GDT
(adhesin tip)

Experimental result
Computational prediction

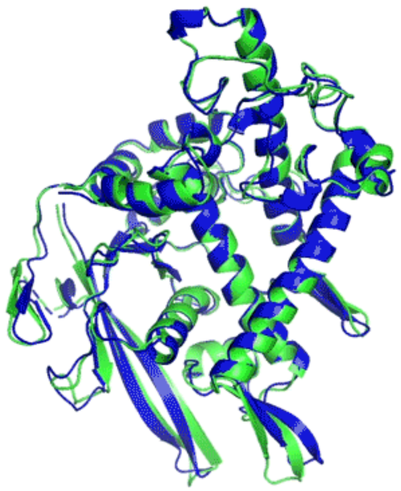
Traditional search based methods

Problems of search algorithm + scoring function approach

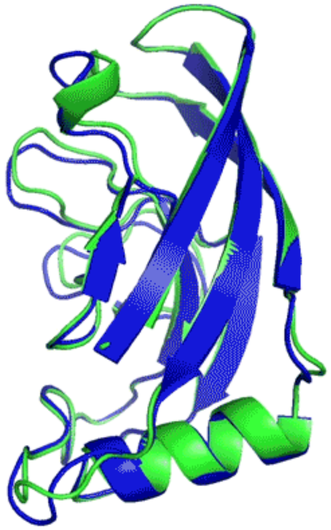
- fail to grasp with the vast search space of blind docking
- struggle with, e.g., side chain changes from unbound to bound protein structures
- unable to dock to imperfect computationally generated protein structures

Wong et al. "*Benchmarking AlphaFold-enabled molecular docking predictions for antibiotic discovery.*"

Karelina et al. "*How accurately can one predict drug binding modes using AlphaFold models?*"



T1037 / 6vr4
90.7 GDT
(RNA polymerase domain)



T1049 / 6y4f
93.3 GDT
(adhesin tip)

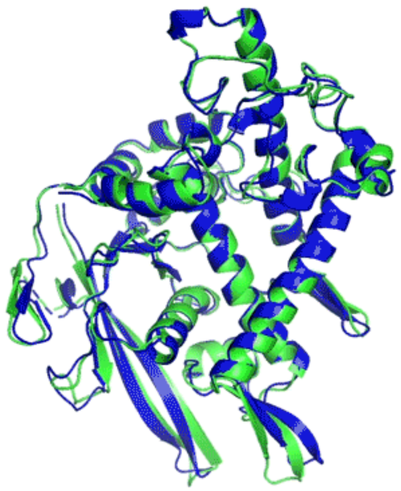
Experimental result
Computational prediction

Traditional search based methods

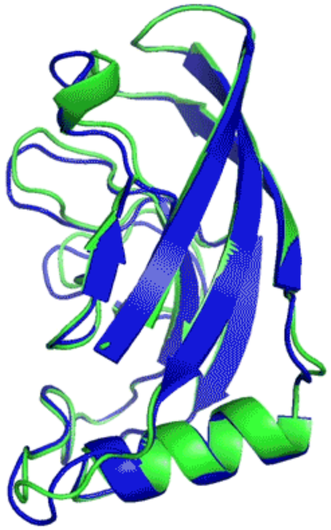
Problems of search algorithm + scoring function approach

- fail to grasp with the vast search space of blind docking
- struggle with, e.g., side chain changes from unbound to bound protein structures
- unable to dock to imperfect computationally generated protein structures

What can deep learning do for docking?



T1037 / 6vr4
90.7 GDT
(RNA polymerase domain)

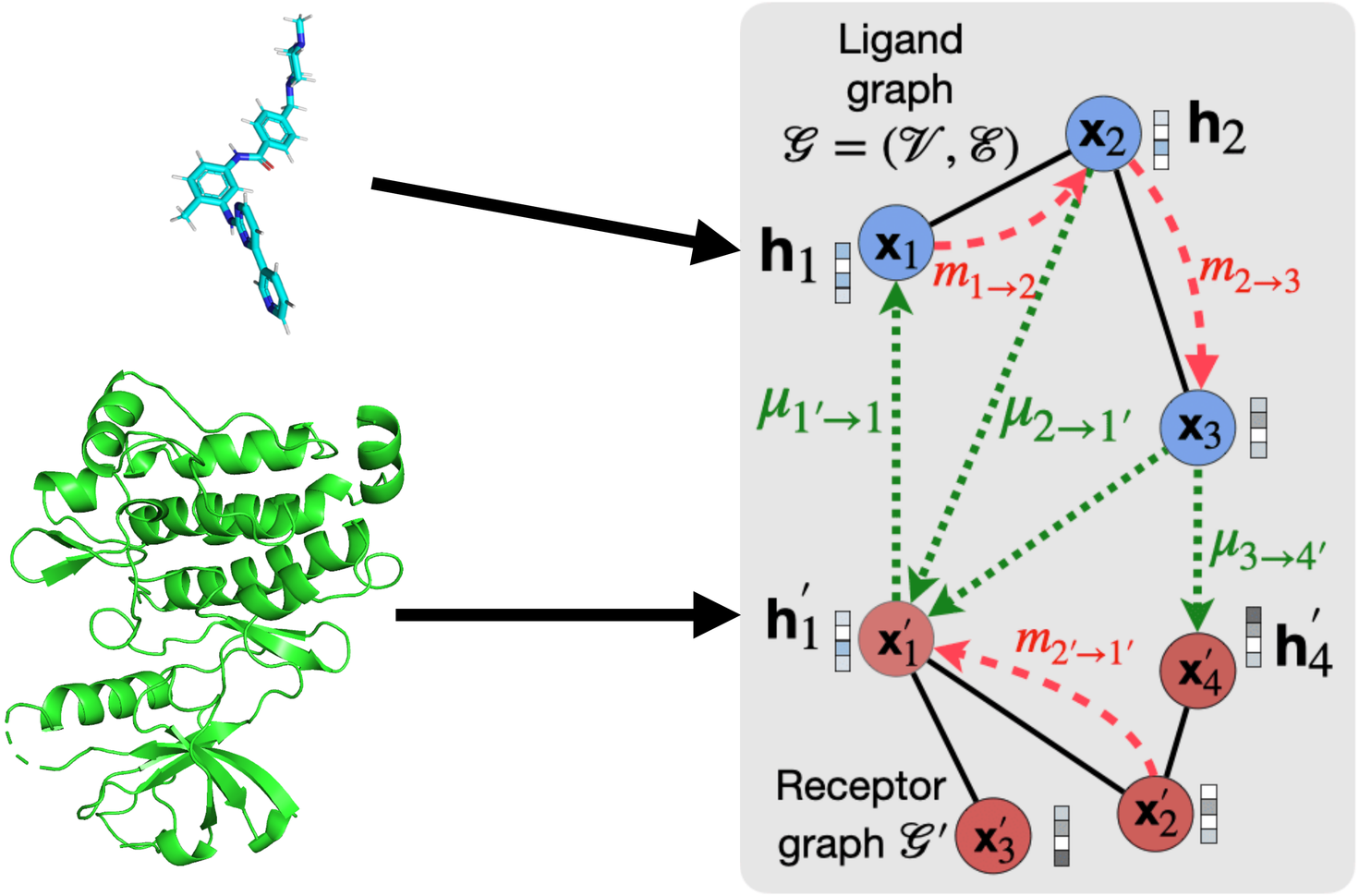


T1049 / 6y4f
93.3 GDT
(adhesin tip)

Experimental result
Computational prediction

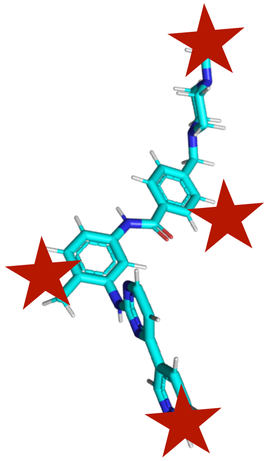
Previous DL approaches

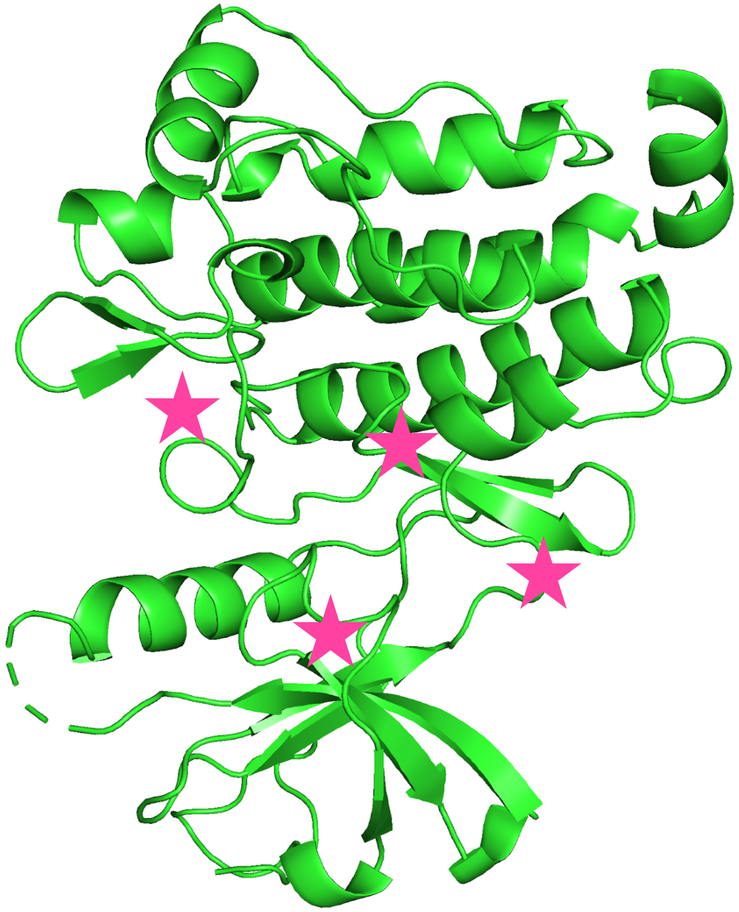
How do they work? E.g. EquiBind



Previous DL approaches

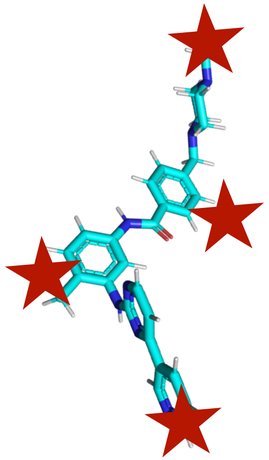
How do they work? E.g. EquiBind

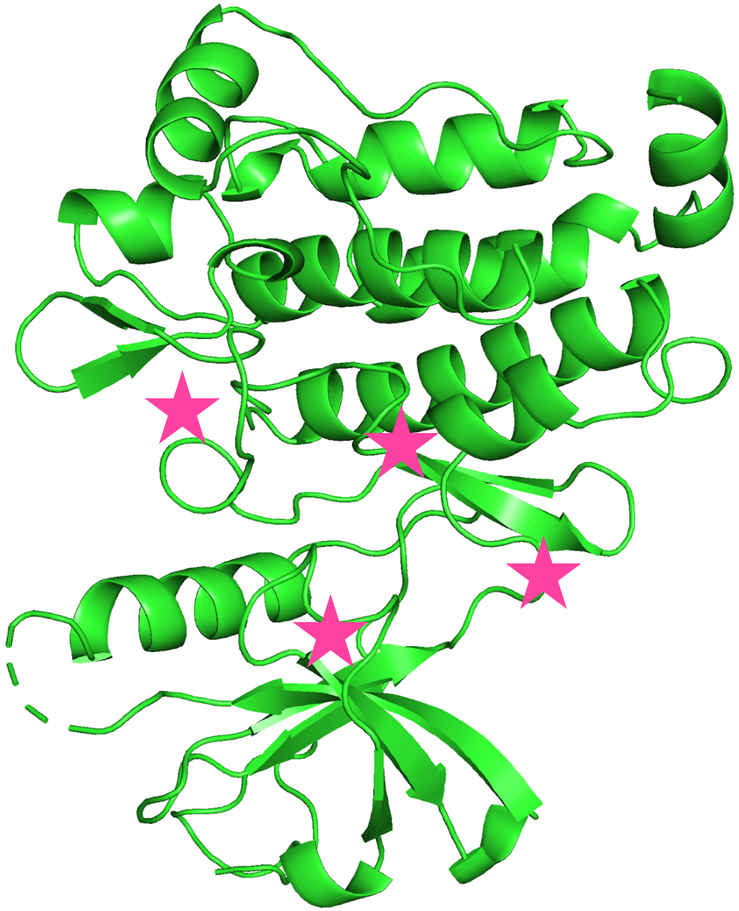




Previous DL approaches

How do they work? E.g. EquiBind



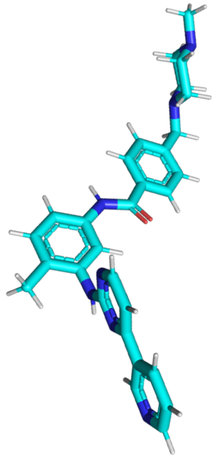


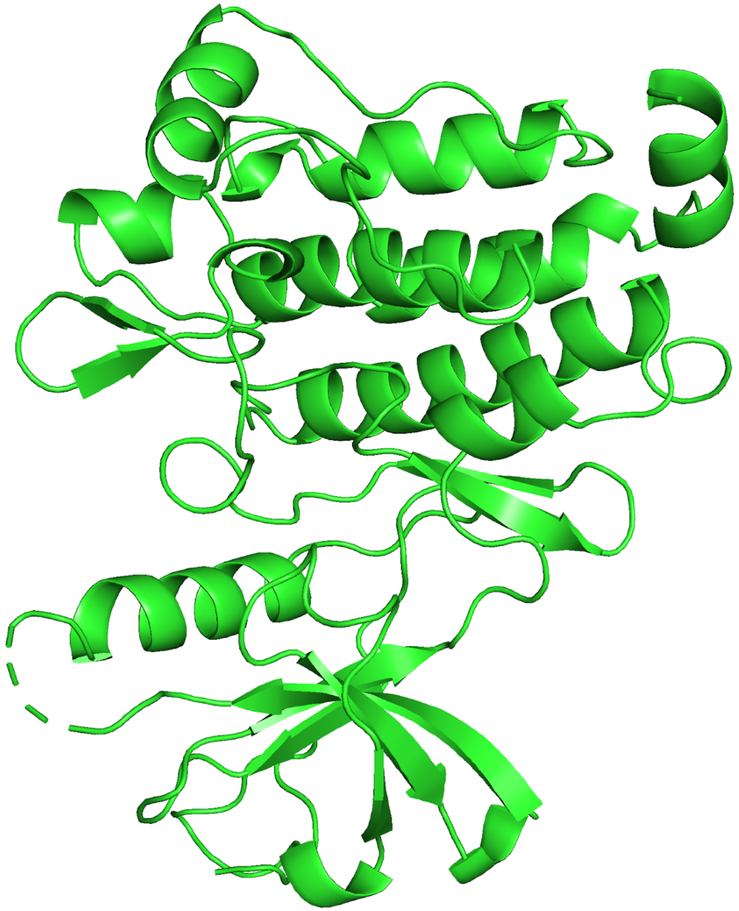
Kabsch algorithm calculates rototranslation to match keypoints

(EquiDock) Ganea et al. 2021, Independent SE(3)-Equivariant Models for End-to-End Rigid Protein Docking

Previous DL approaches

How do they work? E.g. EquiBind

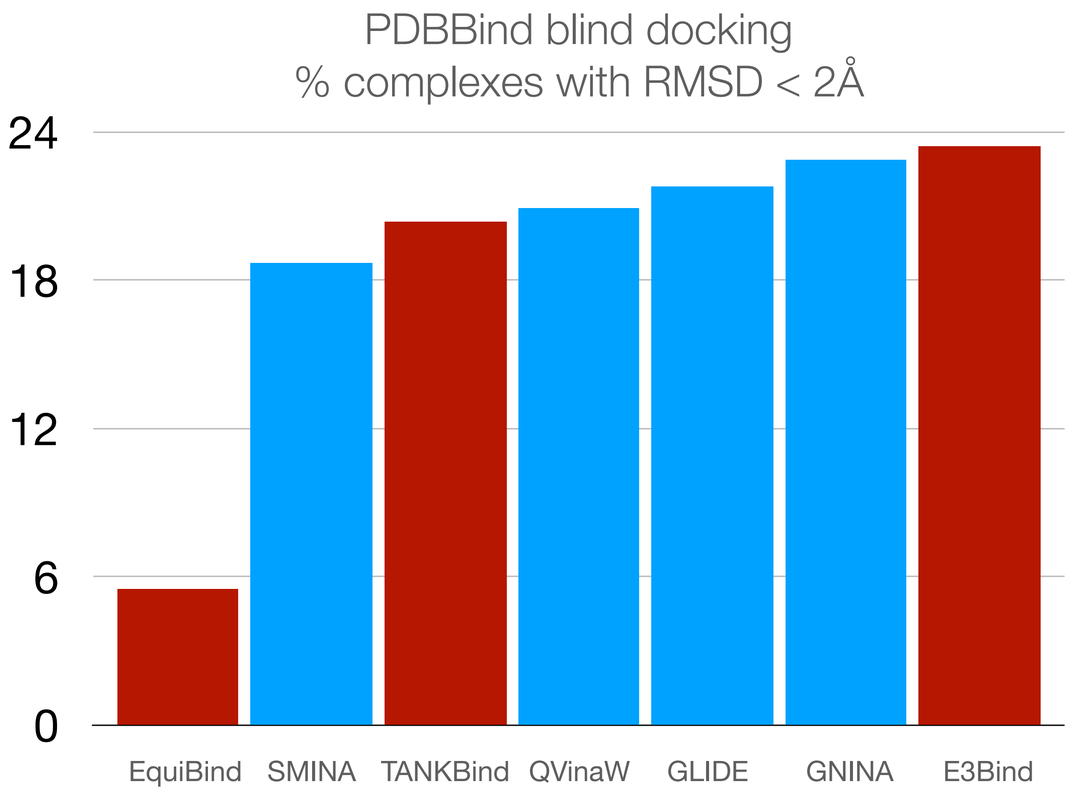




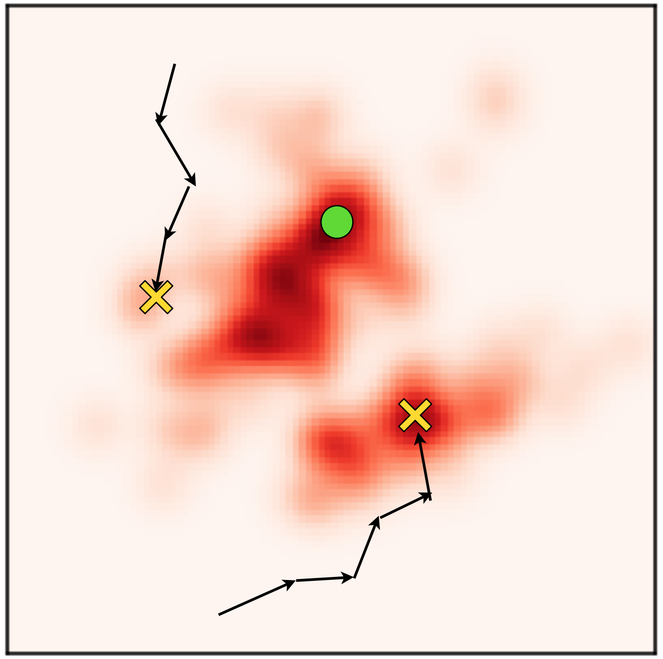
Apply rototranslation to molecule coordinates

Previous DL approaches

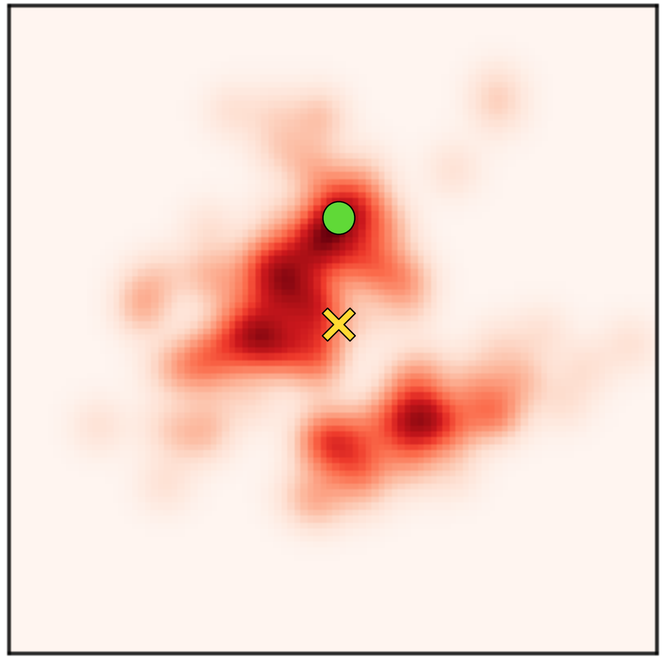
No meaningful advances of SOTA.



Approaches to docking recap



Traditional docking: sampling & optimization over scoring function: **no finite-time guarantees**!

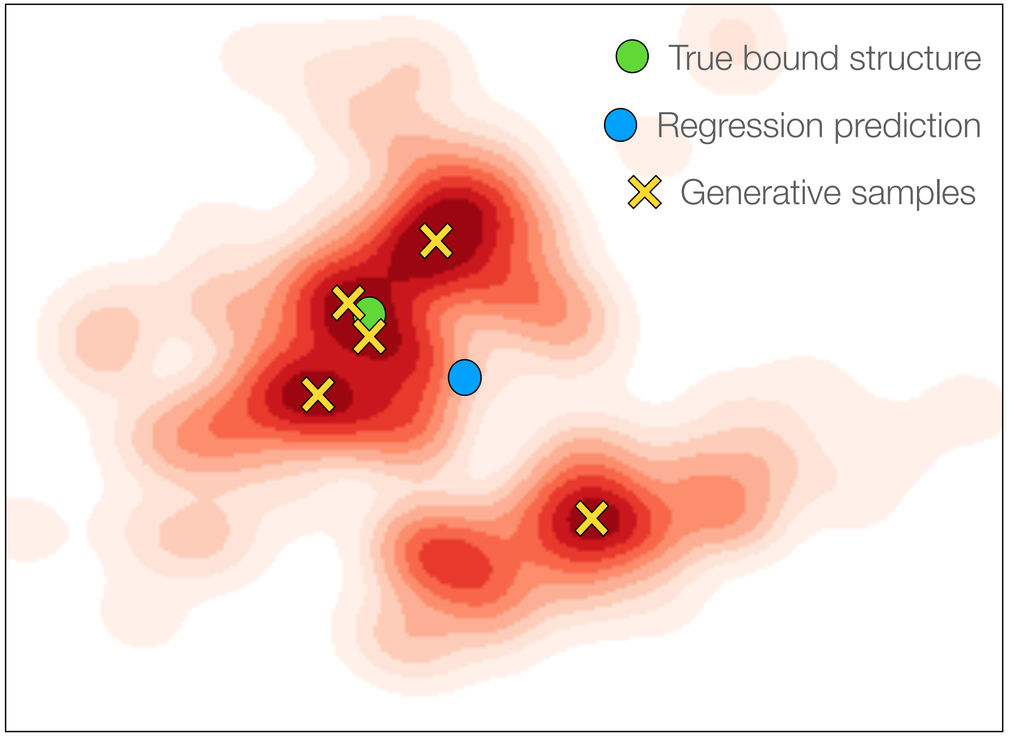


Previous deep learning: poor-quality single prediction with no refinement

Docking as a Generative Modeling Problem

A key paradigm shift from prior deep learning approaches

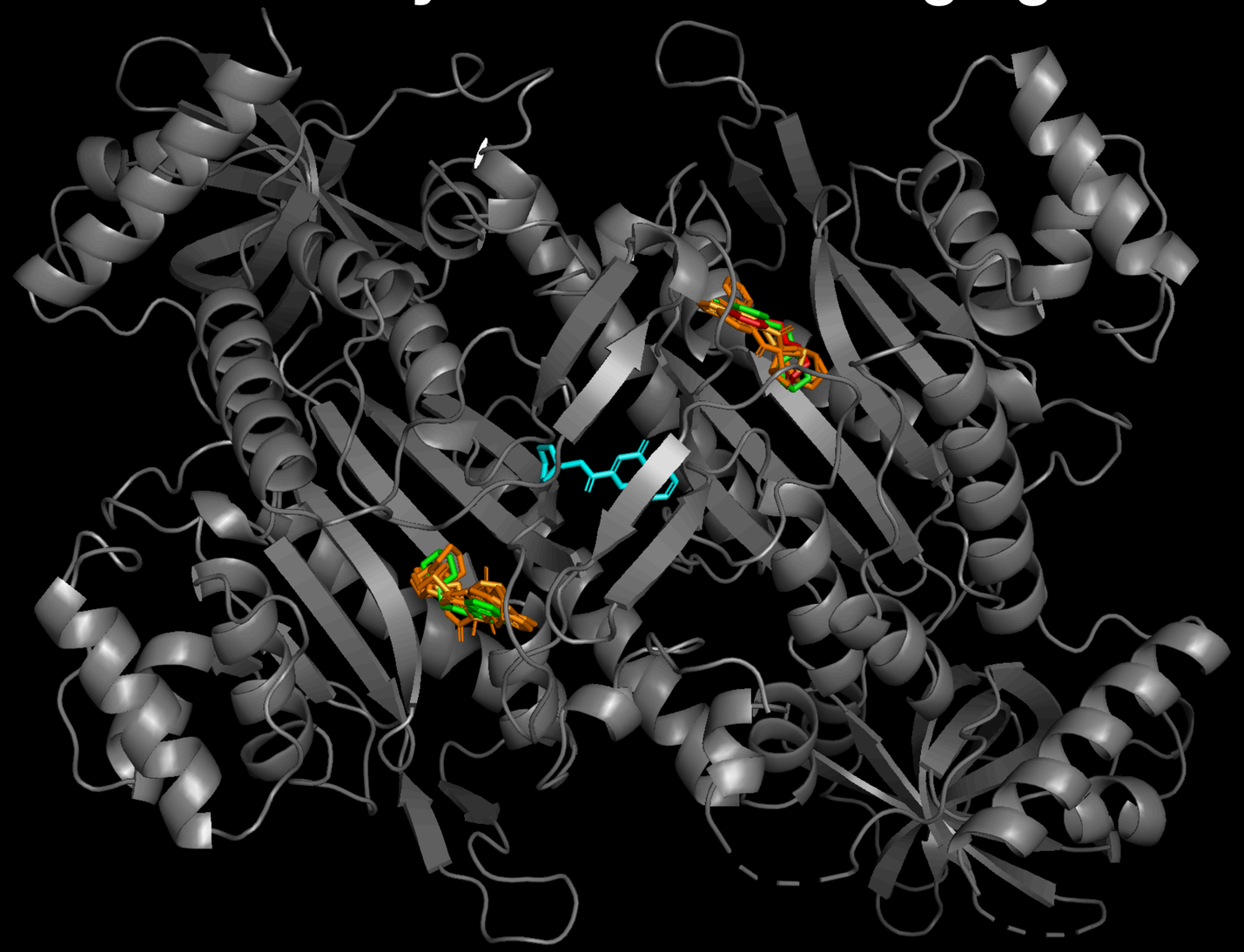
- Docking has significant aleatoric and epistemic uncertainty
- Any method will exhibit uncertainty about correct pose between multiple alternatives
- Regression methods to minimize squared error predict (weighted) mean
- Generative model will populate all/most modes



Docking energy landscape

Regression vs Generation for Docking

Aleatoric uncertainty induces “averaging” effect



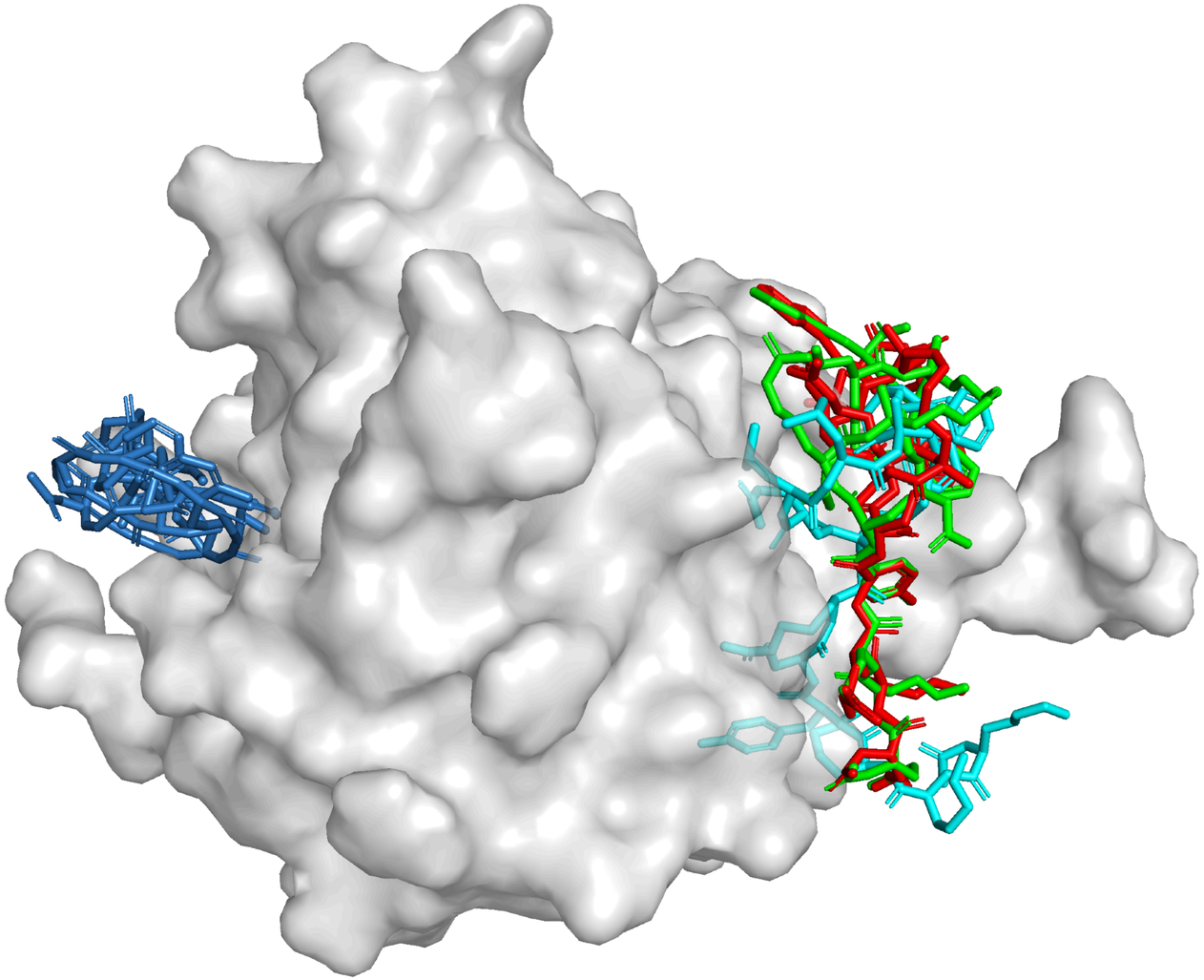
Crystal Structure
EquiBind (regression)
Generative samples
DiffDock top-1

Baragaña et al. PNAS, 2019

PfKRS, drug target in malaria and cryptosporidiosis, complexed with chromone inhibitor

Regression vs Generation for Docking

Model uncertainty is another issue



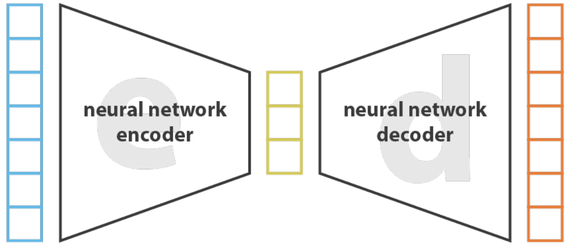
Crystal Structure
EquiBind (regression)
TANKBind (regression)
DiffDock top-1

A Generative Model for Molecular Docking

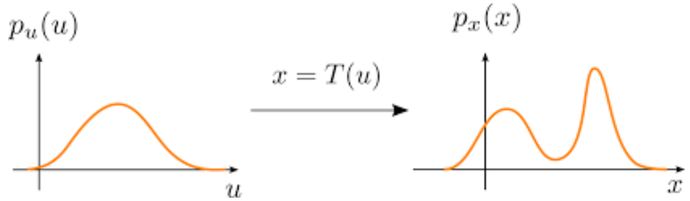
How to make our Generative Model?

ML has developed plenty of options

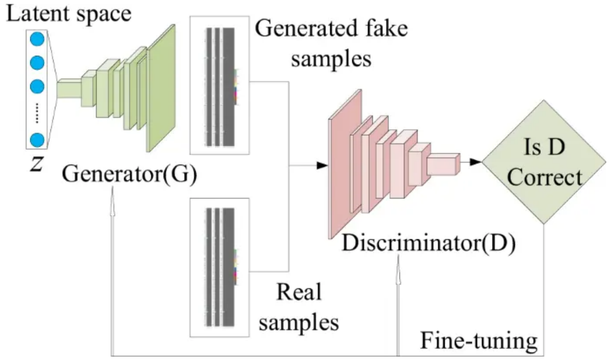
Variational Autoencoders



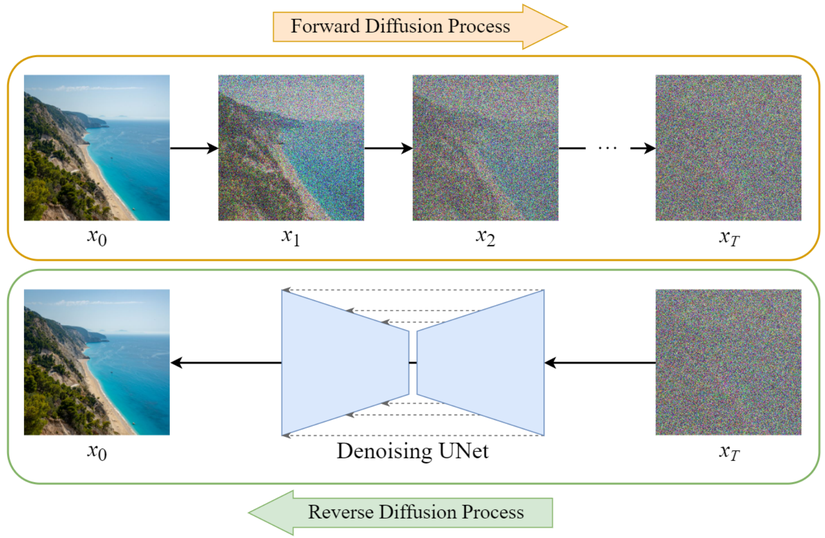
Normalizing Flows



Generative Adversarial Networks

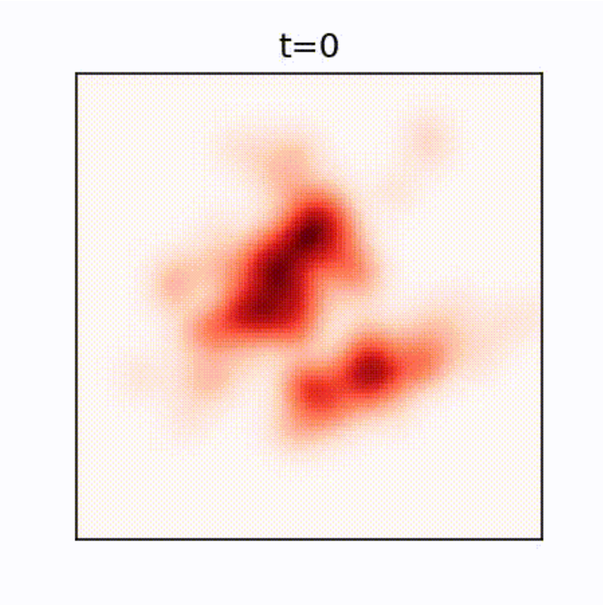


Diffusion Models



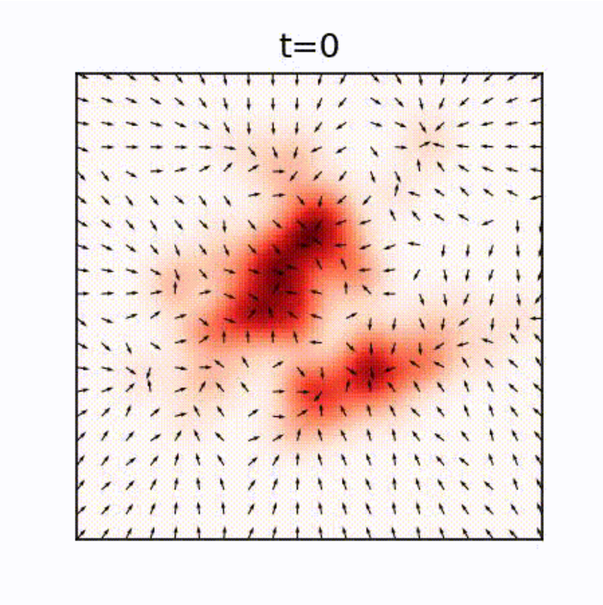
Diffusion Generative Models

Diffusion Generative Models



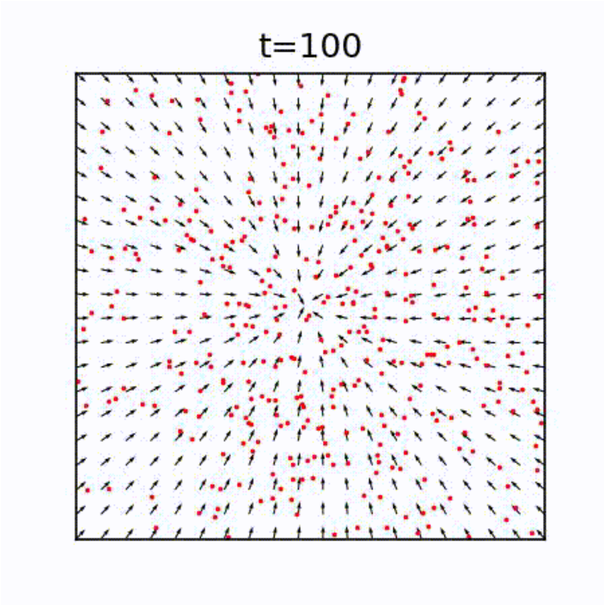
Define the **forward diffusion**

$$d\mathbf{x} = f(\mathbf{x}, t)\, dt + g(t)\, d\mathbf{w}$$



Learn the **score** (gradient of the log density) of the evolving data distribution

$$\mathbf{s}_\theta(\mathbf{x}, t) \approx \nabla_{\mathbf{x}} \log p_t(\mathbf{x})$$



Sample the **reverse diffusion**

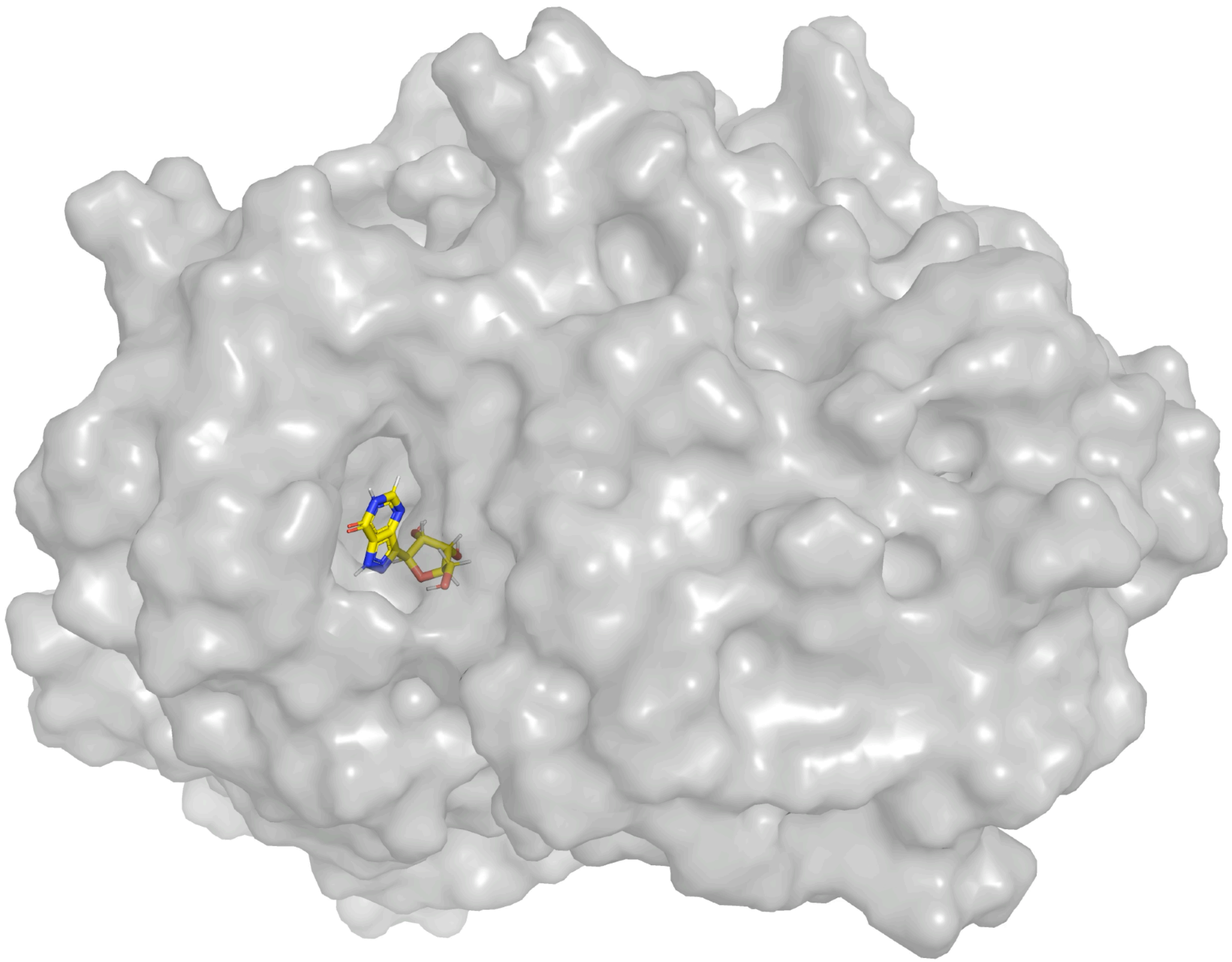
$$d\mathbf{x} = [f(t) - g^2(t)\, \mathbf{s}_\theta(\mathbf{x}, t)]\, dt + g(t)\, d\mathbf{w}$$

[Andersen '82; Song et al '21]

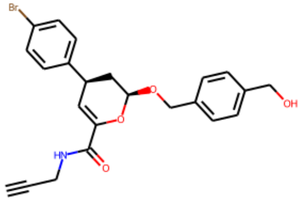
Space of Ligand Poses

A ligand's pose technically is $L \in \mathbb{R}^{3n}$

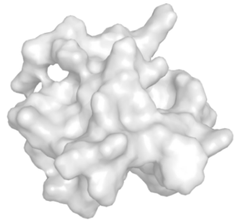
...but docking involves far fewer degrees of freedom

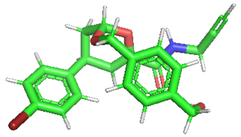


Space of Ligand Poses



Embed in 3D
with RDKit





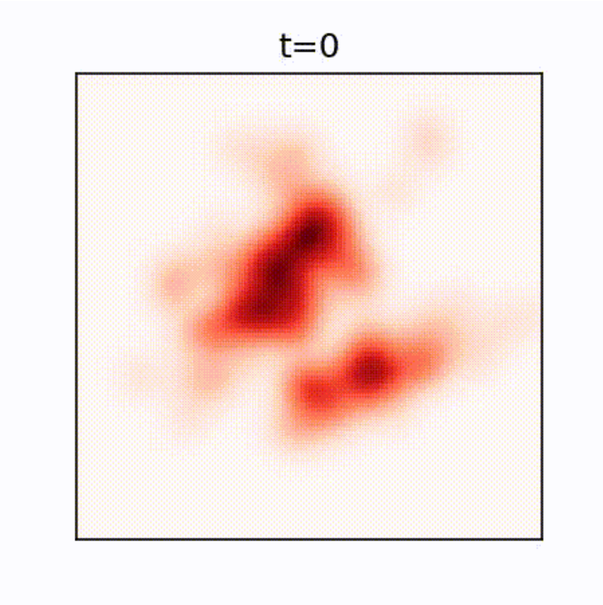
Docking
rigid body motion &
torsional flexibility

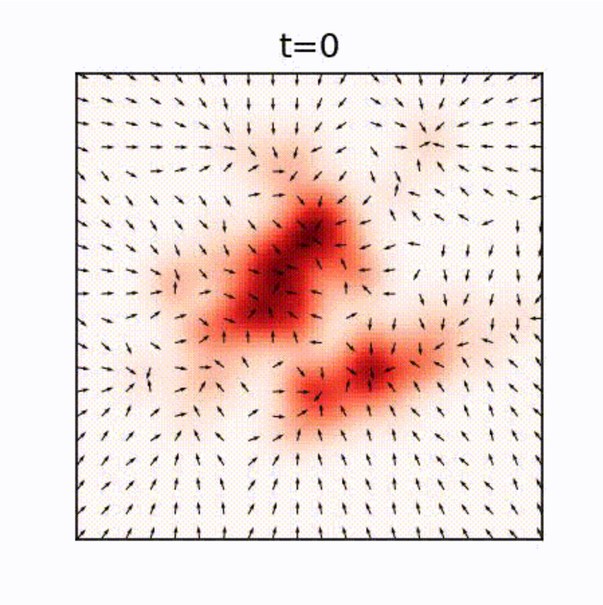
Ligand pose described by

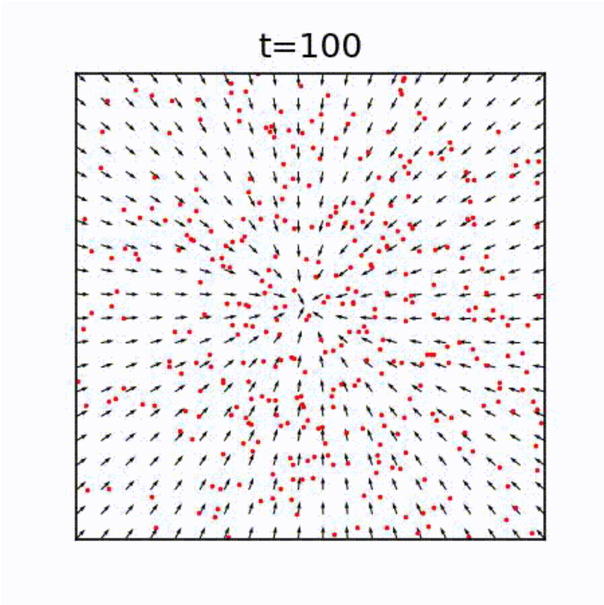
(1) Local structures	(2) Position	(3) Orientation	(4) Torsion angles
Bond lengths Bond angles Chirality Ring structures	Rigid body motion		Torsional flexibility
	Find the pocket	Fit in the pocket	

Keep local structures fixed: diffuse over m+6 dim. submanifold

Diffusion Generative Models







Define the **forward diffusion**

$$d\mathbf{x} = f(\mathbf{x}, t)\, dt + g(t)\, d\mathbf{w}$$

Learn the **score** (gradient of the log density) of the evolving data distribution

$$\mathbf{s}_\theta(\mathbf{x}, t) \approx \nabla_{\mathbf{x}} \log p_t(\mathbf{x})$$

Sample the **reverse diffusion**

$$d\mathbf{x} = [f(t) - g^2(t)\, \mathbf{s}_\theta(\mathbf{x}, t)]\, dt + g(t)\, d\mathbf{w}$$

[Andersen '82; Song et al '21]

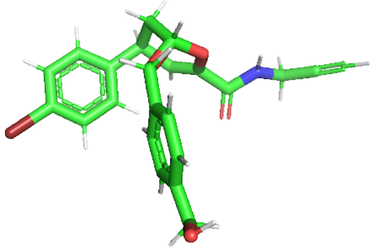
Mapping to the Product Space

Point on ligand pose manifold “parameterized” by:

Position $\in \mathbb{R}^3$
Orientation $\in SO(3)$
Torsions $\in \mathbb{T}^m$

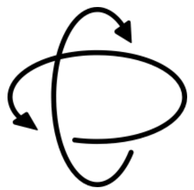
“Diffeomorphic” to product space $\mathbb{R}^3 \times SO(3) \times \mathbb{T}^m$

Need to map **displacements** on the product space to **changes** of pose.

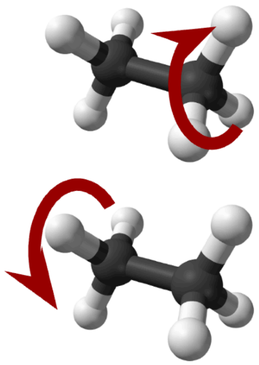


→ Translation **p**
Rotation **ω**
Torsion **τ** = ($\tau_1 \ldots \tau_m$)

→ **?**

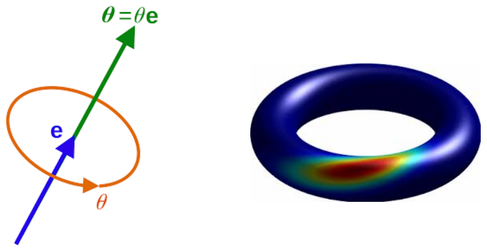


Rotation:
around center of mass

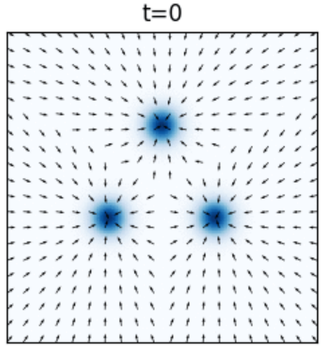


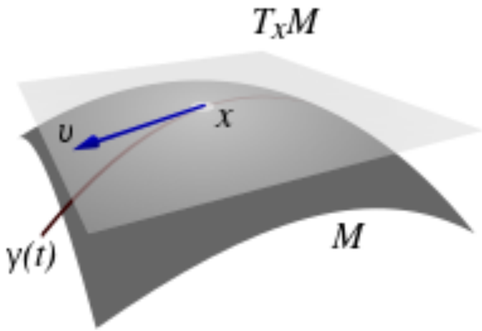
Torsion:
post-torsion RMSD alignment
→ no linear or angular momentum

Product Space Diffusion



Diffusion generative modeling works on manifolds [de Bortoli et al, '22] …provided the score model predicts in the **tangent space**

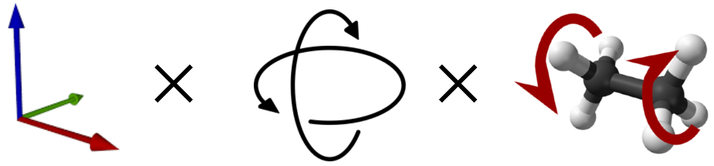




Space	$\mathbb{R}^3$ (position)	SO(3) (orientations)	$\mathbb{T}^m$ (torsion angles)
Tangent space	$\mathbb{R}^3$ (translation vectors)	$\mathbb{R}^3$ (rotation vectors)	$\mathbb{R}^m$ (torsion updates)
Heat kernel	Normal	IGSO(3)	Wrapped normal
Stationary dist.	Normal	Uniform	Uniform
SE(3) symmetry	Equivariant	Equivariant	Invariant

…and that we can:

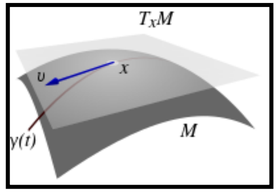
1. sample the heat kernel for arbitrary t
2. compute its score
3. sample from the stationary distribution (t = T)



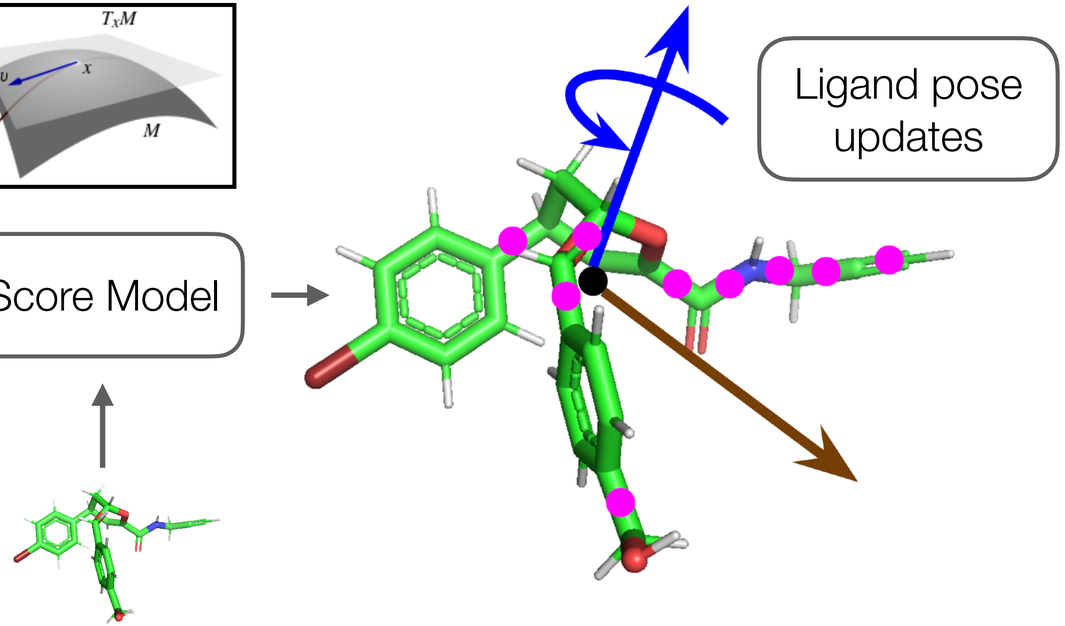
Steps × **Turns** × **Twists**

Score Model

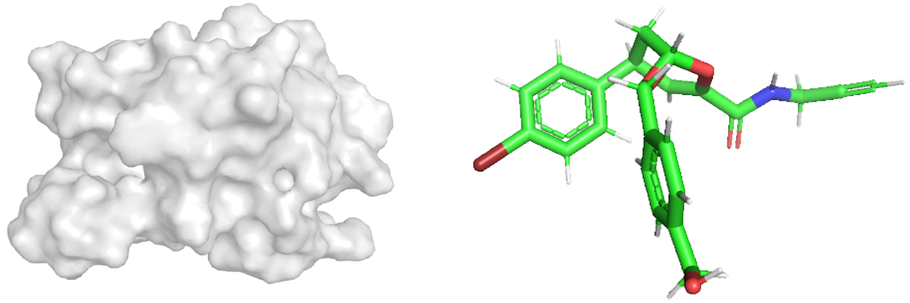
Space	$\mathbb{R}^3$ (position)	SO(3) (orientations)	$\mathbb{T}^m$ (torsion angles)
Tangent space	$\mathbb{R}^3$ (translation vec.)	$\mathbb{R}^3$ (rotation vectors)	$\mathbb{R}^m$ (torsion updates)
SE(3) symmetry	Equivariant	Equivariant	Invariant



Score Model

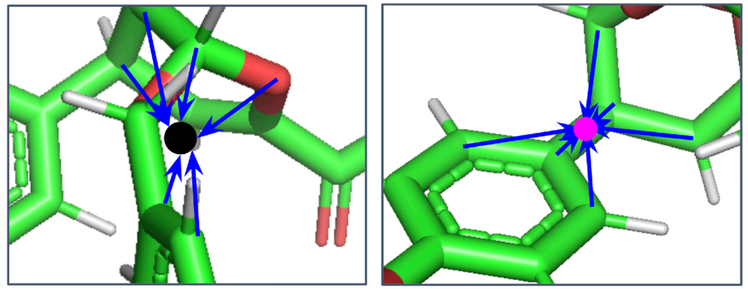


e3nn [Geiger & Smidt]

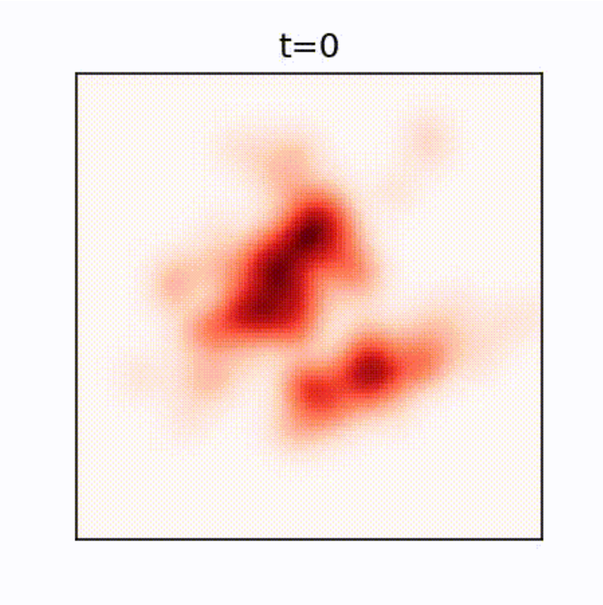


K ×

E(3)NN tensor product conv.

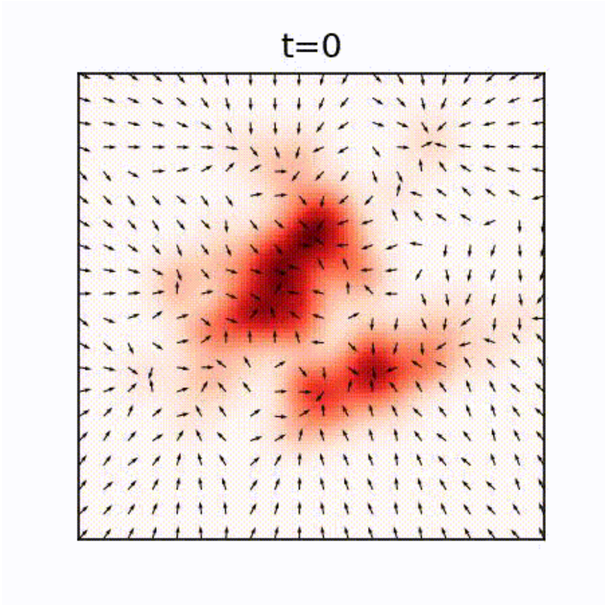


Diffusion Generative Models



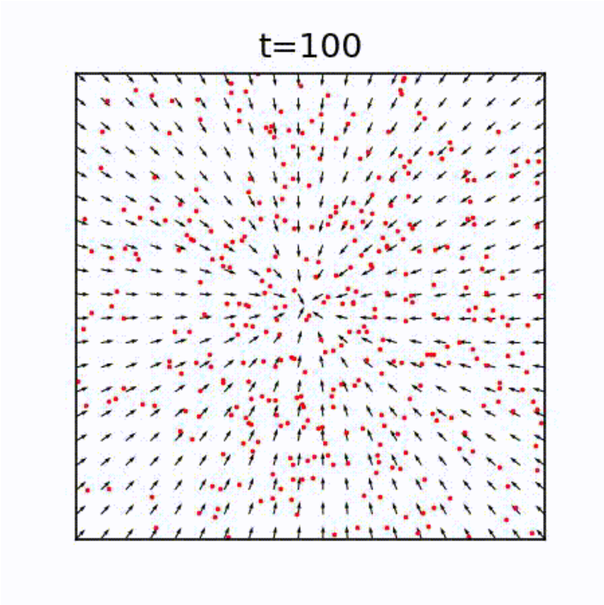
Define the **forward diffusion**

$$d\mathbf{x} = f(\mathbf{x}, t)\, dt + g(t)\, d\mathbf{w}$$



Learn the **score** (gradient of the log density) of the evolving data distribution

$$\mathbf{s}_\theta(\mathbf{x}, t) \approx \nabla_{\mathbf{x}} \log p_t(\mathbf{x})$$



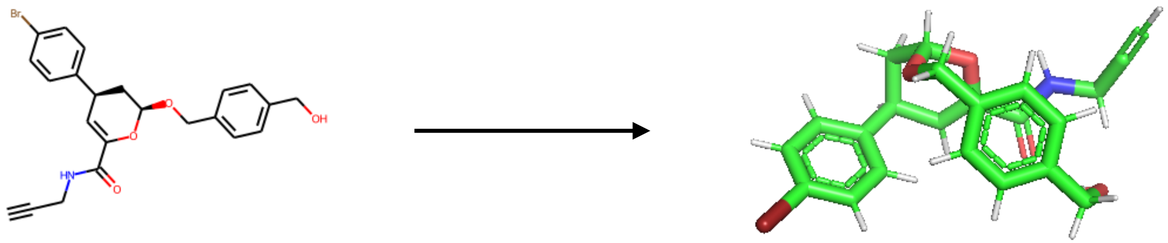
Sample the **reverse diffusion**

$$d\mathbf{x} = [f(t) - g^2(t)\, \mathbf{s}_\theta(\mathbf{x}, t)]\, dt + g(t)\, d\mathbf{w}$$

[Andersen '82; Song et al '21]

Sampling

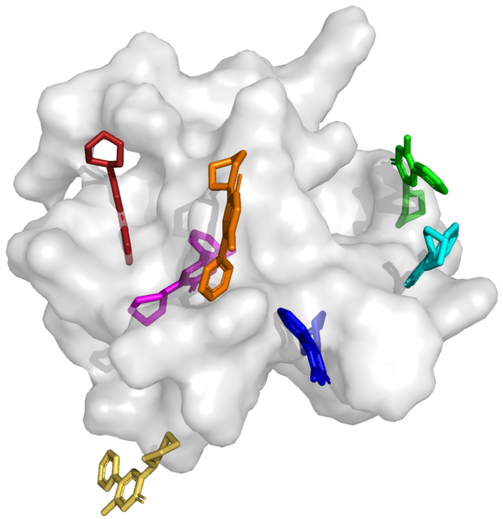
1. **Embed with RDKit**
2. Sample N random poses
3. Simulate reverse diffusion
4. Rank and select top M poses



Sampling

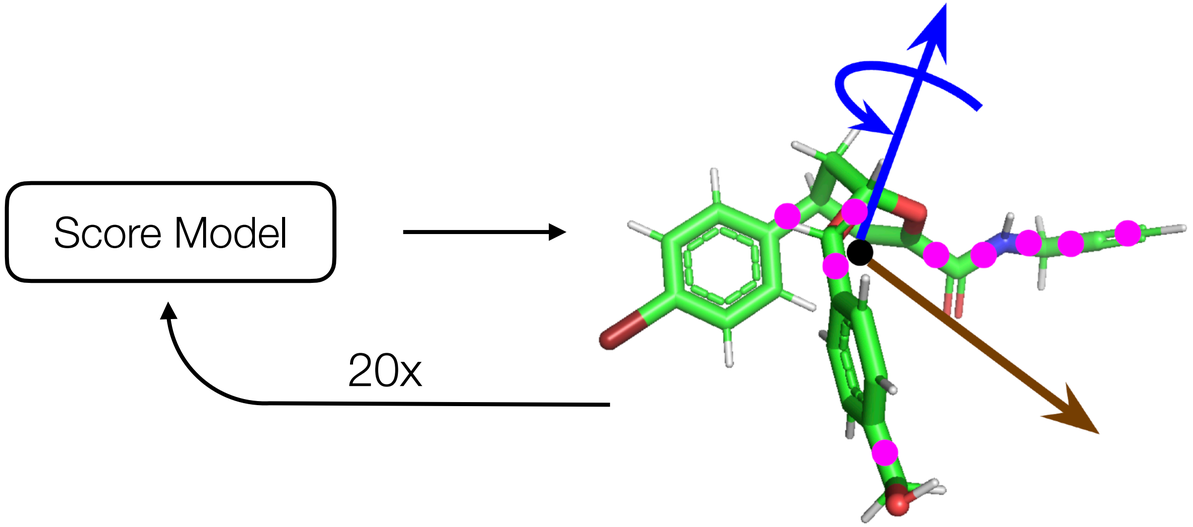
1. Embed with RDKit
2. **Sample N random poses**
3. Simulate reverse diffusion
4. Rank and select top M poses

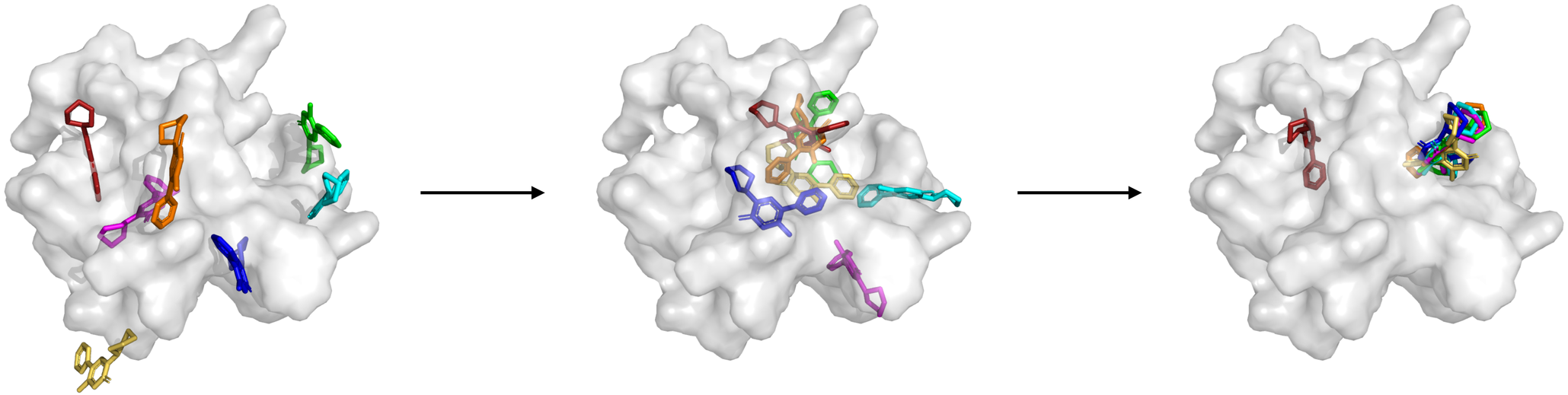
Space	$\mathbb{R}^3$ (position)	SO(3) (orientations)	$\mathbb{T}^m$ (torsion angles)
Stationary distribution	Normal	Uniform	Uniform



Sampling

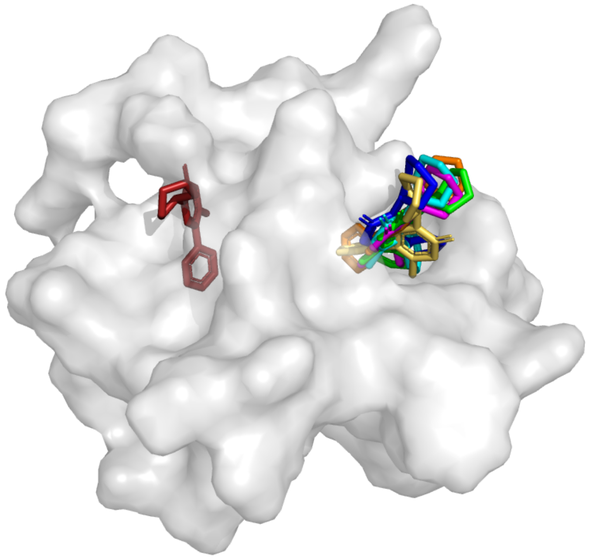
1. Embed with RDKit
2. Sample N random poses
3. **Simulate reverse diffusion**
4. Rank and select top M poses



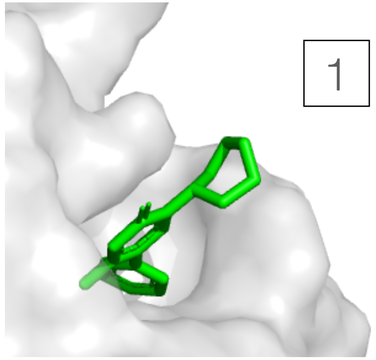


Sampling

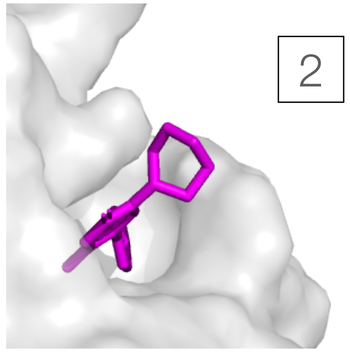
1. Embed with RDKit
2. Sample N random poses
3. Simulate reverse diffusion
4. **Rank and select top M poses**



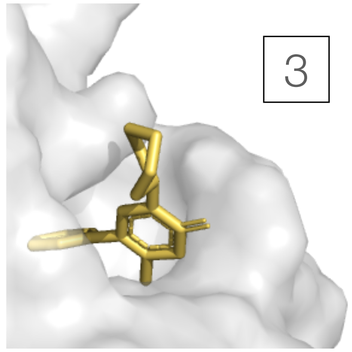
Confidence Model



Score: +1.3

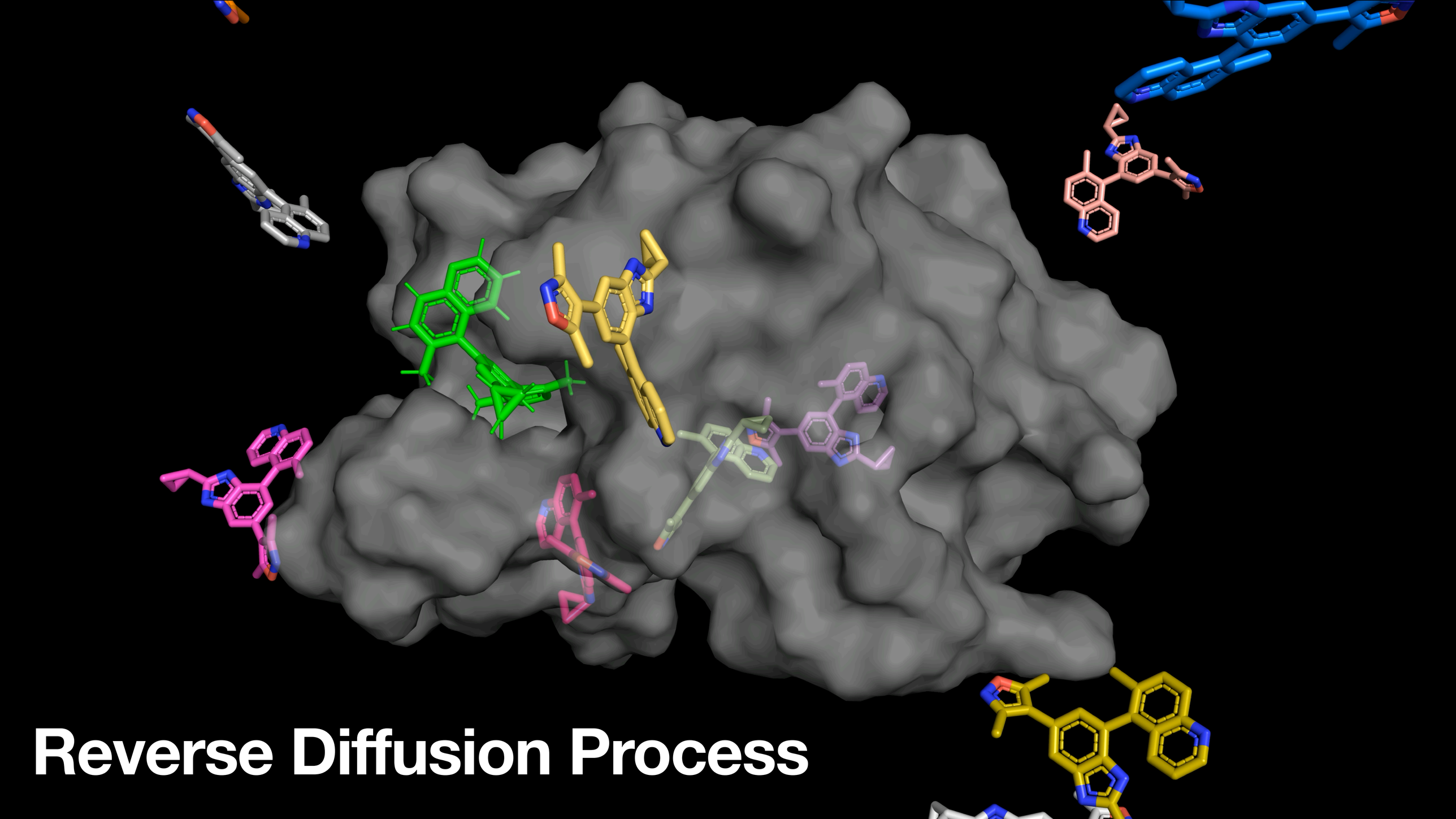


Score: +1.1



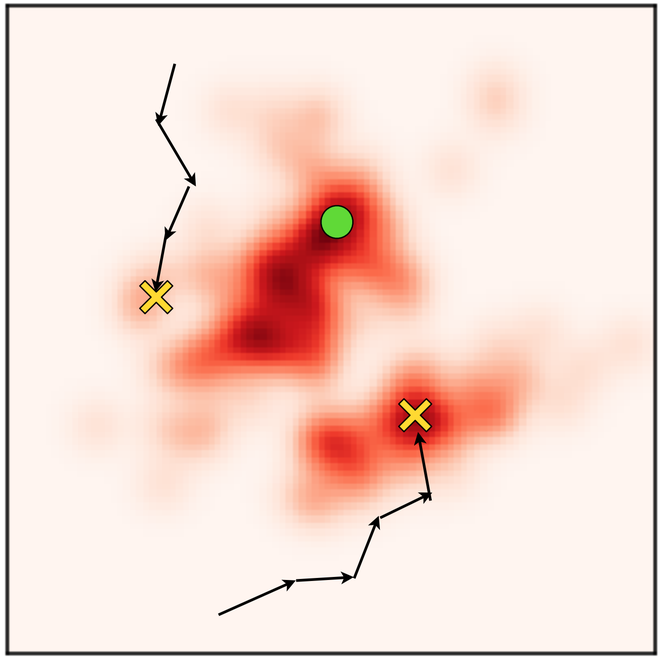
Score: +0.8

...

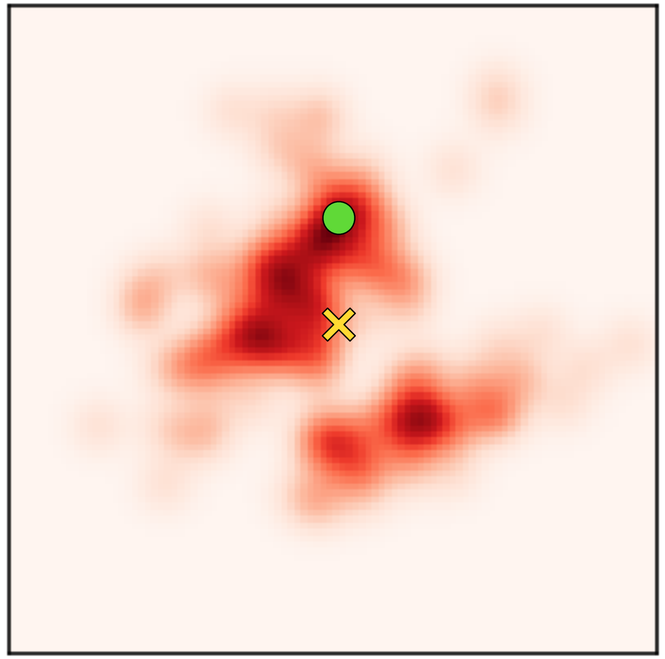


Reverse Diffusion Process

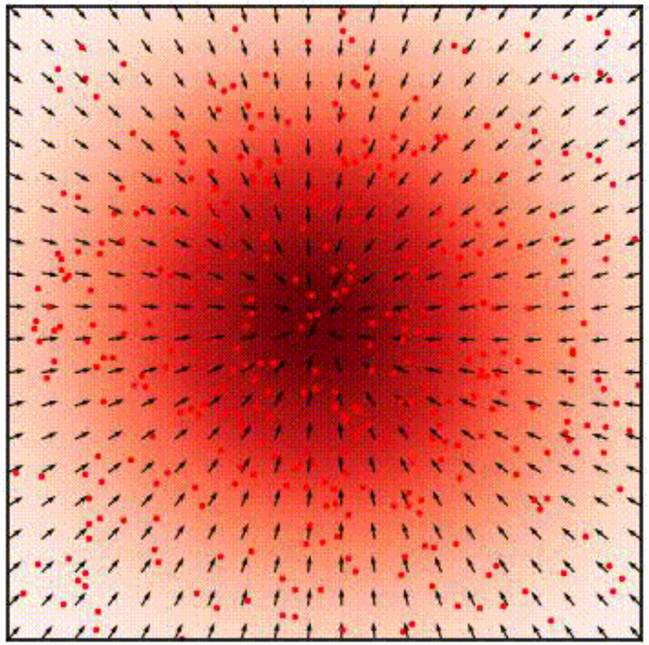
Approaches to docking recap



Traditional docking: sampling & optimization over scoring function: **no finite-time guarantees**!



Previous deep learning: poor-quality single prediction and no refinement



Diffusion: sample from **nonconvex** density in **finite time** via a **time-evolving** vector field

Results

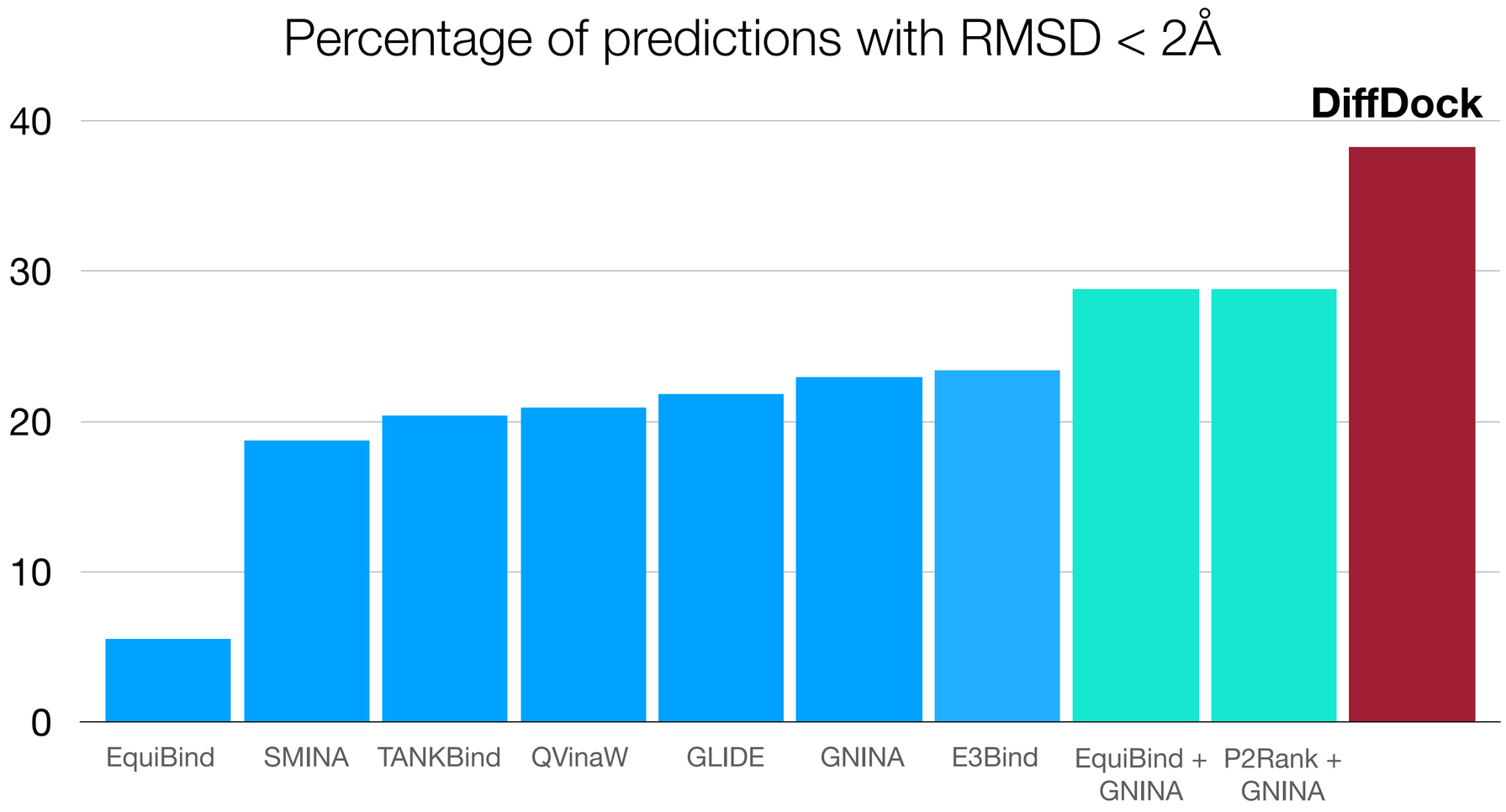
Standard benchmark PDBBind

19k experimentally determined structures of small molecules + proteins

Baselines: search-based and deep learning

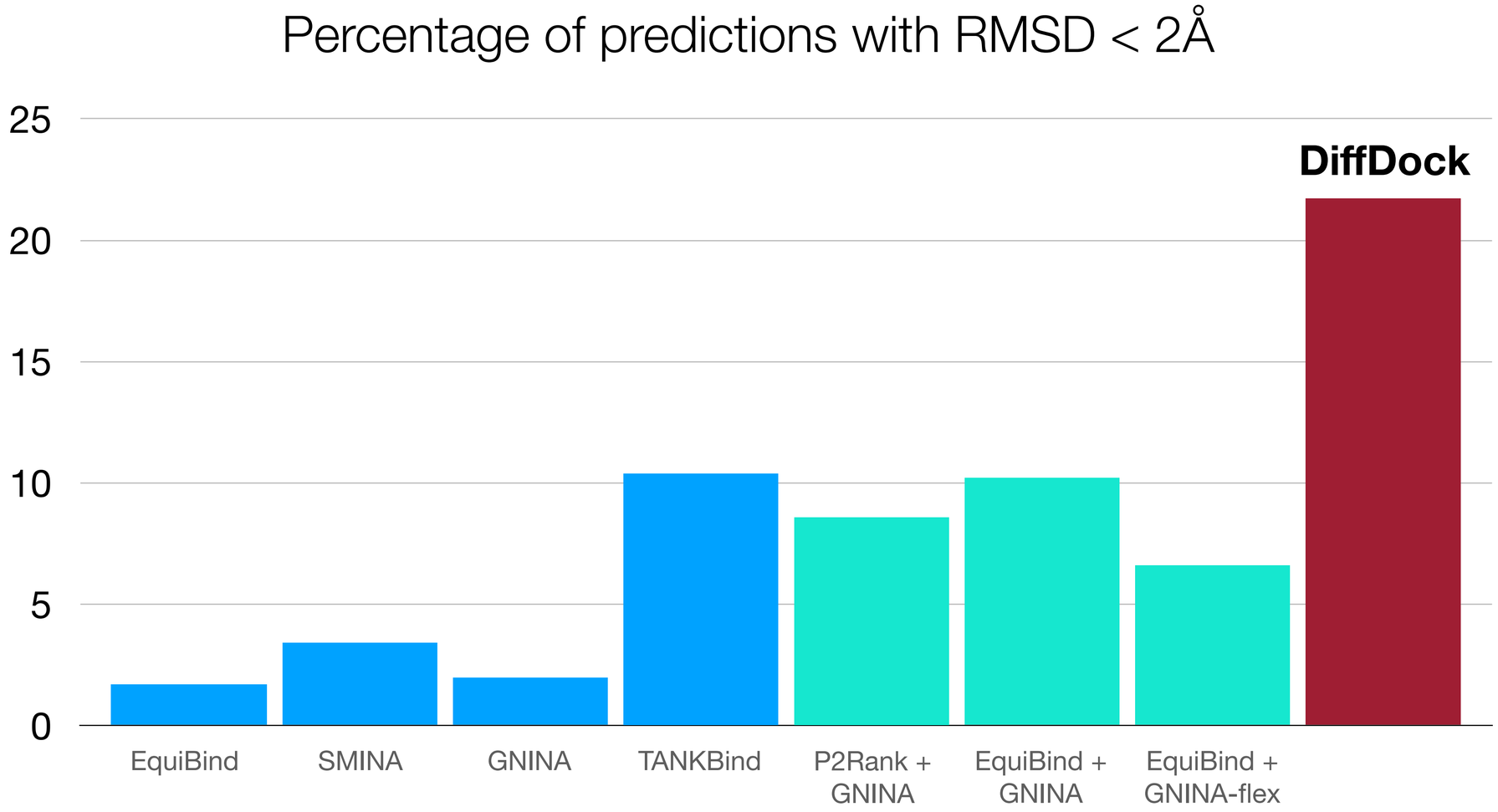
GNINA	SMINA	QuickVina-W	GLIDE	EquiBind	TankBind
McNutt et al. 2021	Koes et al. 2013	Hassan et al. 2017	Schrödinger. Release 2021-4	Stärk et al. 2022	Lu et al. 2022

Blind docking on holo-proteins



Outperform search-based, deep learning, and pocket prediction + search-based methods

Blind docking on predicted structures



Retains significantly higher accuracy on ESMFold structures

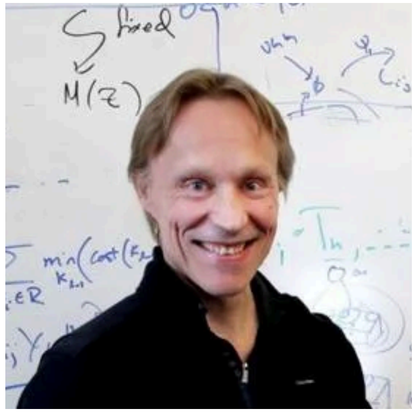
DiffDock

Diffusion Steps, Twists, and Turns for Molecular Docking

Gabriele Corso* Hannes Stärk* Bowen Jing* Regina Barzilay Tommi Jaakkola







All links in our GitHub: https://github.com/gcorso/DiffDock