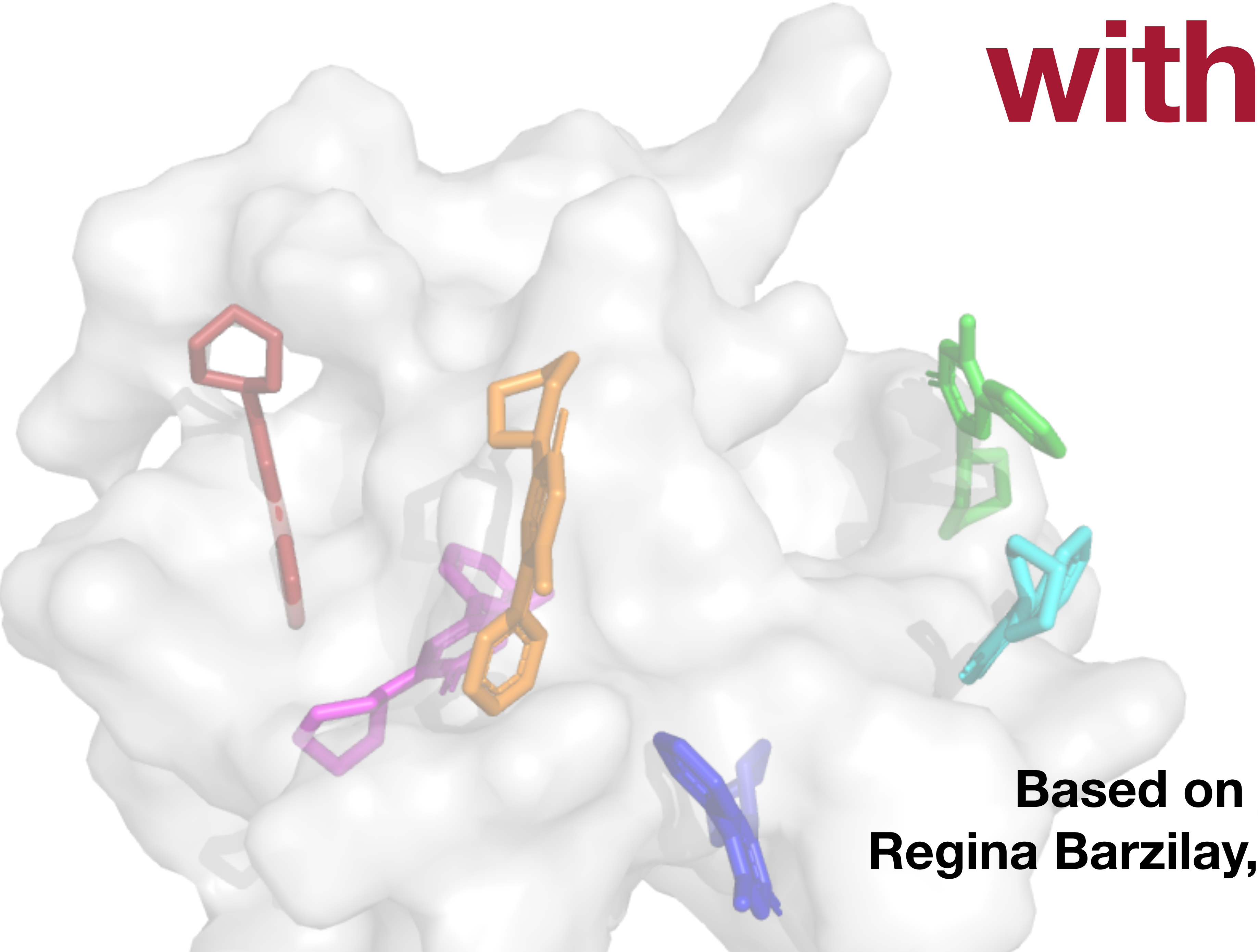
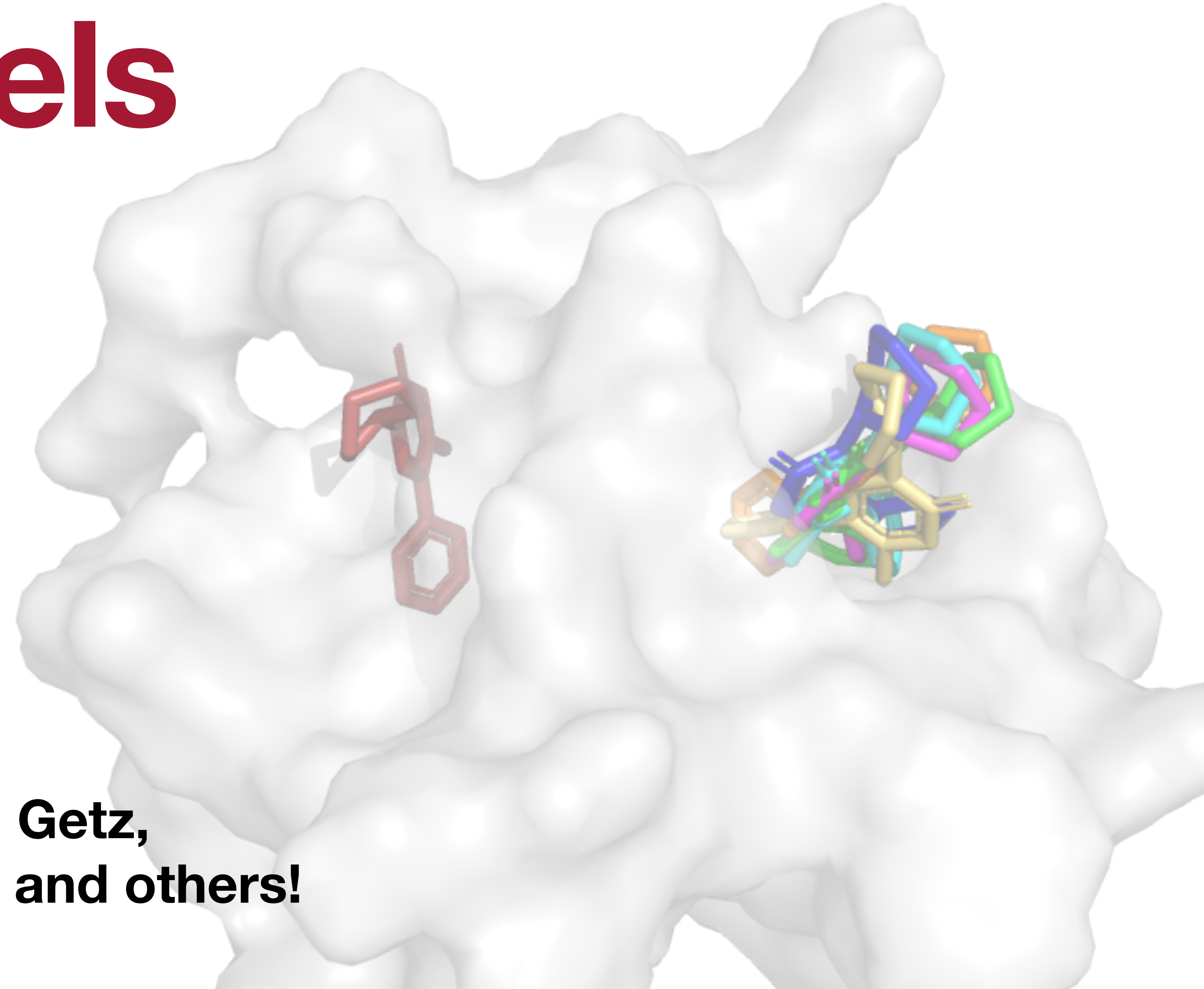
DiffDock & FlexDock

Advancing Molecular Docking with Generative Models

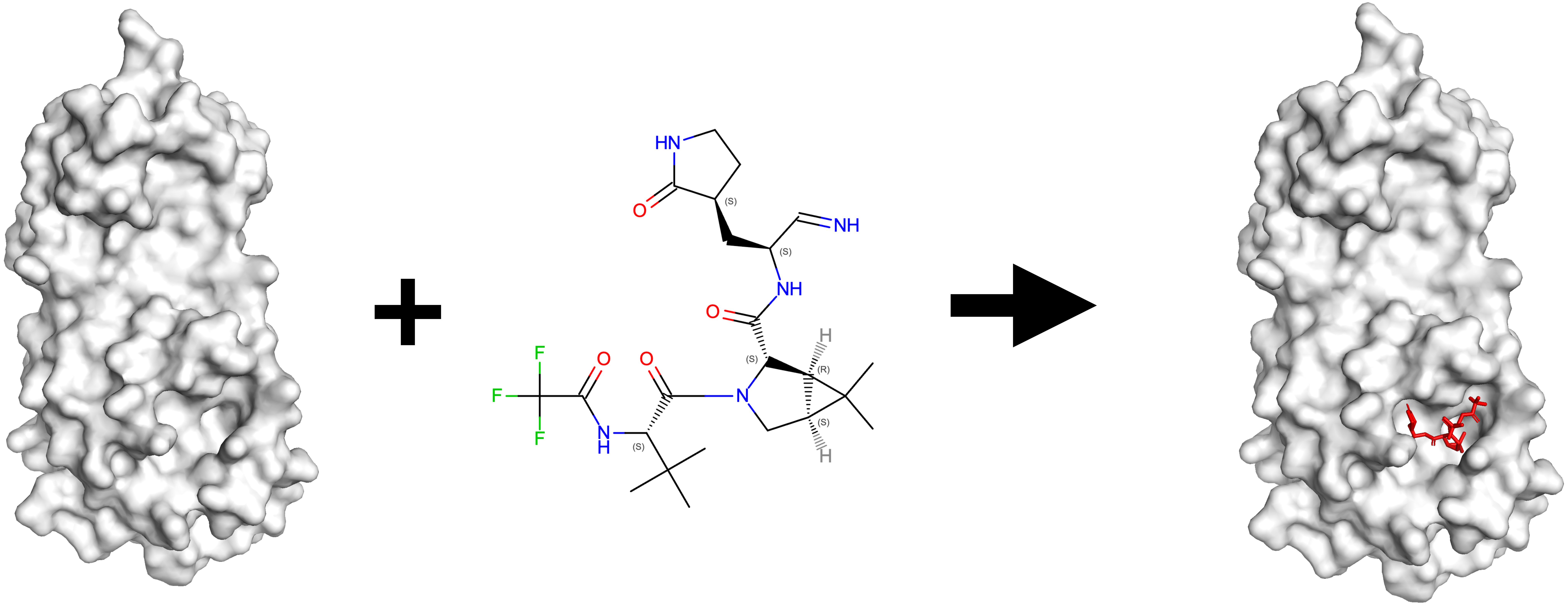
Gabriele Corso





Based on joint work with Vignesh Ram Somnath, Noah Getz,
Regina Barzilay, Tommi Jaakkola, Hannes Stärk, Bowen Jing and others!

Protein-Ligand Docking



Input: protein structure + molecule

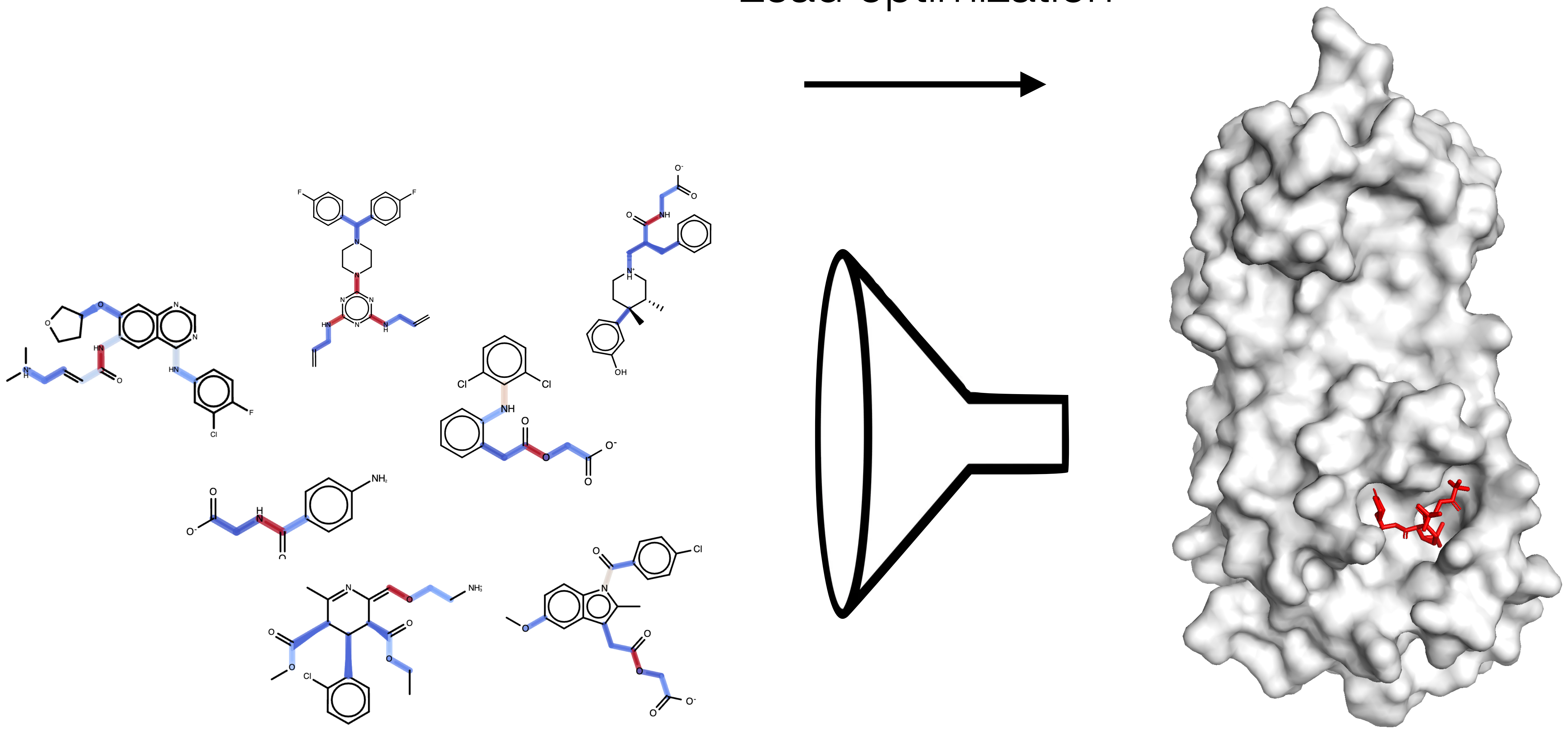
Output: bound structure

Protein-Ligand Docking

Virtual screening

Hit discovery

Lead optimization



Protein-Ligand Docking

Virtual screening

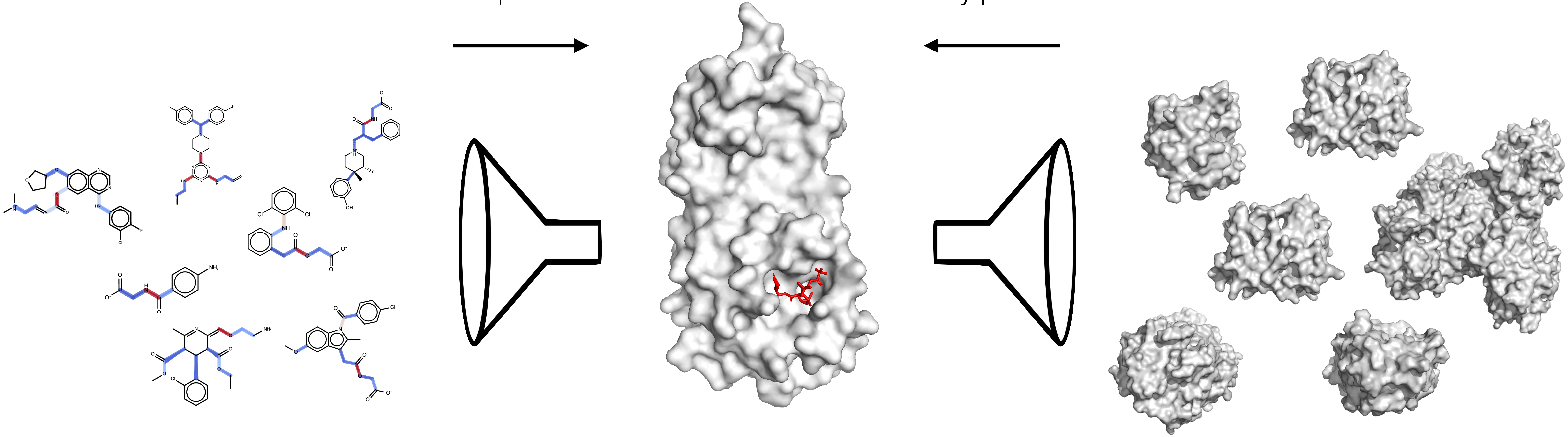
Hit discovery

Lead optimization

Reverse screening

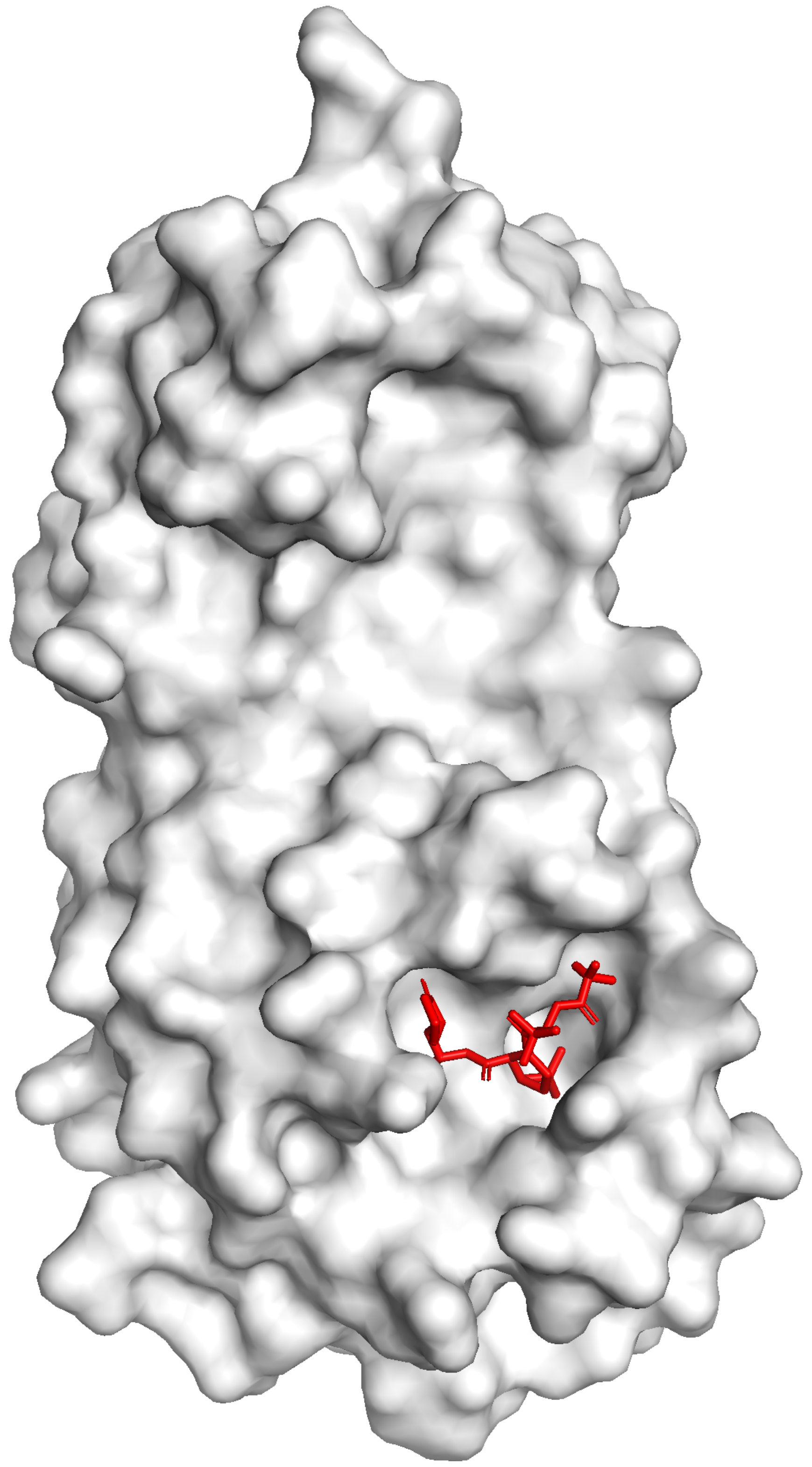
MoA identification

Toxicity prediction



We are NOT doing sampling

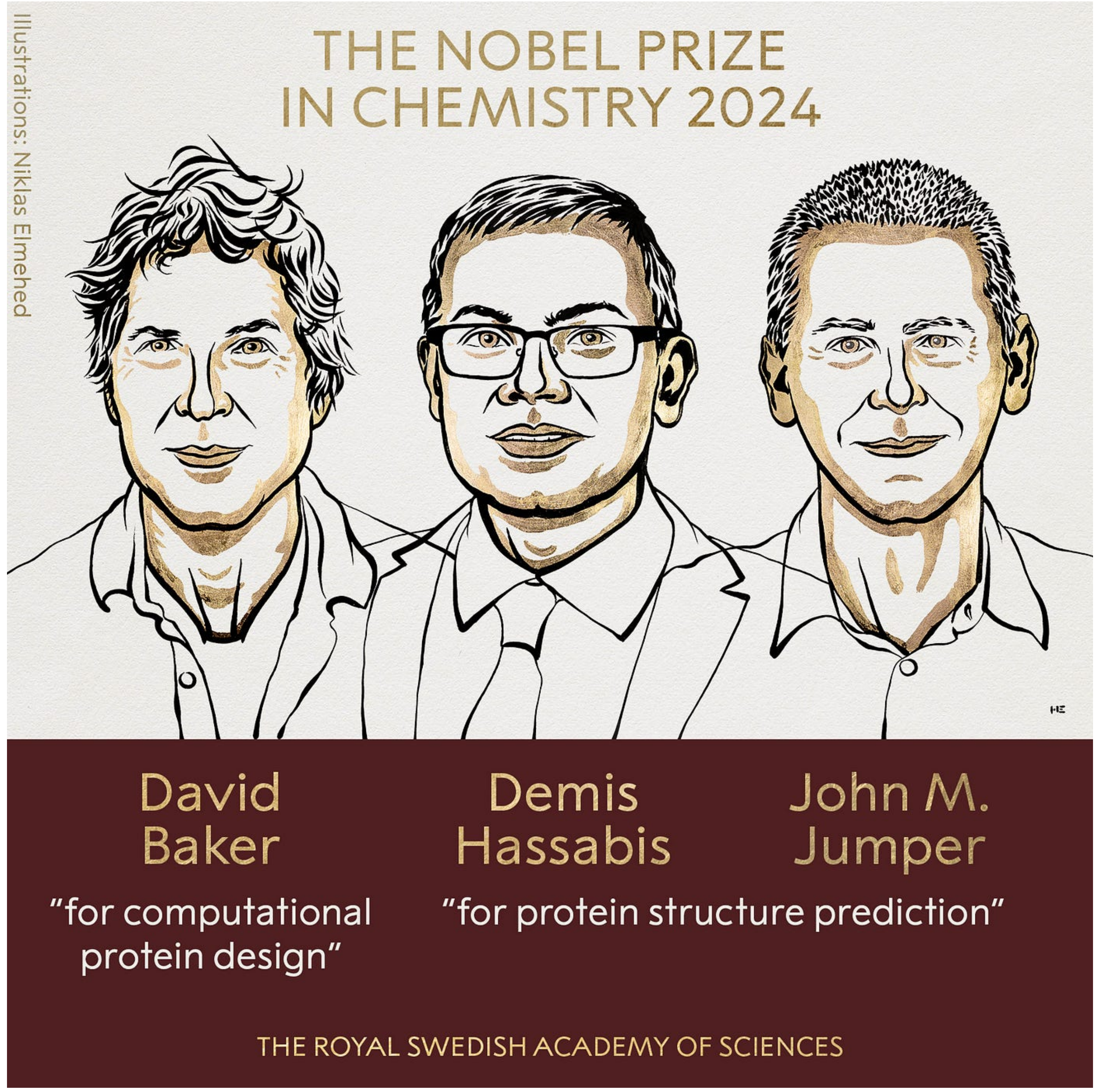
“It’s fake”…



Output: bound structure

We are NOT doing sampling

“It’s fake”… but it is useful



Illustrations: Niklas Elmehed

THE NOBEL PRIZE
IN CHEMISTRY 2024

David
Baker

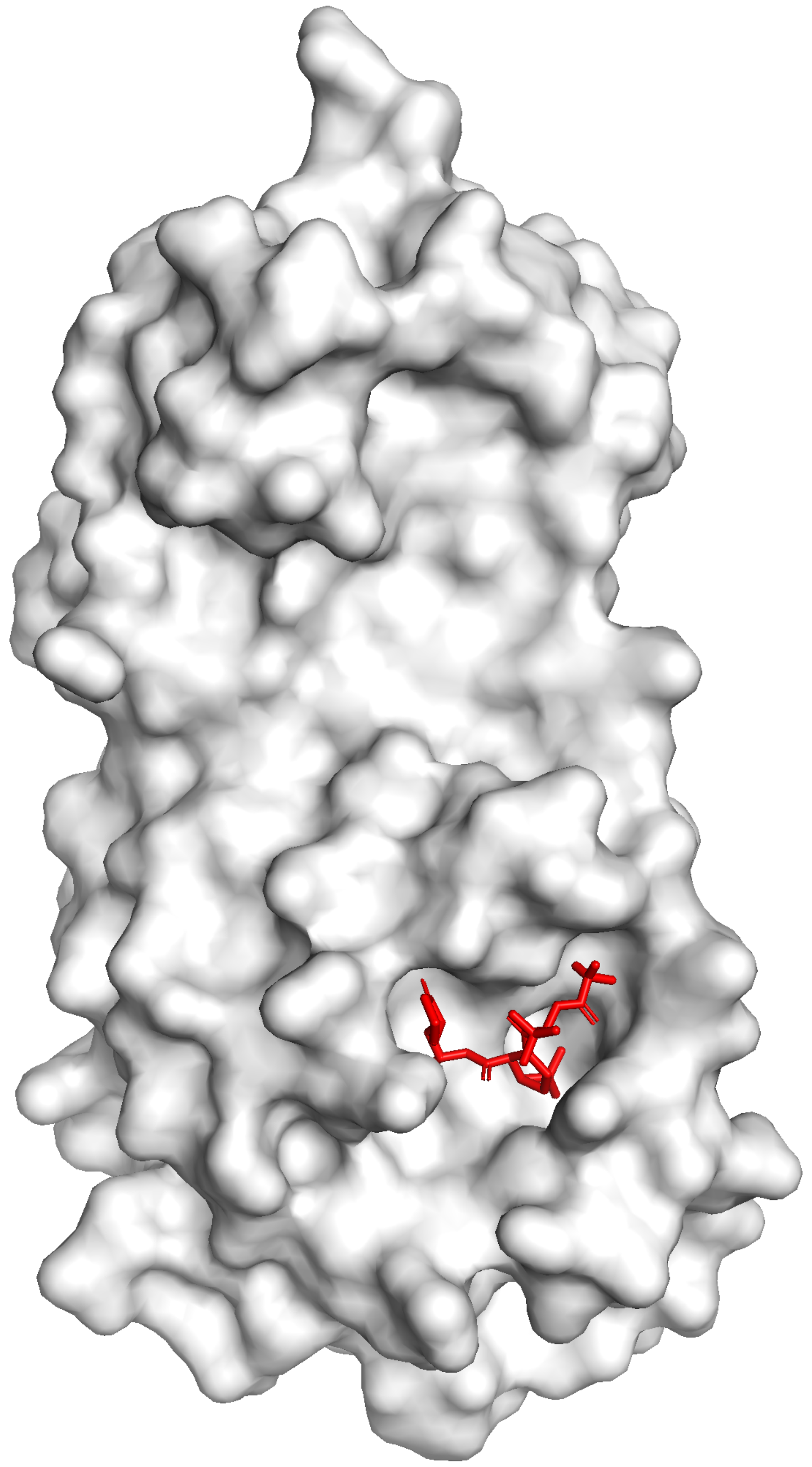
“for computational
protein design”

Demis
Hassabis

John M.
Jumper

“for protein structure prediction”

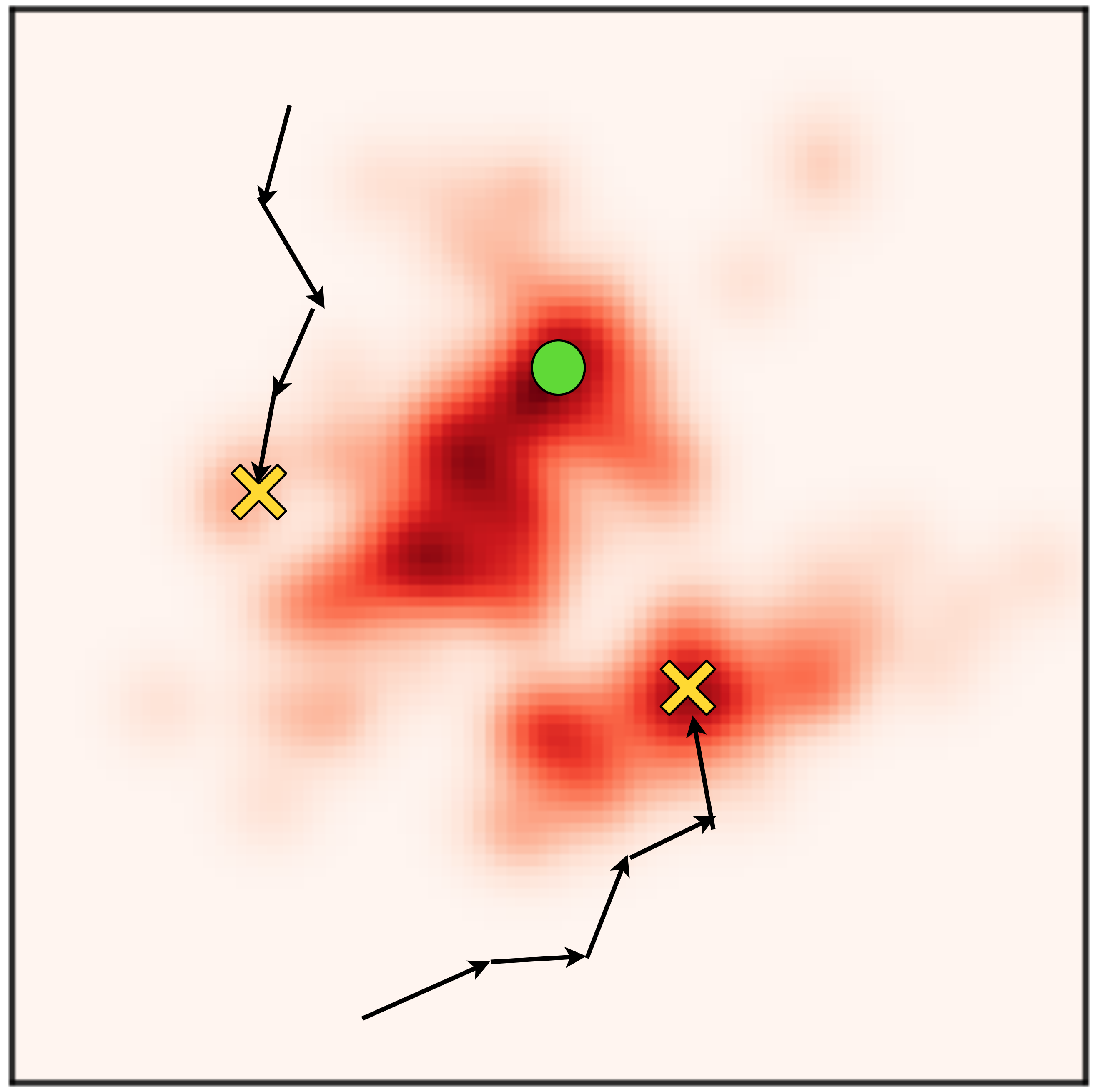
THE ROYAL SWEDISH ACADEMY OF SCIENCES



Output: bound structure

Different Approaches to Docking

Search-based methods

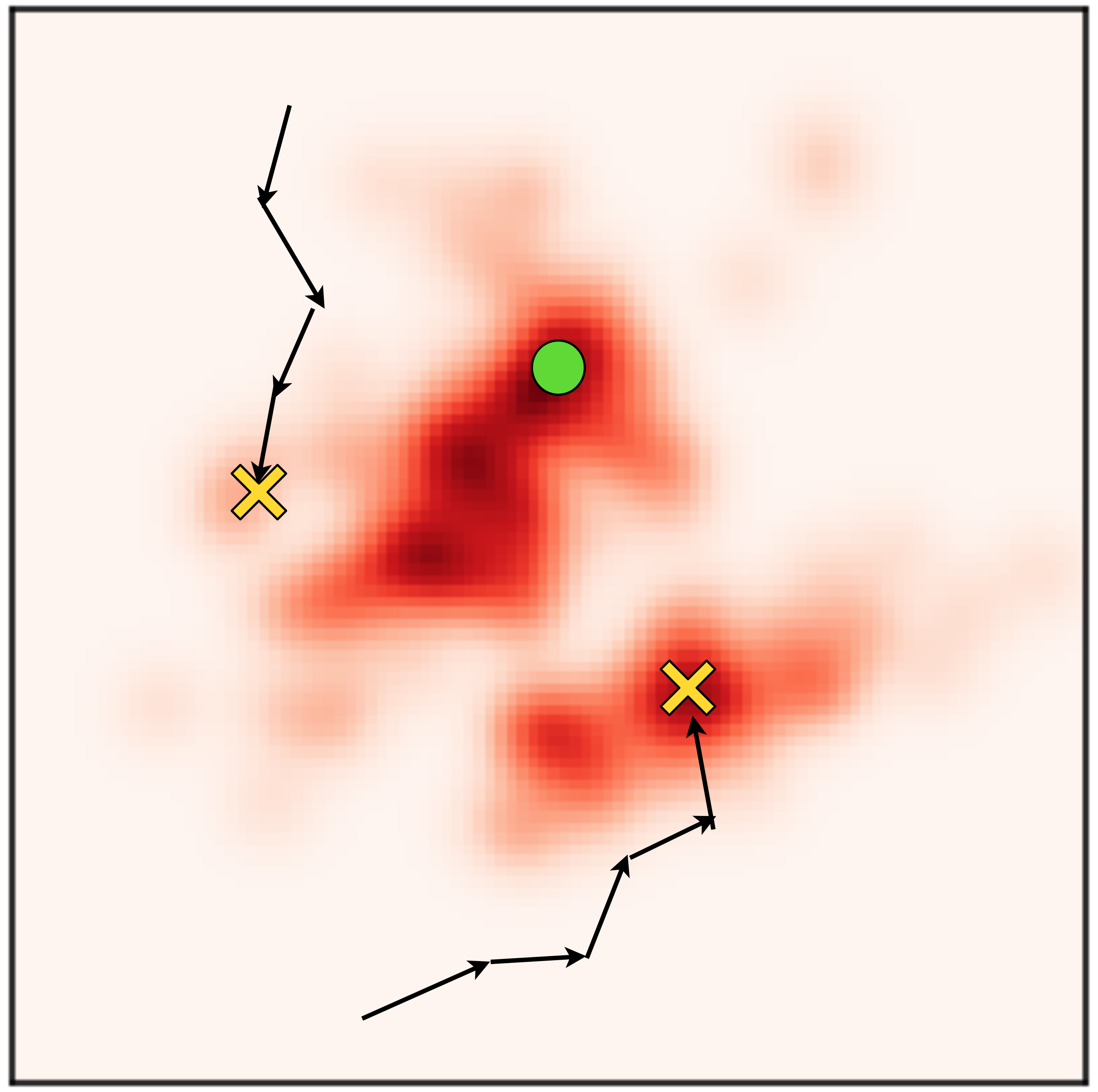


Sampling & optimization
over scoring function

Corso, Stark, Jing, Barzilay, Jaakkola. ICLR 2023

Different Approaches to Docking

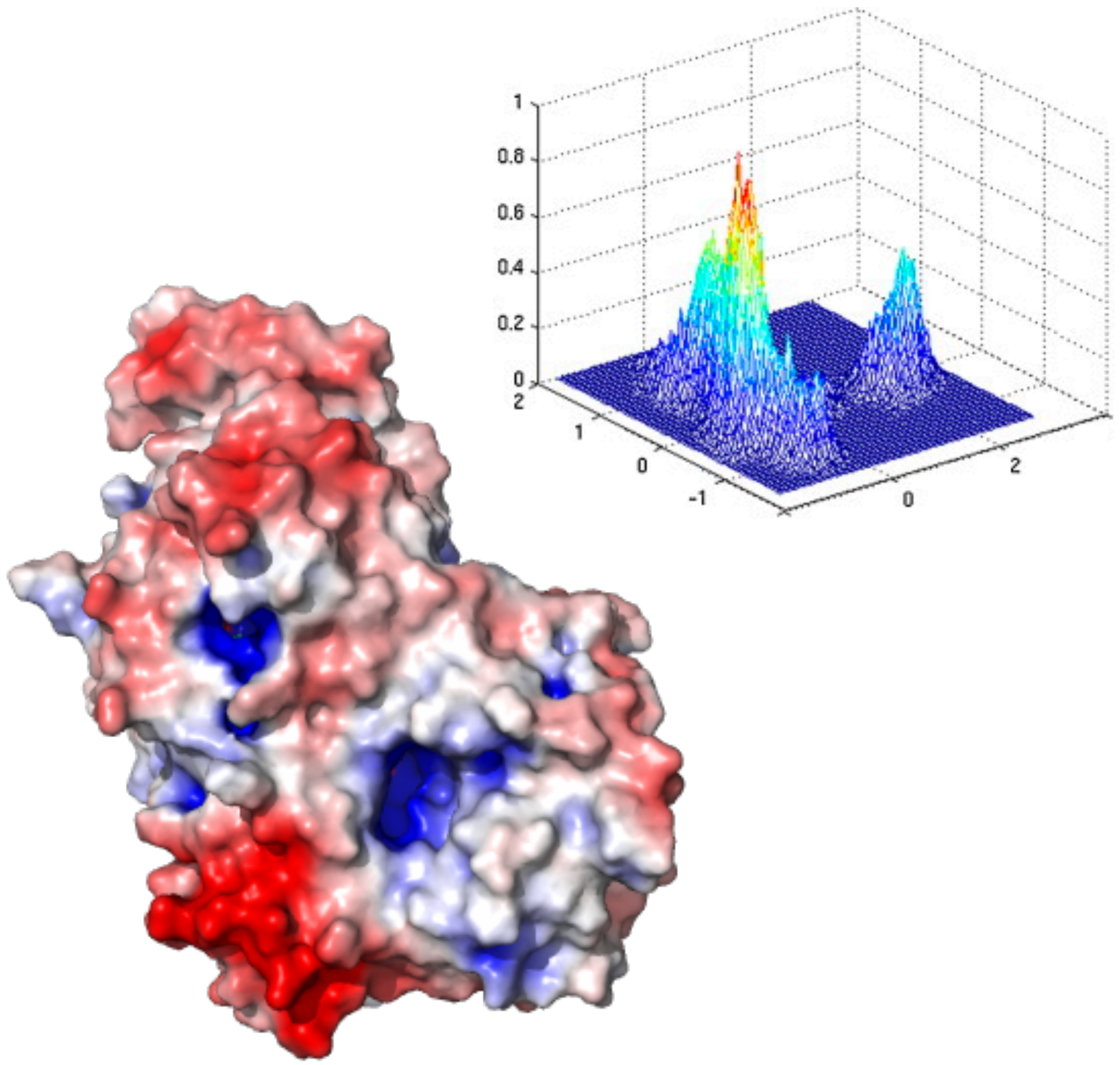
Search-based methods



Sampling & optimization
over scoring function

➜ **no finite-time guarantees**

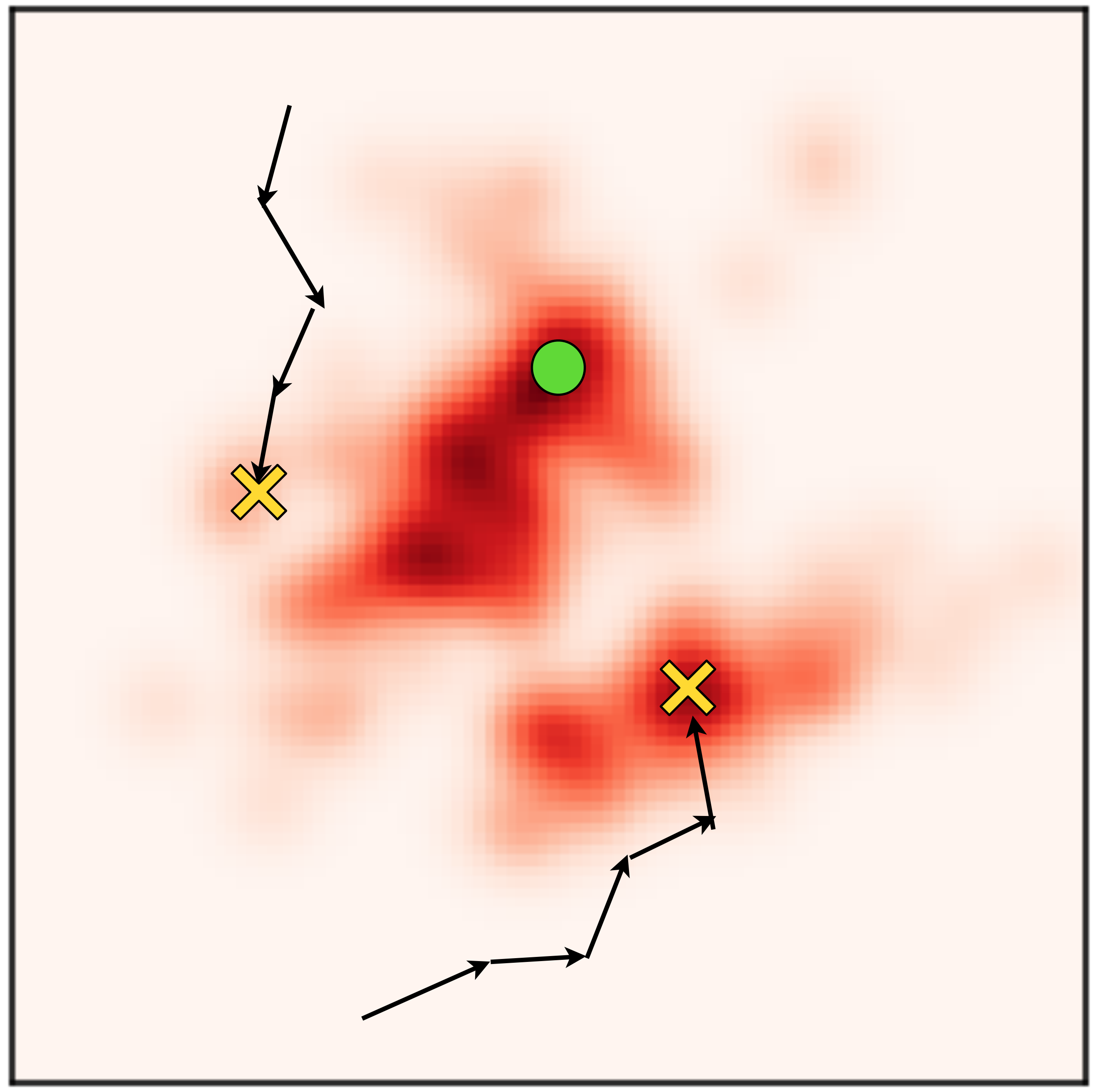
- Fail to grasp with the **vast search space** of blind docking



Corso, Stark, Jing, Barzilay, Jaakkola. ICLR 2023

Different Approaches to Docking

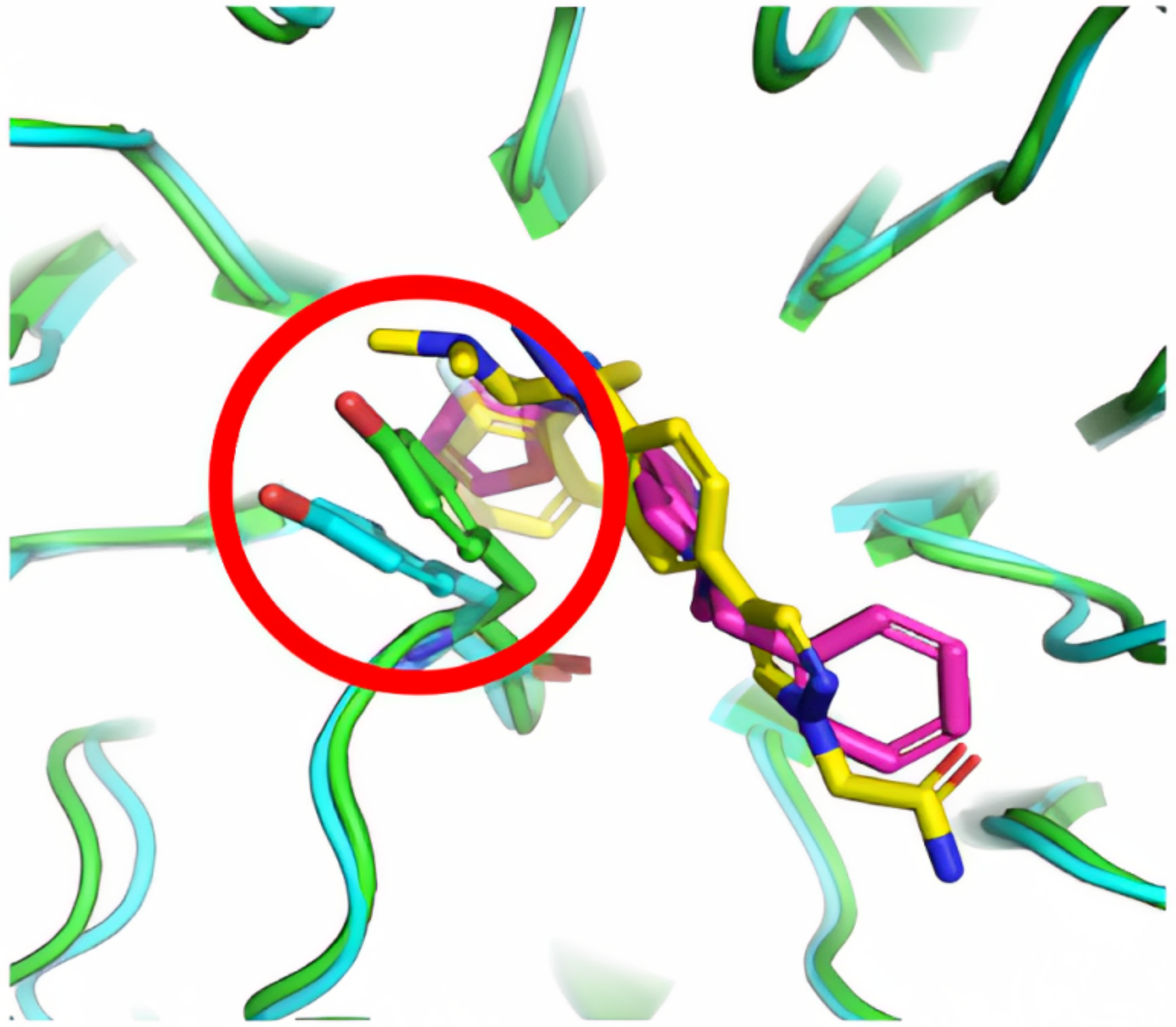
Search-based methods



Sampling & optimization over scoring function

➜ **no finite-time guarantees**

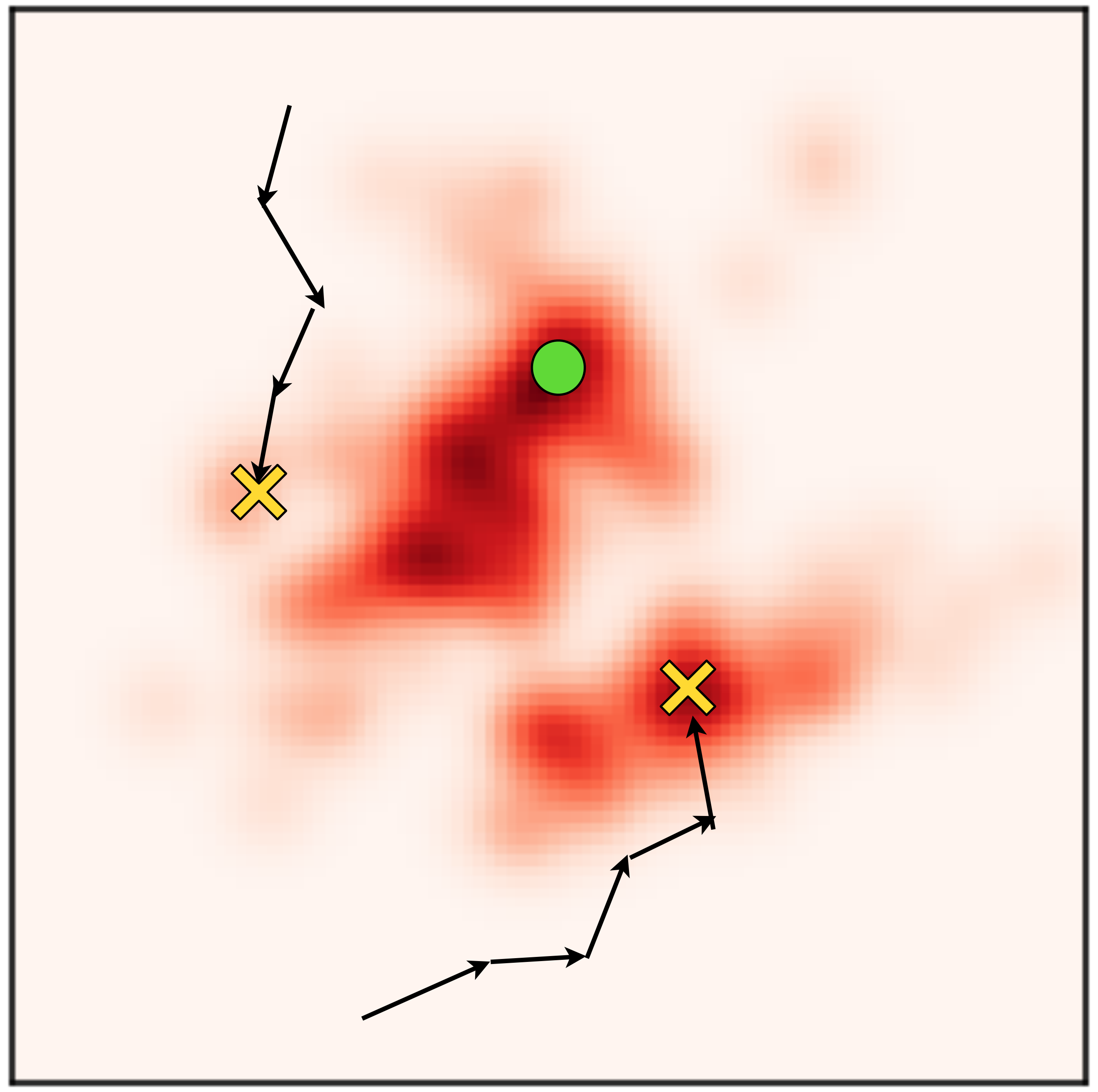
- Fail to grasp with the **vast search space** of blind docking
- **Struggle** with, e.g., side chain **flexibility** from unbound to bound protein structures



Corso, Stark, Jing, Barzilay, Jaakkola. ICLR 2023

Different Approaches to Docking

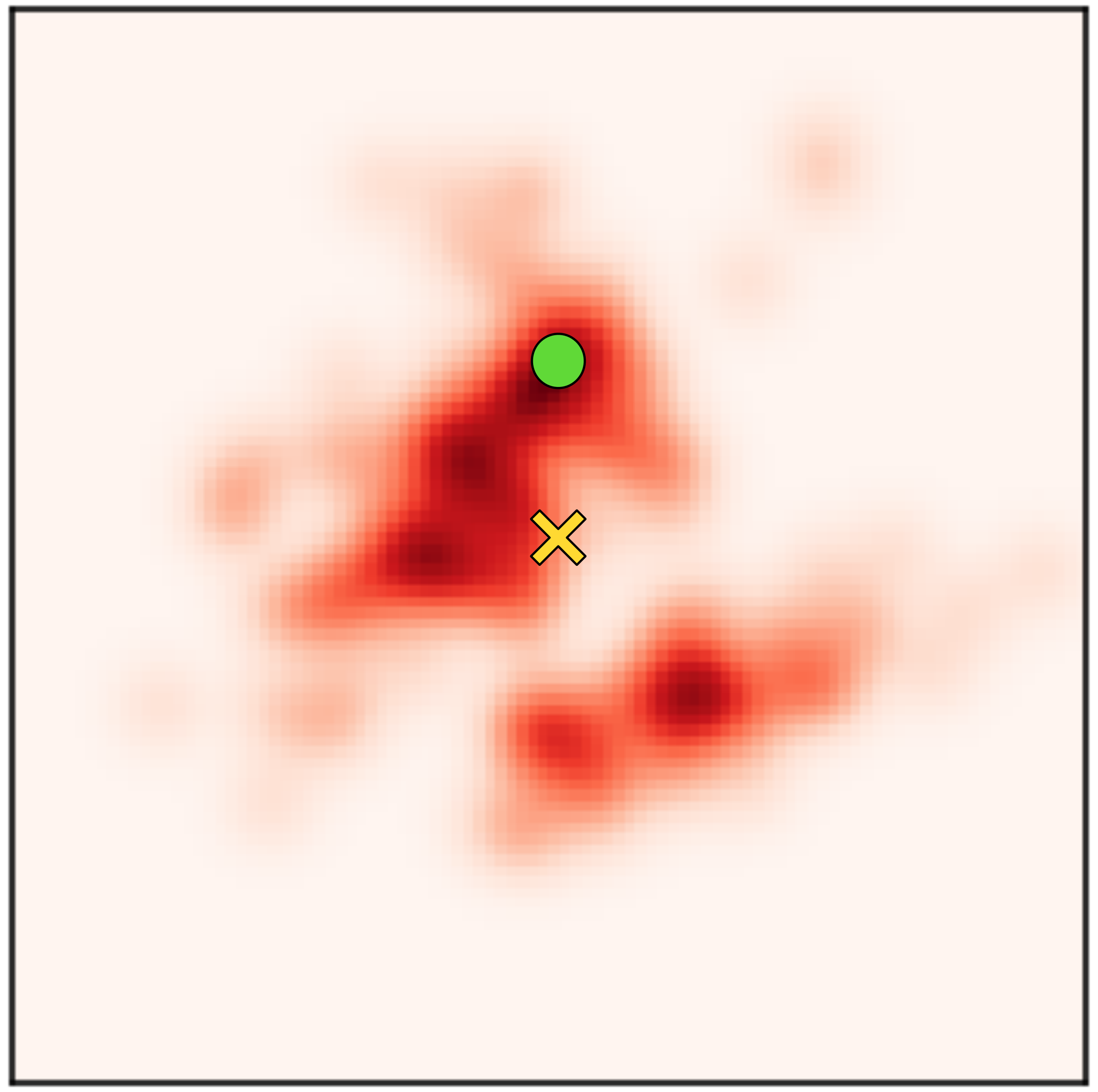
Search-based methods



Sampling & optimization
over scoring function

➜ **no finite-time guarantees**

Regression models

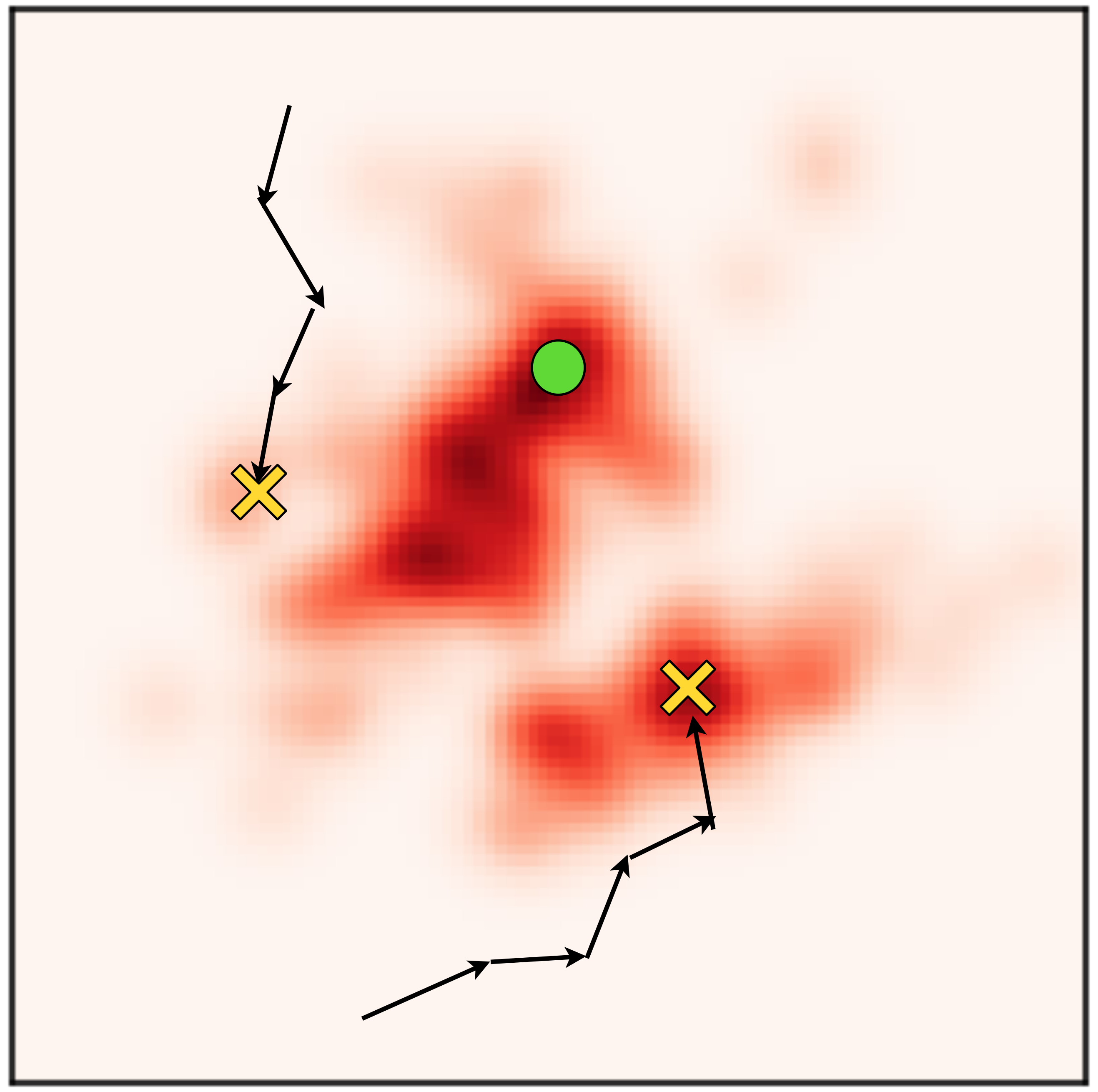


Previous deep learning methods were
based on regression objective

Corso, Stark, Jing, Barzilay, Jaakkola. ICLR 2023

Different Approaches to Docking

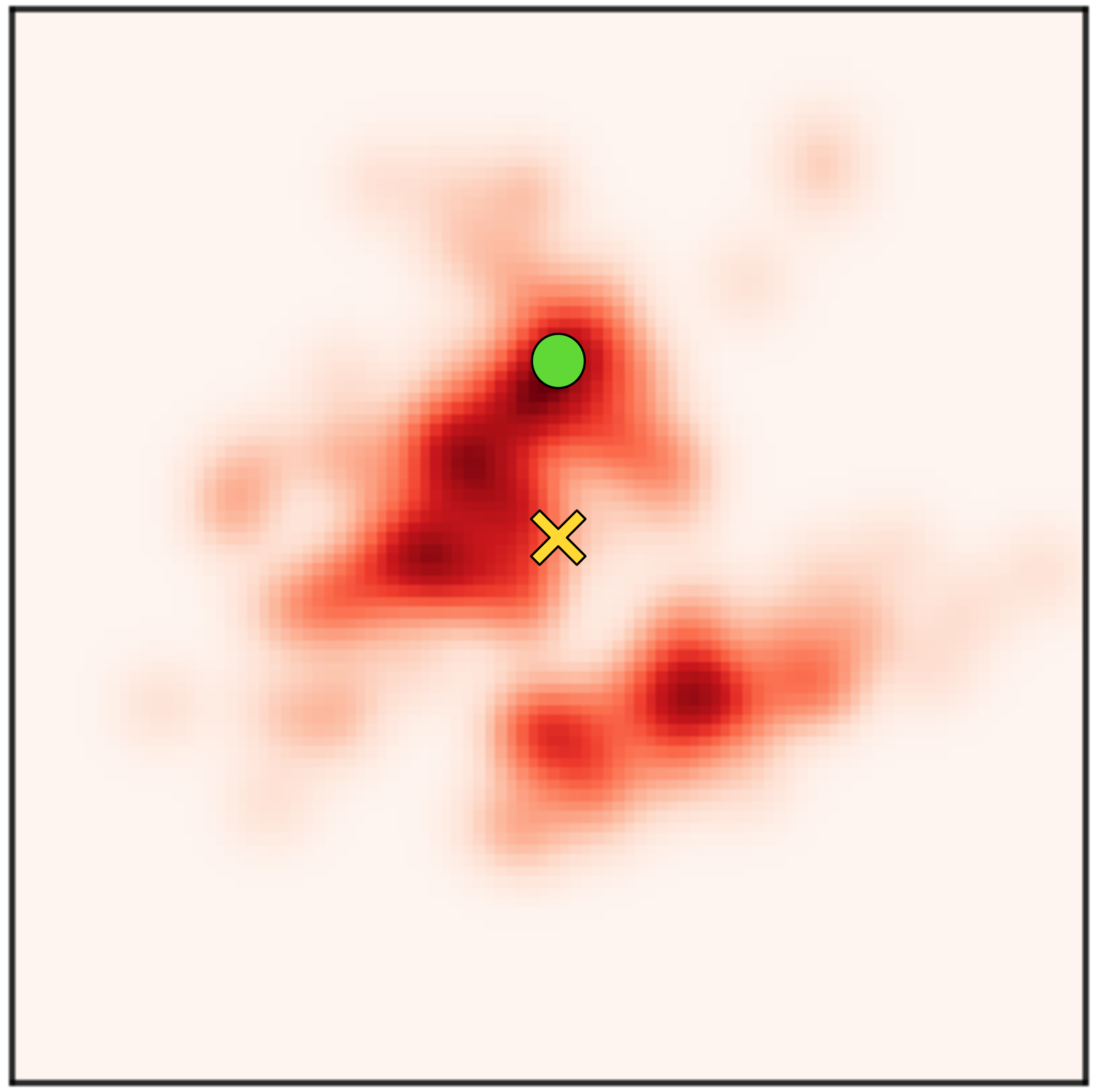
Search-based methods



Sampling & optimization
over scoring function

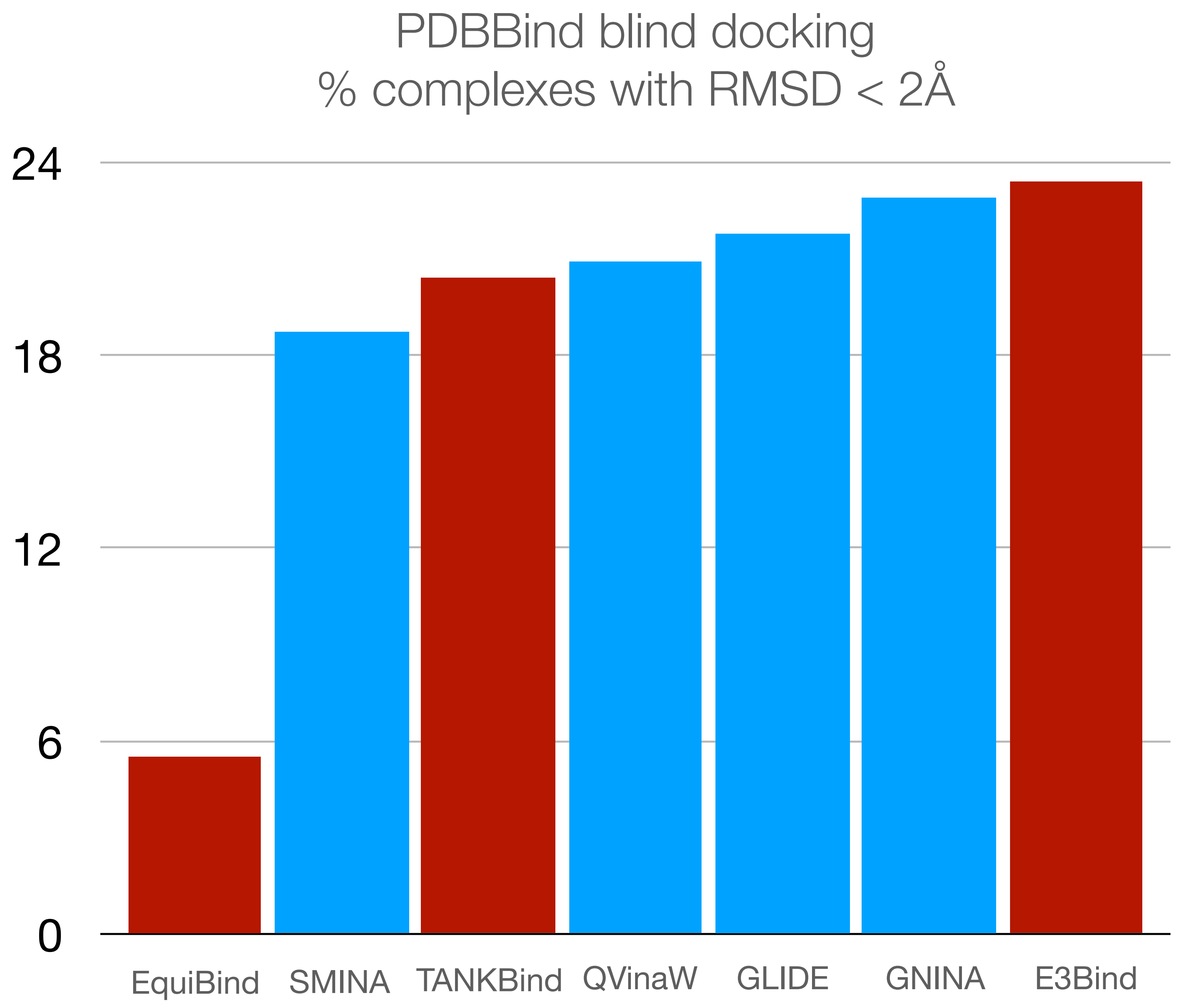
➜ **no finite-time guarantees**

Regression models



Previous deep learning methods were
based on regression objective

➜ **fast but poor-quality predictions**

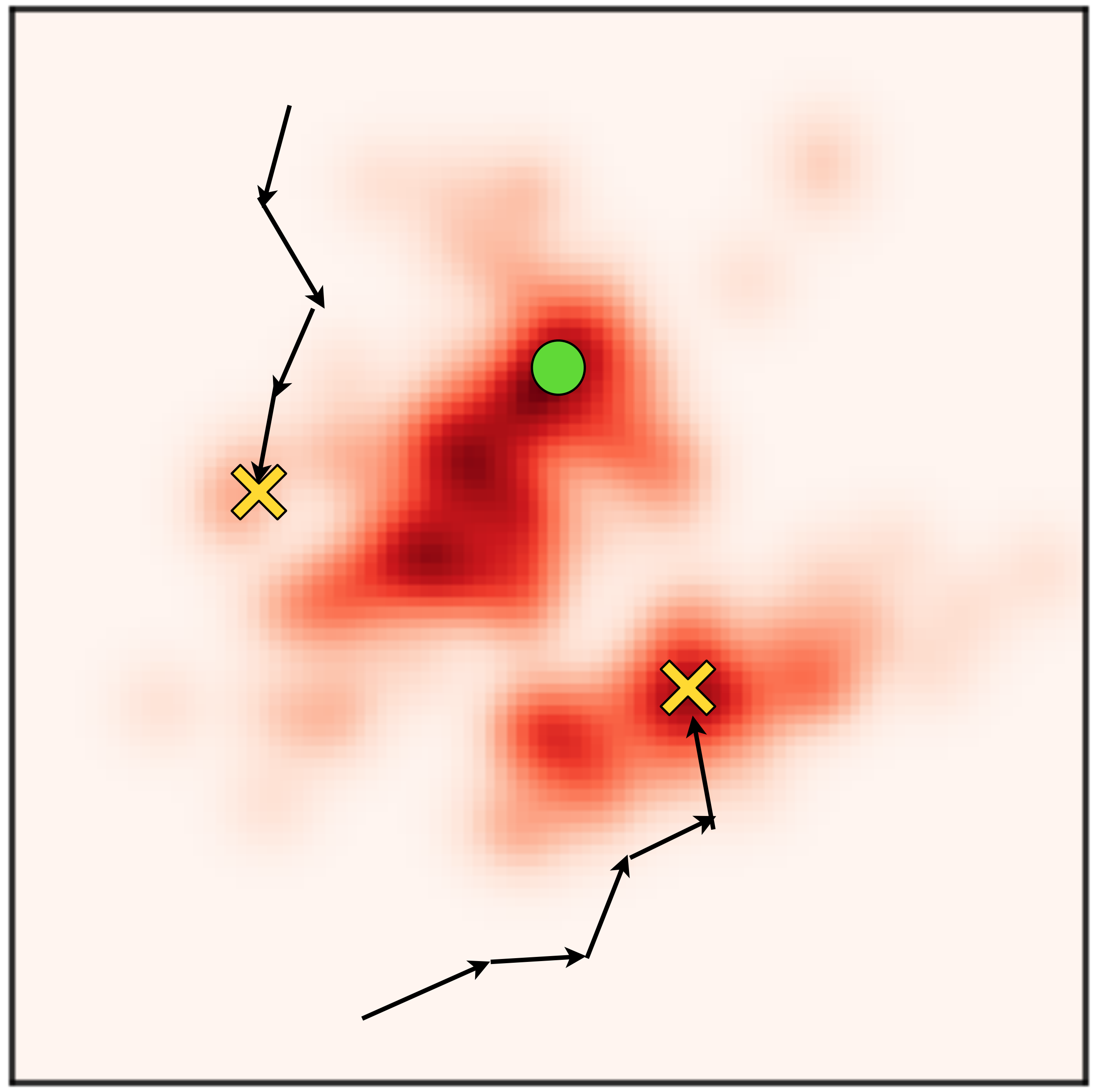


No clear improvement to existing models

Corso, Stark, Jing, Barzilay, Jaakkola. ICLR 2023

Different Approaches to Docking

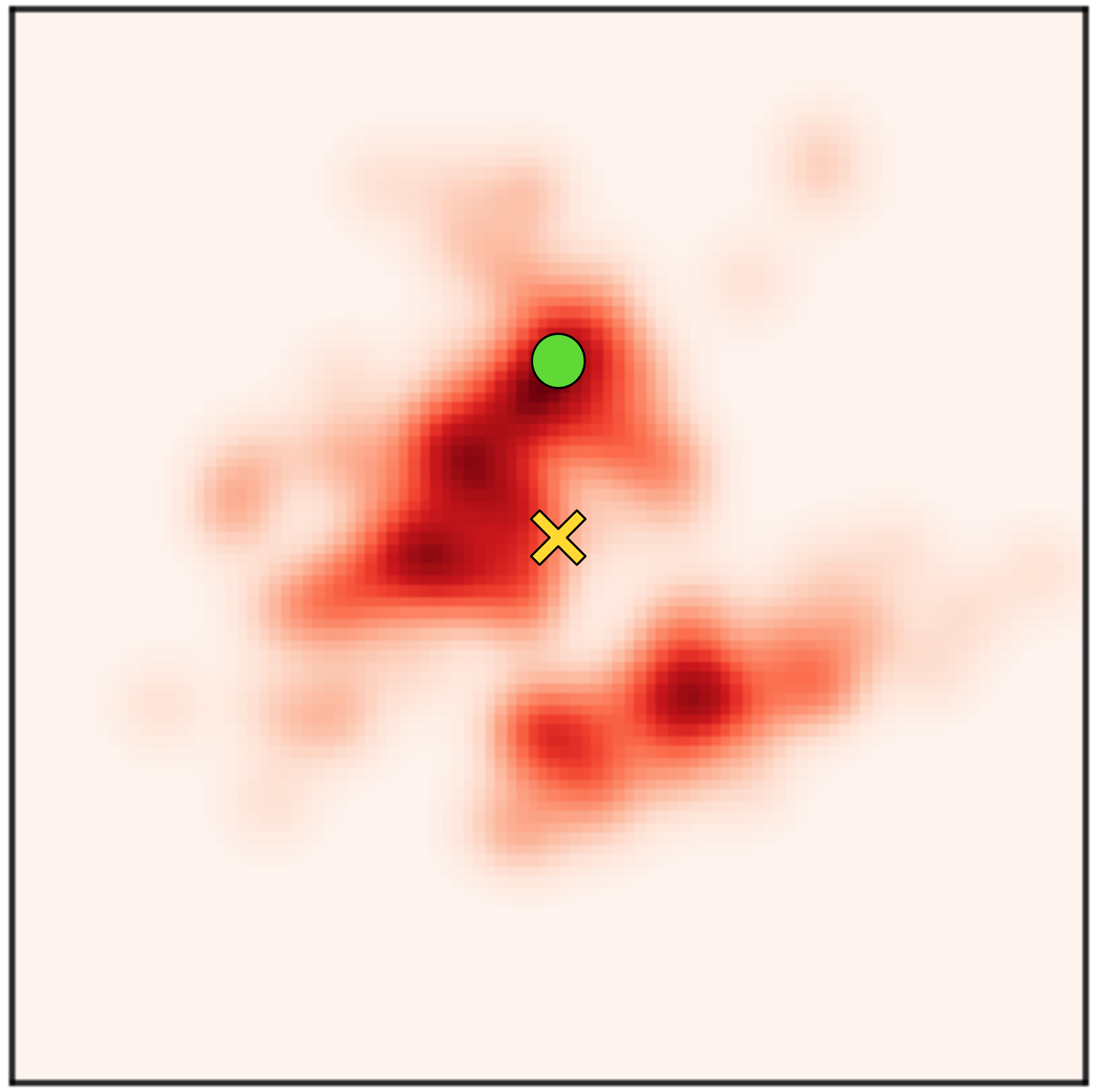
Search-based methods



Sampling & optimization over scoring function

➜ **no finite-time guarantees**

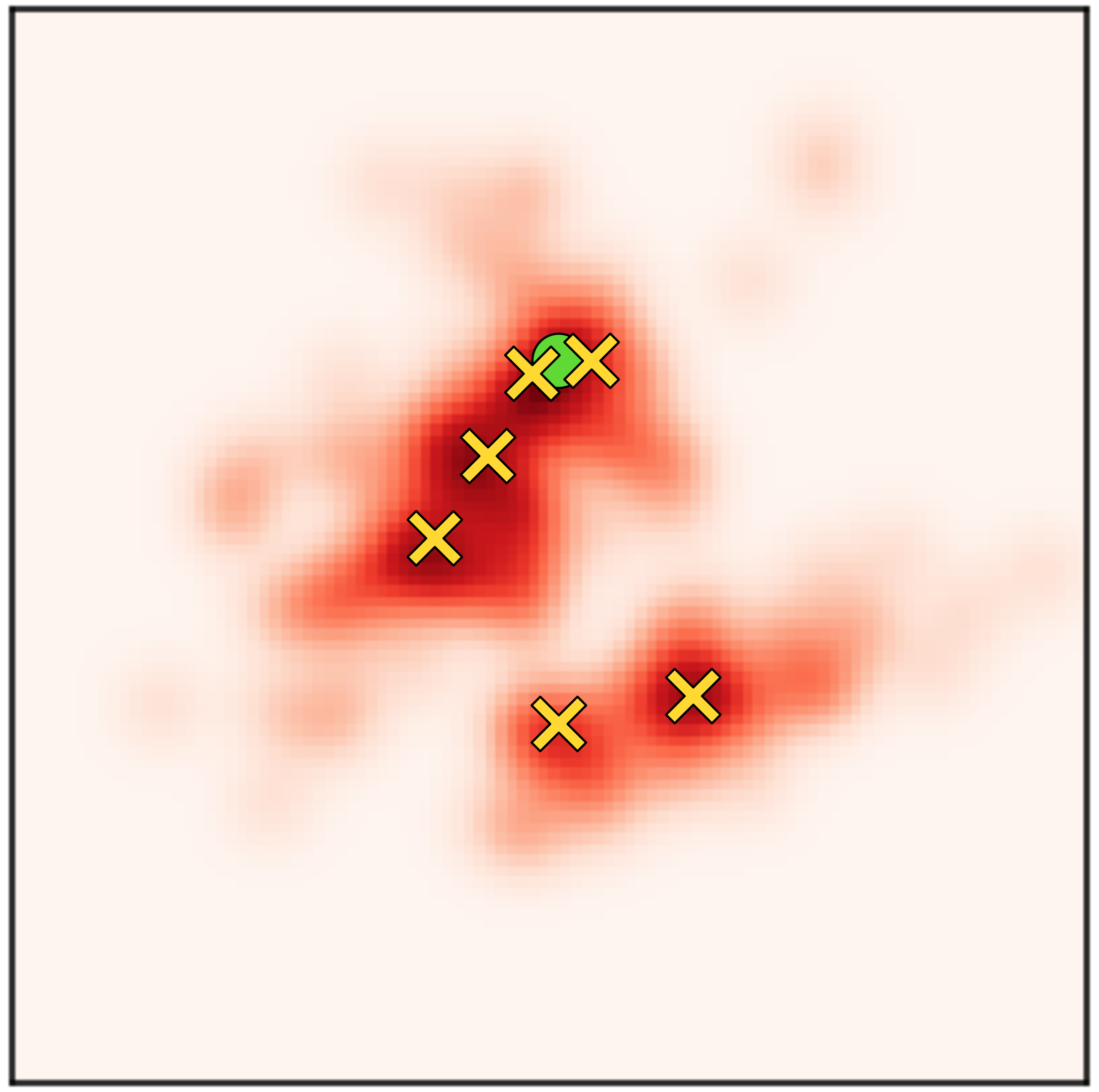
Regression models



Previous deep learning methods were based on regression objective

➜ **fast but poor-quality predictions**

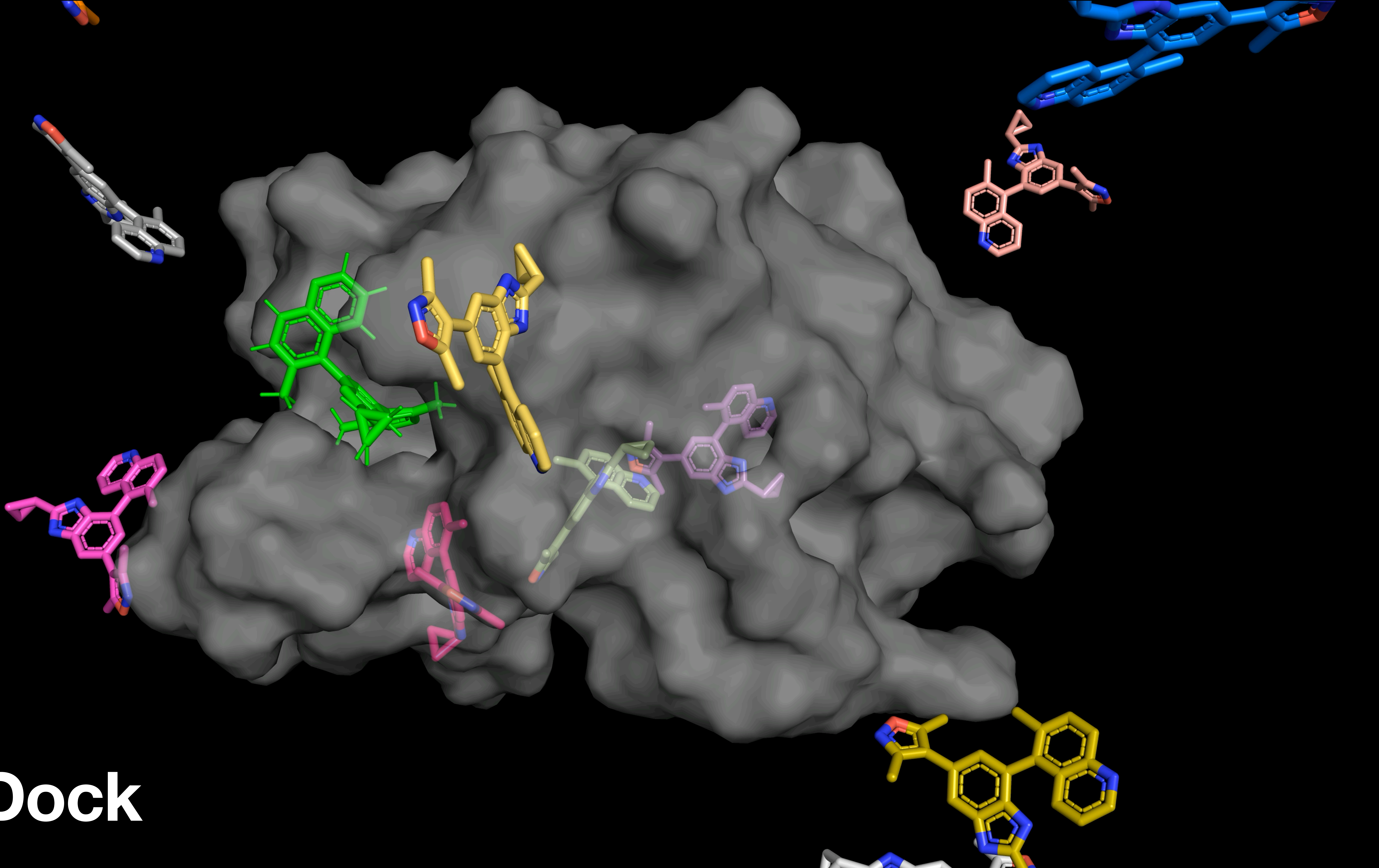
Generative models



Deep generative models with finite time sampling

➜ **correct handling of uncertainty**

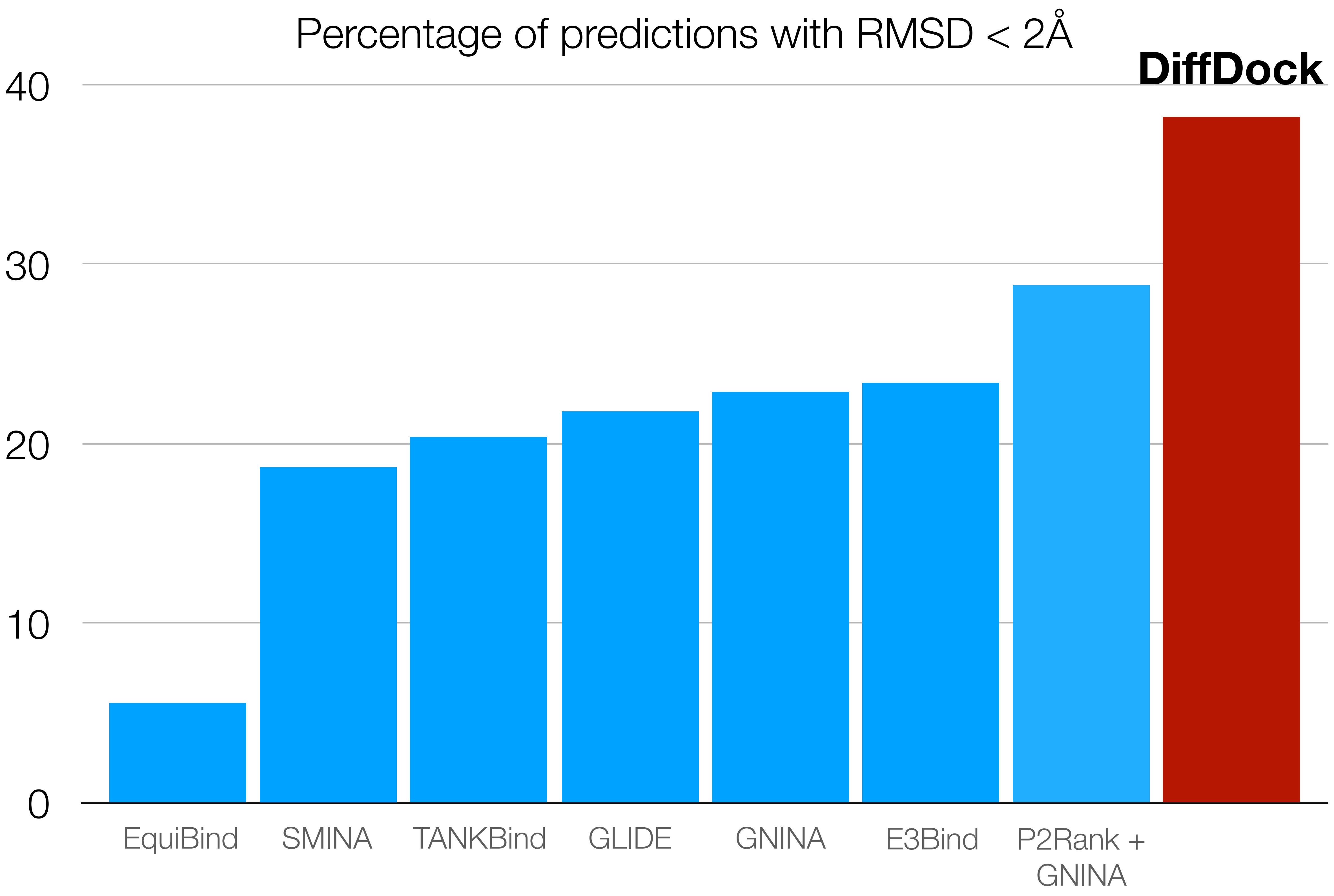
Corso, Stark, Jing, Barzilay, Jaakkola. ICLR 2023



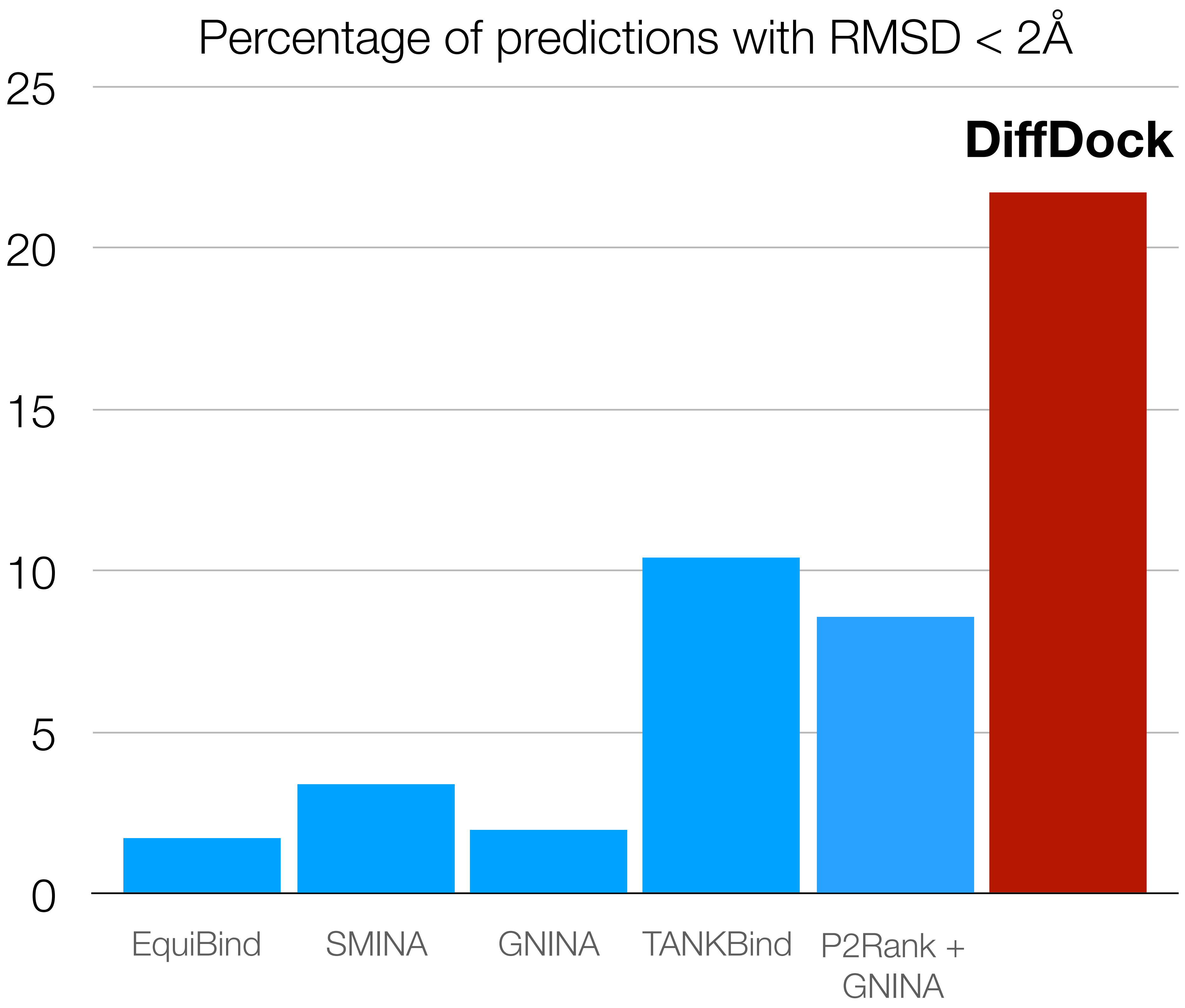
DiffDock

DiffDock

Blind Docking Performance



Holo protein structures



ESMFold structures

Biggest Outstanding Challenges

- **Generalization**: DiffDock struggles when given completely unseen protein classes

- **Receptor flexibility** needs to to be accounted for in order to obtain highly-accurate blind predictions

- **Pose relaxation** is currently required to do some downstream predictions

- No direct **binding affinity** measure

Biggest Outstanding Challenges

- **Generalization**: DiffDock struggles when given completely unseen protein classes

Corso, Deng, Fry, Polizzi, Barzilay, Jaakkola. ICLR 2024

- **Receptor flexibility** needs to to be accounted for in order to obtain highly-accurate blind predictions
- **Pose relaxation** is currently required to do some downstream predictions

Corso, Somnath, Getz, Barzilay, Jaakkola, Krause. Under review.

- No direct **binding affinity** measure

Coming soon!

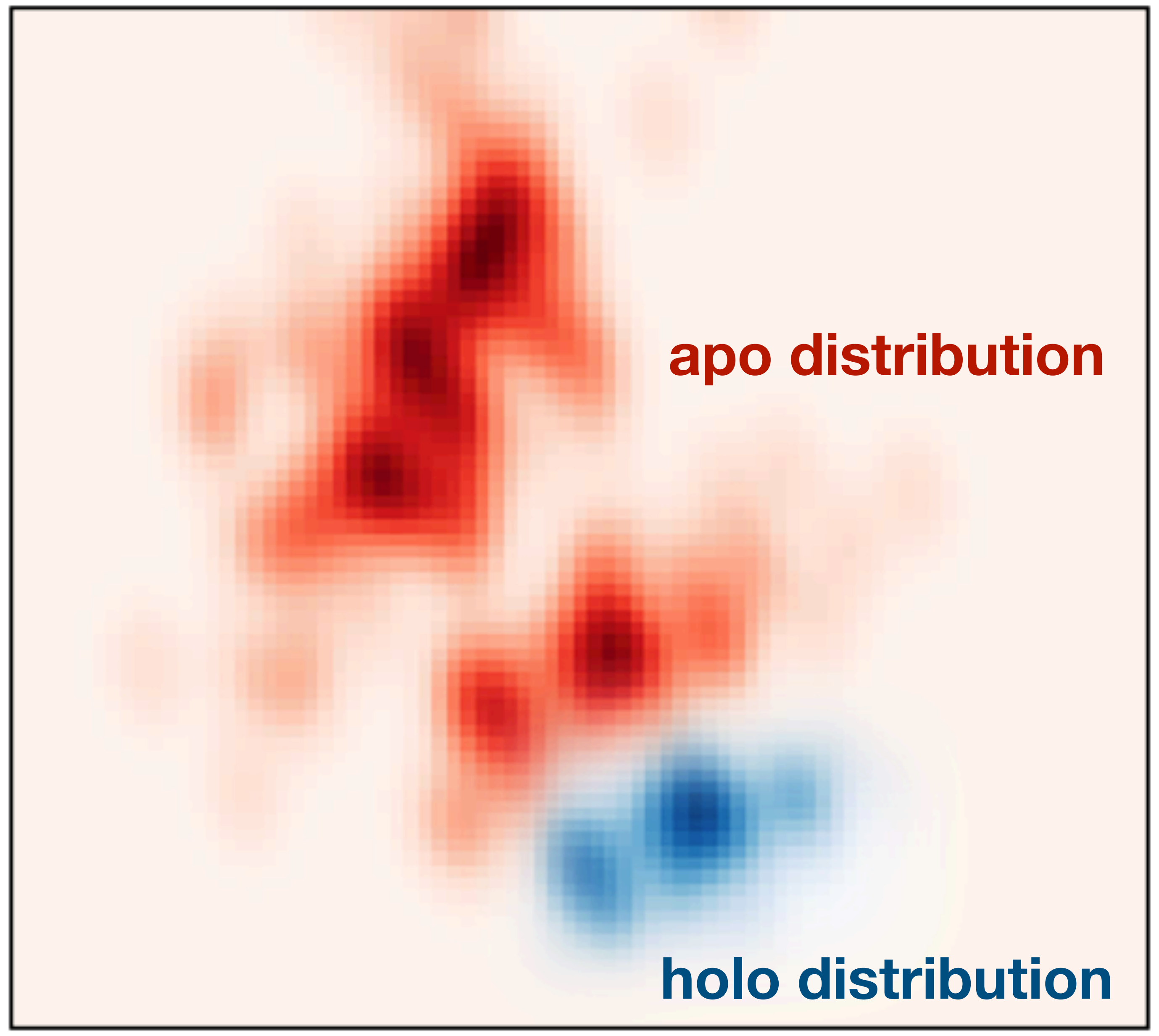
Generative Modeling for Flexible Docking

Flexible docking involves also predicting the conformational change of the protein from the apo (unbound) to holo (bound) state

Generative Modeling for Flexible Docking

Flexible docking involves also predicting the conformational change of the protein from the apo (unbound) to holo (bound) state

We can frame flexible docking as the process of mapping the distribution of apo protein structures to that of holo structures bound to a given ligand.



Flow Matching

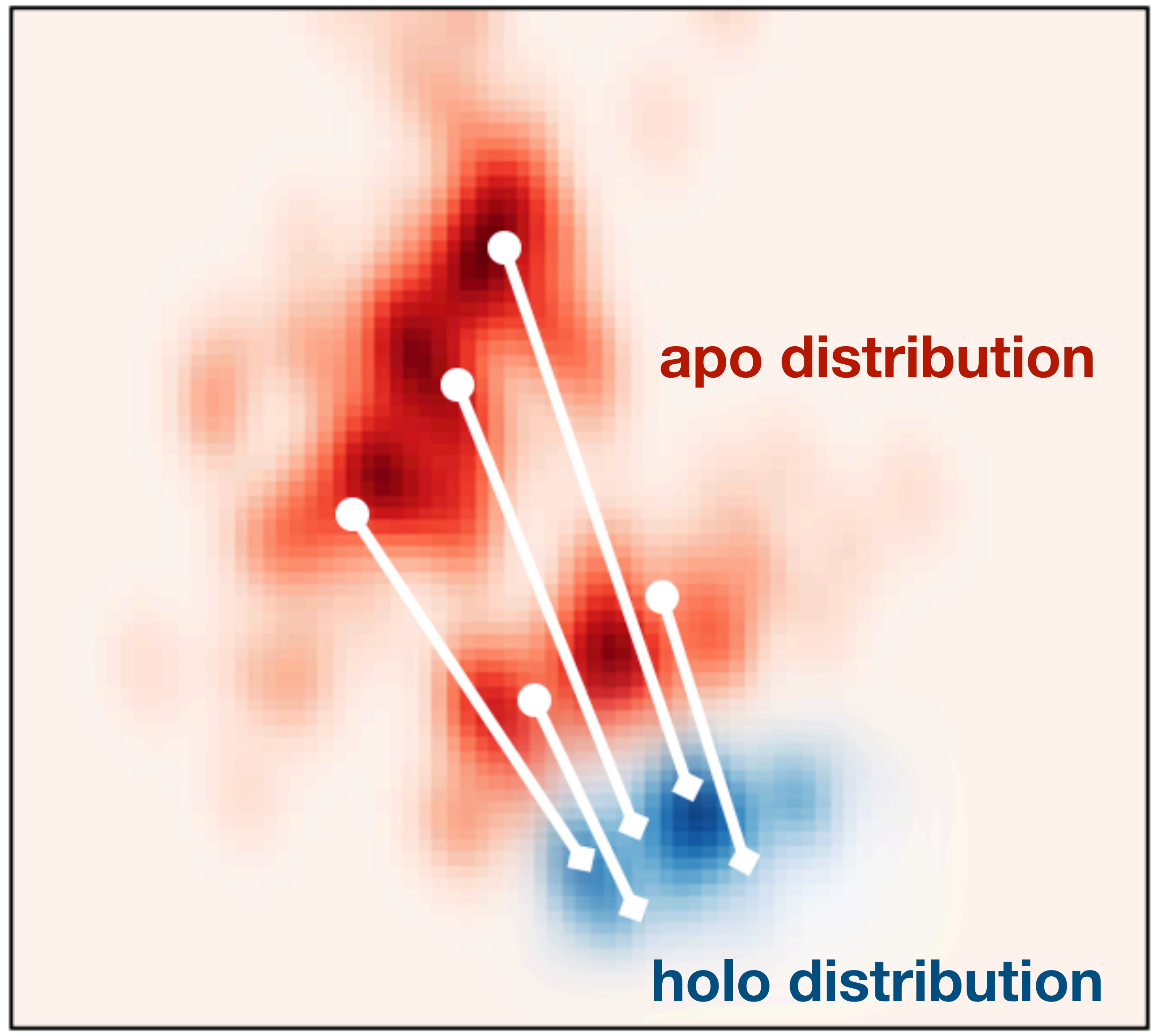
FM Sampling process

1. Sample from $x_0 \sim q_0$
2. Flow x_0 to x_1

FM Objective

$$\min_{\theta} \; \mathbb{E}_{t,(\mathbf{x}_0,\mathbf{x}_1)\sim q}\left[\|v_t(\mathbf{x}_t;\theta) - u_t(\mathbf{x}_t \,|\, \mathbf{x}_1)\|^2\right]$$

where q has marginals q_0 and q_1.



Flow Matching

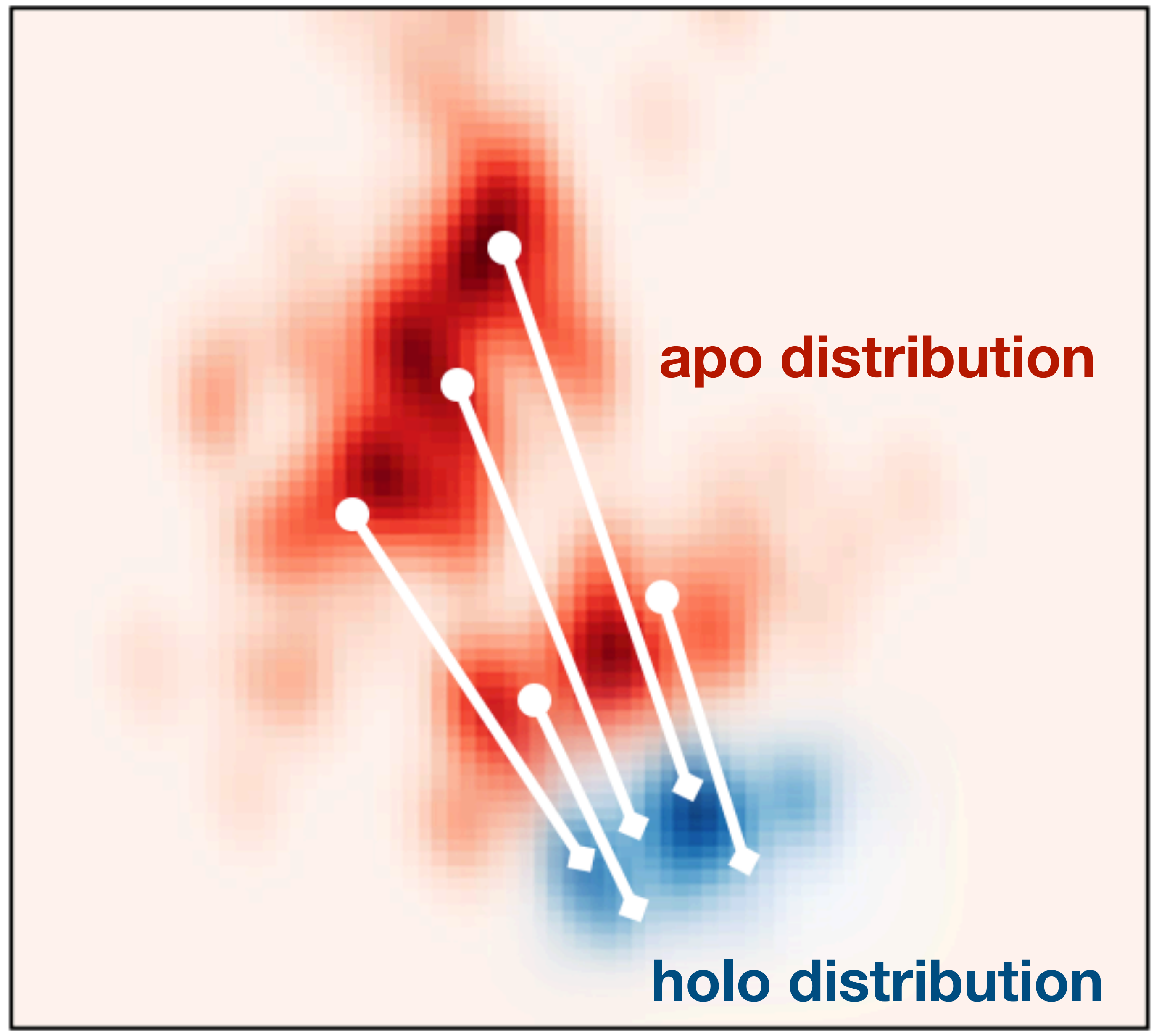
FM Sampling process

1. Sample from $x_0 \sim q_0$
2. Flow x_0 to x_1

FM Objective

$$\min_{\theta} \mathbb{E}_{t,(\mathbf{x}_0,\mathbf{x}_1)\sim q} \left[\| v_t(\mathbf{x}_t;\theta) - u_t(\mathbf{x}_t \,|\, \mathbf{x}_1) \|^2 \right]$$

where q has marginals q_0 and q_1.



Problem: flow matching imposes very complex transport problem resulting in high (Wasserstein) approximation errors.

Unbalanced Flow Matching

Idea: relaxing marginal preservation condition of flow matching we can define much easier transport problems

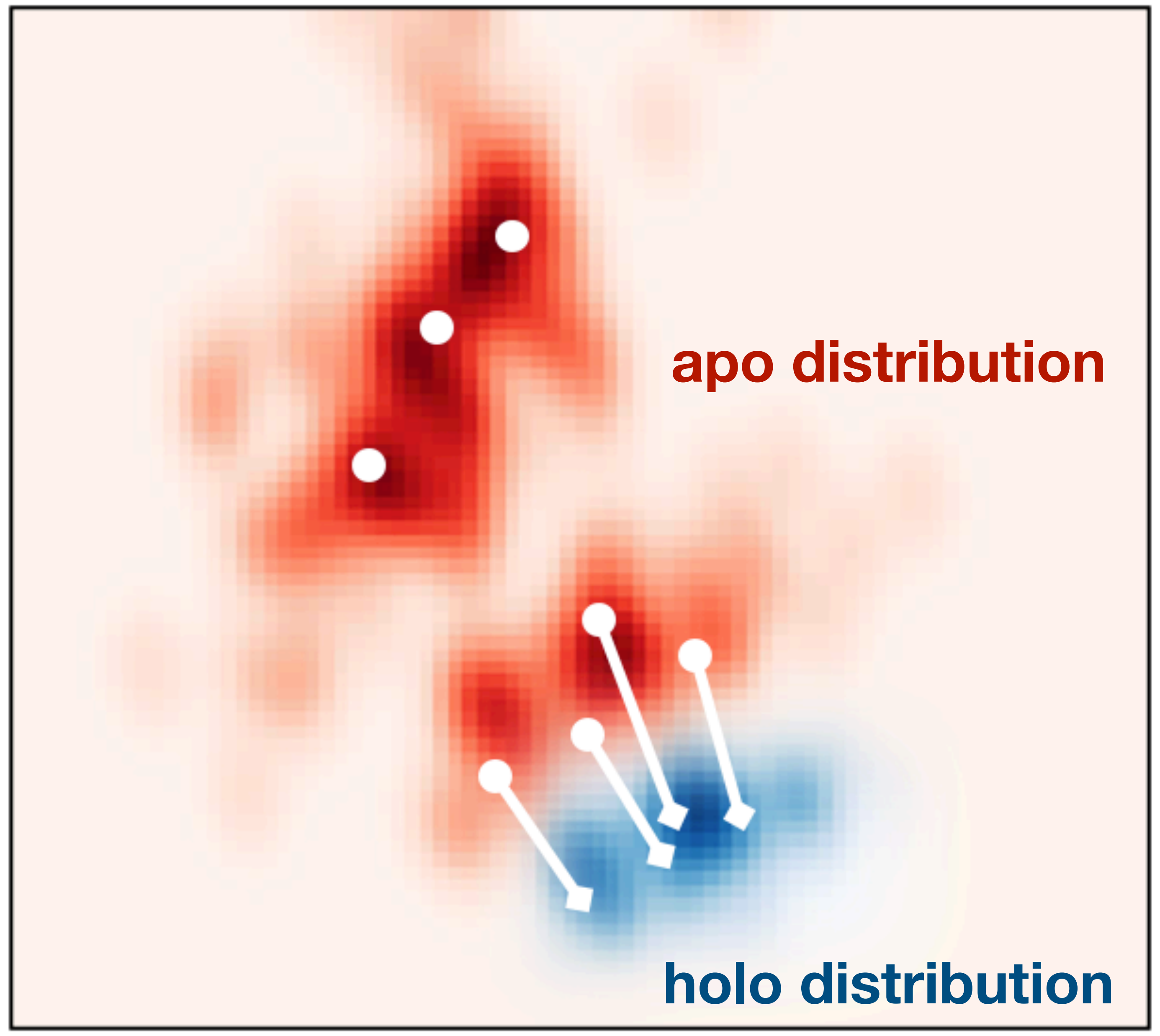
Corso, Somnath, Getz, Barzilay, Jaakkola, Krause. Under review.

Unbalanced Flow Matching

Idea: relaxing marginal preservation condition of flow matching we can define much easier transport problems

Unbalanced FM Sampling process

1. Sample from $x_0 \sim q_0$
2. Flow x_0 to x_1
3. Accept x_1 or return to 1



Corso, Somnath, Getz, Barzilay, Jaakkola, Krause. Under review.

Unbalanced Flow Matching

Idea: relaxing marginal preservation condition of flow matching we can define much easier transport problems

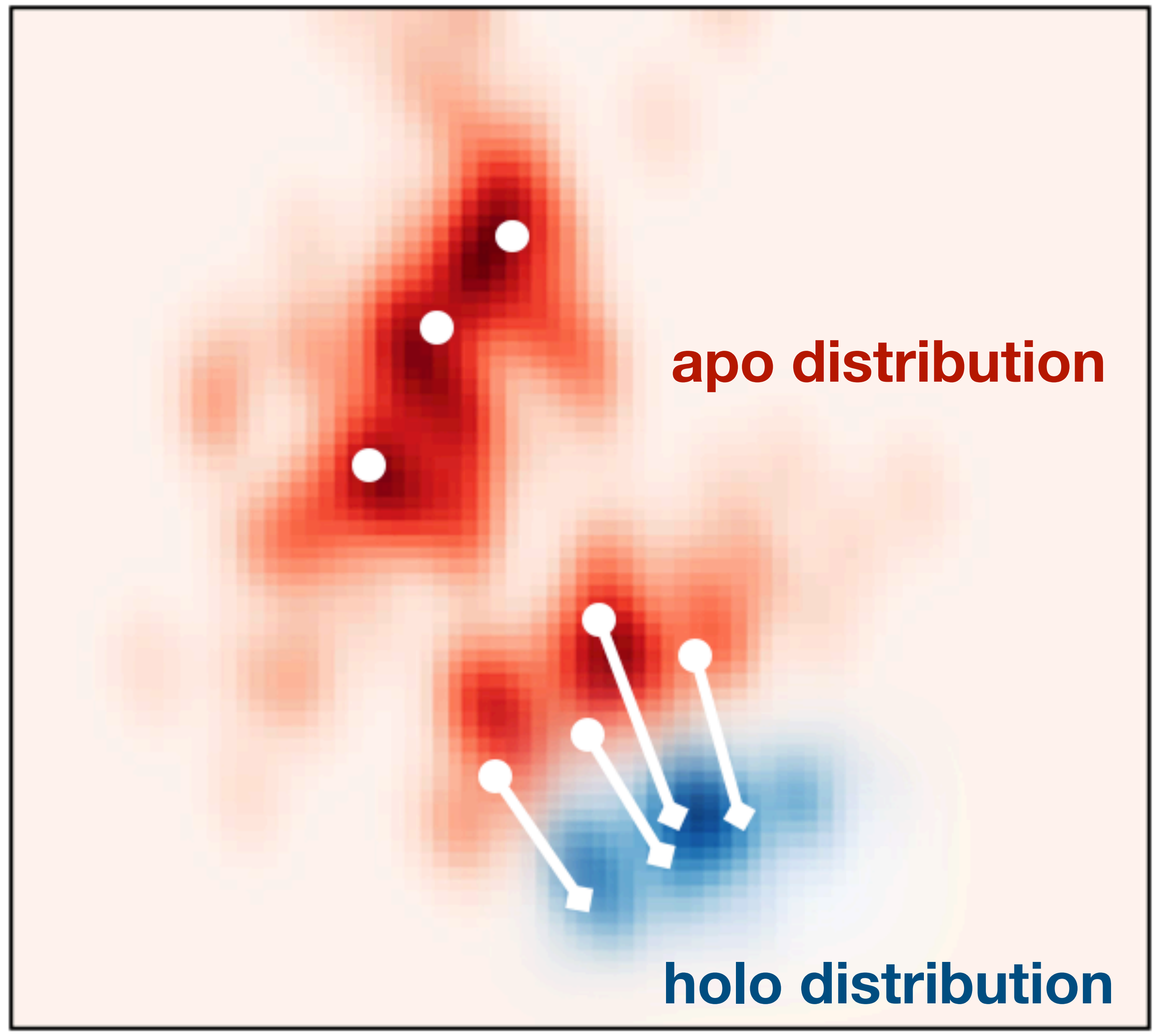
Unbalanced FM Sampling process

1. Sample from $x_0 \sim q_0$
2. Flow x_0 to x_1
3. Accept x_1 or return to 1

Unbalanced FM Objective

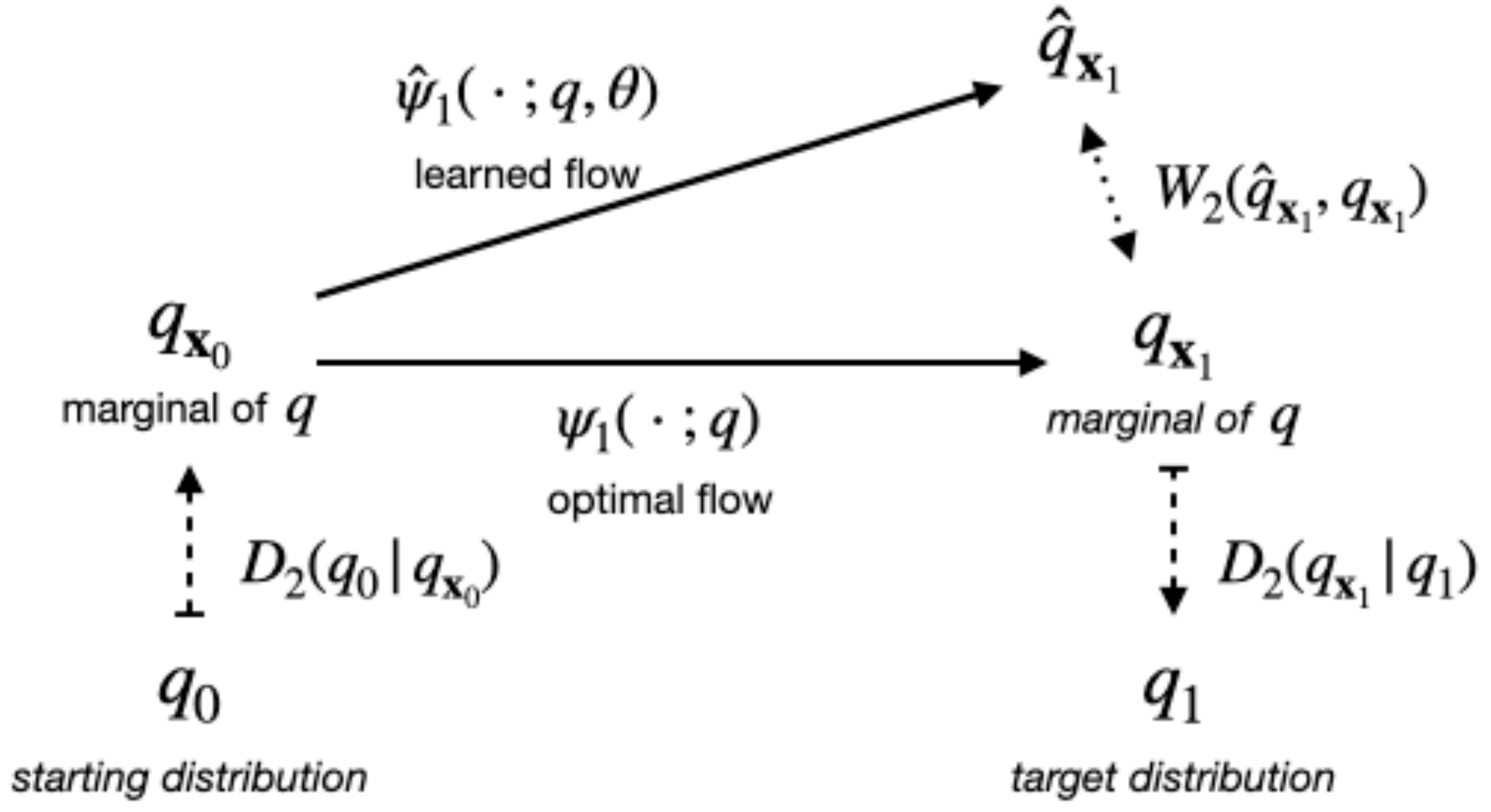
$$\min_{q,\theta} \; \alpha \, \mathbb{E}_{t,(\mathbf{x}_0,\mathbf{x}_1)\sim q}\left[\|v_t(\mathbf{x}_t;\theta) - u_t(\mathbf{x}_t\,|\,\mathbf{x}_1)\|^2\right] + D_2(q_0\,|\,q_{\mathbf{x}_0}) + D_2(q_{\mathbf{x}_1}\,|\,q_1)$$

with arbitrary coupling distribution q with marginals $q_{\mathbf{x}_0}$ and $q_{\mathbf{x}_1}$.



Corso, Somnath, Getz, Barzilay, Jaakkola, Krause. Under review.

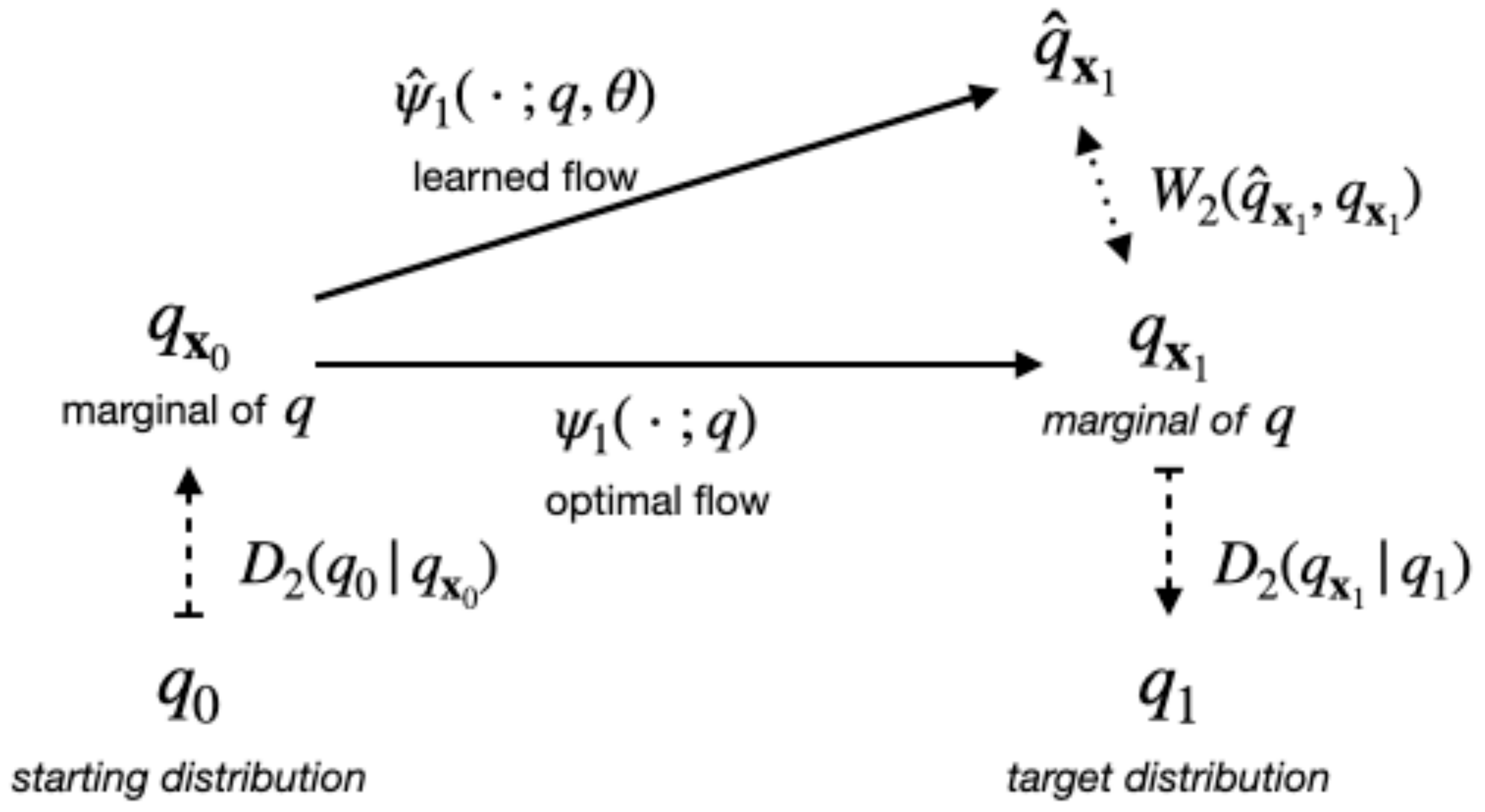
Efficiency vs Approximation



We can show that the UFM objective is a bound on the approximation error vs sampling efficiency tradeoff.

$$\mathscr{L}_{UFM}(q,\theta) = \alpha\, \mathbb{E}_{t,(\mathbf{x}_0,\mathbf{x}_1)\sim q}\left[\|v_t(\mathbf{x}_t;\theta) - u_t(\mathbf{x}_t\,|\,\mathbf{x}_1)\|^2\right] + D_2(q_0\,|\,q_{\mathbf{x}_0}) + D_2(q_{\mathbf{x}_1}\,|\,q_1)$$

Efficiency vs Approximation



We can show that the UFM objective is a bound on the approximation error vs sampling efficiency tradeoff.

$$\mathscr{L}_{UFM}(q, \theta) = \alpha\, \underline{\mathbb{E}_{t,(\mathbf{x}_0,\mathbf{x}_1)\sim q}\left[\|v_t(\mathbf{x}_t; \theta) - u_t(\mathbf{x}_t \,|\, \mathbf{x}_1)\|^2\right]} + D_2(q_0 \,|\, q_{\mathbf{x}_0}) + D_2(q_{\mathbf{x}_1} \,|\, q_1)$$

Proposition (Benton et al., 2023): under appropriate assumptions the approximation error of the learned flow is bounded by FM objective:

$$W_2^2(\hat{q}_{\mathbf{x}_1}(\,\cdot\,|\,\theta), q_{\mathbf{x}_1}) \leq L^2 \cdot \mathbb{E}_{t,q}\left[\|v_t(\mathbf{x}_t; \theta) - u_t(\mathbf{x}_t \,|\, \mathbf{x}_1)\|^2\right]$$

Efficiency vs Approximation

$$\begin{tikzcd}
 & & \hat{q}_{\mathbf{x}_1} \arrow[d, dashed, leftrightarrow, "W_2(\hat{q}_{\mathbf{x}_1}, q_{\mathbf{x}_1})"] \\
\begin{array}{c} q_{\mathbf{x}_0} \\ \text{marginal of } q \end{array} \arrow[urr, "\hat{\psi}_1(\,\cdot\,; q, \theta)", "\text{learned flow}"'] \arrow[rr, "\psi_1(\,\cdot\,; q)"', "\text{optimal flow}"] & & \begin{array}{c} q_{\mathbf{x}_1} \\ \text{marginal of } q \end{array} \arrow[d, dashed, "D_2(q_{\mathbf{x}_1} \,|\, q_1)"] \\
\begin{array}{c} q_0 \\ \text{starting distribution} \end{array} \arrow[u, dashed, "D_2(q_0 \,|\, q_{\mathbf{x}_0})"'] & & \begin{array}{c} q_1 \\ \text{target distribution} \end{array}
\end{tikzcd}$$

We can show that the UFM objective is a bound on the approximation error vs sampling efficiency tradeoff.

$$\mathscr{L}_{UFM}(q, \theta) = \underbrace{\alpha\, \mathbb{E}_{t,(\mathbf{x}_0,\mathbf{x}_1)\sim q}\left[\|v_t(\mathbf{x}_t; \theta) - u_t(\mathbf{x}_t \,|\, \mathbf{x}_1)\|^2\right]} + \underbrace{D_2(q_0 \,|\, q_{\mathbf{x}_0}) + D_2(q_{\mathbf{x}_1} \,|\, q_1)}$$

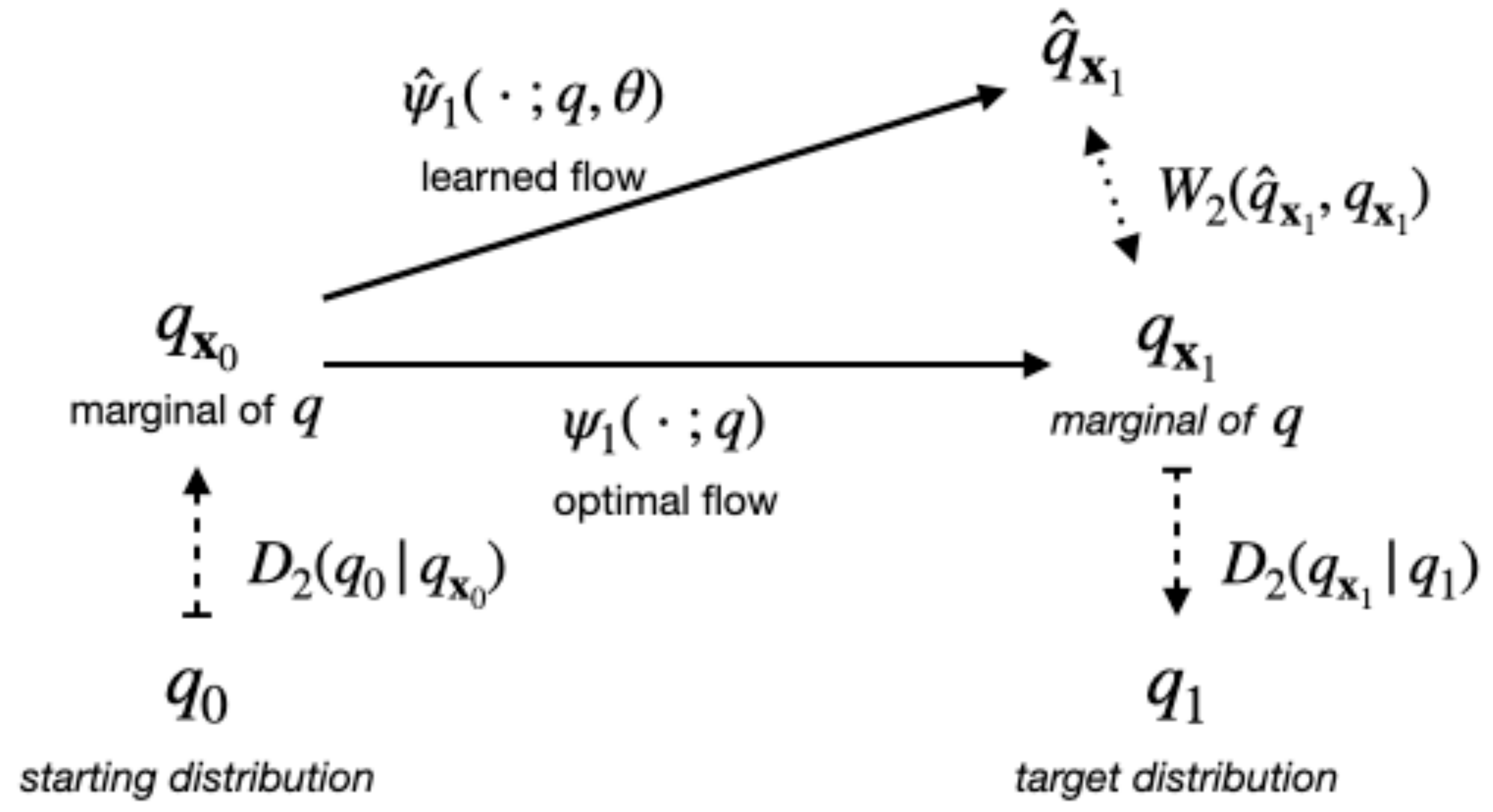
Proposition (Benton et al., 2023): under appropriate assumptions the approximation error of the learned flow is bounded by FM objective:

$$W_2^2(\hat{q}_{\mathbf{x}_1}(\,\cdot\,|\,\theta), q_{\mathbf{x}_1}) \leq L^2 \cdot \mathbb{E}_{t,q}\left[\|v_t(\mathbf{x}_t; \theta) - u_t(\mathbf{x}_t \,|\, \mathbf{x}_1)\|^2\right]$$

Proposition: ESS*, for sampling q_1 when having access to samples of q_0 and a perfectly trained unbalanced flow with coupling q is bounded by:

$$\text{ESS}^*(q) \geq \exp\left[-D_2(q_0 \,|\, q_{\mathbf{x}_0}) - D_2(q_{\mathbf{x}_1} \,|\, q_1)\right]$$

Efficiency vs Approximation



We can show that the UFM objective is a bound on the approximation error vs sampling efficiency tradeoff.

$$\mathscr{L}_{UFM}(q, \theta) = \alpha\, \underbrace{\mathbb{E}_{t,(\mathbf{x}_0,\mathbf{x}_1)\sim q}\left[\|v_t(\mathbf{x}_t;\theta) - u_t(\mathbf{x}_t|\mathbf{x}_1)\|^2\right]} + \underbrace{D_2(q_0 | q_{\mathbf{x}_0}) + D_2(q_{\mathbf{x}_1} | q_1)}$$

Proposition (Benton et al., 2023): under appropriate assumptions the approximation error of the learned flow is bounded by FM objective:

$$W_2^2(\hat{q}_{\mathbf{x}_1}(\cdot\,|\,\theta), q_{\mathbf{x}_1}) \le L^2 \cdot \mathbb{E}_{t,q}\left[\|v_t(\mathbf{x}_t;\theta) - u_t(\mathbf{x}_t|\mathbf{x}_1)\|^2\right]$$

Proposition: ESS*, for sampling q_1 when having access to samples of q_0 and a perfectly trained unbalanced flow with coupling q is bounded by:

$$\mathrm{ESS}^*(q) \ge \exp\left[-D_2(q_0 | q_{\mathbf{x}_0}) - D_2(q_{\mathbf{x}_1} | q_1)\right]$$

$$\beta\; \underbrace{W_2^2(\hat{q}_{\mathbf{x}_1}(\cdot\,|\,\theta), q_{\mathbf{x}_1})}_{\text{Approximation error}} - \underbrace{\log \mathrm{ESS}^*(q)}_{\text{Sampling efficiency}} \le \mathscr{L}_{\mathrm{UFM}}$$

Unbalanced FM optimization

$$\mathscr{L}_{UFM}(q,\theta) = \alpha\, \mathbb{E}_{t,(\mathbf{x}_0,\mathbf{x}_1)\sim q}\left[\|v_t(\mathbf{x}_t;\theta) - u_t(\mathbf{x}_t\,|\,\mathbf{x}_1)\|^2\right] + D_2(q_0\,|\,q_{\mathbf{x}_0}) + D_2(q_{\mathbf{x}_1}\,|\,q_1)$$

Unbalanced FM optimization

$$\begin{aligned}\mathscr{L}_{UFM}(q,\theta) = {} & \alpha\, \mathbb{E}_{t,(\mathbf{x}_0,\mathbf{x}_1)\sim q}\left[\|v_t(\mathbf{x}_t;\theta) - u_t(\mathbf{x}_t \,|\, \mathbf{x}_1)\|^2\right] + D_2(q_0 \,|\, q_{\mathbf{x}_0}) + D_2(q_{\mathbf{x}_1} \,|\, q_1) \\ \leq {} & \mathbb{E}_{(\mathbf{x}_0,\mathbf{x}_1)\sim q}[C(\mathbf{x}_0,\mathbf{x}_1)] + D_2(q_0 \,|\, q_{\mathbf{x}_0}) + D_2(q_{\mathbf{x}_1} \,|\, q_1) \triangleq \mathrm{UOT}(q_0,q_1)\end{aligned}$$

Unbalanced FM optimization

$$\mathscr{L}_{UFM}(q,\theta) = \ \alpha\, \mathbb{E}_{t,(\mathbf{x}_0,\mathbf{x}_1)\sim q}\left[\|v_t(\mathbf{x}_t;\theta) - u_t(\mathbf{x}_t\,|\,\mathbf{x}_1)\|^2\right] + D_2(q_0\,|\,q_{\mathbf{x}_0}) + D_2(q_{\mathbf{x}_1}\,|\,q_1)$$

$$\leq \ \mathbb{E}_{(\mathbf{x}_0,\mathbf{x}_1)\sim q}[C(\mathbf{x}_0,\mathbf{x}_1)] + D_2(q_0\,|\,q_{\mathbf{x}_0}) + D_2(q_{\mathbf{x}_1}\,|\,q_1) \triangleq \text{UOT}(q_0,q_1)$$

The UFM objective can be bound by the Unbalanced OT objective which suggests set of families to choose q from.

Unbalanced FM optimization

$$\mathscr{L}_{UFM}(q,\theta) = \alpha\, \mathbb{E}_{t,(\mathbf{x}_0,\mathbf{x}_1)\sim q}\left[\|v_t(\mathbf{x}_t;\theta) - u_t(\mathbf{x}_t\,|\,\mathbf{x}_1)\|^2\right] + D_2(q_0\,|\,q_{\mathbf{x}_0}) + D_2(q_{\mathbf{x}_1}\,|\,q_1)$$

$$\leq \mathbb{E}_{(\mathbf{x}_0,\mathbf{x}_1)\sim q}[C(\mathbf{x}_0,\mathbf{x}_1)] + D_2(q_0\,|\,q_{\mathbf{x}_0}) + D_2(q_{\mathbf{x}_1}\,|\,q_1) \triangleq \text{UOT}(q_0,q_1)$$

The UFM objective can be bound by the Unbalanced OT objective which suggests set of families to choose q from.

Because we only have access to individual samples we choose

$$q(\mathbf{x}_0,\mathbf{x}_1) = q_0(\mathbf{x}_0)\; q_1(\mathbf{x}_1)\; \mathbb{1}_{\|\mathbf{x}_0-\mathbf{x}_1\|<C}$$

Unbalanced FM optimization

$$\mathscr{L}_{UFM}(q,\theta) = \alpha\, \mathbb{E}_{t,(\mathbf{x}_0,\mathbf{x}_1)\sim q}\left[\|v_t(\mathbf{x}_t;\theta) - u_t(\mathbf{x}_t\,|\,\mathbf{x}_1)\|^2\right] + D_2(q_0\,|\,q_{\mathbf{x}_0}) + D_2(q_{\mathbf{x}_1}\,|\,q_1)$$

$$\leq \mathbb{E}_{(\mathbf{x}_0,\mathbf{x}_1)\sim q}[C(\mathbf{x}_0,\mathbf{x}_1)] + D_2(q_0\,|\,q_{\mathbf{x}_0}) + D_2(q_{\mathbf{x}_1}\,|\,q_1) \triangleq \text{UOT}(q_0,q_1)$$

The UFM objective can be bound by the Unbalanced OT objective which suggests set of families to choose q from.

Because we only have access to individual samples we choose

$$q(\mathbf{x}_0,\mathbf{x}_1) = q_0(\mathbf{x}_0)\; q_1(\mathbf{x}_1)\; \mathbb{1}_{\|\mathbf{x}_0-\mathbf{x}_1\|<C}$$

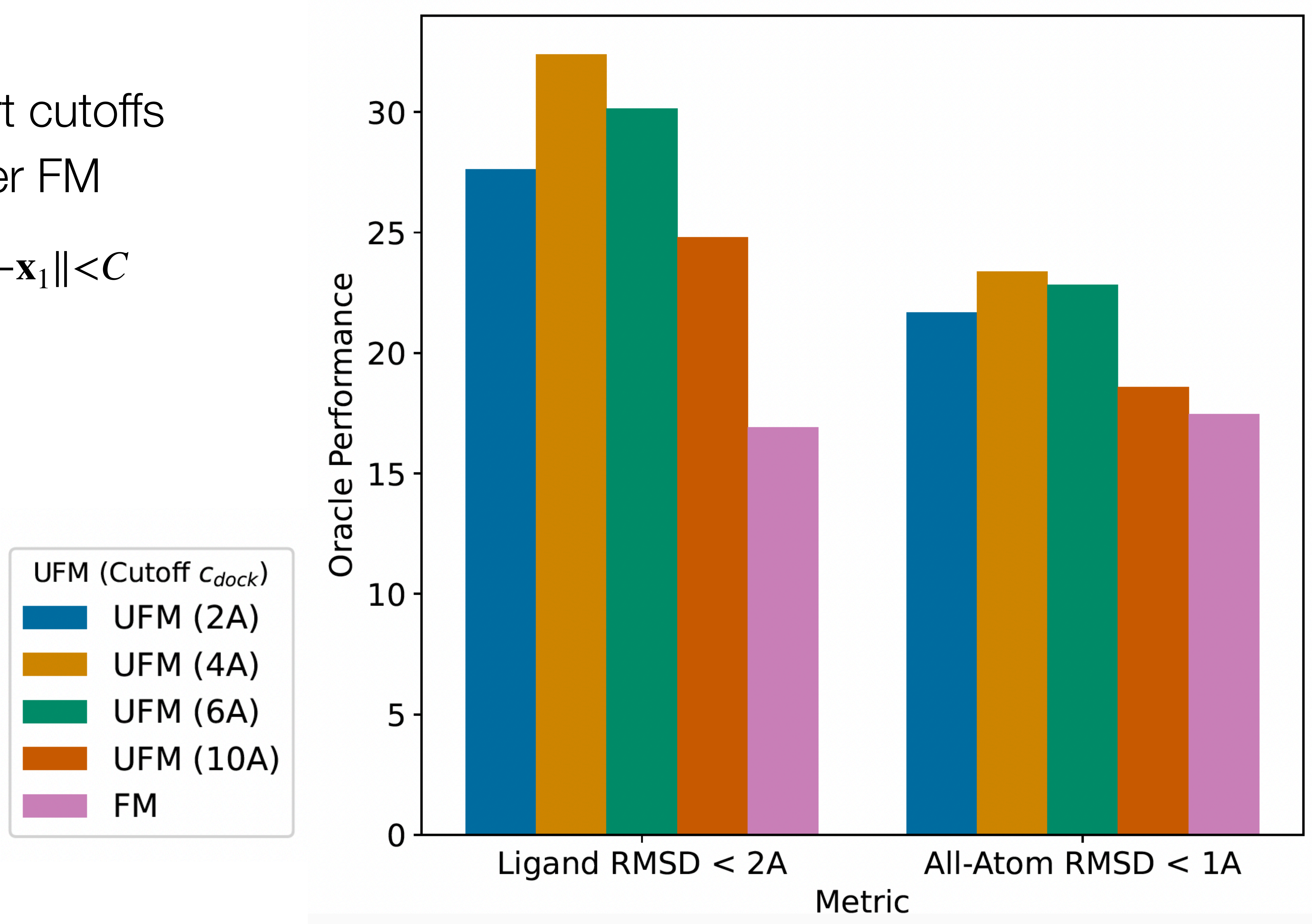
Then, given q, the UFM objective boils down to Flow Matching:

$$\min_{\theta}\; \mathbb{E}_{t,(\mathbf{x}_0,\mathbf{x}_1)\sim q}\left[\|v_t(\mathbf{x}_t;\theta) - u_t(\mathbf{x}_t\,|\,\mathbf{x}_1)\|^2\right]$$

Flexible Docking with Unbalanced FM

Choosing q with different transport cutoffs highlights the value of UFM over FM

$$q(\mathbf{x}_0, \mathbf{x}_1) = q_0(\mathbf{x}_0)\ q_1(\mathbf{x}_1)\ \mathbb{1}_{\|\mathbf{x}_0-\mathbf{x}_1\|<C}$$



Biggest Outstanding Challenges

- **Generalization**: DiffDock struggles when given completely unseen protein classes

Corso, Deng, Fry, Polizzi, Barzilay, Jaakkola. ICLR 2024

- **Receptor flexibility** needs to to be accounted for in order to obtain highly-accurate blind predictions
- **Pose relaxation** is currently required to do some downstream predictions

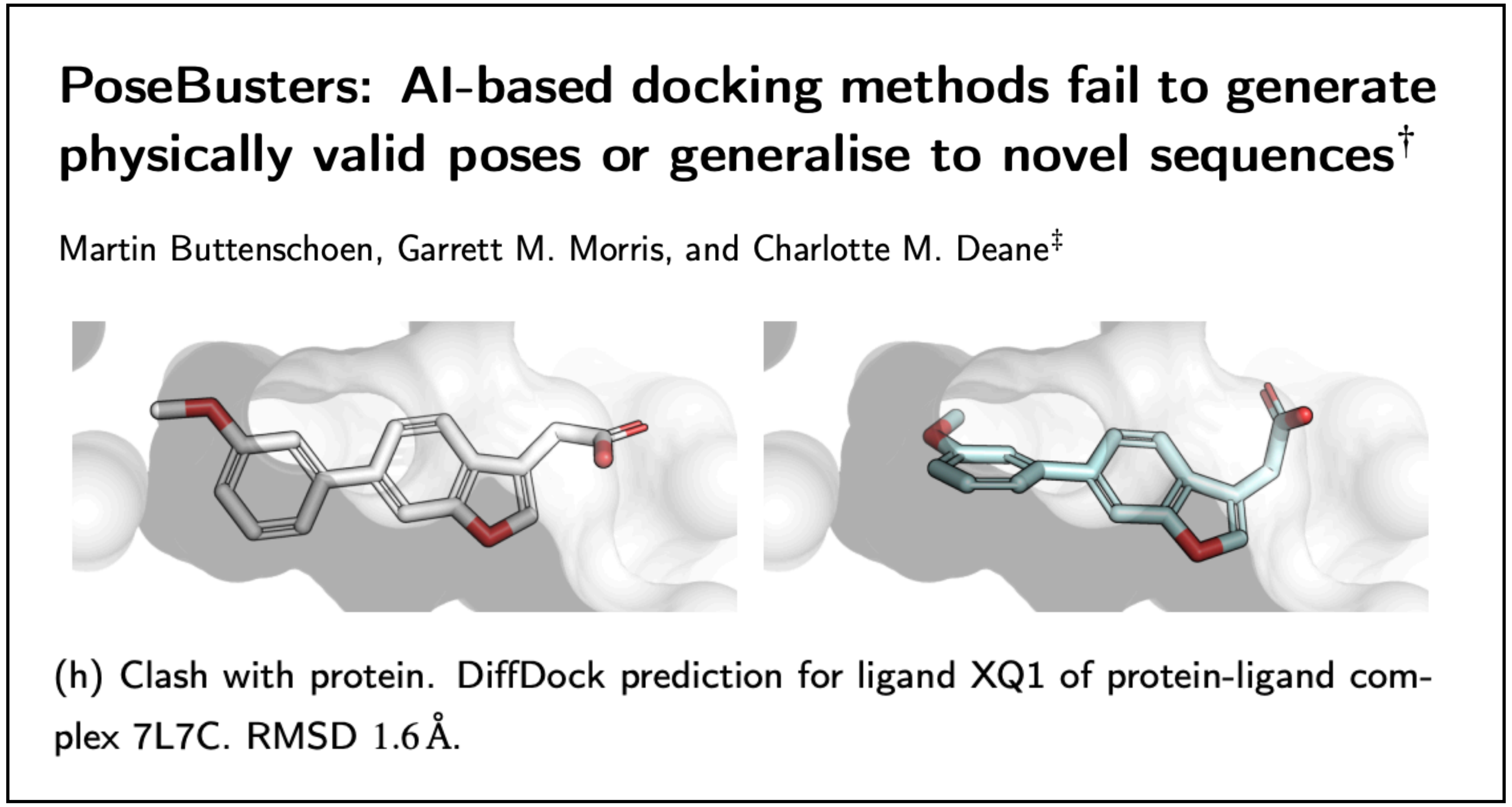
Corso, Somnath, Getz, Barzilay, Jaakkola, Krause. Under review.

- No direct **binding affinity** measure

Coming soon!

Pose relaxation

Although docking is typically framed as trying to obtain poses as close as possible to crystal structure, the “physicality” of the poses is also important.



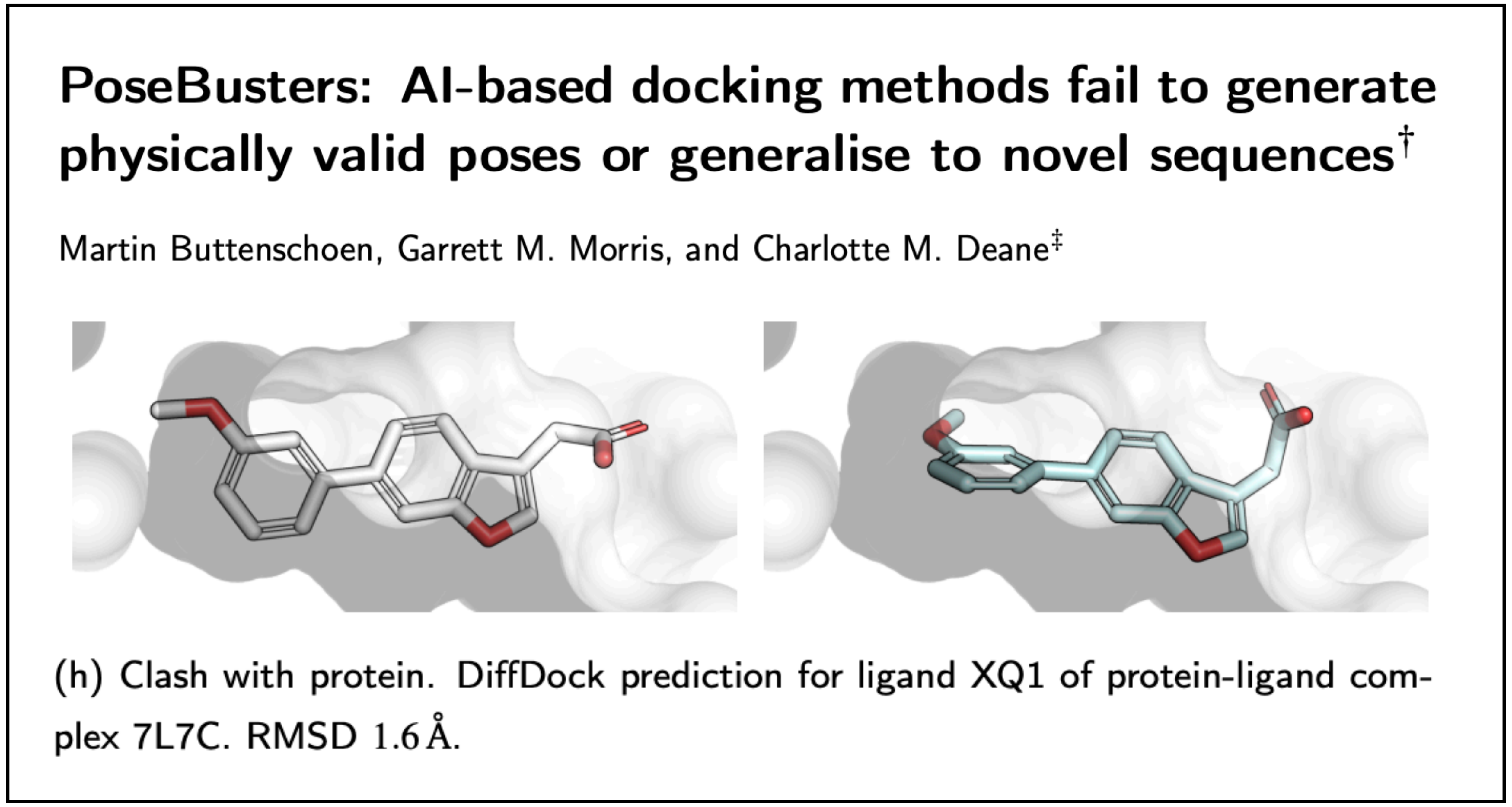
PoseBusters: AI-based docking methods fail to generate physically valid poses or generalise to novel sequences[†]

Martin Buttenschoen, Garrett M. Morris, and Charlotte M. Deane[‡]

(h) Clash with protein. DiffDock prediction for ligand XQ1 of protein-ligand complex 7L7C. RMSD 1.6 Å.

Pose relaxation

Although docking is typically framed as trying to obtain poses as close as possible to crystal structure, the “physicality” of the poses is also important.

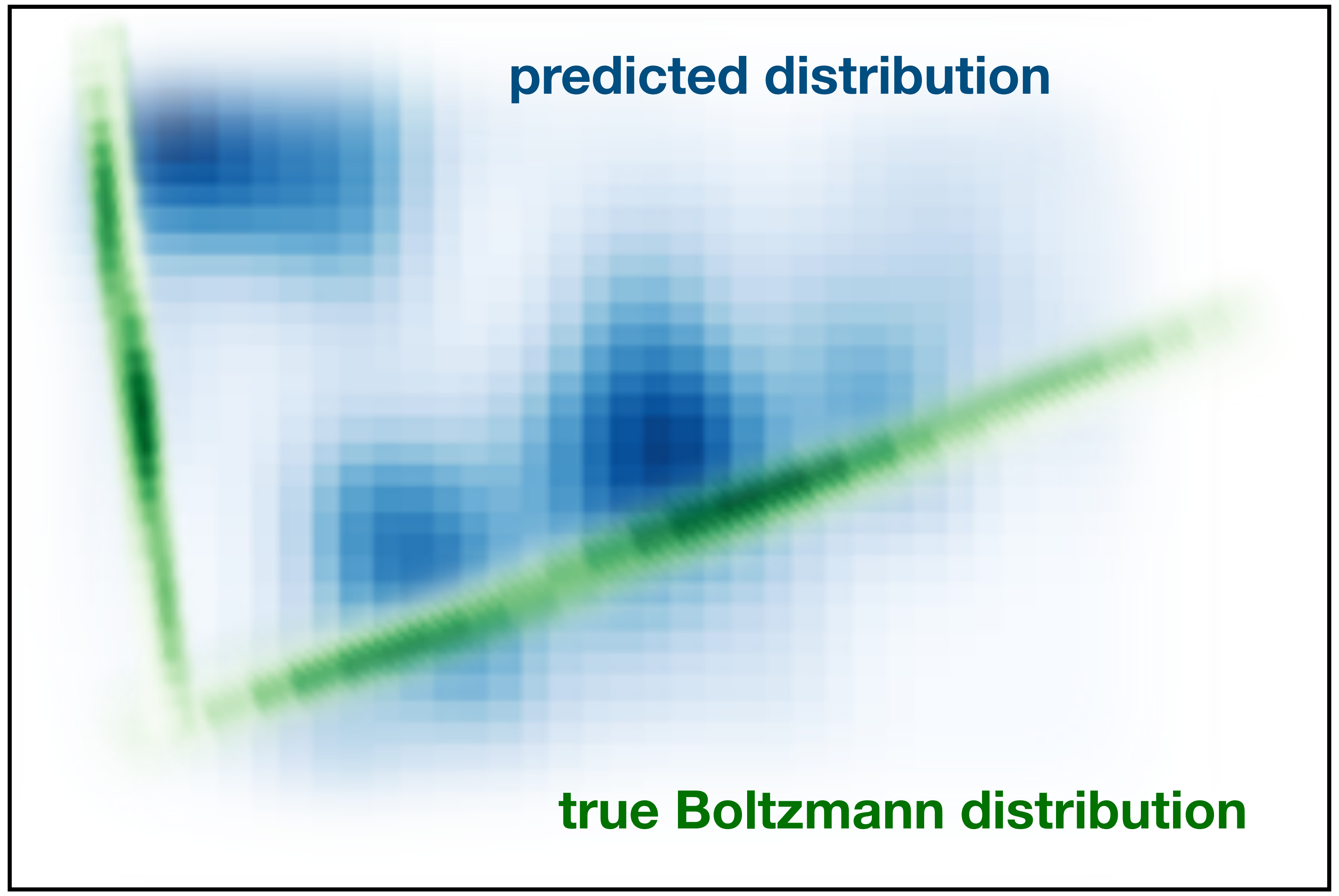


PoseBusters: AI-based docking methods fail to generate physically valid poses or generalise to novel sequences[†]

Martin Buttenschoen, Garrett M. Morris, and Charlotte M. Deane[‡]

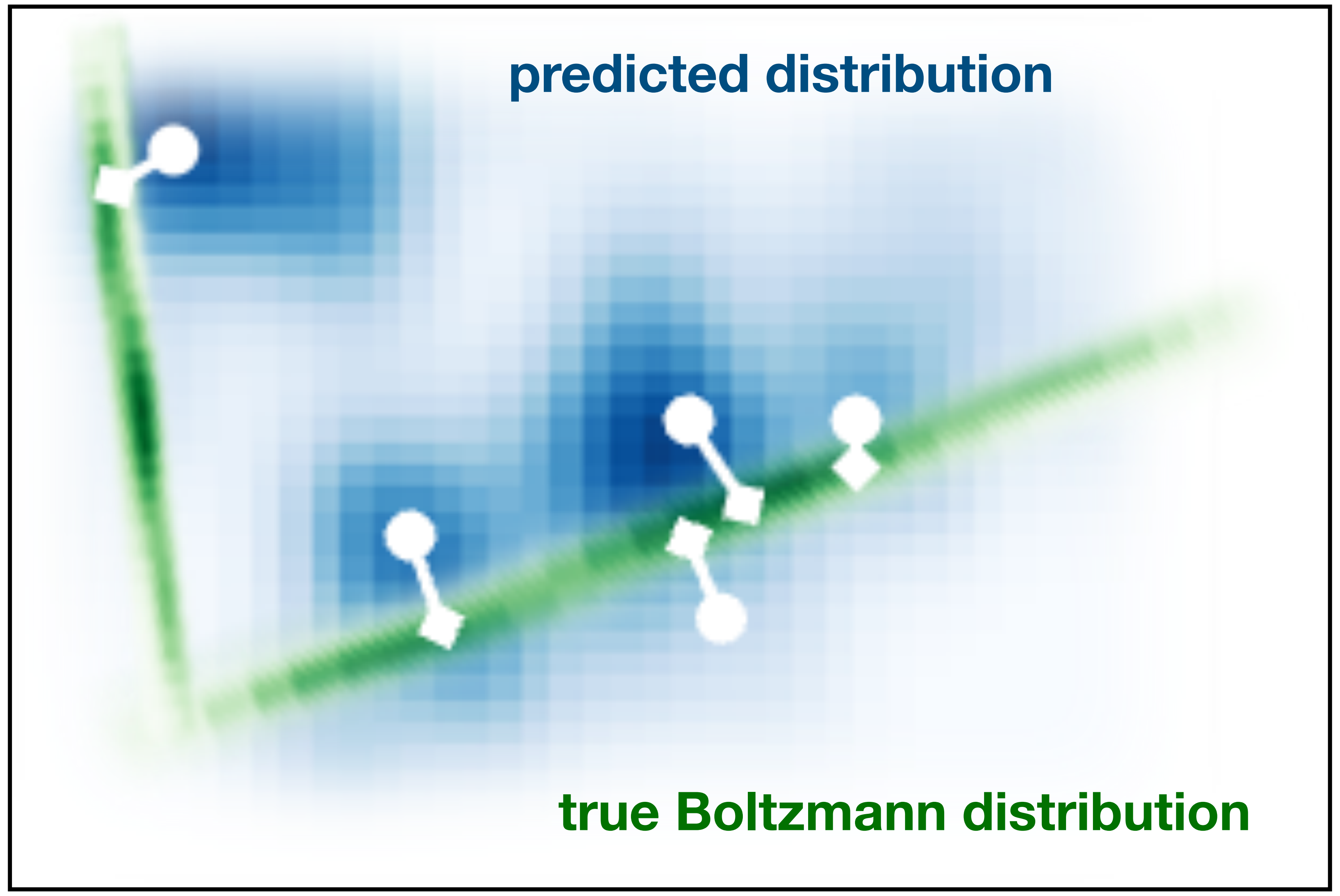
(h) Clash with protein. DiffDock prediction for ligand XQ1 of protein-ligand complex 7L7C. RMSD 1.6 Å.

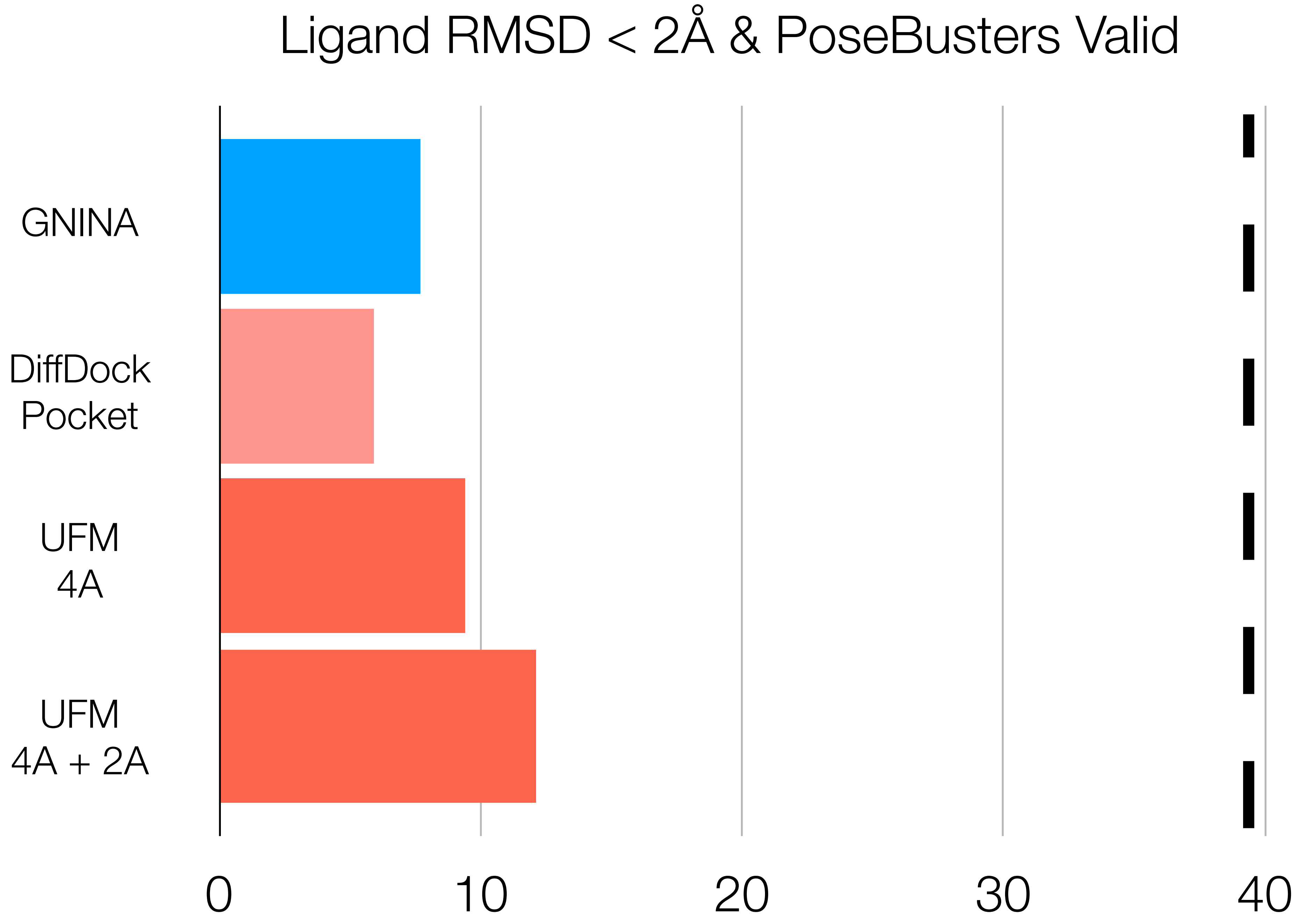
Pose relaxation: refine the structural conformation to find a more energetically favorable



Pose relaxation with Unbalanced FM

Applying “vanilla” Unbalanced FM improves the performance but it is still far from optimal due to vast scale disparity of different degrees of freedom





Energy Loss

To incentivize the model to preserve physicality also in very narrow degrees of freedom we would want an objective like reverse KL.

Energy Loss

To incentivize the model to preserve physicality also in very narrow degrees of freedom we would want an objective like reverse KL.

However, reverse KL (e.g. training Boltzmann Generators with reverse KL) has a few challenges:

1. Requires invertible transformation
2. Requires back propagating through full flow
3. Loss (energy) is very unstable

Energy Loss

To incentivize the model to preserve physicality also in very narrow degrees of freedom we would want an objective like reverse KL.

However, reverse KL (e.g. training Boltzmann Generators with reverse KL) has a few challenges:

1. Requires invertible transformation
2. Requires back propagating through full flow
3. Loss (energy) is very unstable

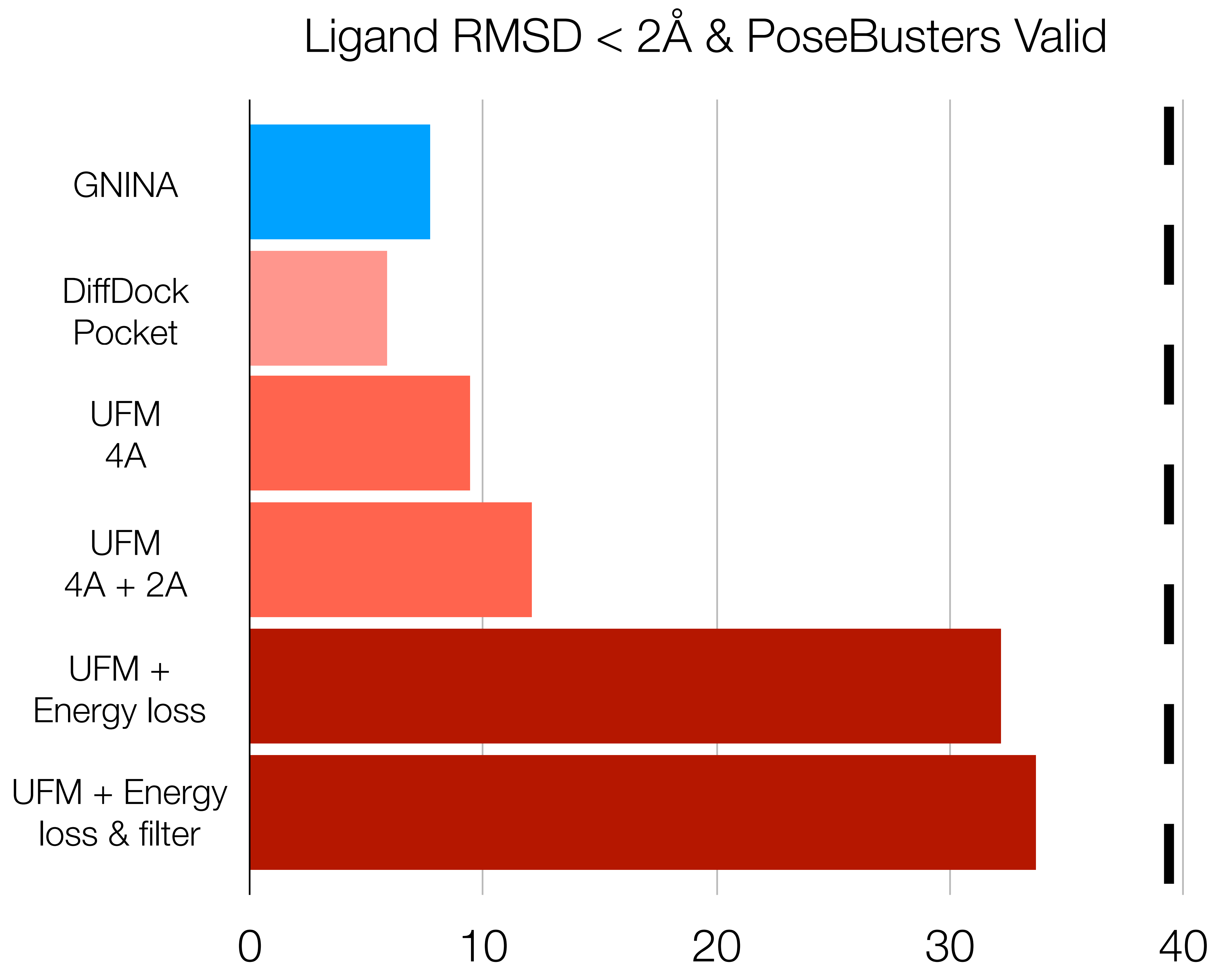
$$\mathscr{L}_{\text{energy}} = \begin{cases} \sum_{i,j} \max\left(\|\hat{\mathbf{x}}_1^{(i)} - \hat{\mathbf{x}}_1^{(j)}\| - U_{i,j}, 0\right) + \max\left(L_{i,j} - \|\hat{\mathbf{x}}_1^{(i)} - \hat{\mathbf{x}}_1^{(j)}\|, 0\right) & \text{if } t > 1 - \epsilon \\ 0 & \text{otherwise} \end{cases}$$

Energy Loss

To incentivize the model to preserve physicality also in very narrow degrees of freedom we would want an objective like reverse KL.

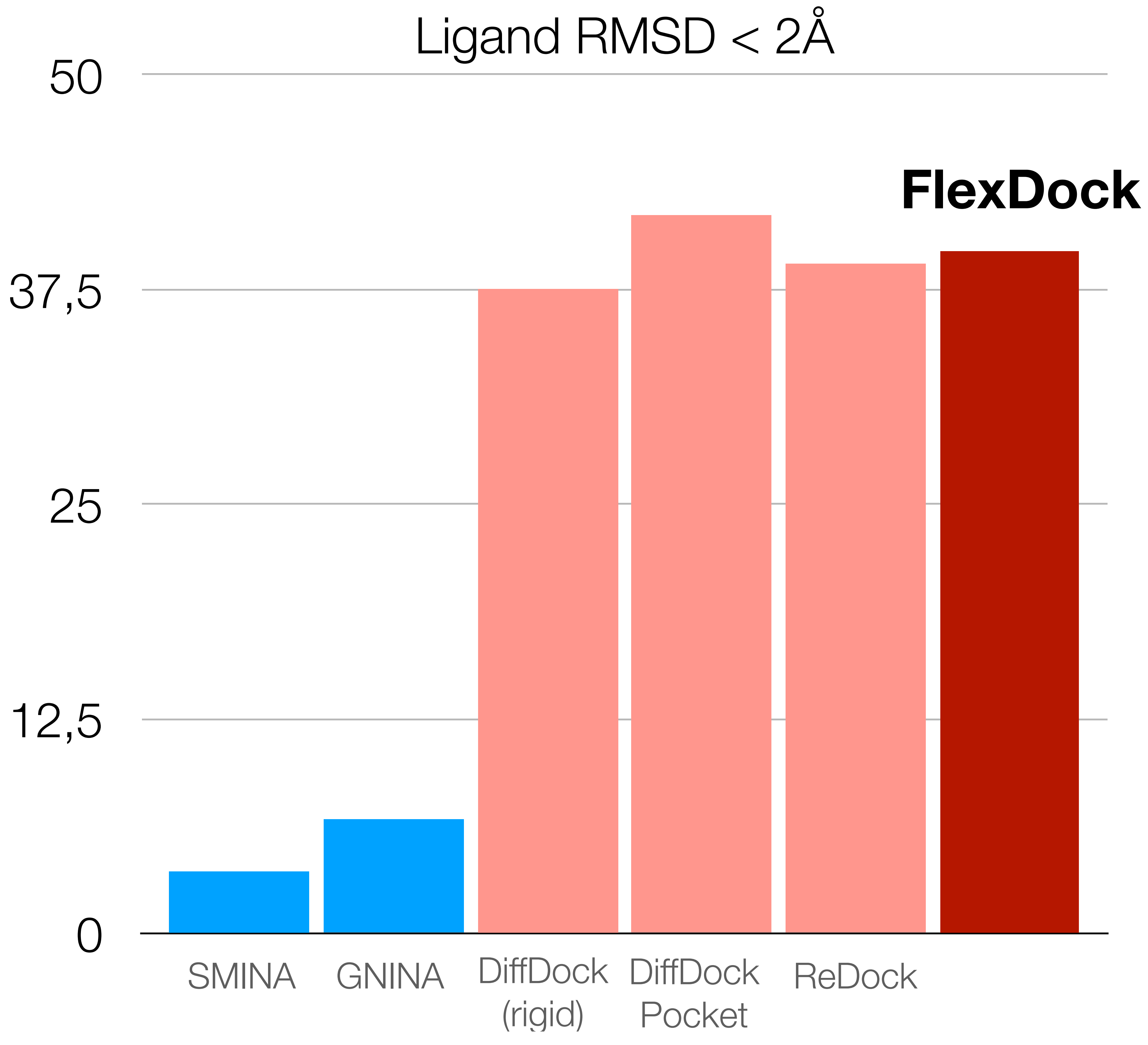
However, reverse KL (e.g. training Boltzmann Generators with reverse KL) has a few challenges:

1. Requires invertible transformation
2. Requires back propagating through full flow
3. Loss (energy) is very unstable

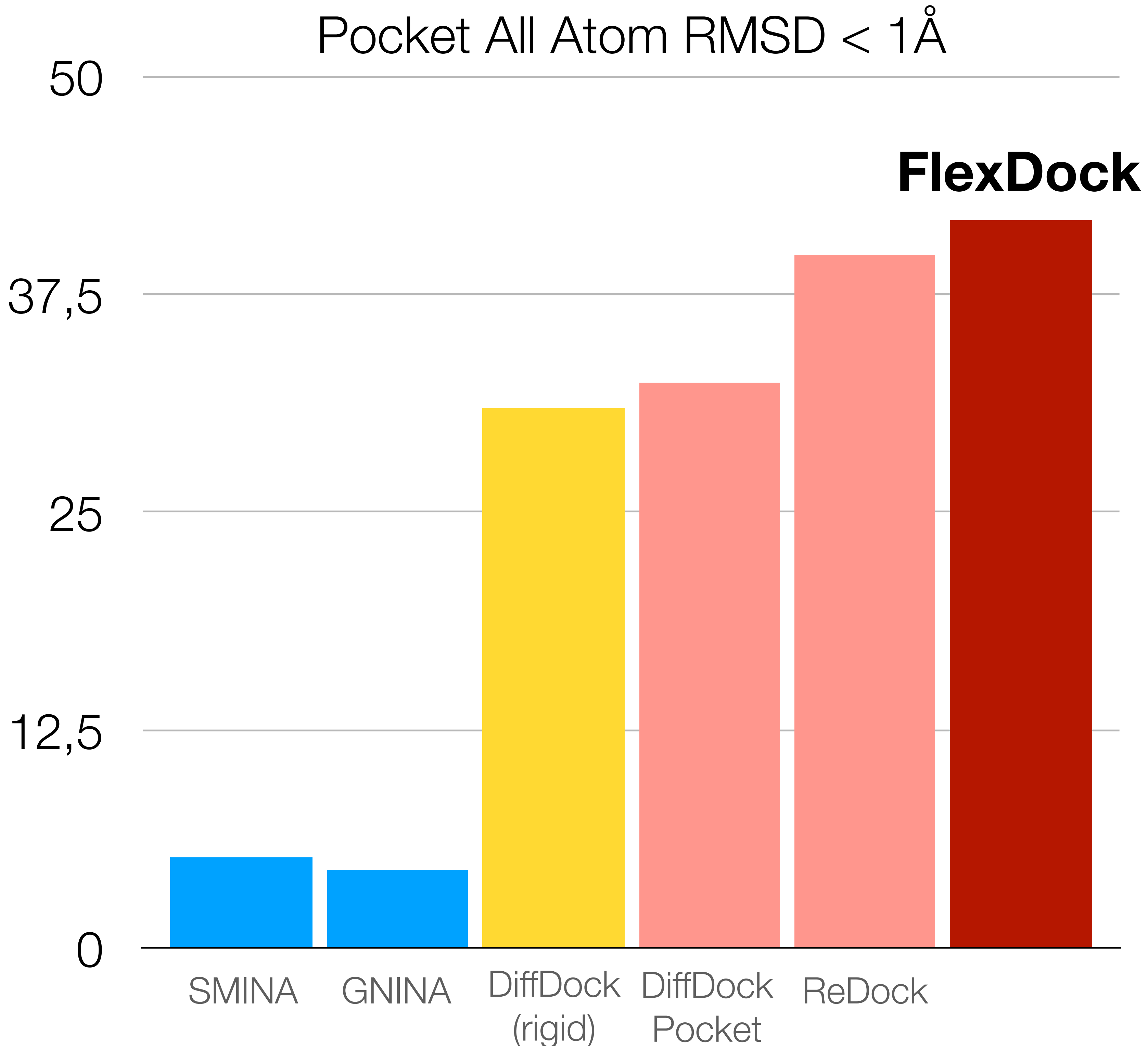


$$\mathscr{L}_{\text{energy}} = \begin{cases} \sum_{i,j} \max\left(\|\hat{\mathbf{x}}_1^{(i)} - \hat{\mathbf{x}}_1^{(j)}\| - U_{i,j}, 0\right) + \max\left(L_{i,j} - \|\hat{\mathbf{x}}_1^{(i)} - \hat{\mathbf{x}}_1^{(j)}\|, 0\right) & \text{if } t > 1 - \epsilon \\ 0 \quad \text{otherwise} & \end{cases}$$

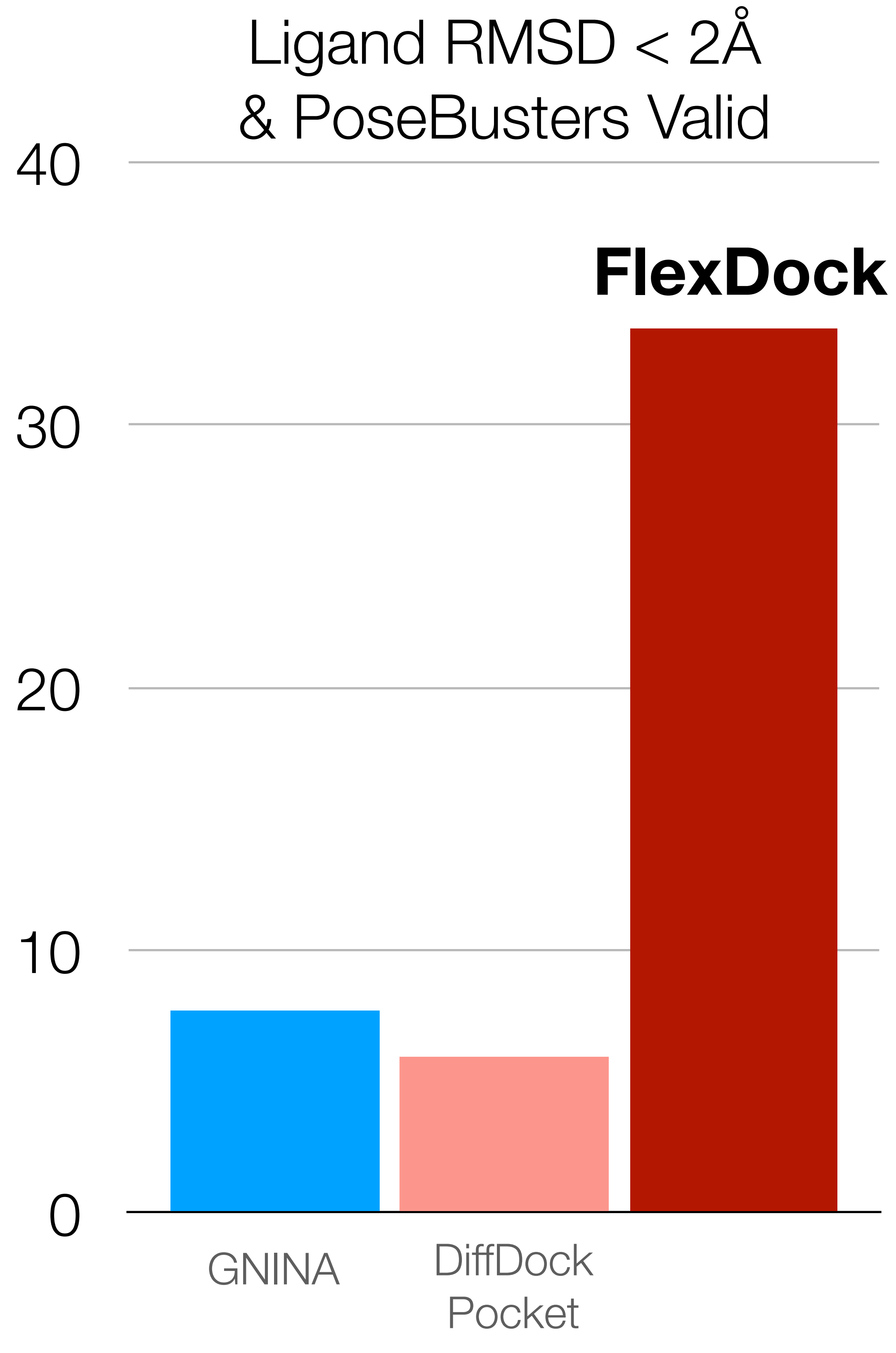
Pocket-based Flexible Docking



Ligand accuracy



Receptor accuracy



Pose quality

MIT CSAIL

JAMEEL CLINIC

MIT MLPDS

DEFENSE THREAT REDUCTION AGENCY
UNITED STATES OF AMERICA

Thank You!

Collaborators:

Tommi Jaakkola
Regina Barzilay
Vignesh Ram Somnath
Noah Getz
Andreas Krause
Hannes Stärk
Bowen Jing

Resources:

DiffDock

Paper: arxiv.org/abs/2210.01776
Code: github.com/gcorso/DiffDock

Unbalanced FM

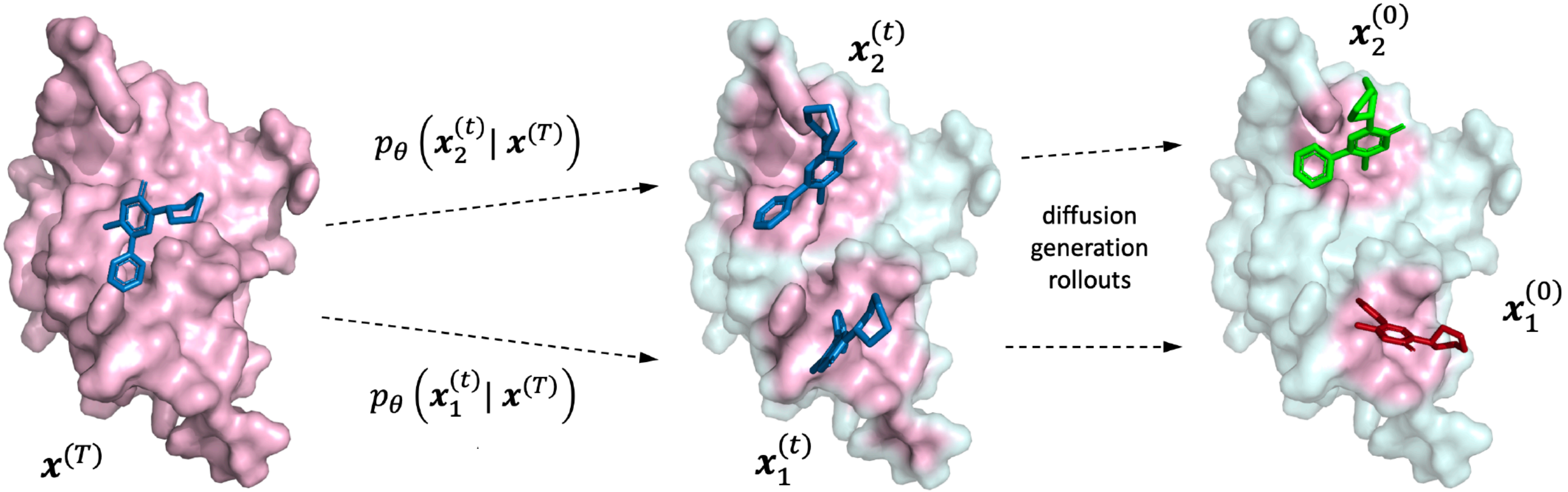
Preprint and code soon!
Or just ask me ;)

Contact me:

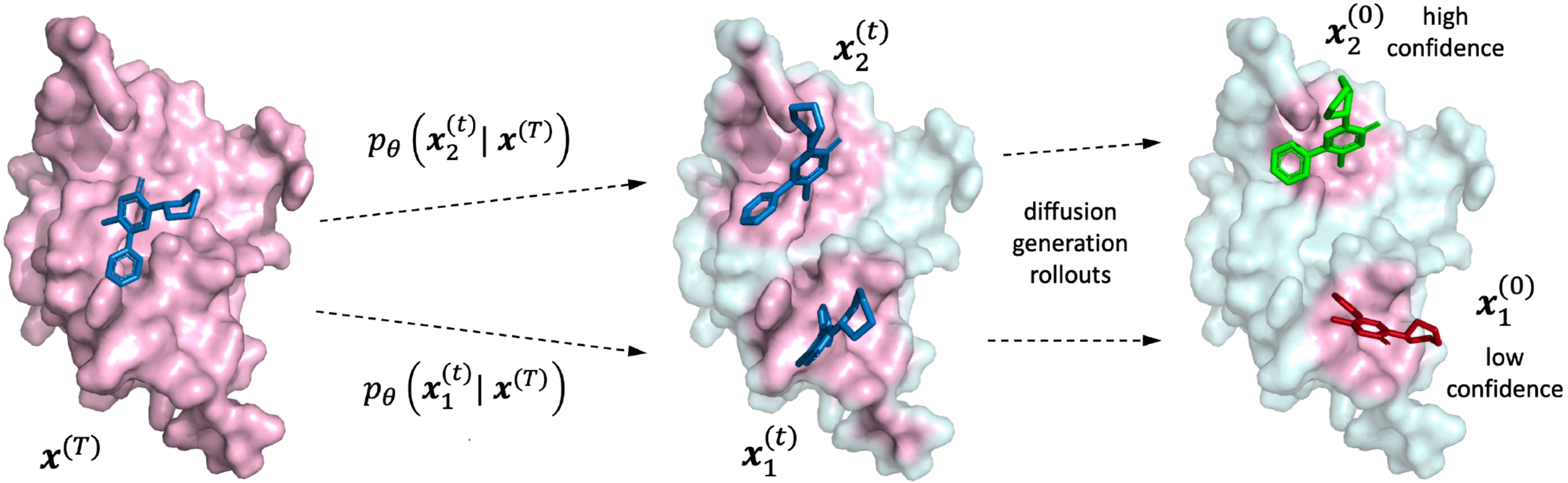
✉ **gcorso@mit.edu**

🐦 **@GabriCorso**

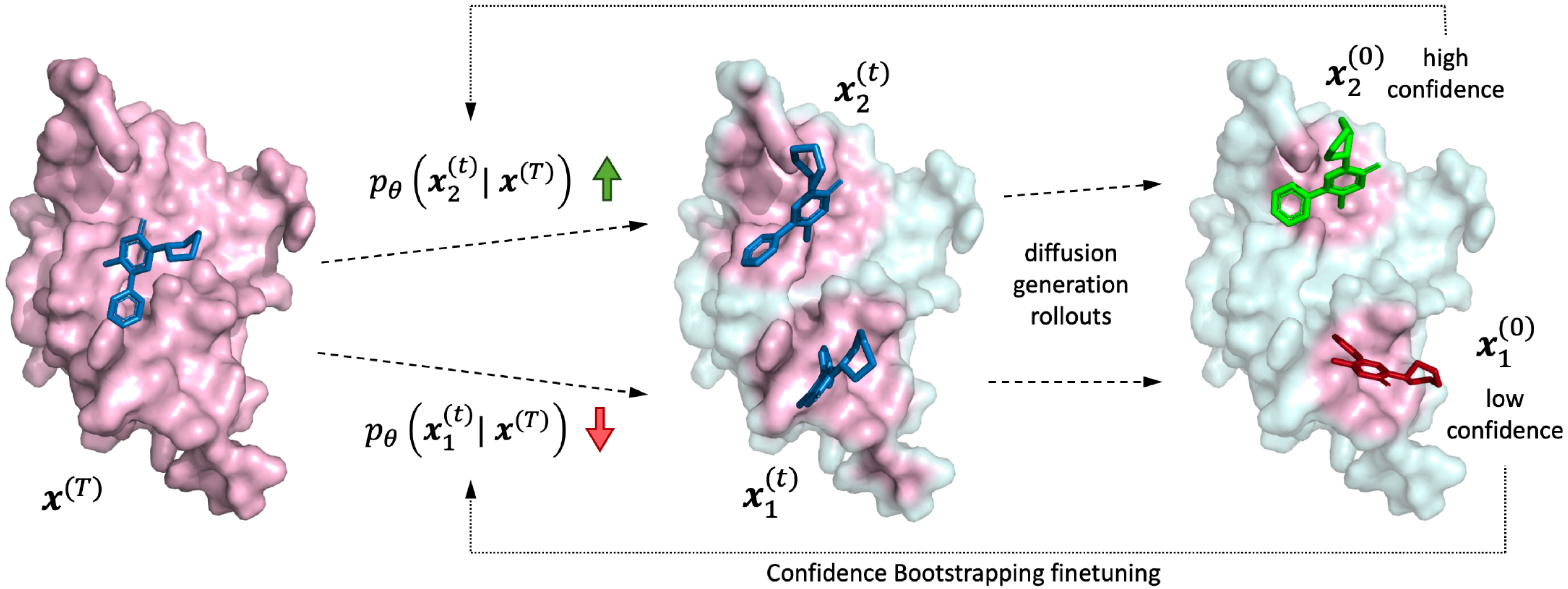
Confidence Bootstrapping



Confidence Bootstrapping


$x^{(T)}$
$p_\theta\left(x_2^{(t)} \mid x^{(T)}\right)$
$p_\theta\left(x_1^{(t)} \mid x^{(T)}\right)$
$x_2^{(t)}$
$x_1^{(t)}$
diffusion generation rollouts
$x_2^{(0)}$
high confidence
$x_1^{(0)}$
low confidence

Confidence Bootstrapping



$x^{(T)}$
$p_\theta\left(x_2^{(t)} | x^{(T)}\right)$ ↑
$p_\theta\left(x_1^{(t)} | x^{(T)}\right)$ ↓
$x_2^{(t)}$
$x_1^{(t)}$
diffusion generation rollouts
$x_2^{(0)}$ high confidence
$x_1^{(0)}$ low confidence
Confidence Bootstrapping finetuning

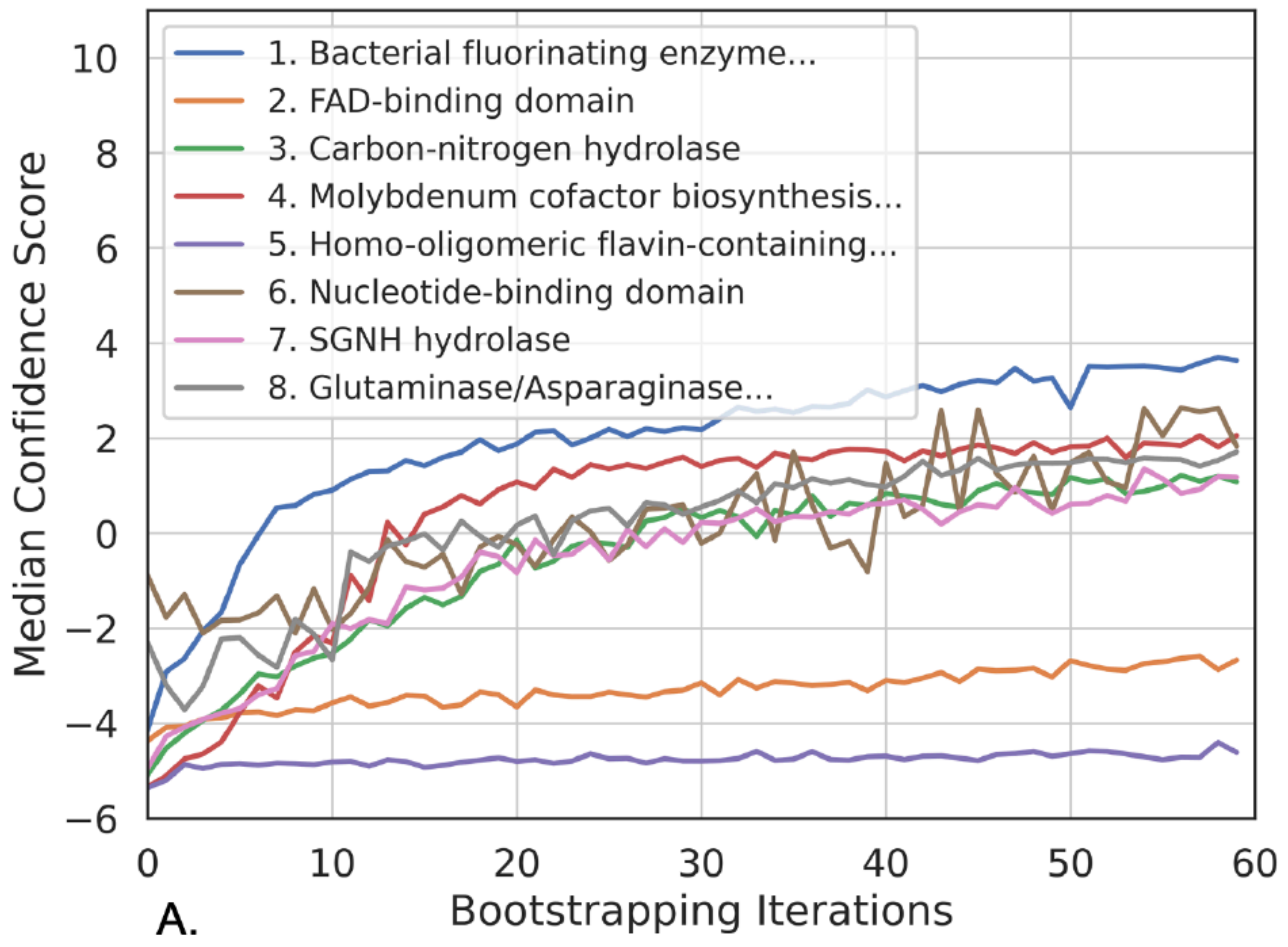
Finetuning on specific protein classes

- We validate the effectiveness of Confidence Bootstrapping by fine-tuning DiffDock to work well on protein classes with no binding structural data is available in training set

Finetuning on specific protein classes

- We validate the effectiveness of Confidence Bootstrapping by fine-tuning DiffDock to work well on protein classes with no binding structural data is available in training set

- As expected the confidence of generated samples increases over iterations

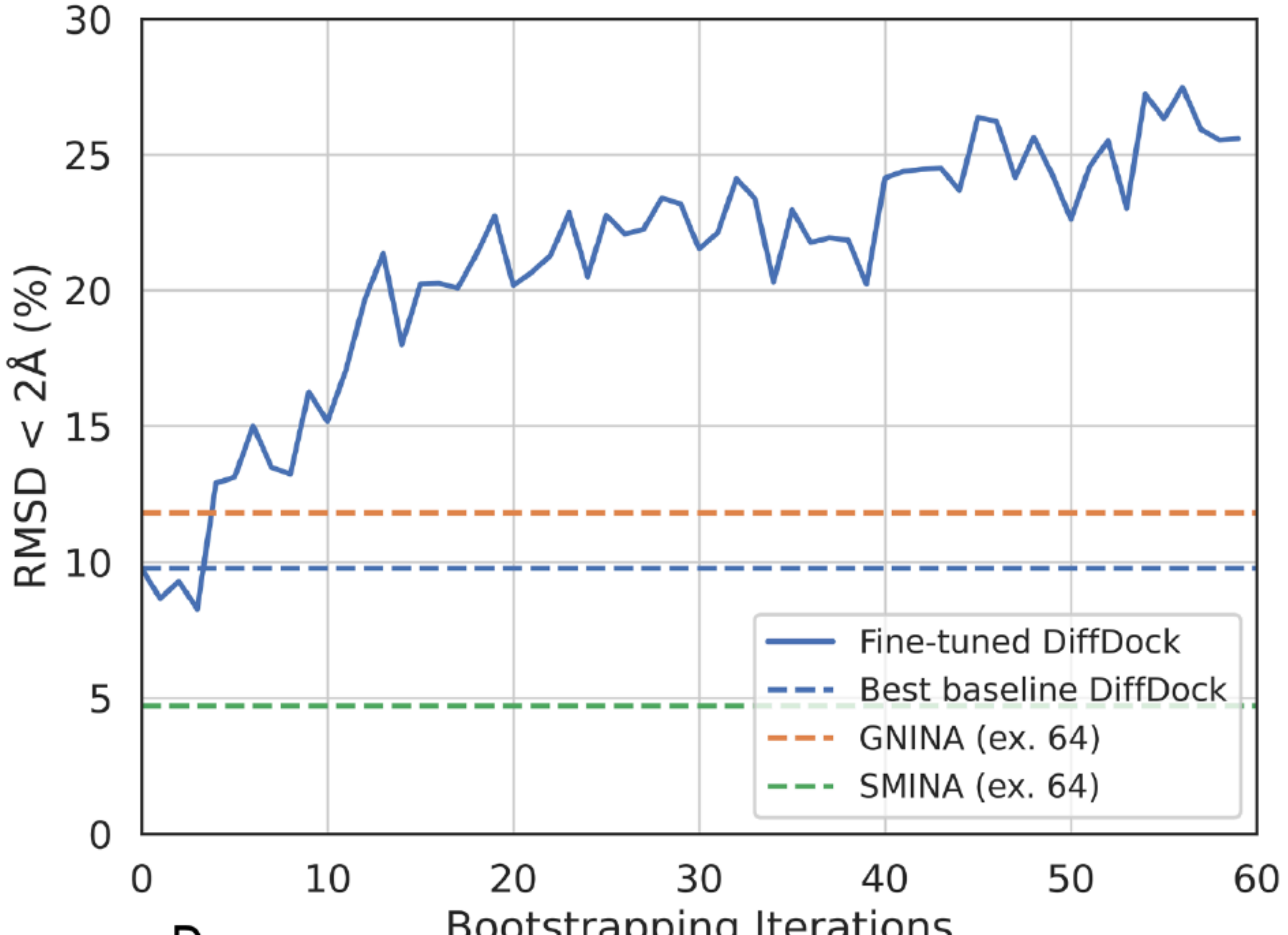


Finetuning on specific protein classes

- We validate the effectiveness of Confidence Bootstrapping by fine-tuning DiffDock to work well on protein classes with no binding structural data is available in training set

- As expected the confidence of generated samples increases over iterations

- On average this translates in significant improvements in docking accuracy



D.

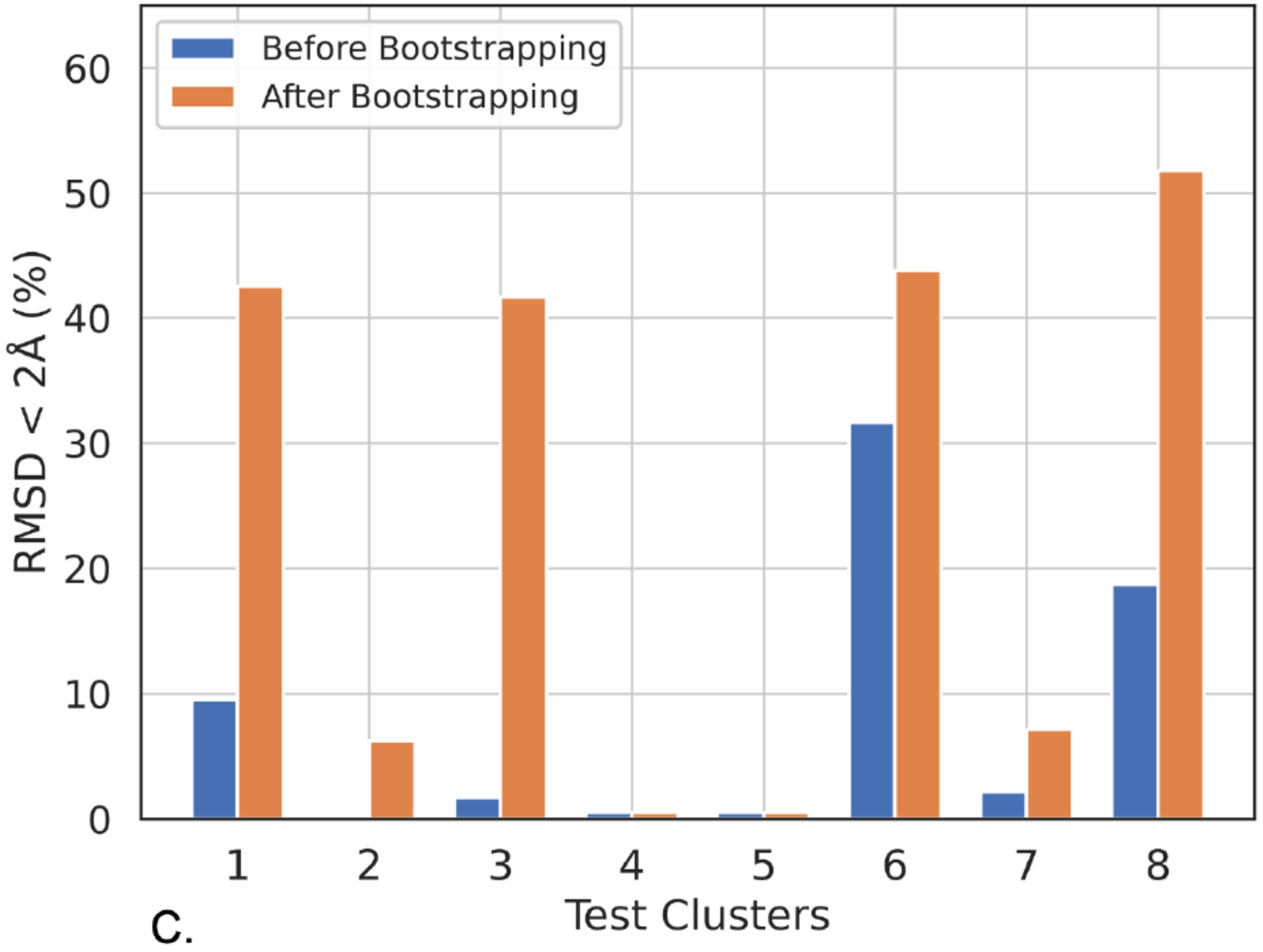
Finetuning on specific protein classes

- We validate the effectiveness of Confidence Bootstrapping by fine-tuning DiffDock to work well on protein classes with no binding structural data is available in training set

- As expected the confidence of generated samples increases over iterations

- On average this translates in significant improvements in docking accuracy

- The performance on individual clusters present interesting insights



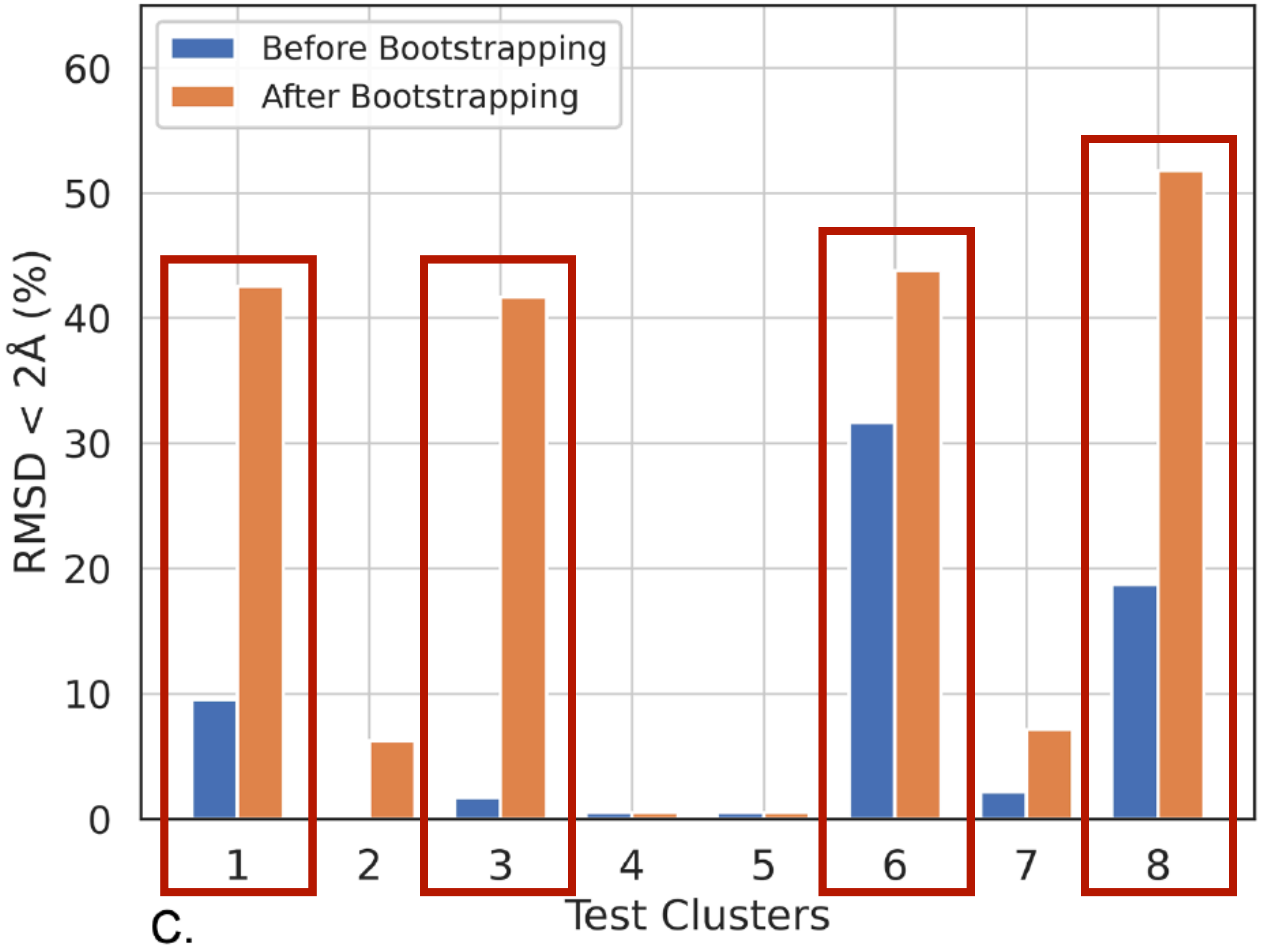
Finetuning on specific protein classes

- We validate the effectiveness of Confidence Bootstrapping by fine-tuning DiffDock to work well on protein classes with no binding structural data is available in training set

- As expected the confidence of generated samples increases over iterations

- On average this translates in significant improvements in docking accuracy

- The performance on individual clusters present interesting insights



For many protein families the model drastically improves docking accuracy

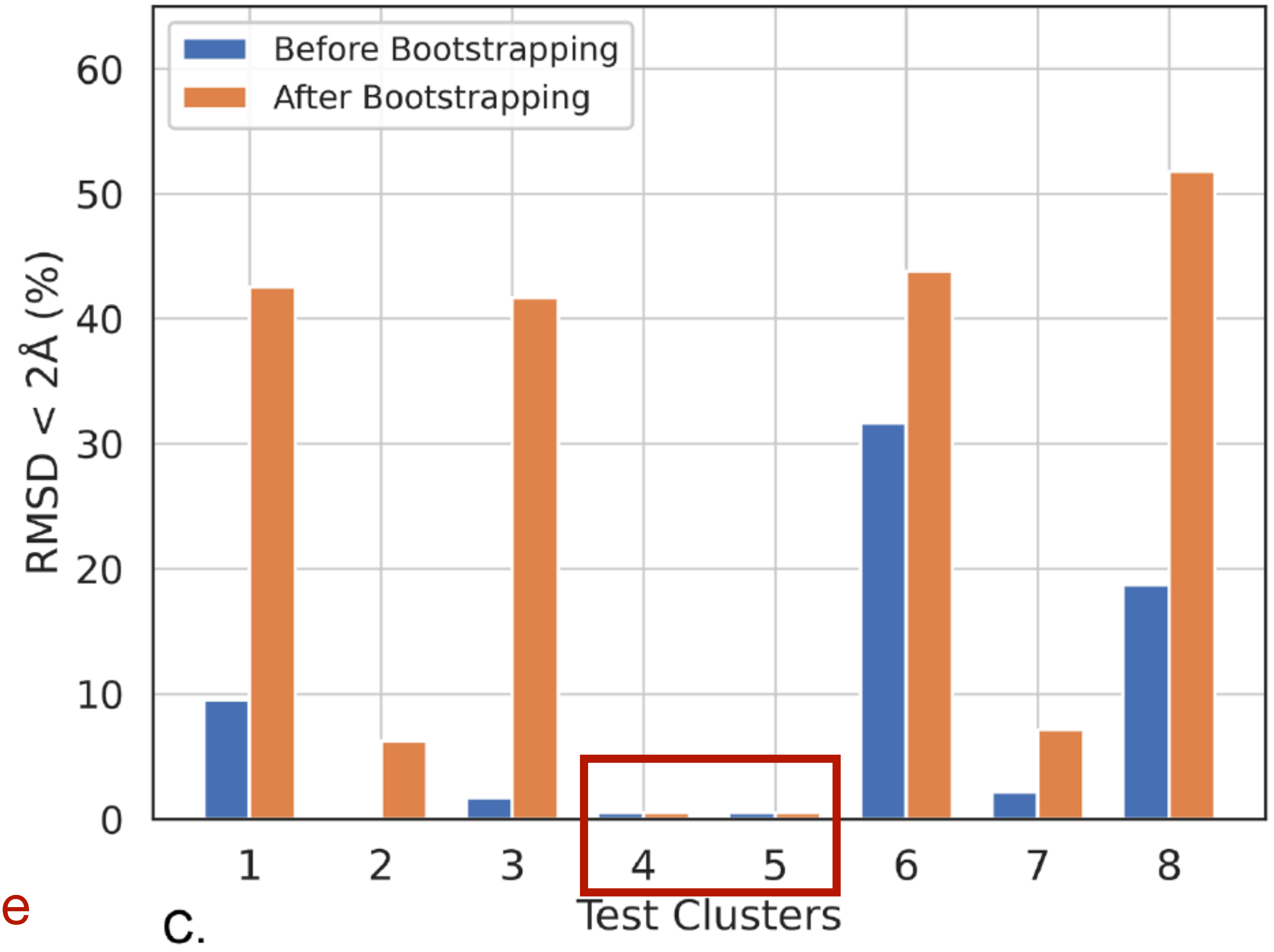
Finetuning on specific protein classes

- We validate the effectiveness of Confidence Bootstrapping by fine-tuning DiffDock to work well on protein classes with no binding structural data is available in training set

- As expected the confidence of generated samples increases over iterations

- On average this translates in significant improvements in docking accuracy

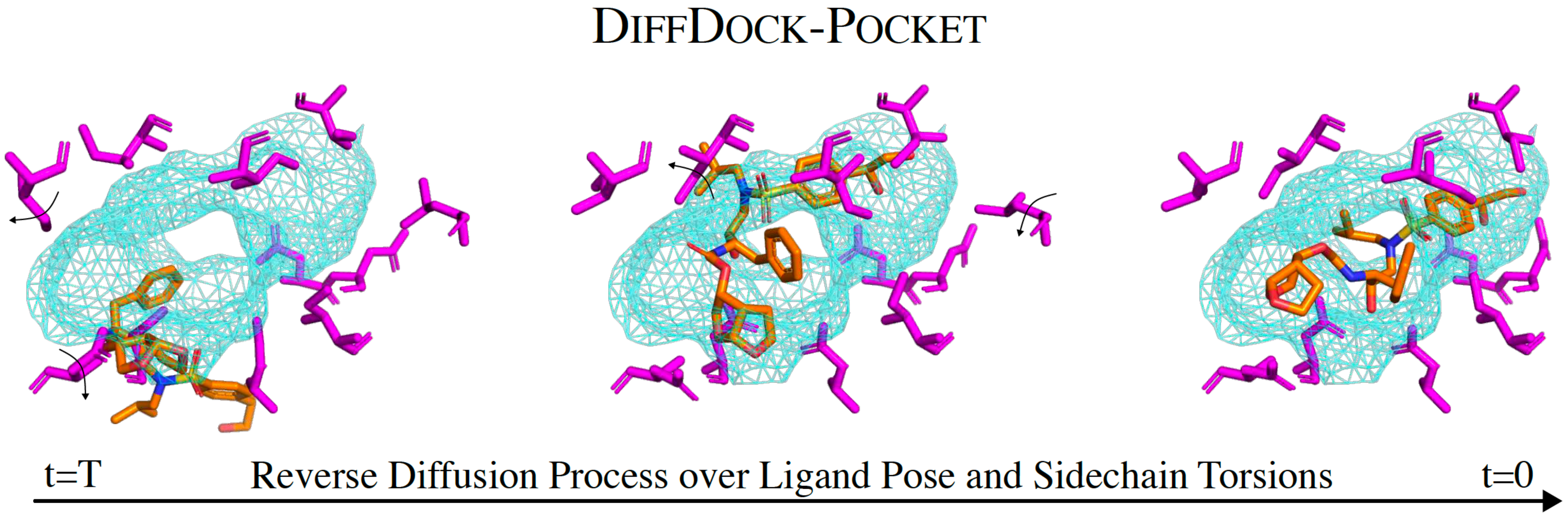
- The performance on individual clusters present interesting insights



But for some where the diffusion model had little/no coverage the method has no way of improve

DiffDock-Pocket

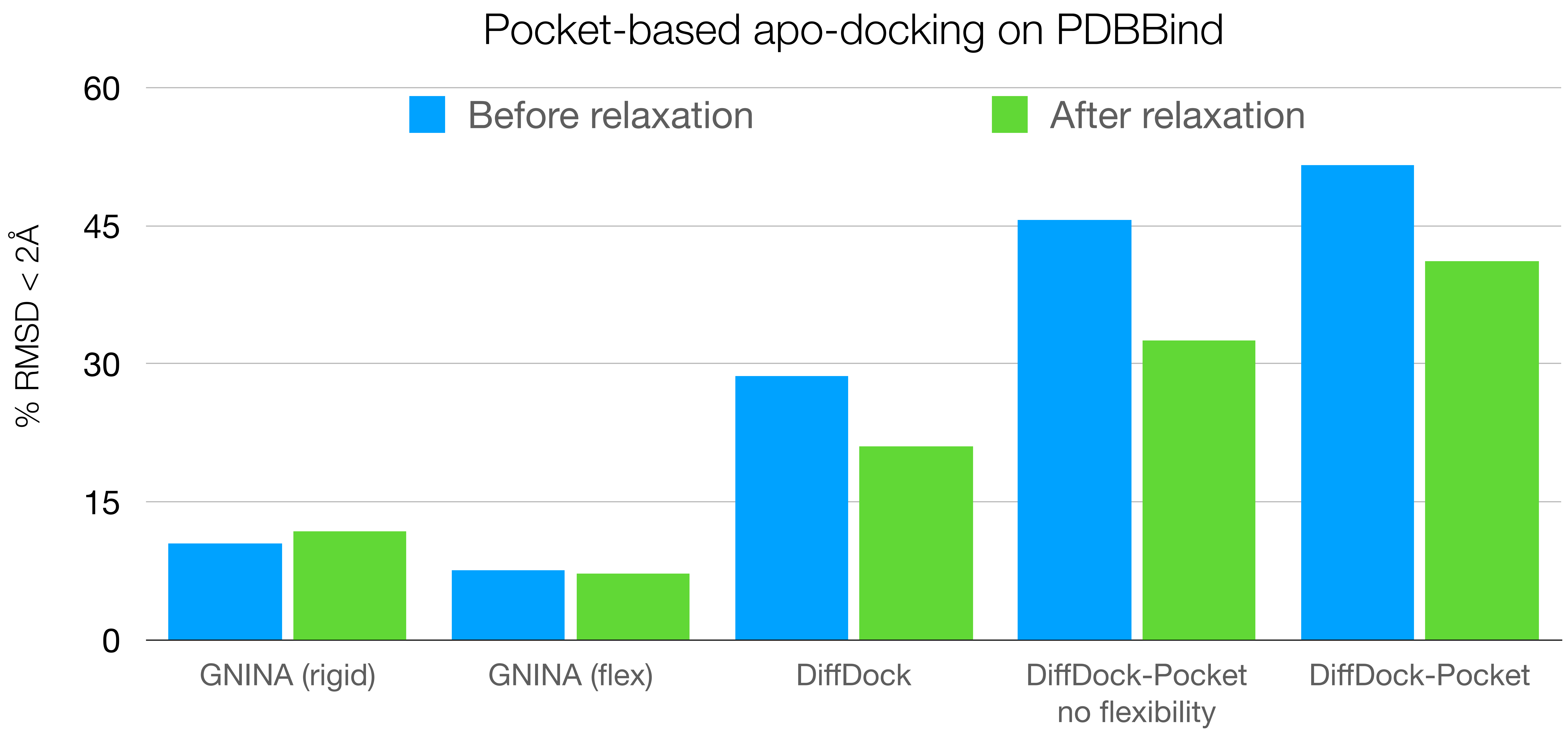
A step towards all-atoms flexible docking


DIFFDOCK-POCKET
t=T
Reverse Diffusion Process over Ligand Pose and Sidechain Torsions
t=0

DiffDock-Pocket: Diffusion for Pocket-Level Docking with Sidechain Flexibility, *Plainer et al.*, Under review

DiffDock-Pocket

A step towards all-atoms flexible docking

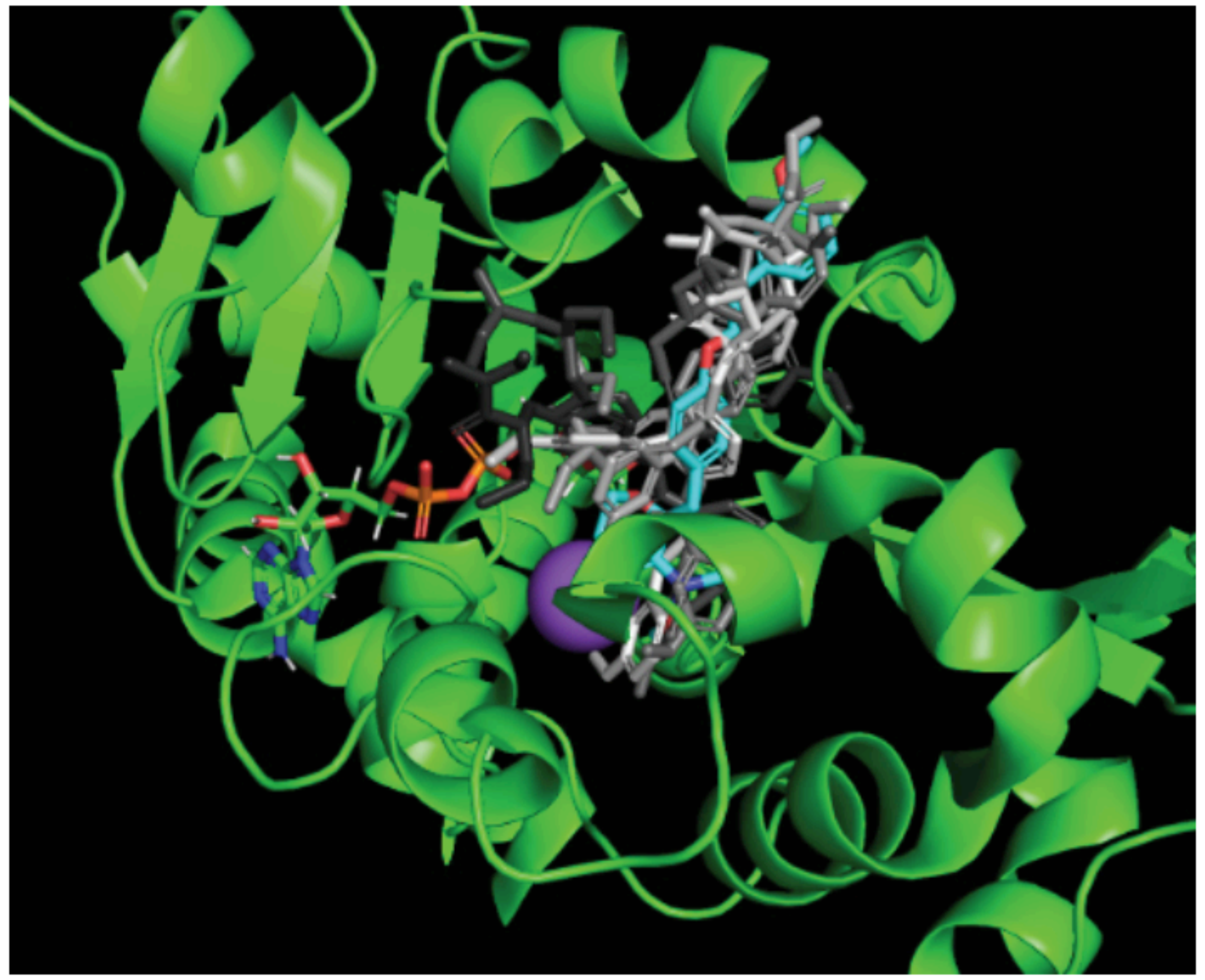


Released very soon!
Stay tuned:

@GabriCorso

DiffDock-Pocket: Diffusion for Pocket-Level Docking with Sidechain Flexibility, *Plainer et al.*, Under review

DiffDock for reverse screening

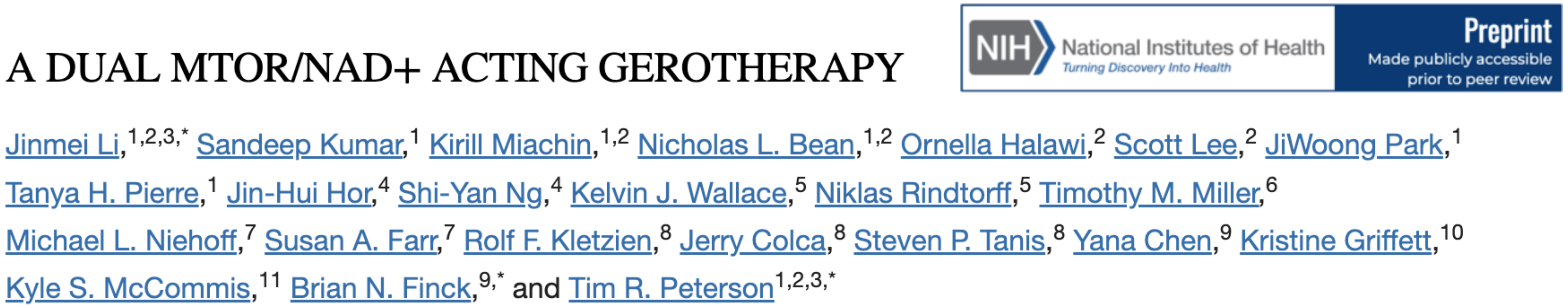


🟩 SIRT3

🟦 BIOIO-1001 (rank 1)

◧ BIOIO-1001 (rank 5 - 40)

🟧 NAD-Ribose


NIH National Institutes of Health
Turning Discovery Into Health
Preprint
Made publicly accessible prior to peer review

A DUAL MTOR/NAD+ ACTING GEROTHERAPY

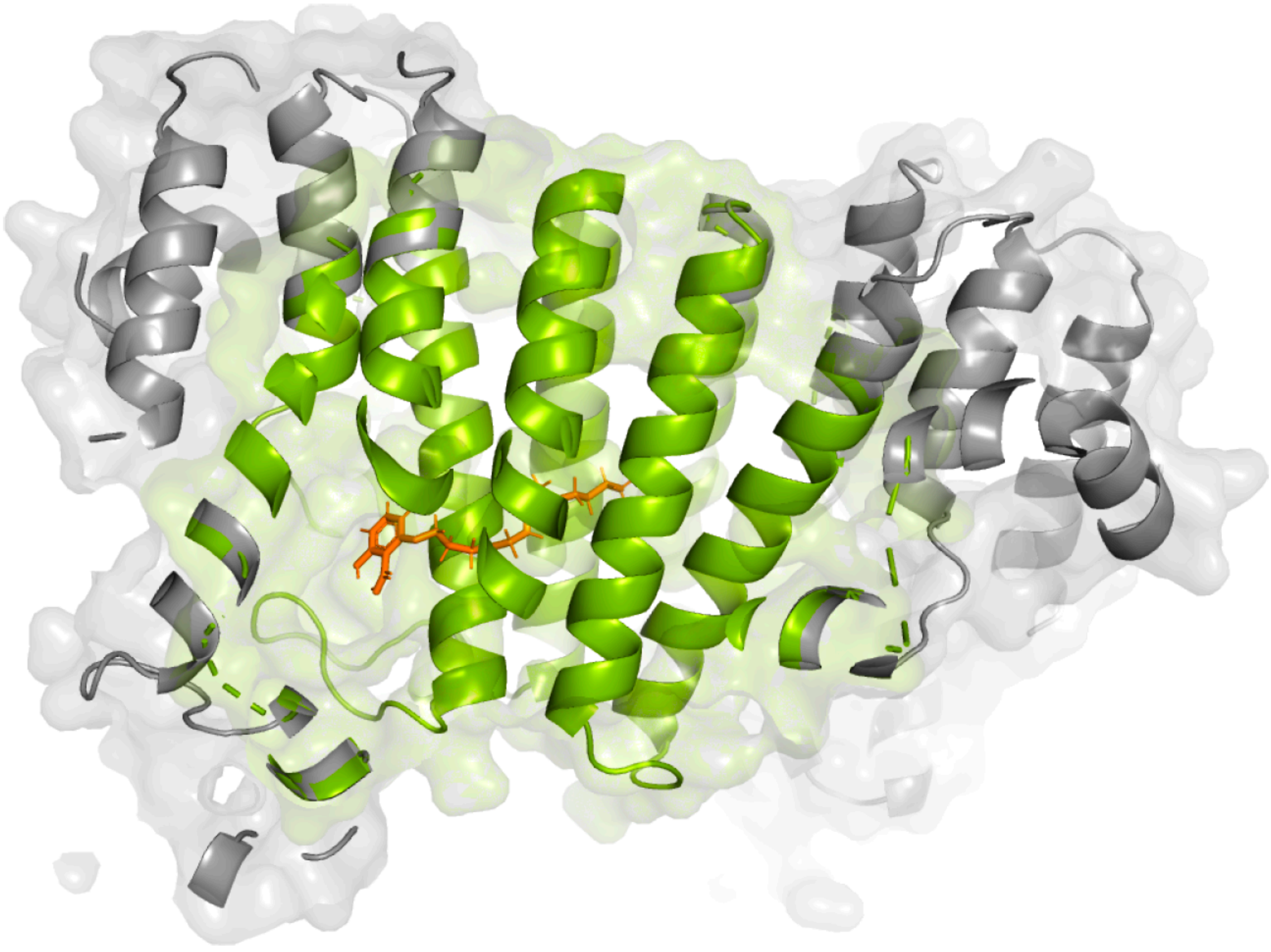
Jinmei Li,[1,2,3,*] Sandeep Kumar,[1] Kirill Miachin,[1,2] Nicholas L. Bean,[1,2] Ornella Halawi,[2] Scott Lee,[2] JiWoong Park,[1] Tanya H. Pierre,[1] Jin-Hui Hor,[4] Shi-Yan Ng,[4] Kelvin J. Wallace,[5] Niklas Rindtorff,[5] Timothy M. Miller,[6] Michael L. Niehoff,[7] Susan A. Farr,[7] Rolf F. Kletzien,[8] Jerry Colca,[8] Steven P. Tanis,[8] Yana Chen,[9] Kristine Griffett,[10] Kyle S. McCommis,[11] Brian N. Finck,[9,*] and Tim R. Peterson[1,2,3,*]

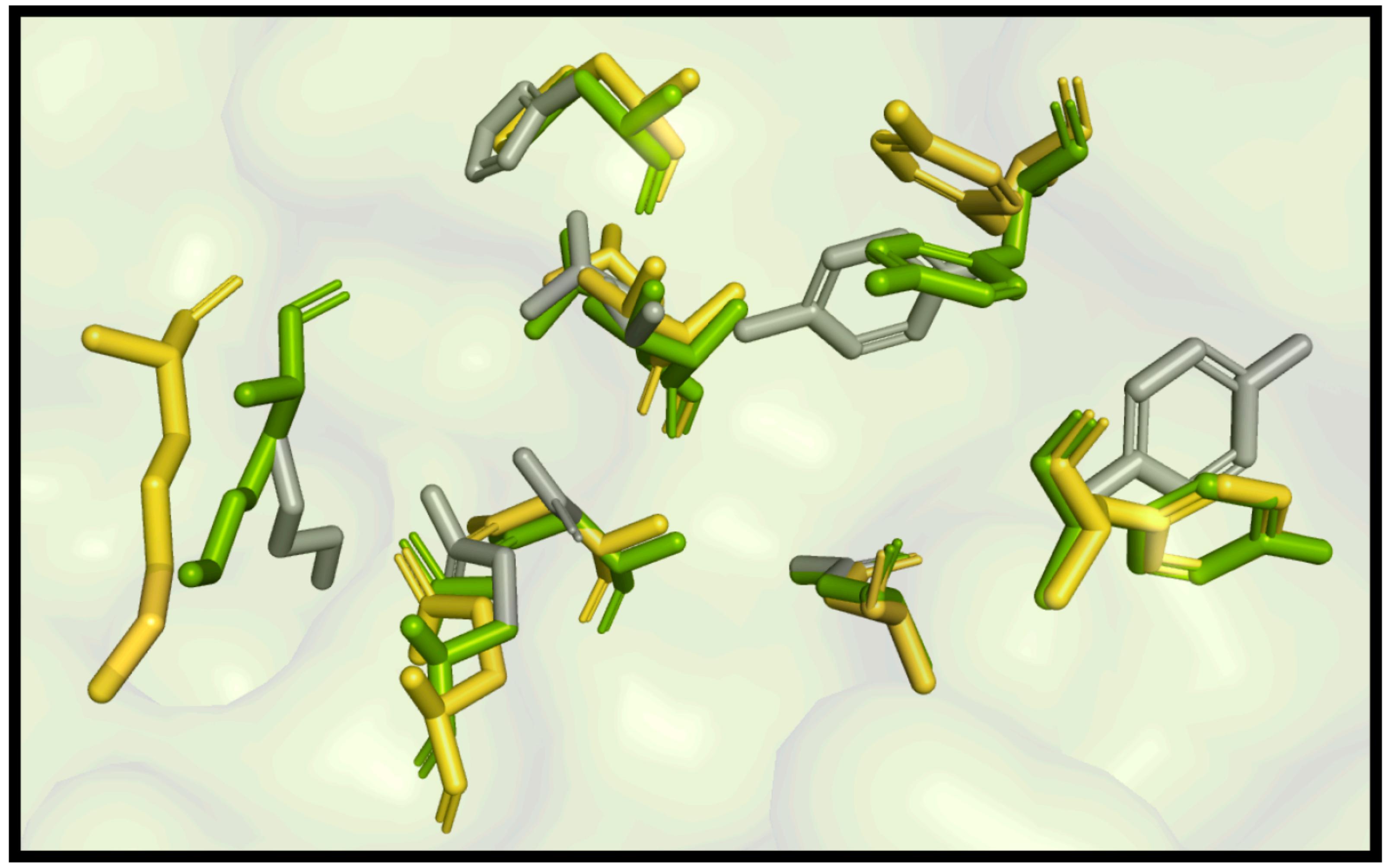
"DiffDock makes drug target identification much more possible. Before one had to do laborious and costly experiments (months to years) with each protein to define the drug docking. But now one can screen many proteins and do the triaging virtually in a day."

Tim R. Peterson
Assistant Professor, Washington University in St. Louis

Used to understand the mechanism of action of a new drug

Pocket-conditioned docking





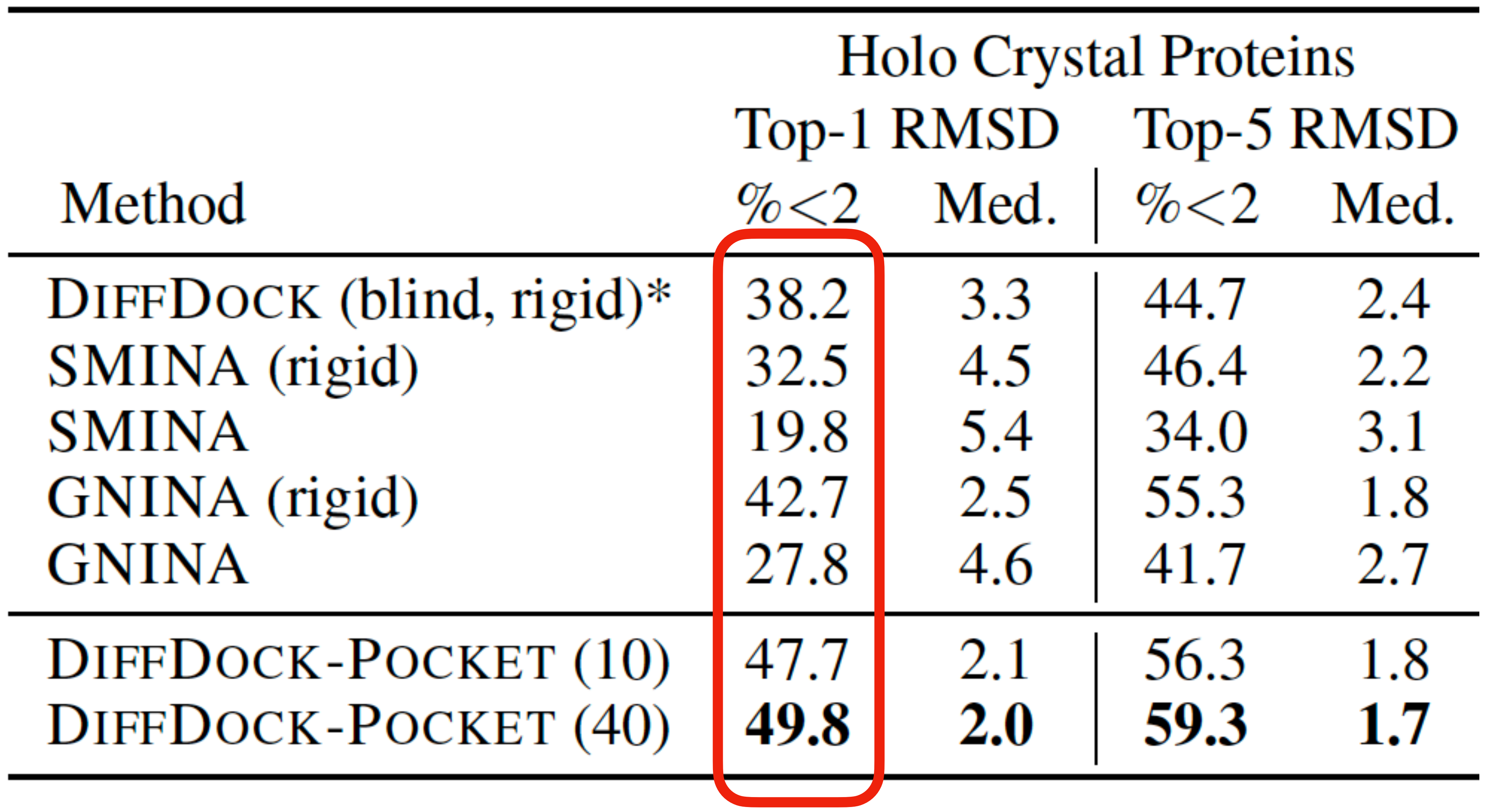
1. Restricted pocket focus

2. Access to full-atomic structures

3. Side-chain torsional flexibility built-into the diffusion process

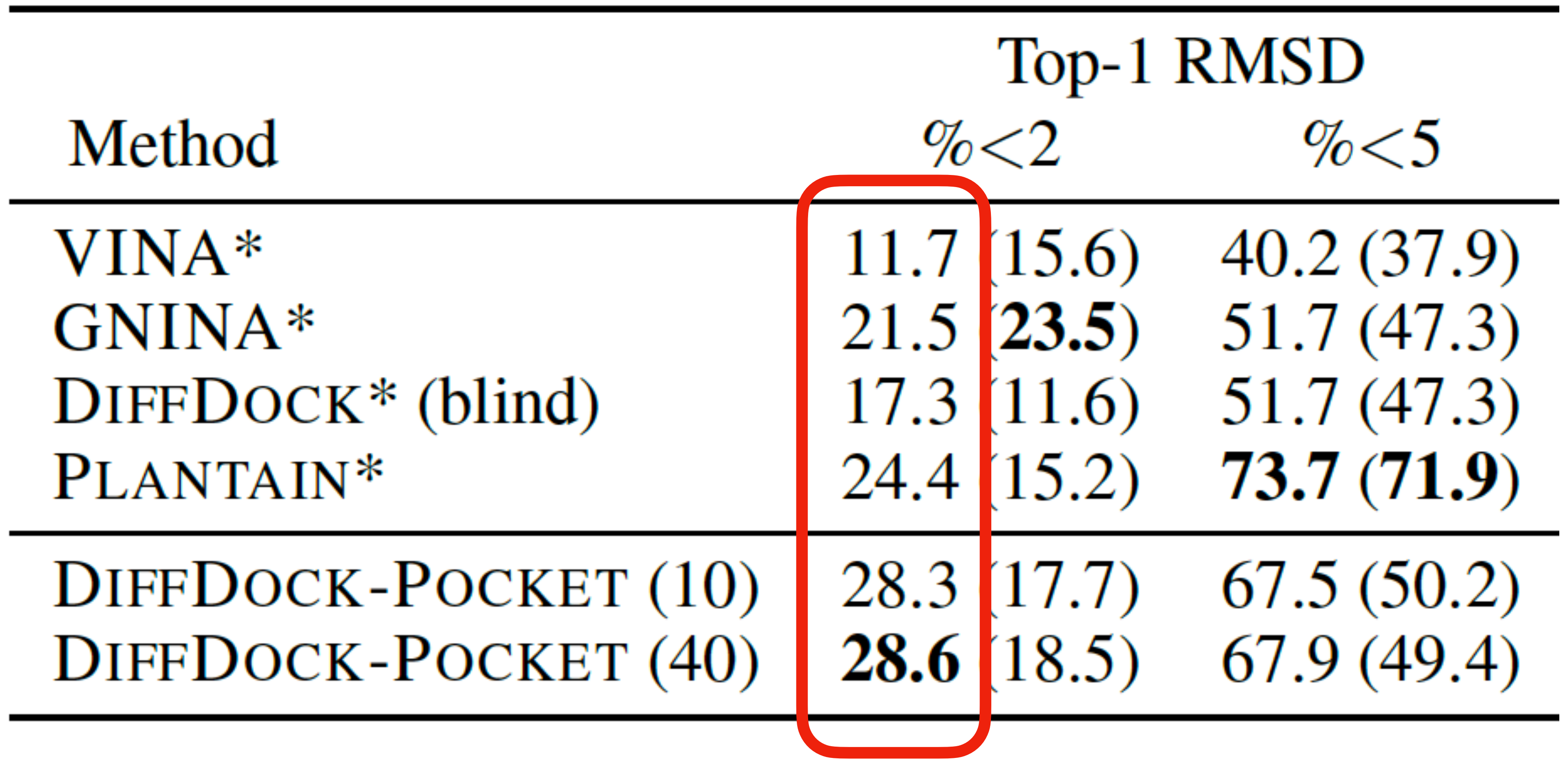
Results

- Holo and cross docking performance on par with best pocket-based methods



	Holo Crystal Proteins			
	Top-1 RMSD		Top-5 RMSD	
Method	%<2	Med.	%<2	Med.
DiffDock (blind, rigid)*	38.2	3.3	44.7	2.4
SMINA (rigid)	32.5	4.5	46.4	2.2
SMINA	19.8	5.4	34.0	3.1
GNINA (rigid)	42.7	2.5	55.3	1.8
GNINA	27.8	4.6	41.7	2.7
DiffDock-Pocket (10)	47.7	2.1	56.3	1.8
DiffDock-Pocket (40)	**49.8**	**2.0**	**59.3**	**1.7**

Holo-docking on PDBBind

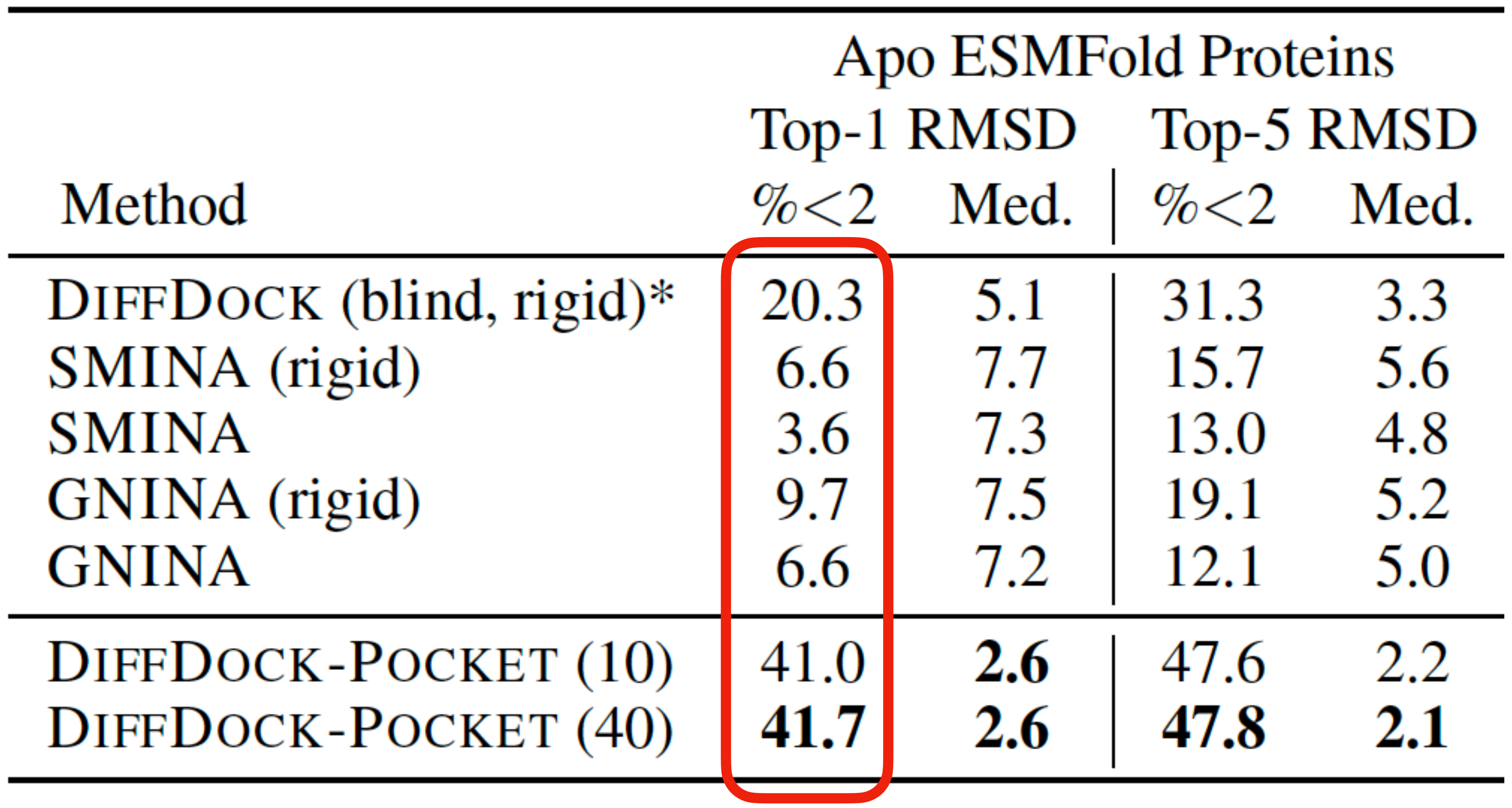


	Top-1 RMSD	
Method	%<2	%<5
VINA*	11.7 (15.6)	40.2 (37.9)
GNINA*	21.5 (**23.5**)	51.7 (47.3)
DiffDock* (blind)	17.3 (11.6)	51.7 (47.3)
Plantain*	24.4 (15.2)	**73.7** (**71.9**)
DiffDock-Pocket (10)	28.3 (17.7)	67.5 (50.2)
DiffDock-Pocket (40)	**28.6** (18.5)	67.9 (49.4)

Cross-docking on unseen proteins from CrossDocked 2020

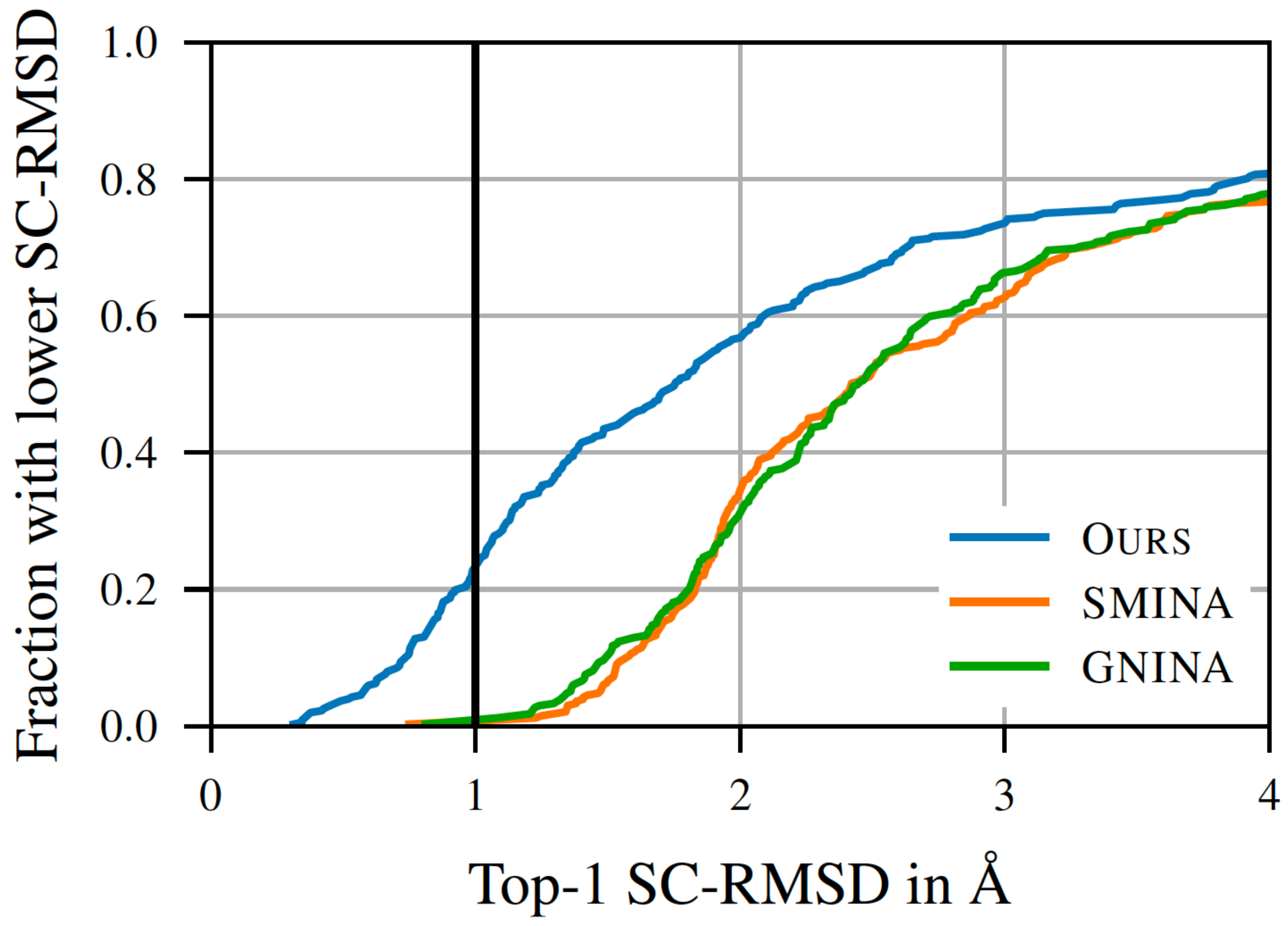
Results

- Holo and cross docking performance on par with best pocket-based methods
- Significantly better apo docking and modeling of receptor flexibility



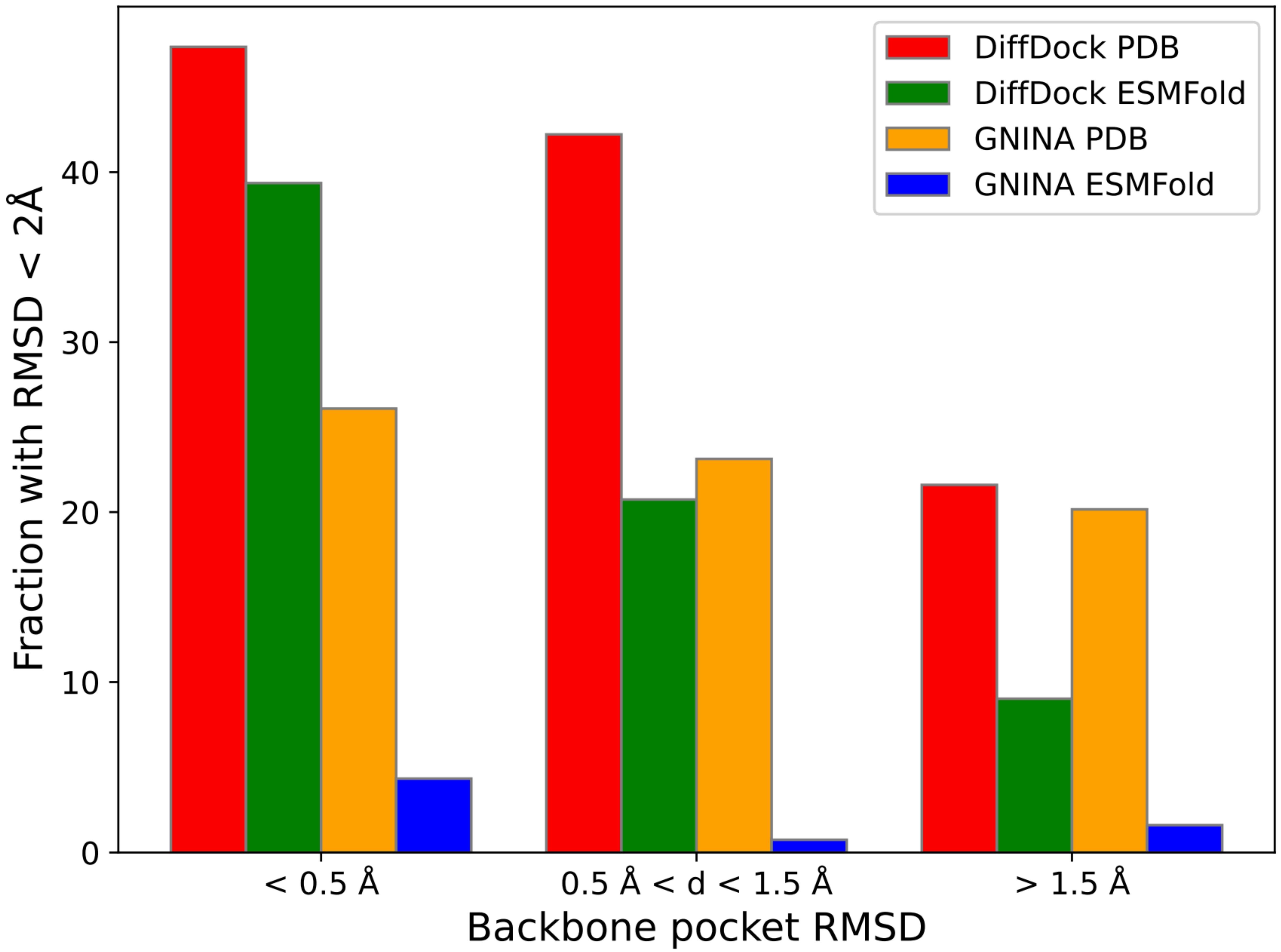
	Apo ESMFold Proteins			
	Top-1 RMSD		Top-5 RMSD	
Method	%<2	Med.	%<2	Med.
DiffDock (blind, rigid)*	20.3	5.1	31.3	3.3
SMINA (rigid)	6.6	7.7	15.7	5.6
SMINA	3.6	7.3	13.0	4.8
GNINA (rigid)	9.7	7.5	19.1	5.2
GNINA	6.6	7.2	12.1	5.0
DiffDock-Pocket (10)	41.0	**2.6**	47.6	2.2
DiffDock-Pocket (40)	**41.7**	**2.6**	**47.8**	**2.1**

Apo-docking on PDBBind



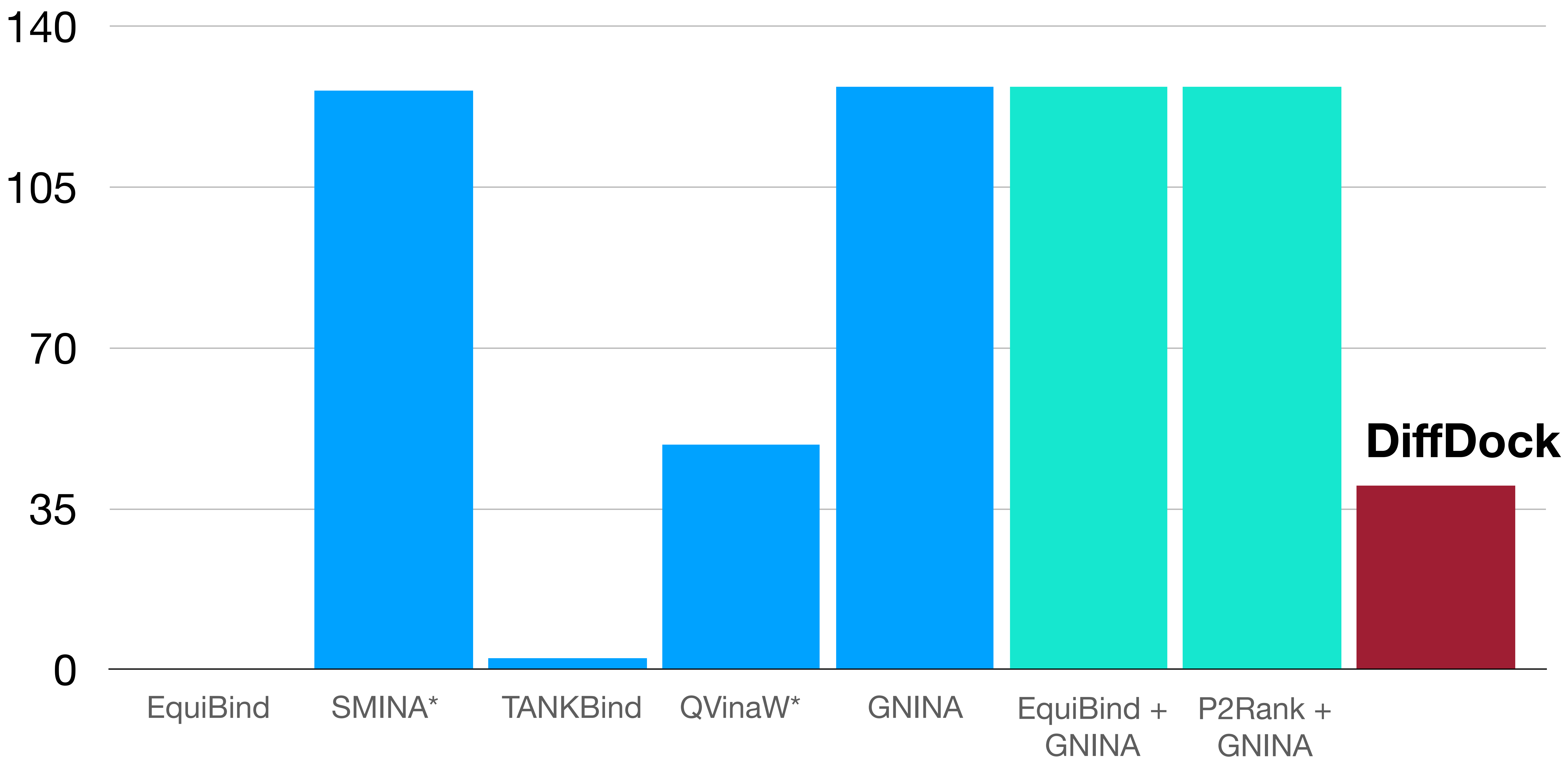
Sidechain RMSD on PDBBind

Performance vs apo precision



Runtime

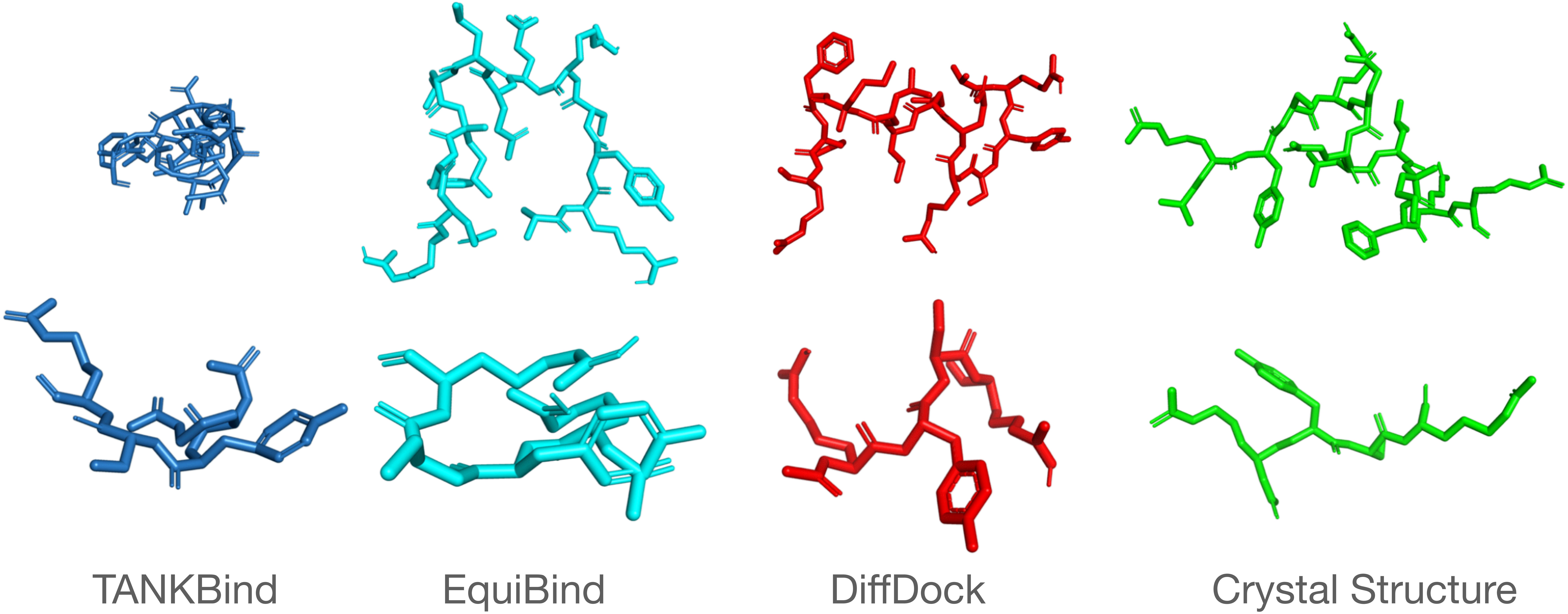
Number of seconds for a single complex



3x faster than the most accurate baseline

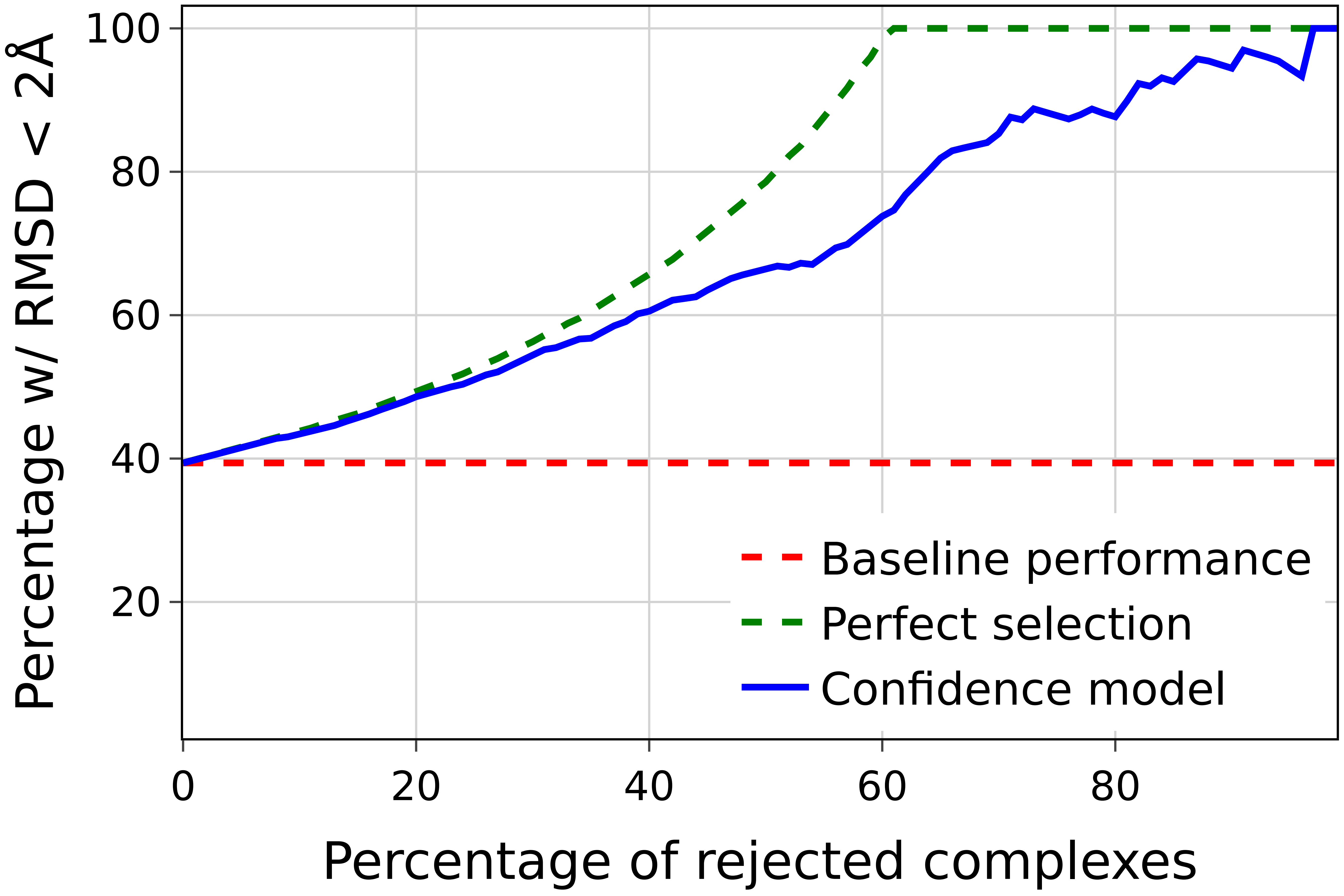
*Ran exclusively on CPU

Physically plausible structures



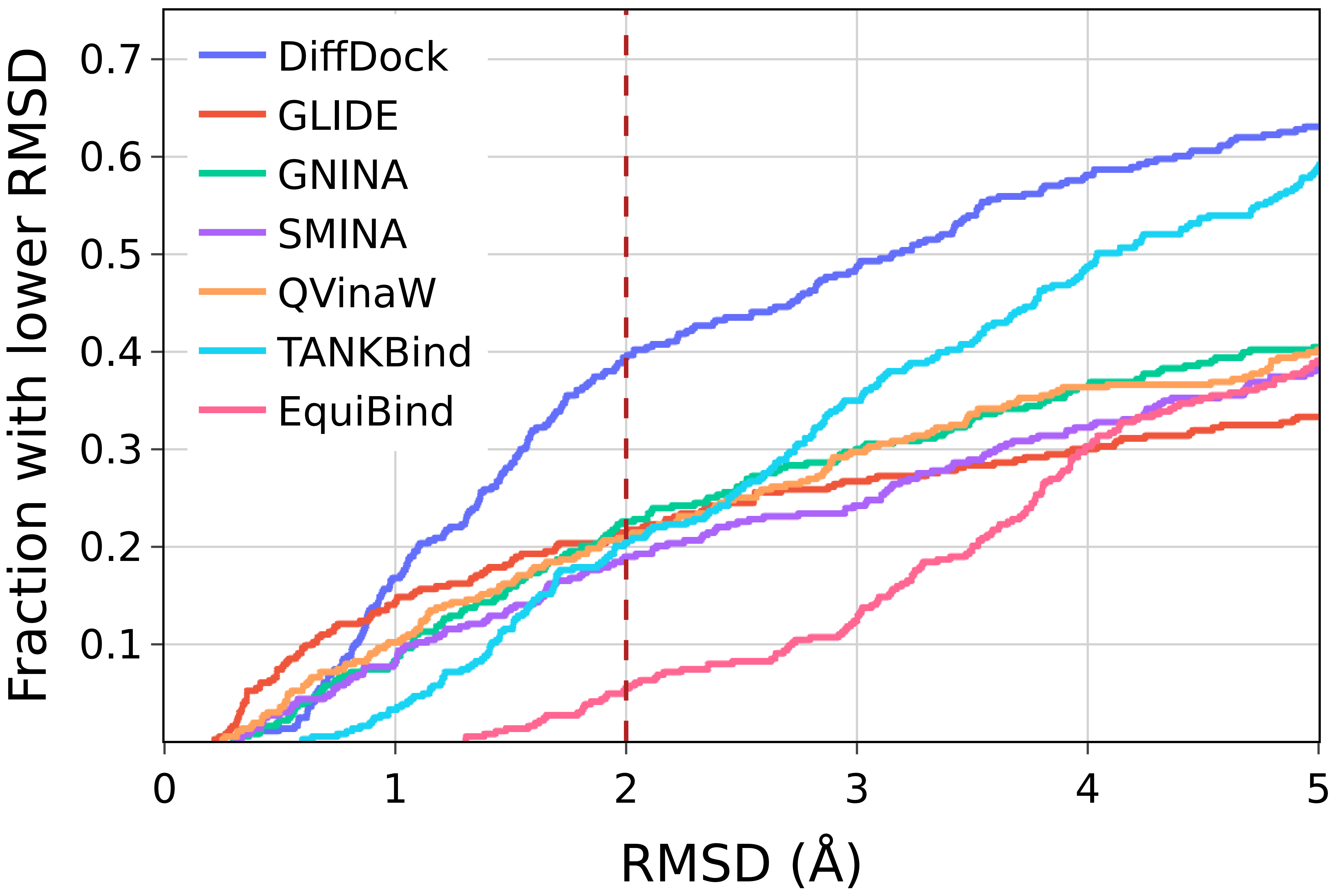
No self intersections unlike previous DL methods

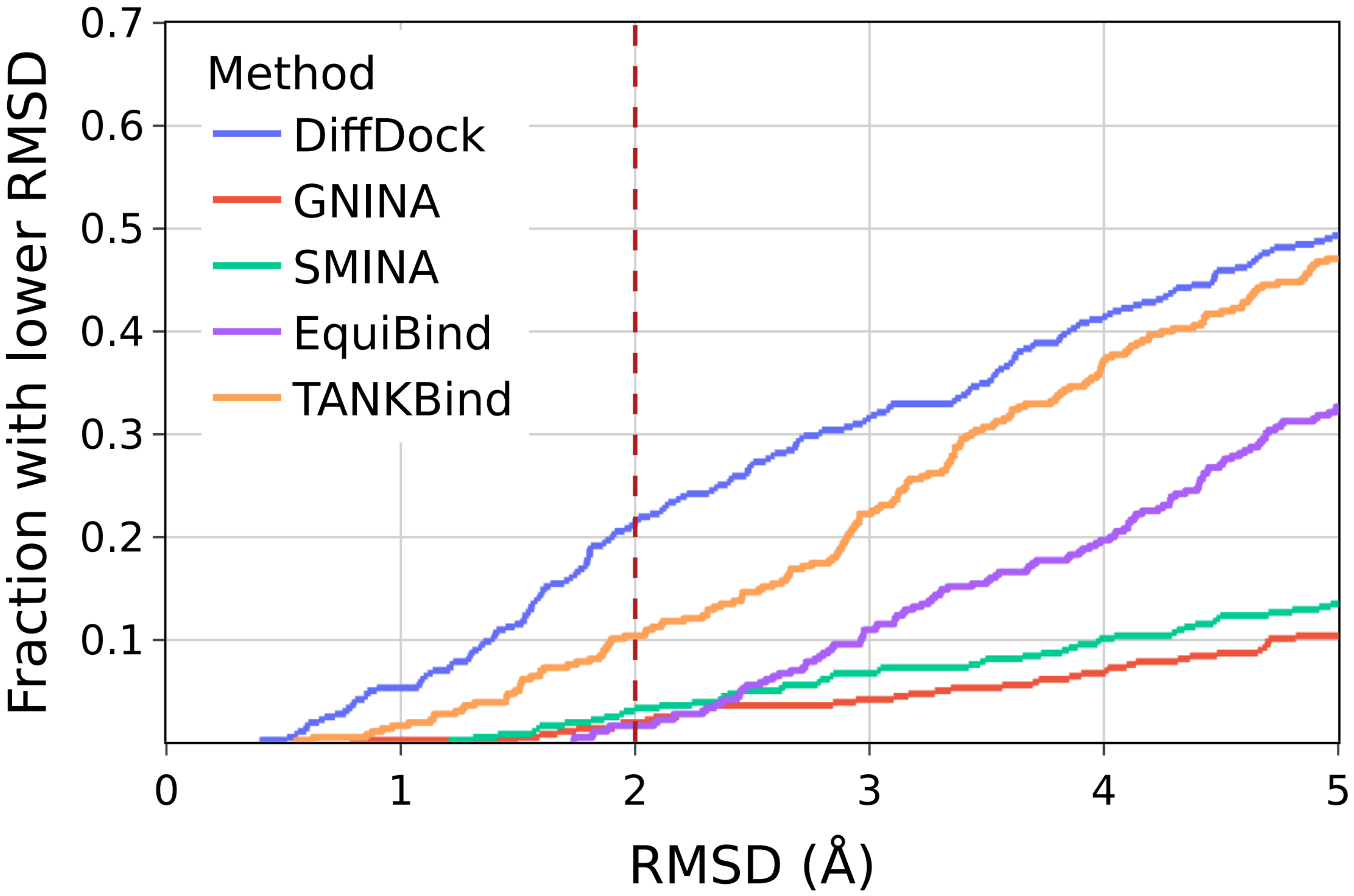
Confidence score quality



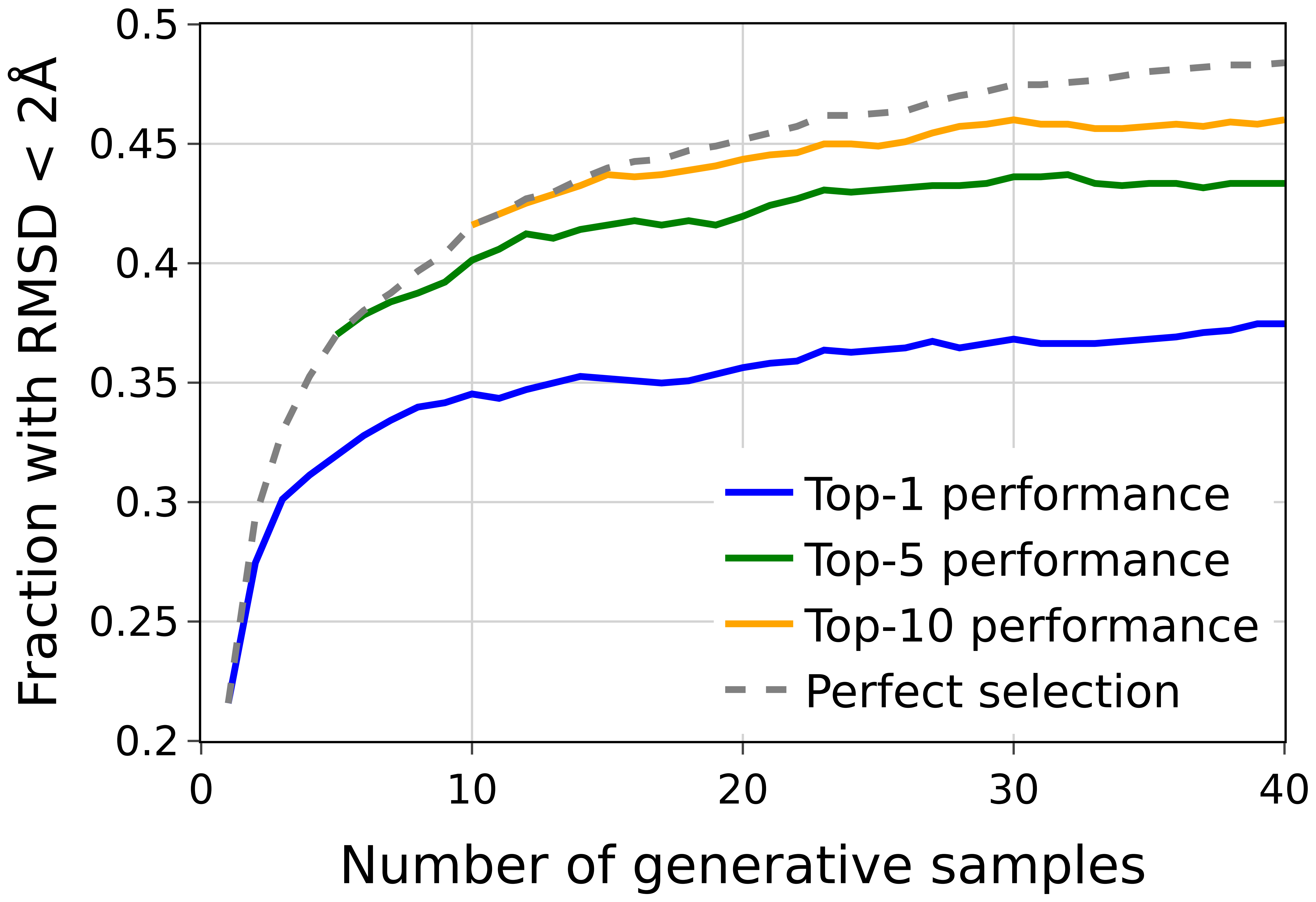
High selective accuracy: valuable information for practitioners

Prediction correctness



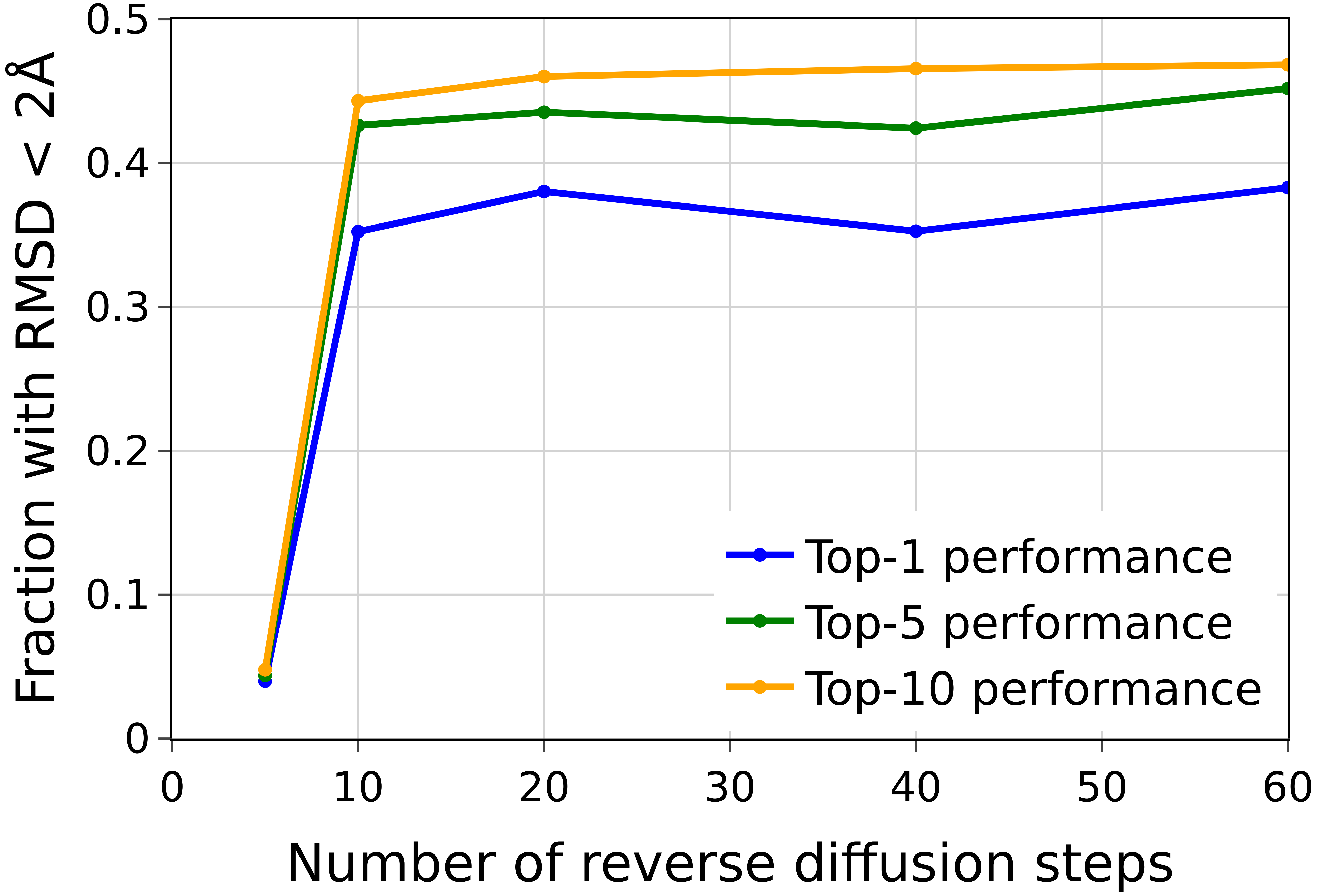


Top-N performance



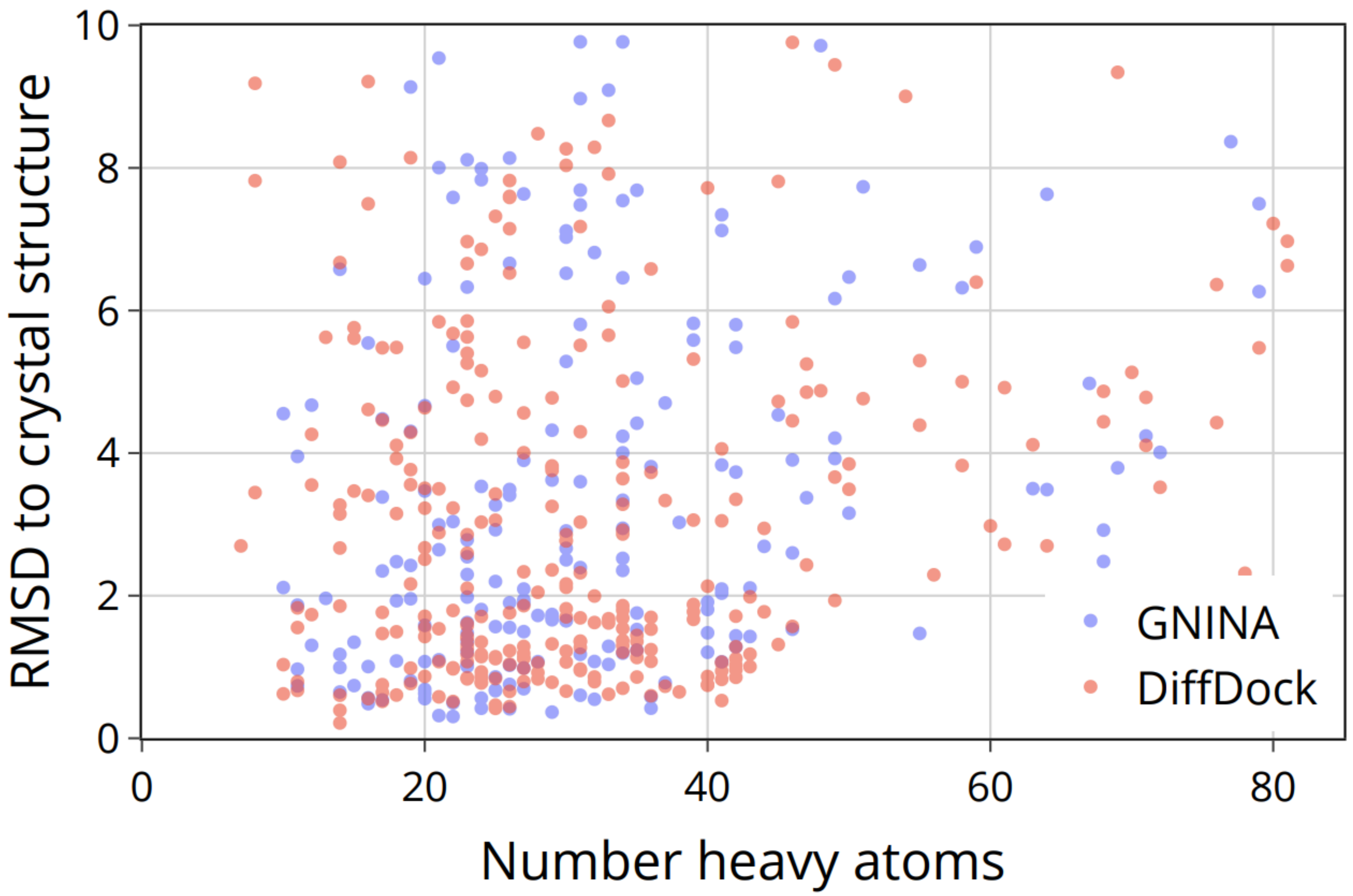
Diverse set of structure predictions

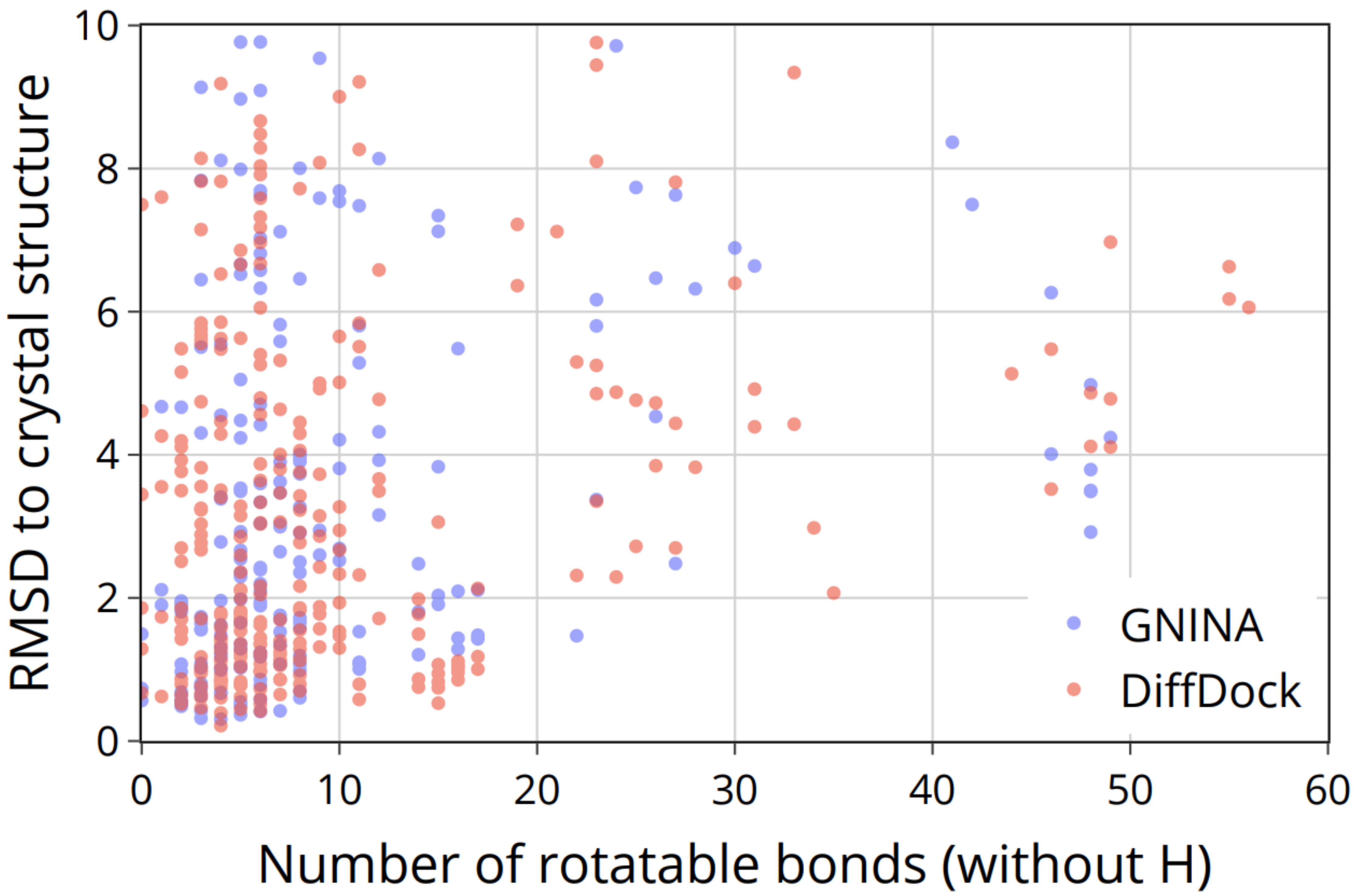
Number of Diffusion Steps



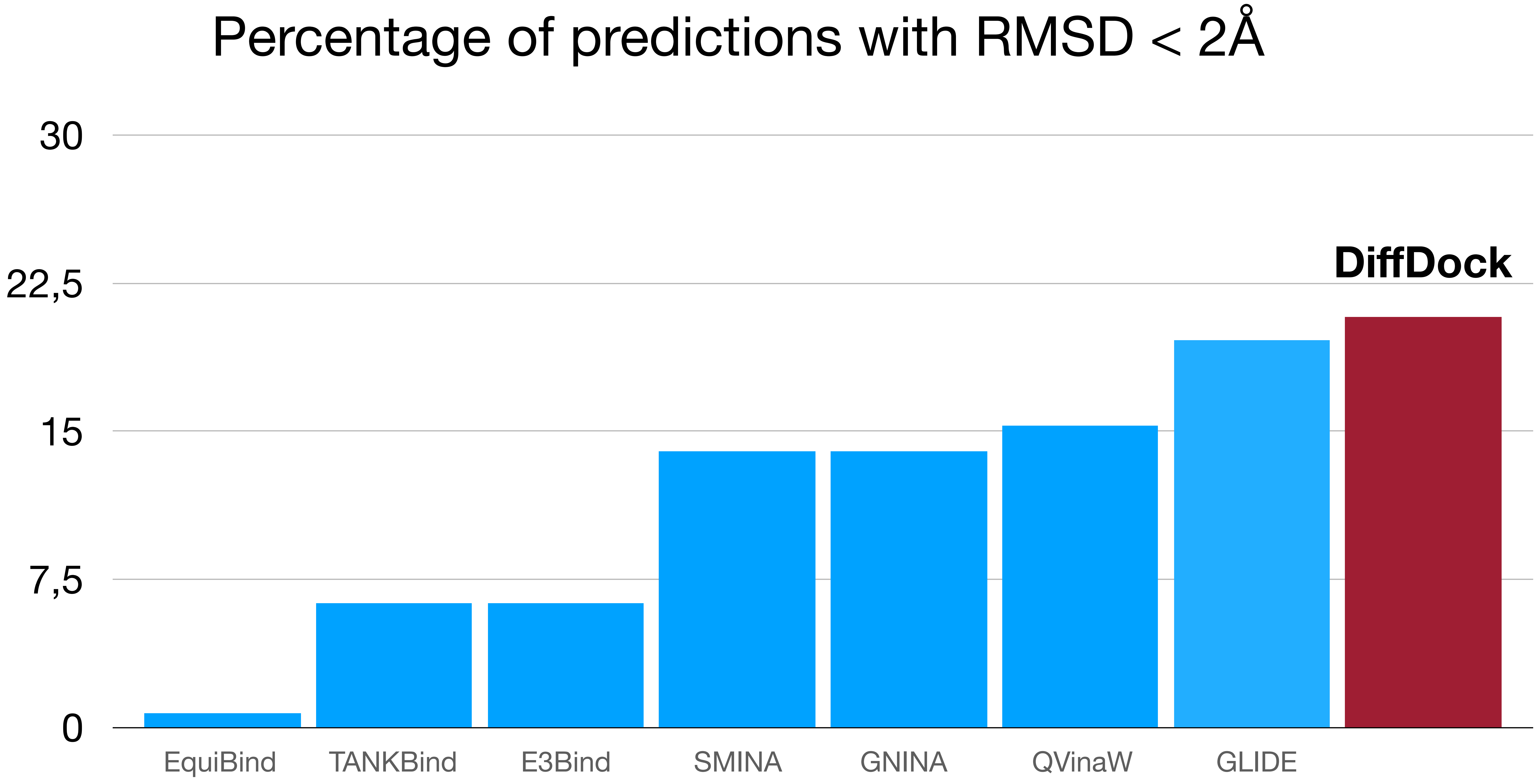
Only 10 steps required for high performance

Performance vs size



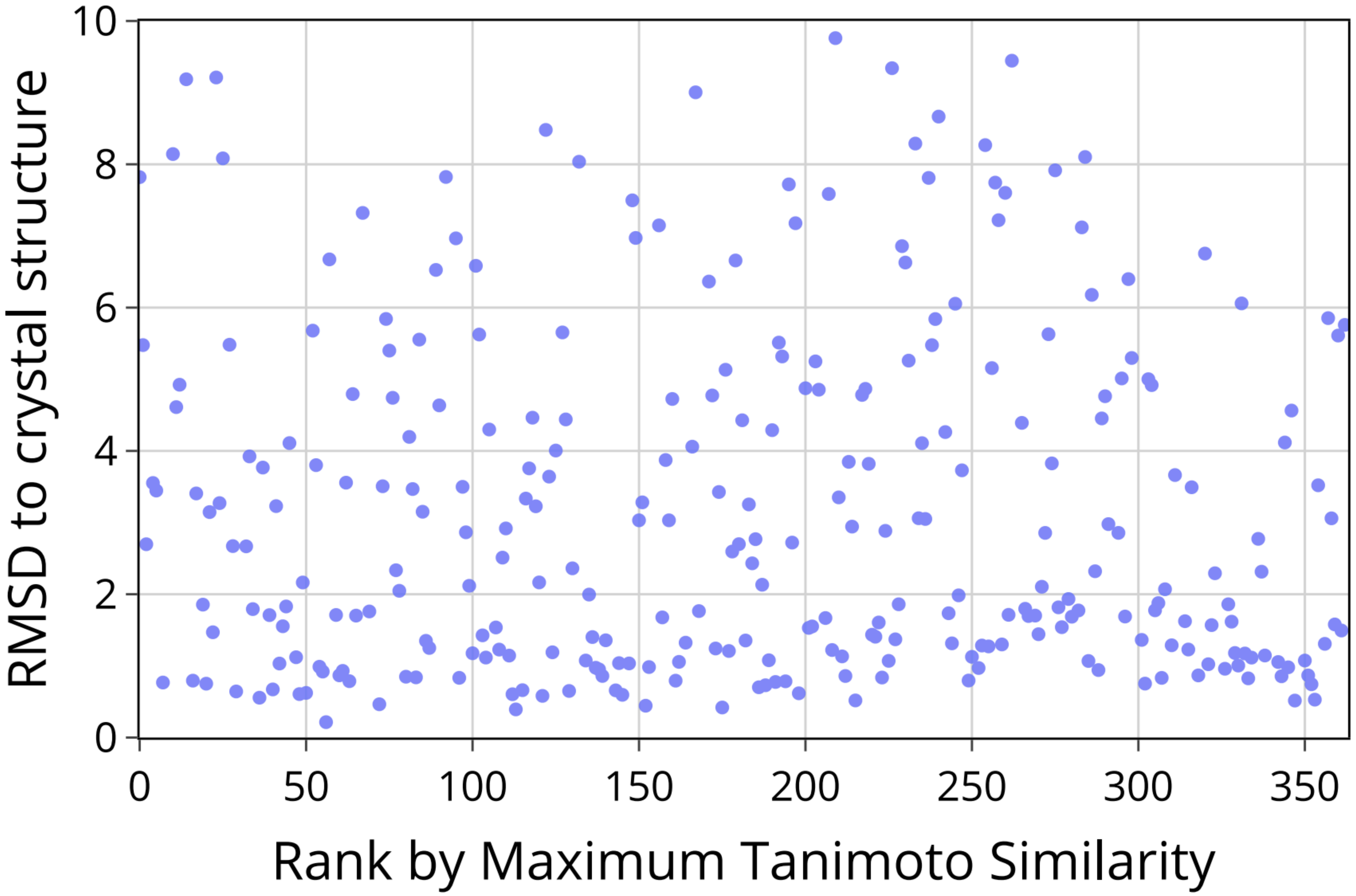


Generalization to unseen receptors

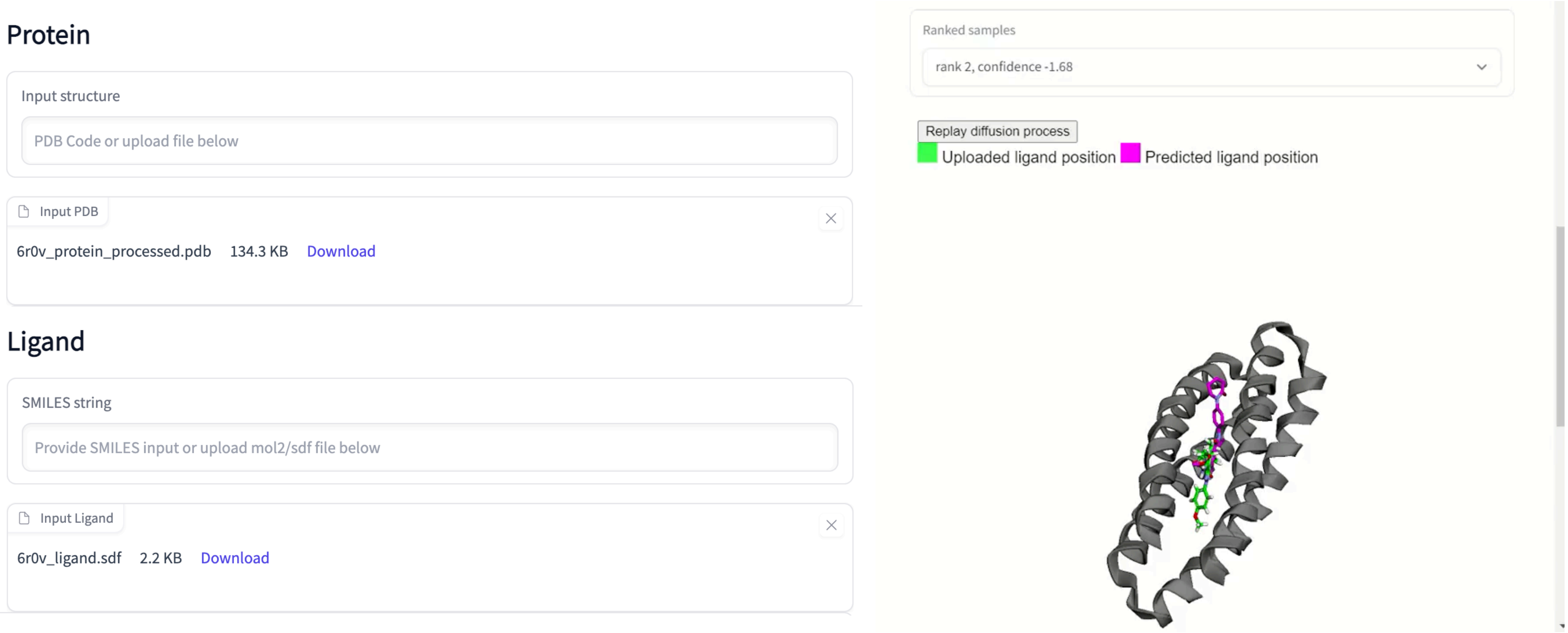


Able to generalize: outperform classical methods

Performance vs similarity



Online Tools: HuggingFace Spaces



Protein

Input structure

PDB Code or upload file below

Input PDB

6r0v_protein_processed.pdb 134.3 KB Download

Ligand

SMILES string

Provide SMILES input or upload mol2/sdf file below

Input Ligand

6r0v_ligand.sdf 2.2 KB Download

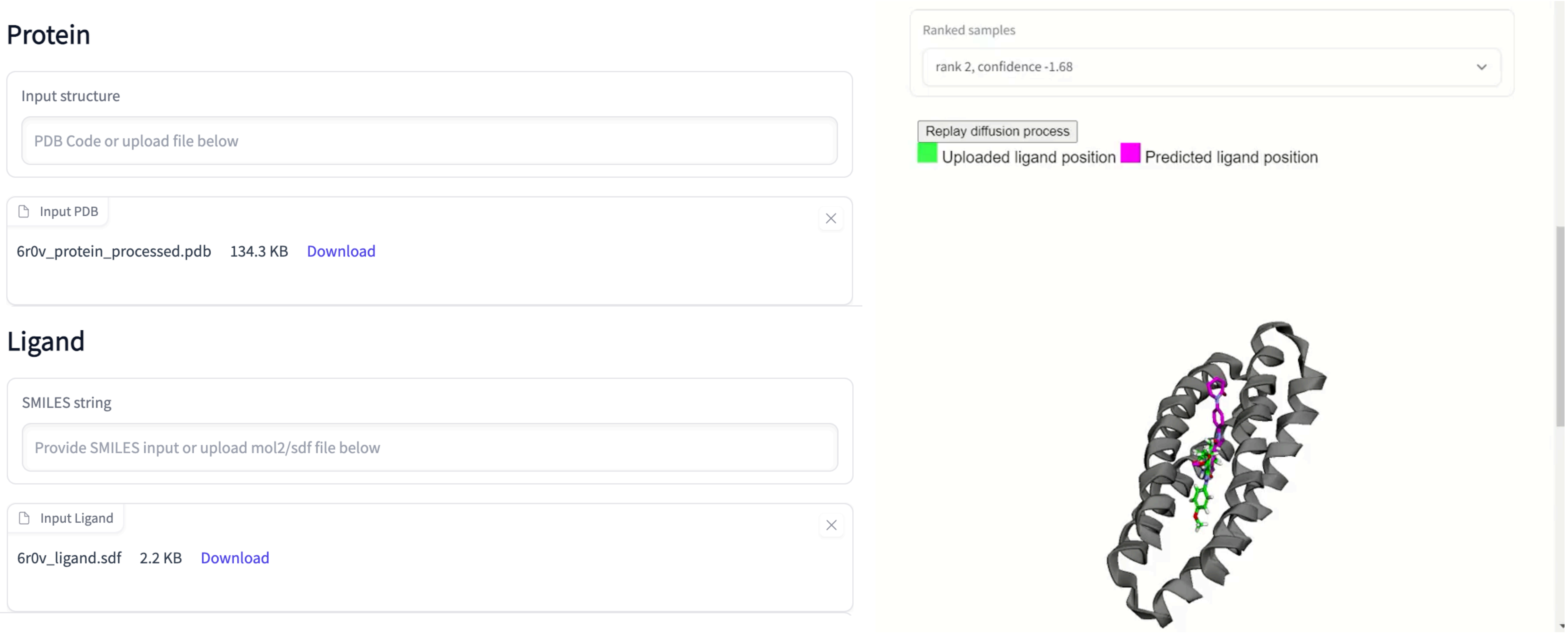
Ranked samples

rank 2, confidence -1.68

Replay diffusion process

🟩 Uploaded ligand position 🟪 Predicted ligand position

Online Tools: HuggingFace Spaces



Protein

Input structure

PDB Code or upload file below

Input PDB

6r0v_protein_processed.pdb 134.3 KB Download

Ligand

SMILES string

Provide SMILES input or upload mol2/sdf file below

Input Ligand

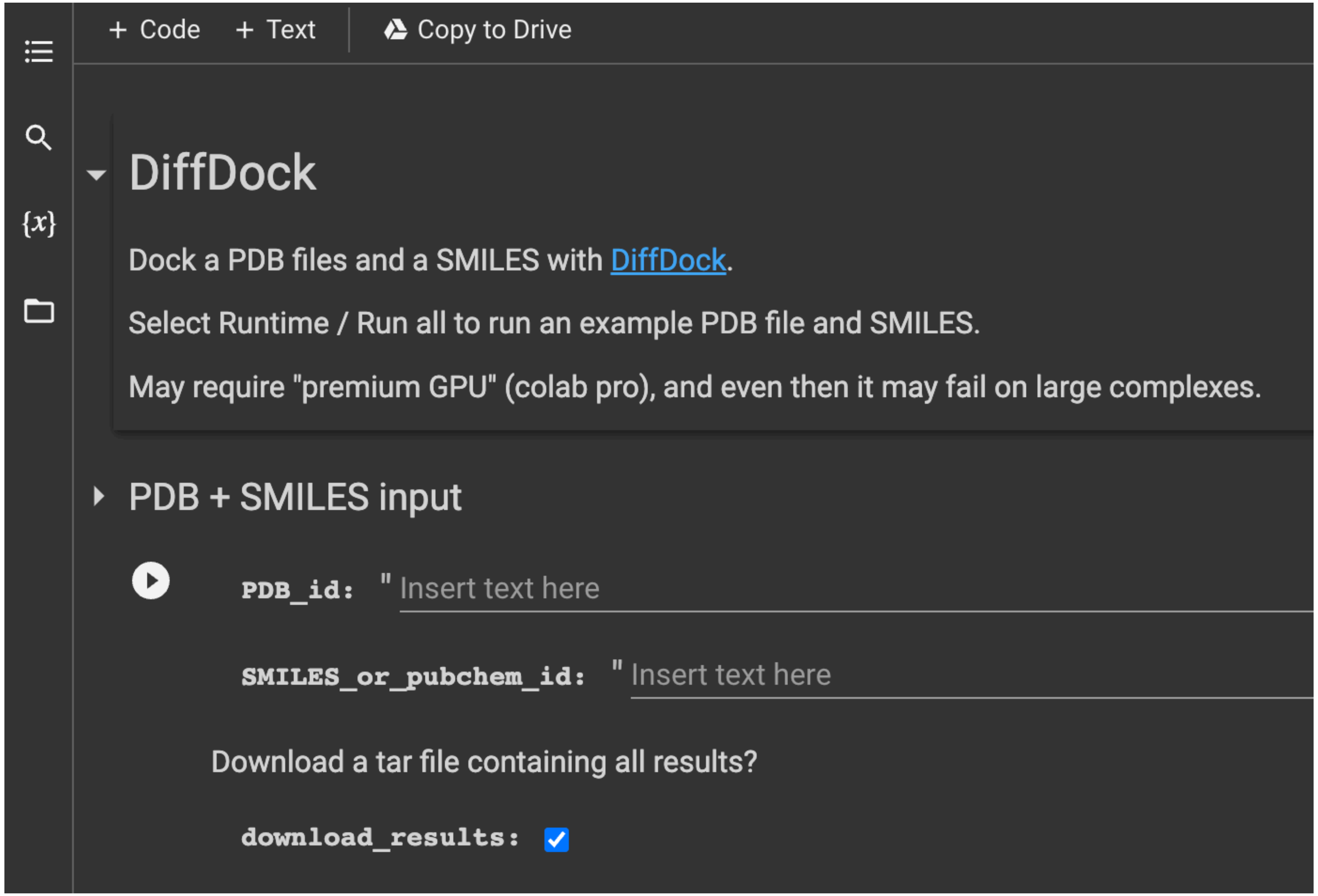
6r0v_ligand.sdf 2.2 KB Download

Ranked samples

rank 2, confidence -1.68

🟩 Uploaded ligand position 🟪 Predicted ligand position

Online Tools: Google Colab



DiffDock

Dock a PDB files and a SMILES with DiffDock.

Select Runtime / Run all to run an example PDB file and SMILES.

May require "premium GPU" (colab pro), and even then it may fail on large complexes.

PDB + SMILES input

PDB_id: " Insert text here

SMILES_or_pubchem_id: " Insert text here

Download a tar file containing all results?

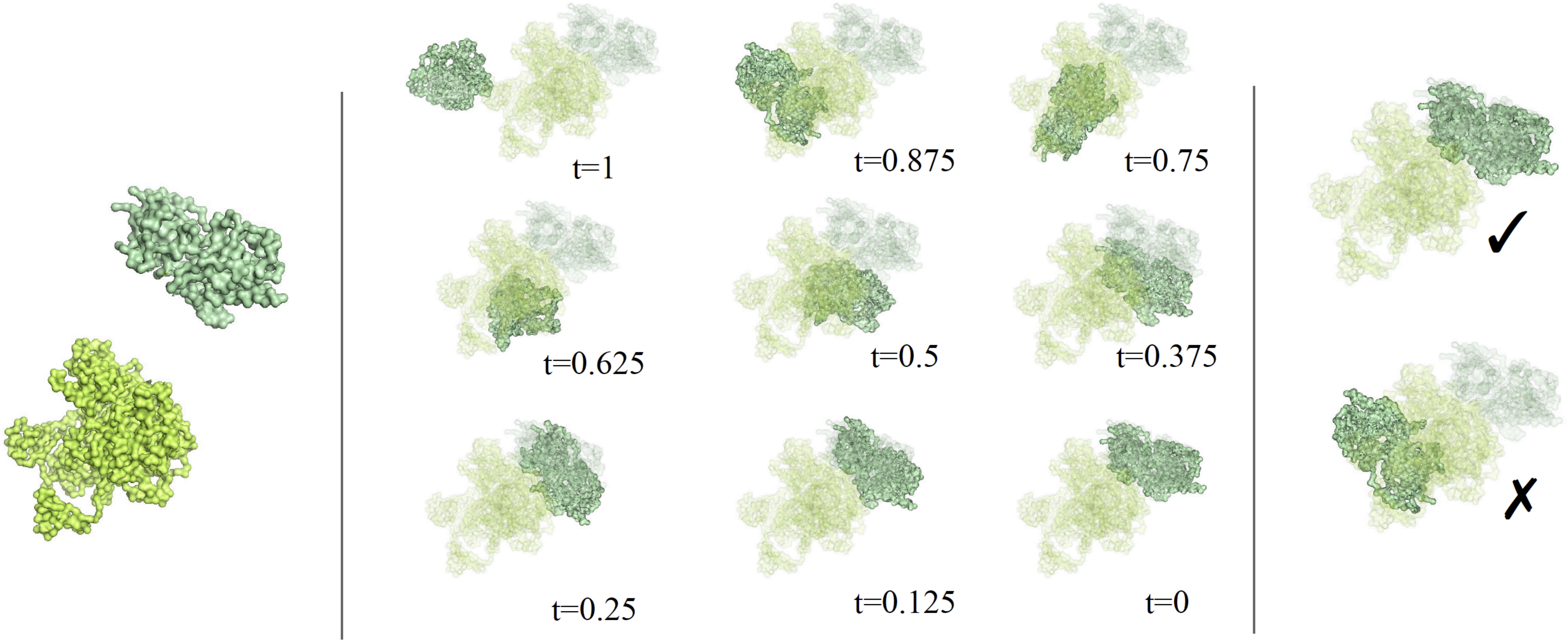
download_results: ☑

Protein-protein docking

Amine Ketata

Cedrik Laue

Ruslan Mammadov



Input: unbound
protein structures

DiffDock-PP
Reverse diffusion process

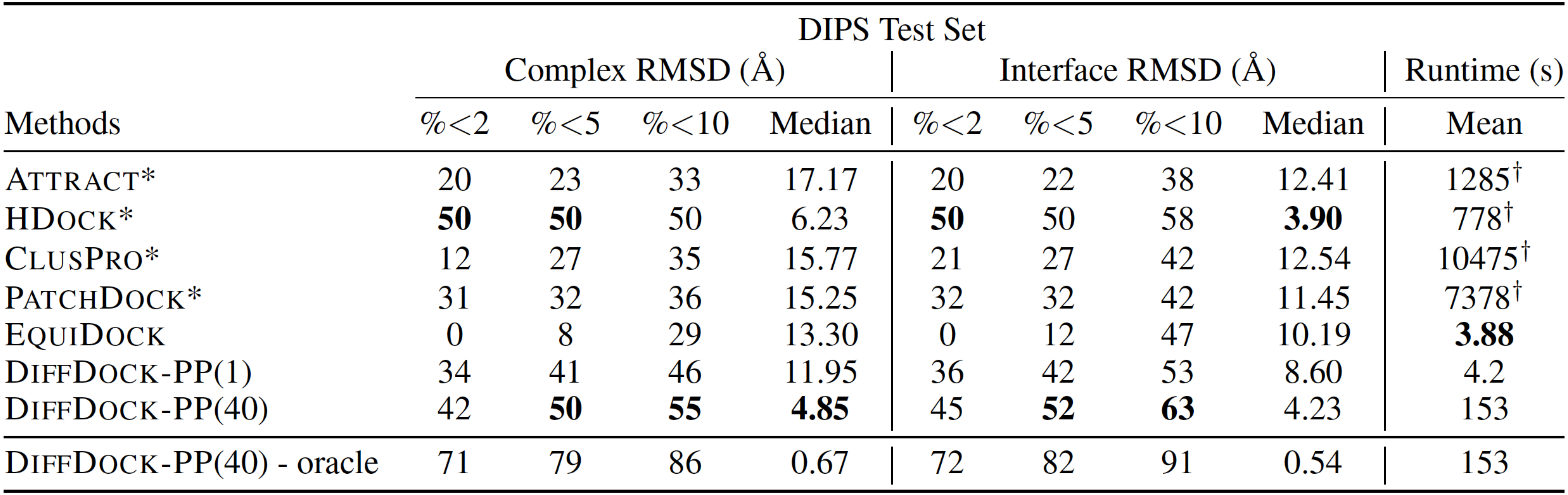
Confidence-based
pose selection

Protein-protein docking

Amine Ketata

Cedrik Laue

Ruslan Mammadov



	DIPS Test Set								
	Complex RMSD (Å)				Interface RMSD (Å)				Runtime (s)
Methods	%<2	%<5	%<10	Median	%<2	%<5	%<10	Median	Mean
Attract*	20	23	33	17.17	20	22	38	12.41	$1285^{\dagger}$
HDock*	**50**	**50**	50	6.23	**50**	50	58	**3.90**	$778^{\dagger}$
ClusPro*	12	27	35	15.77	21	27	42	12.54	$10475^{\dagger}$
PatchDock*	31	32	36	15.25	32	32	42	11.45	$7378^{\dagger}$
EquiDock	0	8	29	13.30	0	12	47	10.19	**3.88**
DiffDock-PP(1)	34	41	46	11.95	36	42	53	8.60	4.2
DiffDock-PP(40)	42	**50**	**55**	**4.85**	45	**52**	**63**	4.23	153
DiffDock-PP(40) - oracle	71	79	86	0.67	72	82	91	0.54	153