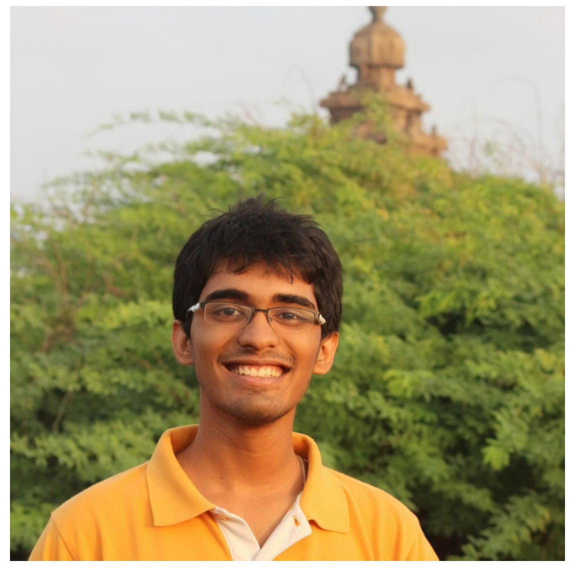
FlexDock

Composing Unbalanced Flows for Flexible Docking and Relaxation

*Gabriele Corso**

*Vignesh Ram Somnath**



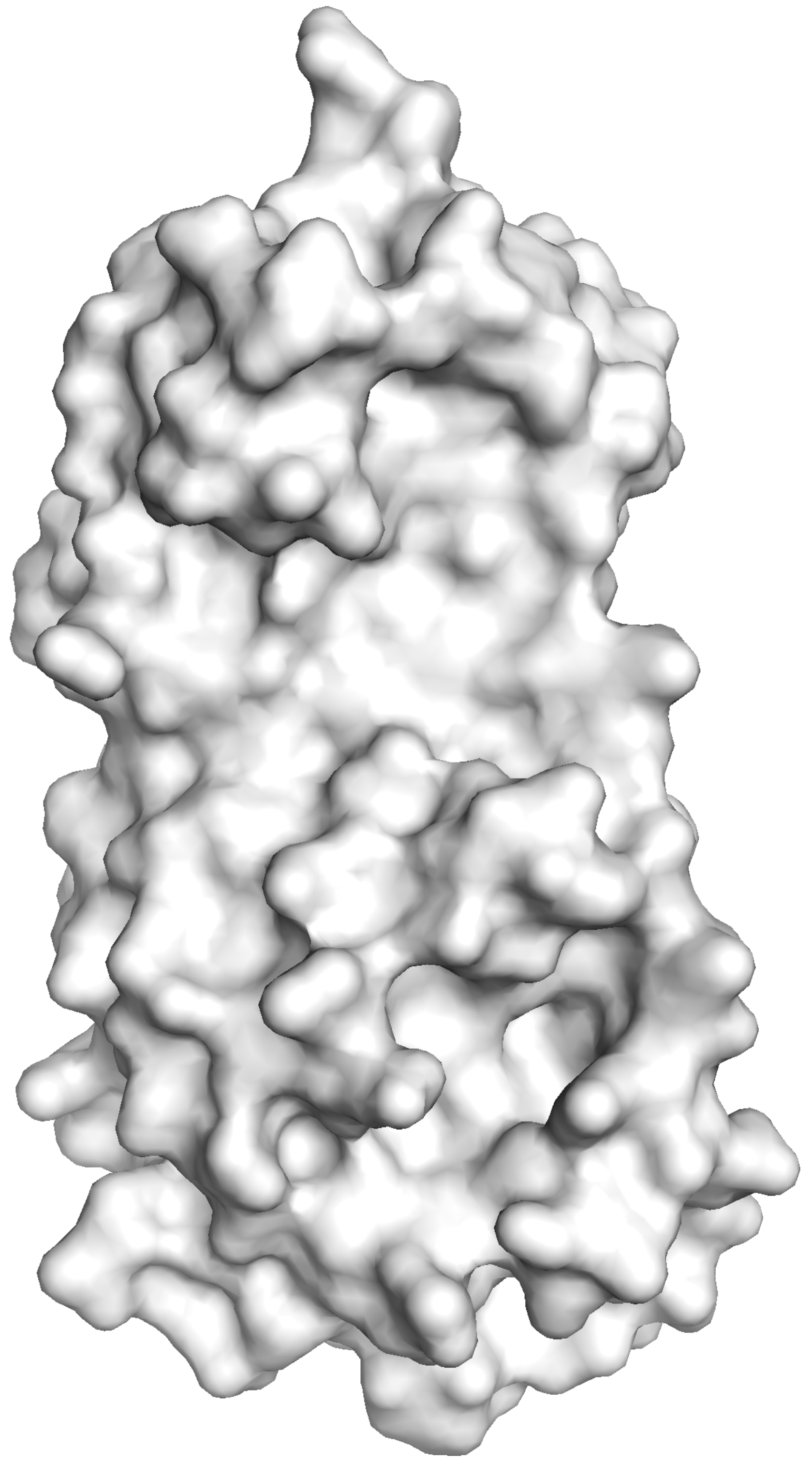
*Noah Getz**

Regina Barzilay

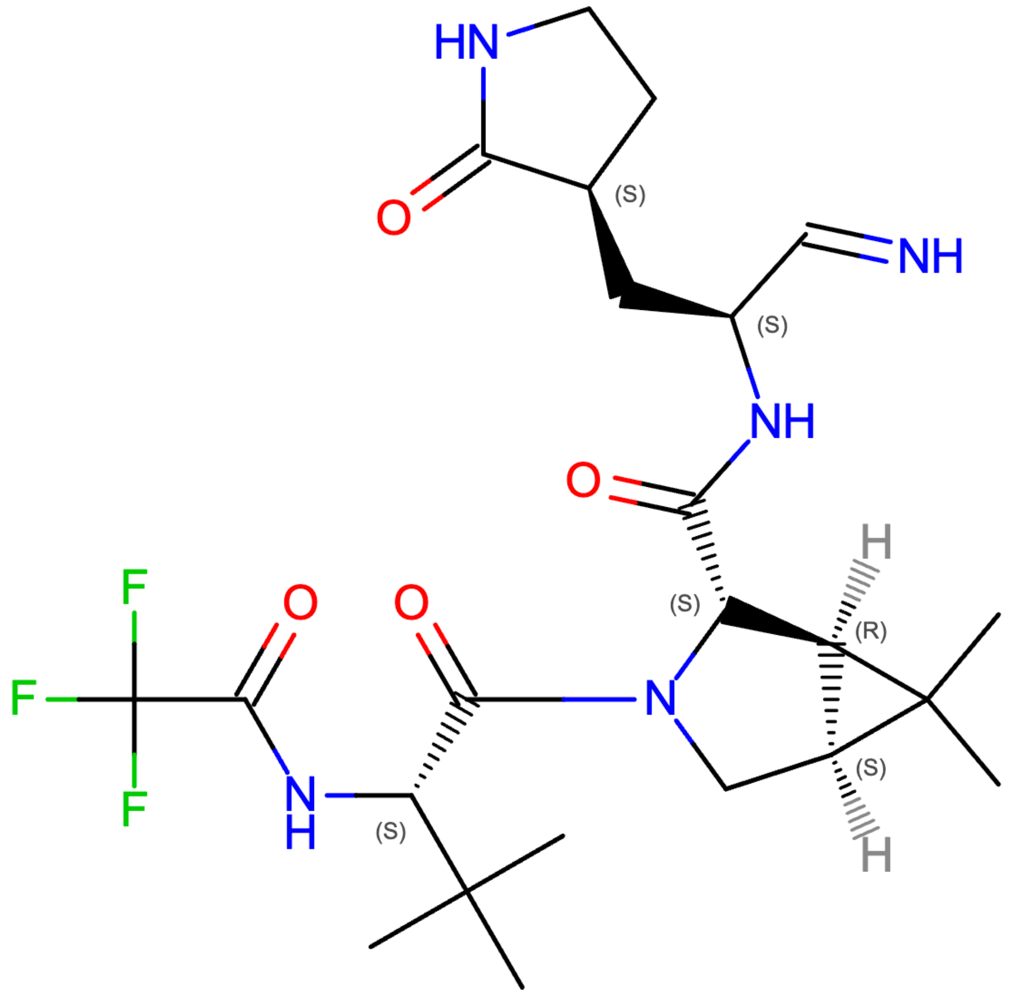
Tommi Jaakkola

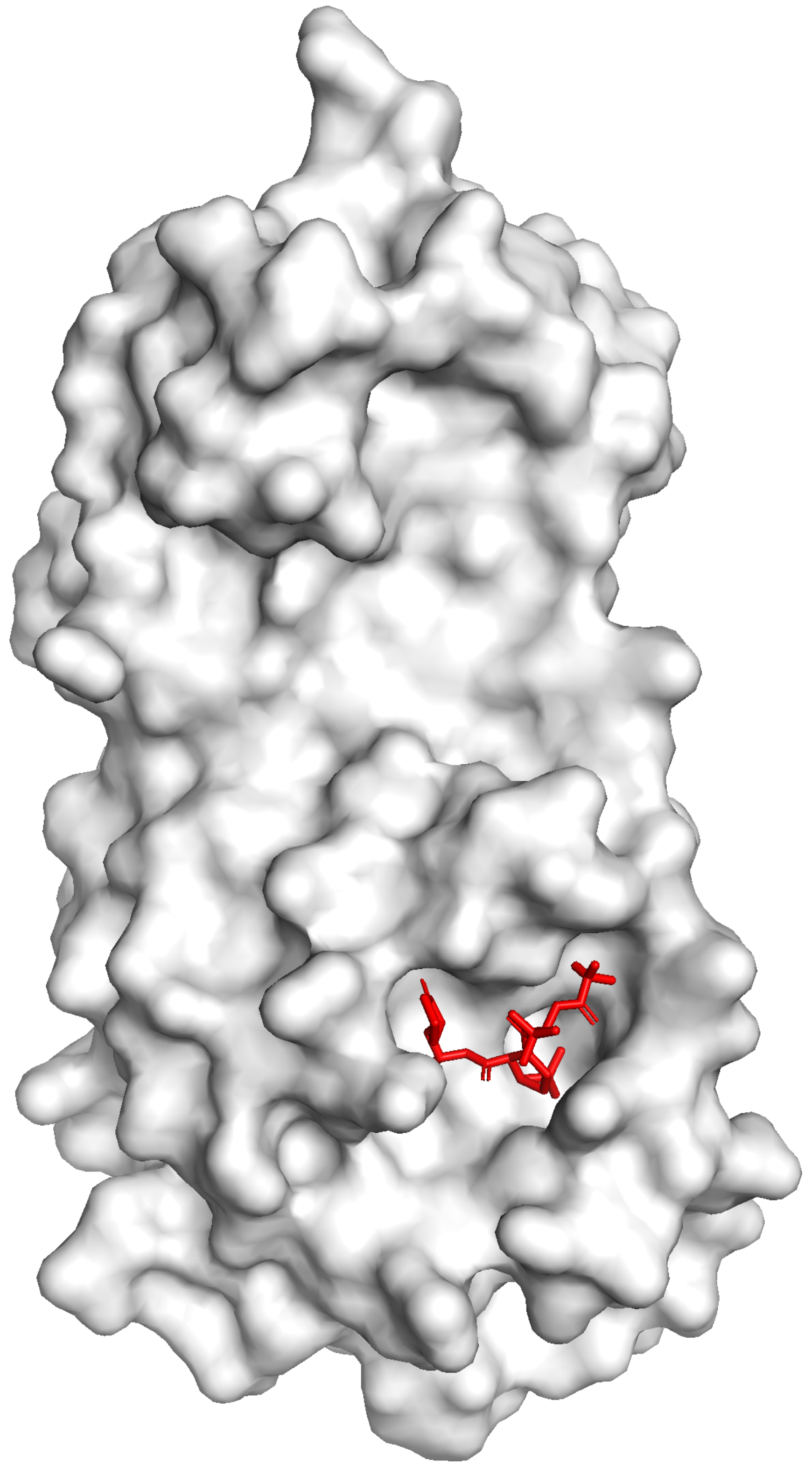
Andreas Krause

Protein-Ligand Docking



+





Input: protein structure + molecule **Output: bound structure**

Generative Models for Docking

- Model the inherent epistemic and aleatoric uncertainty associated with the docking problem

Generative Models for Docking

- Model the inherent epistemic and aleatoric uncertainty associated with the docking problem

Main Drawbacks:

- Typically assume the proteins have a fixed structure
- Generate poses that fails one or more physical plausibility checks

Addressed in this Work

Accounting for Protein Flexibility

Co-Folding: Predict the bound structure of protein and small molecule from scratch

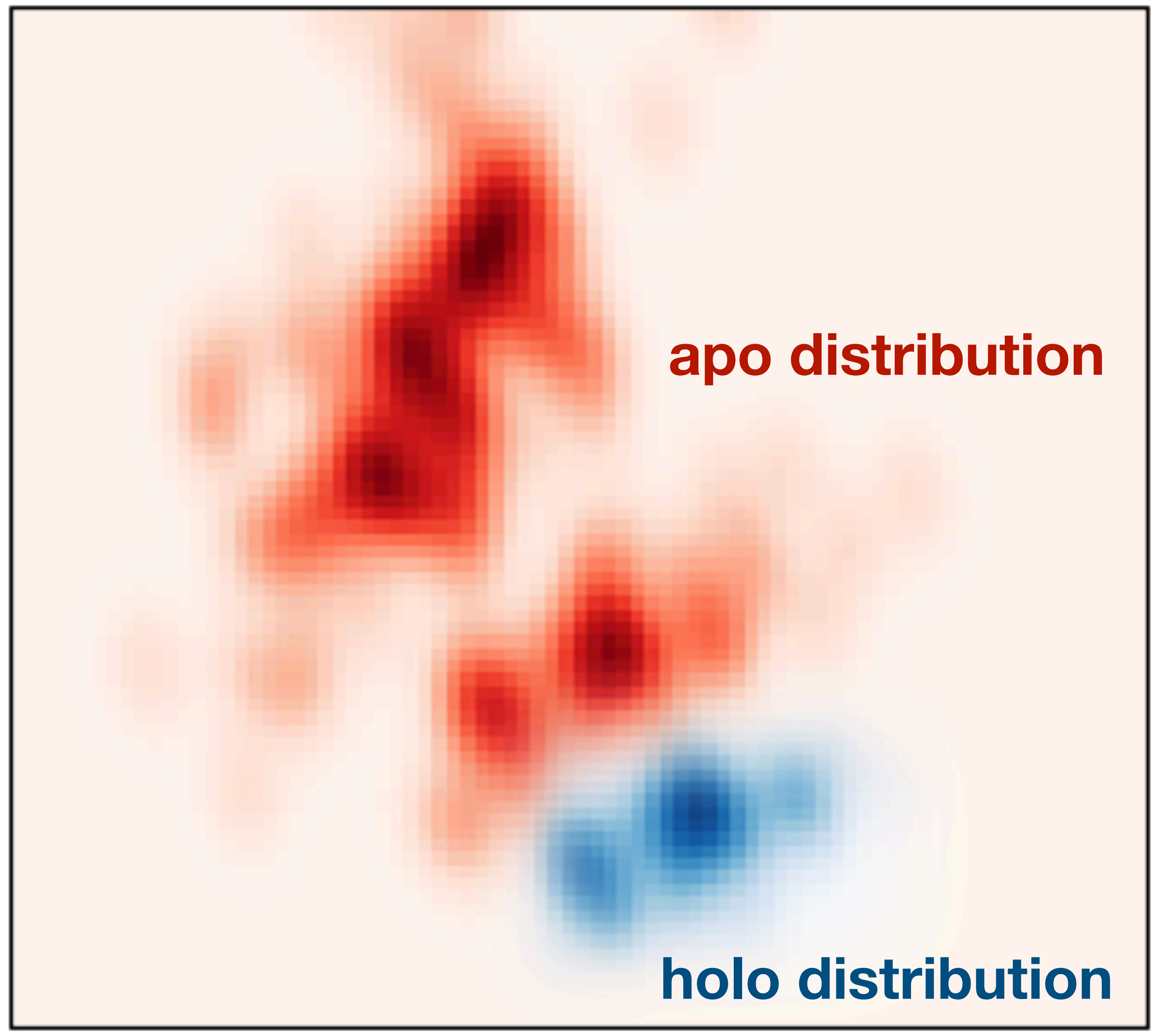
- Model has to largely re-learn protein folding, with consequently slow inference

Flexible Docking: Model the limited structural transformation between unbound & bound proteins

- Search-based methods struggle to efficiently account for additional degrees of freedom
- Diffusion models need to refold local pockets entirely, with poor accuracy and non-physical poses

Generative Modeling for Flexible Docking

We 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

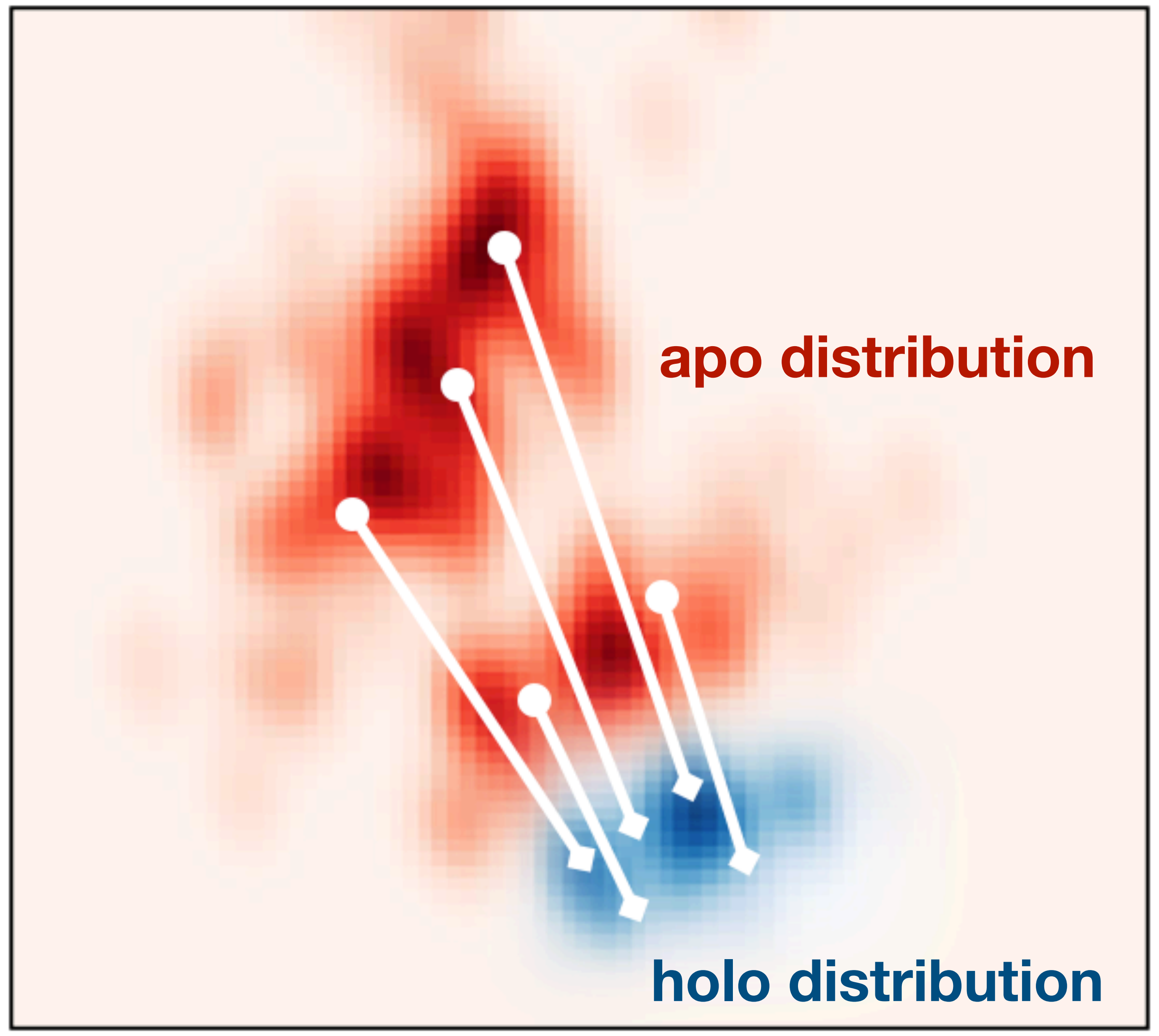
FM Sampling process

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

FM Objective

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

where q has marginals q_0 and q_1.



Flow Matching

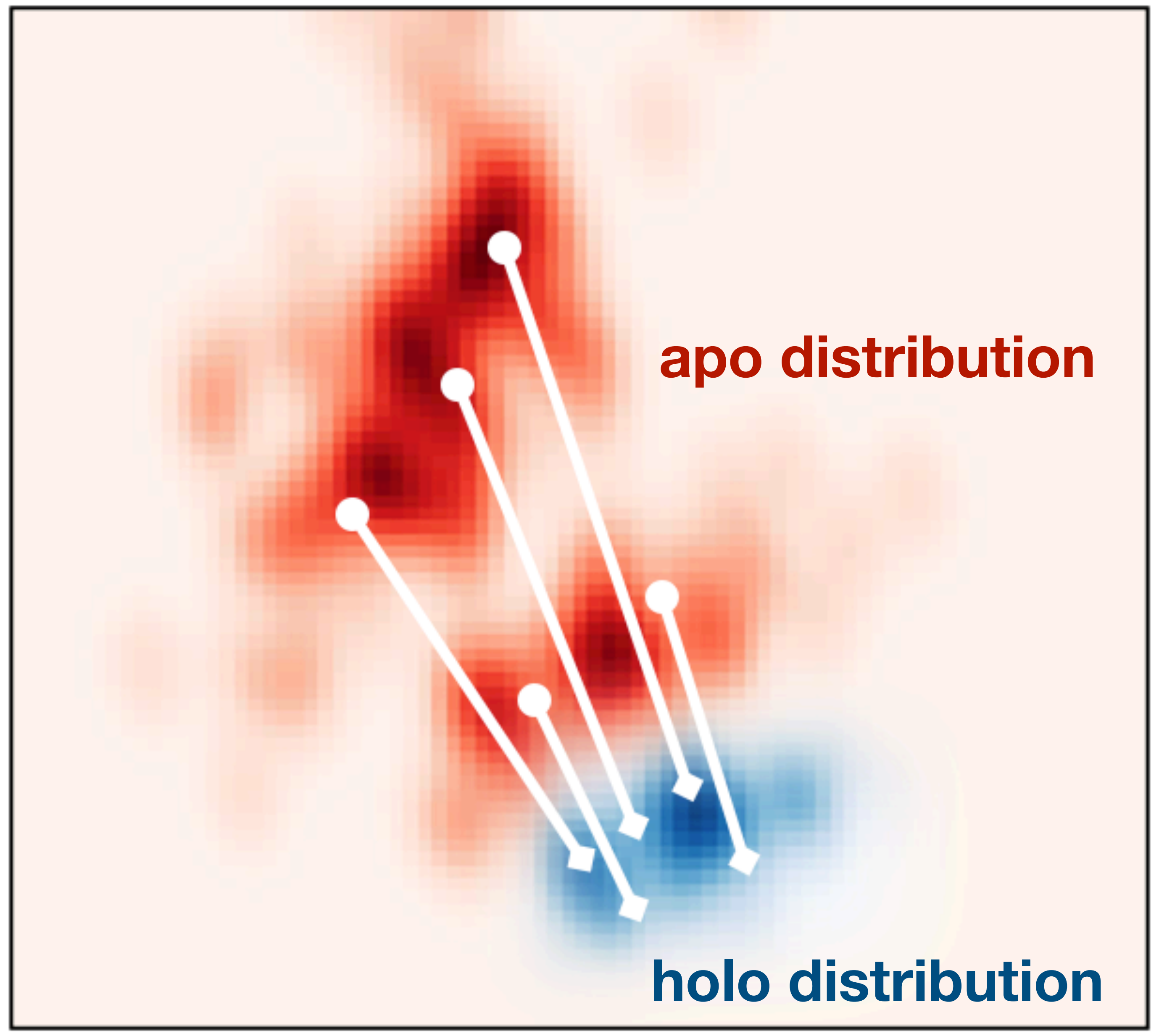
FM Sampling process

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

FM Objective

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

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

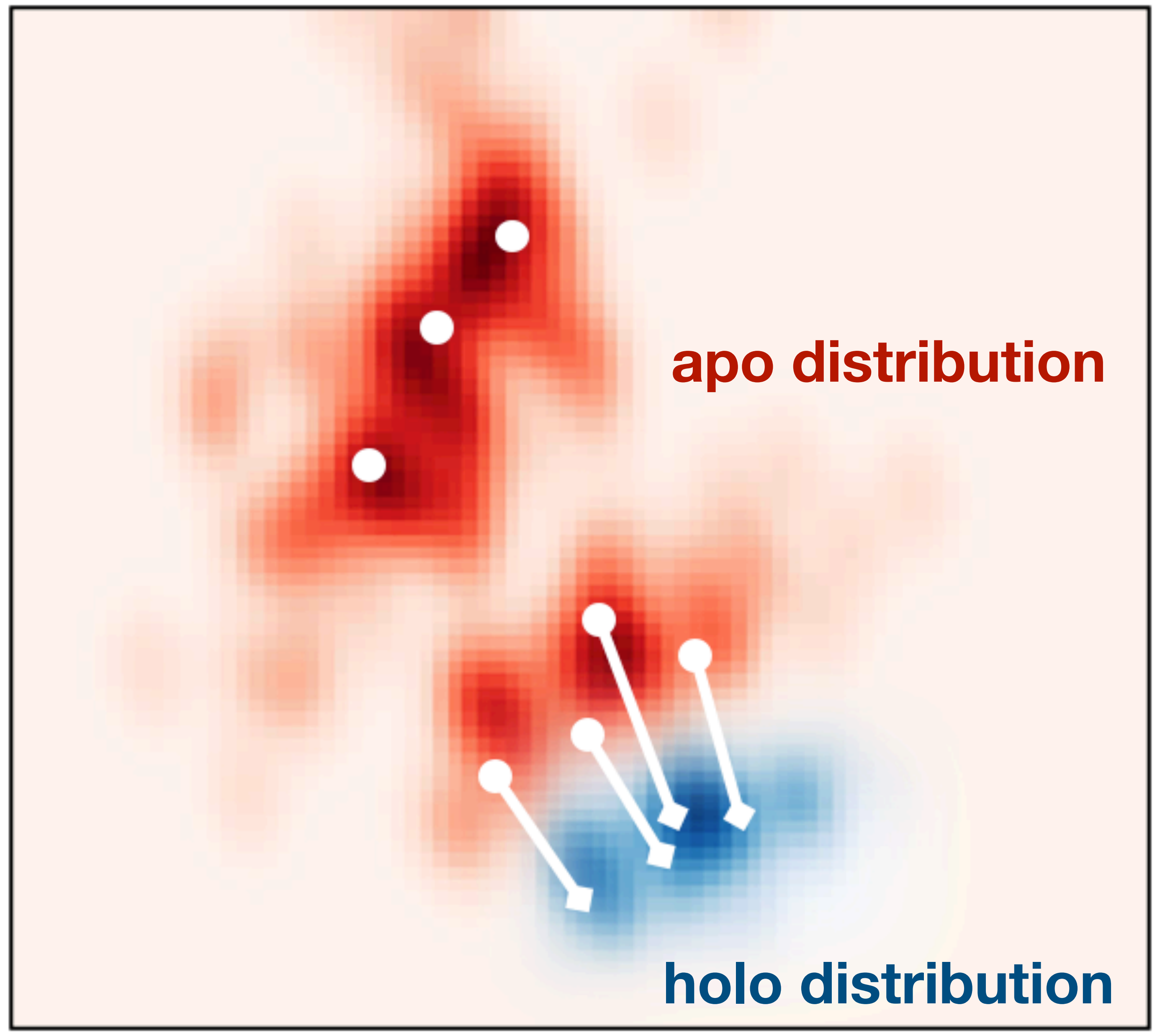
Idea: relaxing marginal preservation condition of flow matching we can define much easier transport problems

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

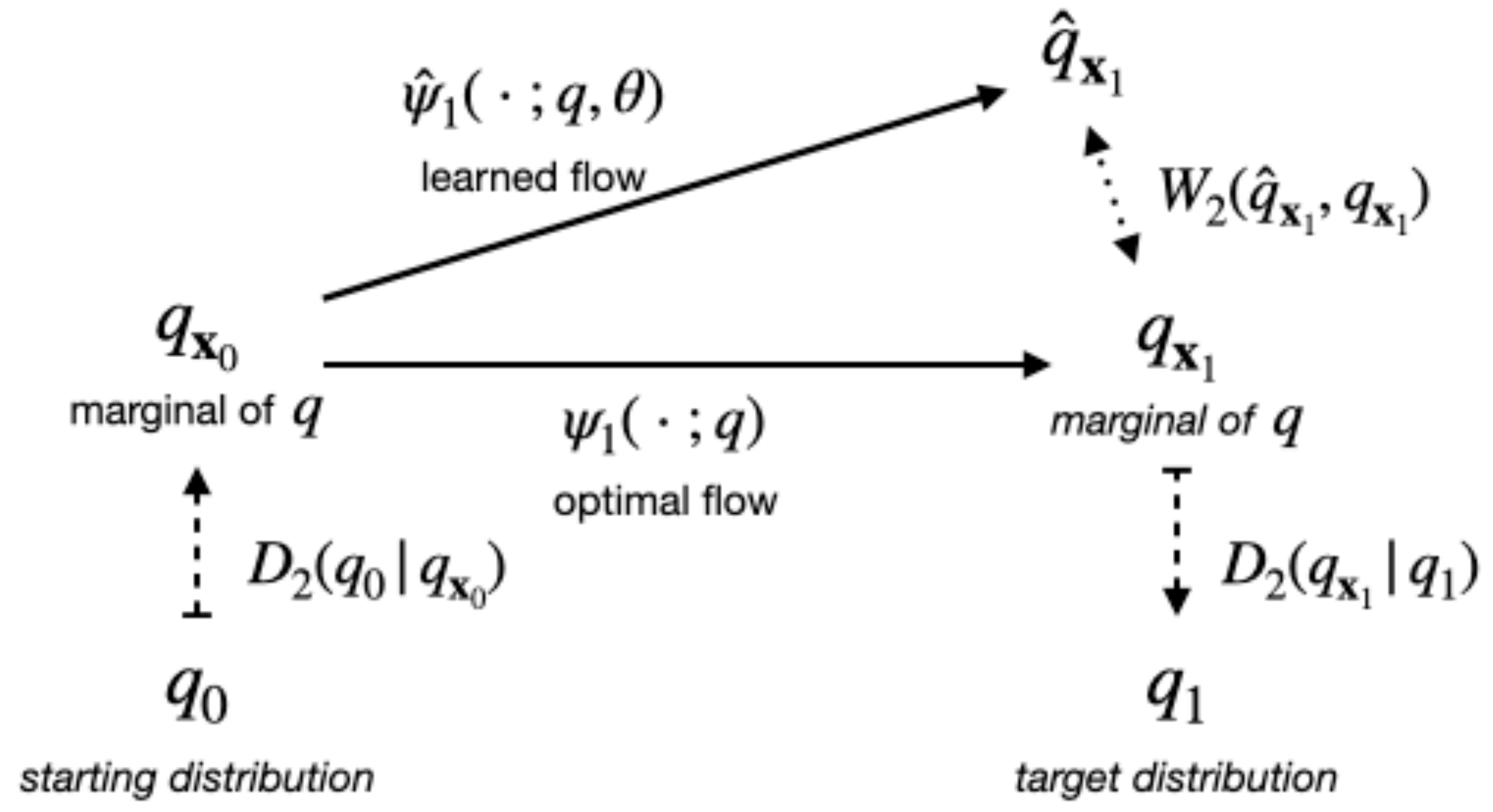


Unbalanced FM Objective

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

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

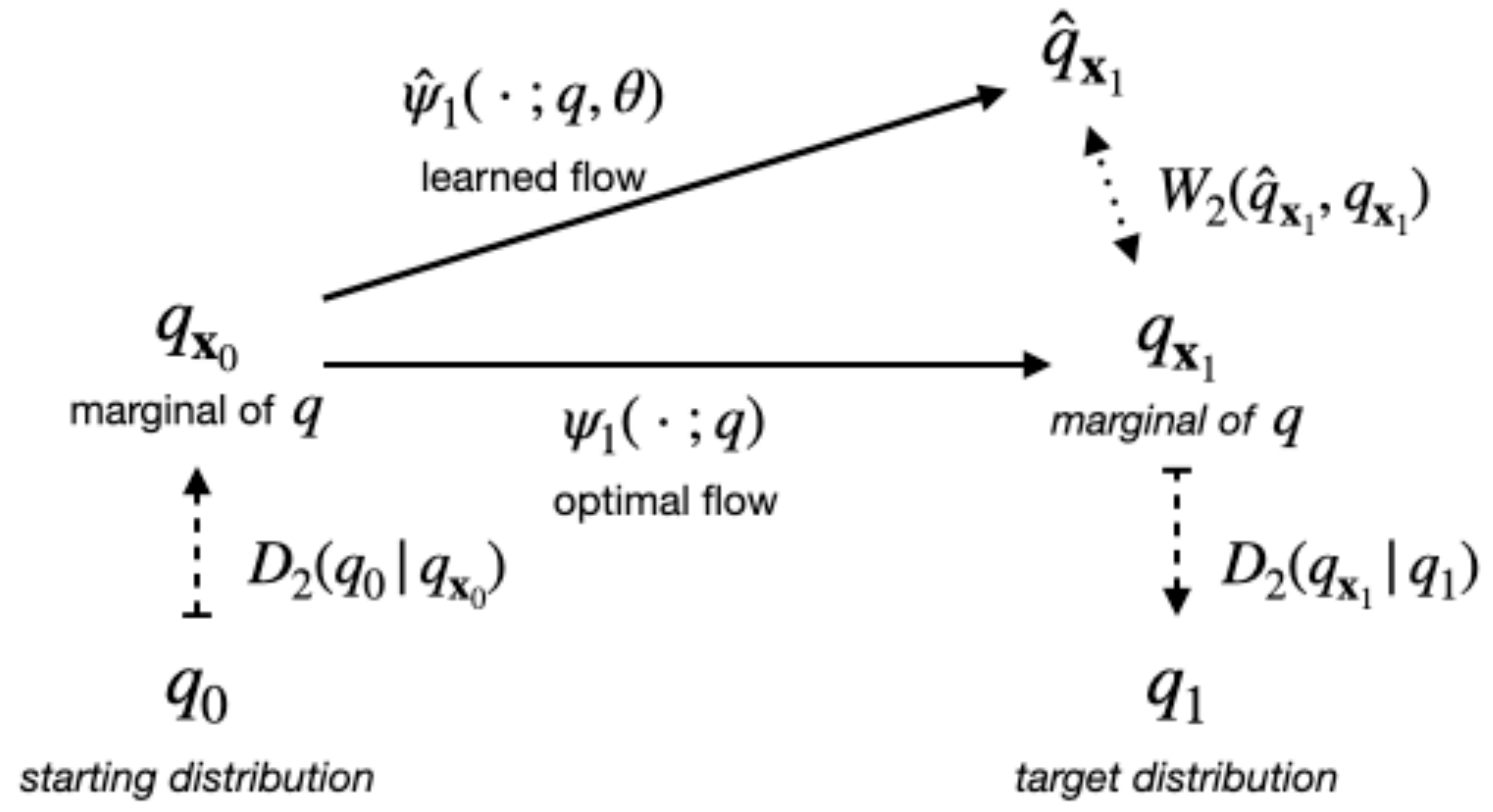
Efficiency vs Approximation



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



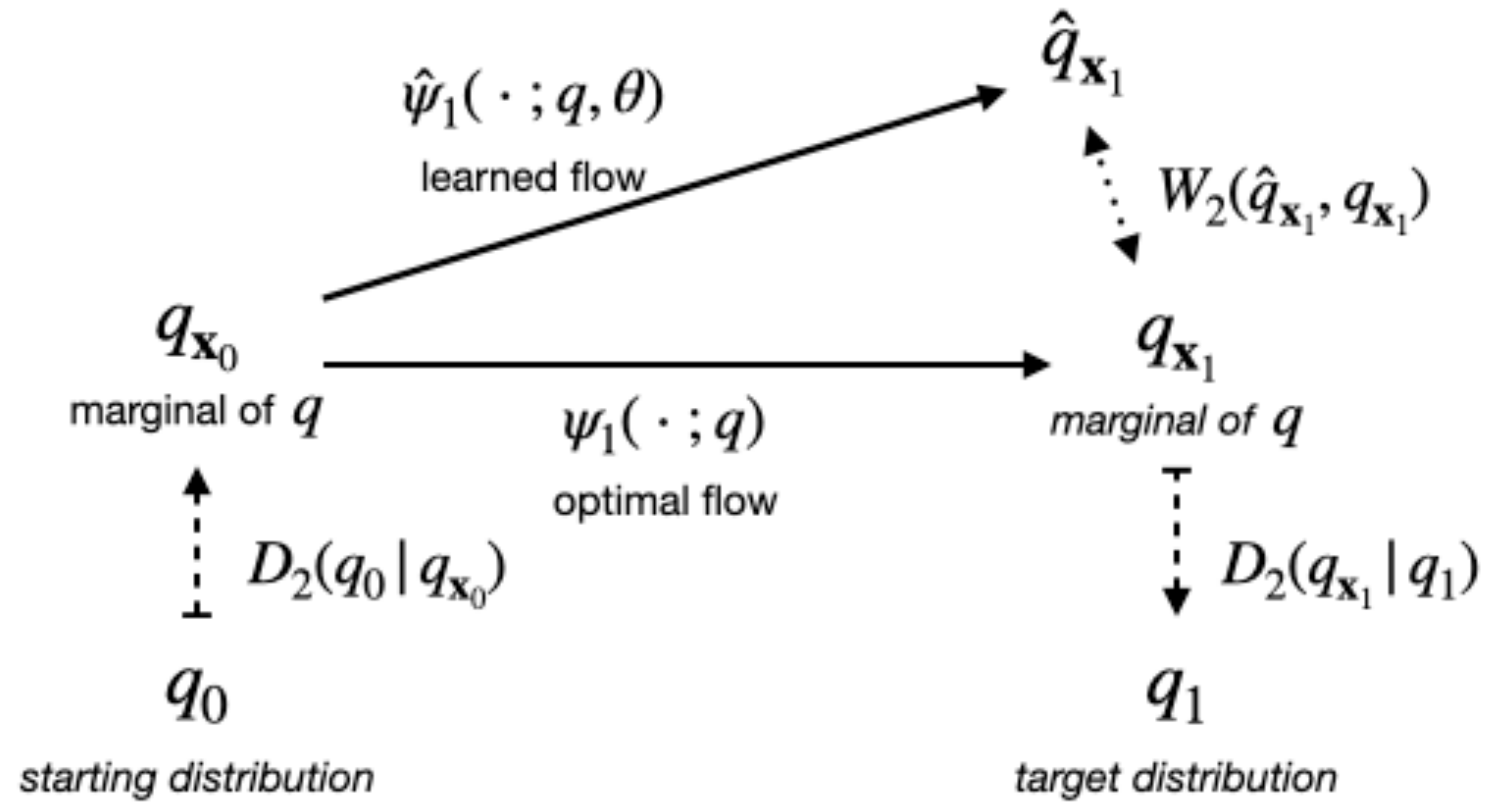
We can show that the UFM objective is a bound on the approximation error vs sampling efficiency tradeoff.

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

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

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

Efficiency vs Approximation



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

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

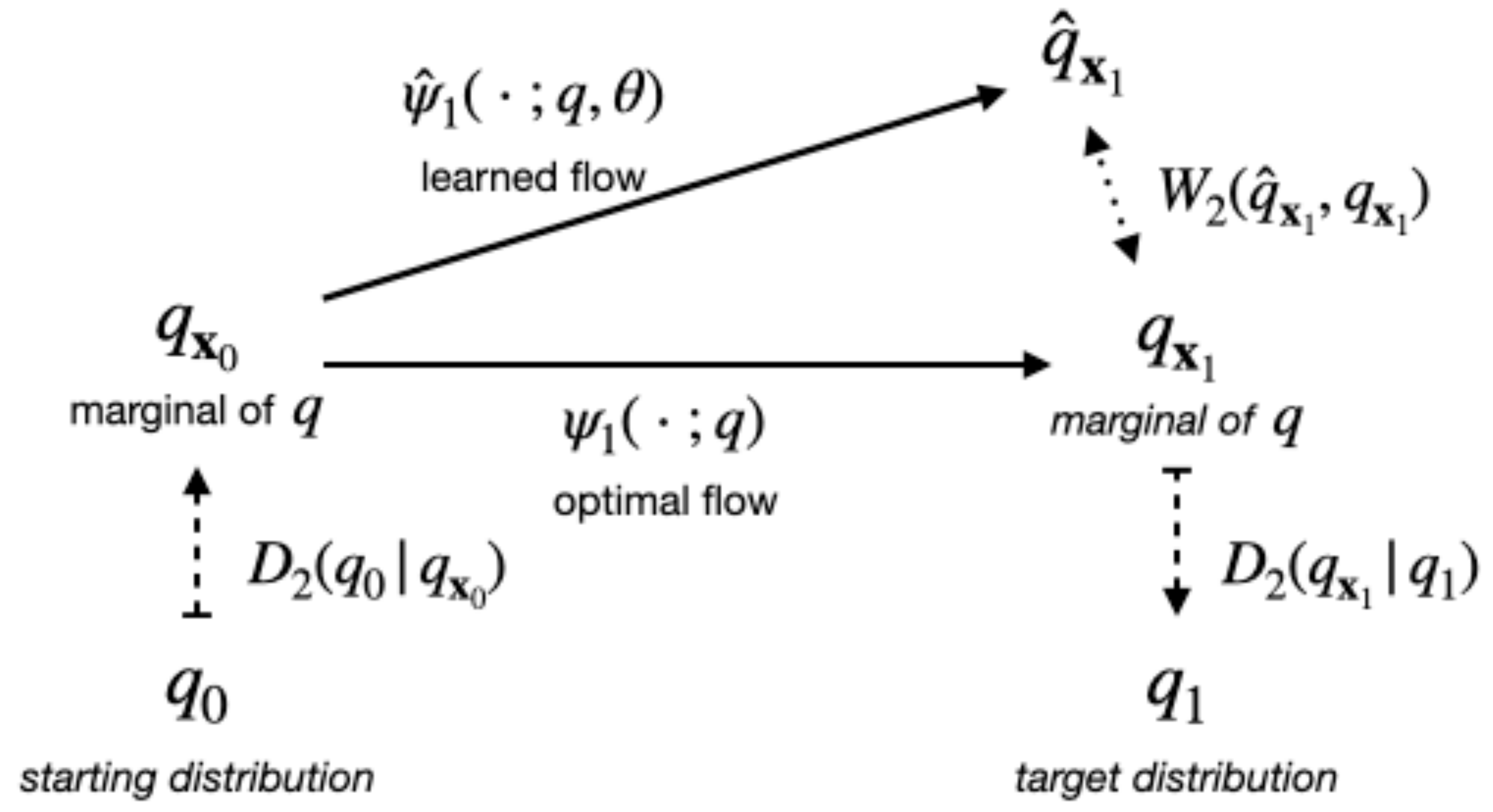
Proposition (Benton et al., 2023): under appropriate assumptions the approximation error of the learned flow is bounded by FM objective:

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

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.

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

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

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

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:

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

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

Optimizing the Objective

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

Optimizing the Objective

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

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

Optimizing the Objective

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

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

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

Optimizing the Objective

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

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

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

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

Optimizing the Objective

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

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

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

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

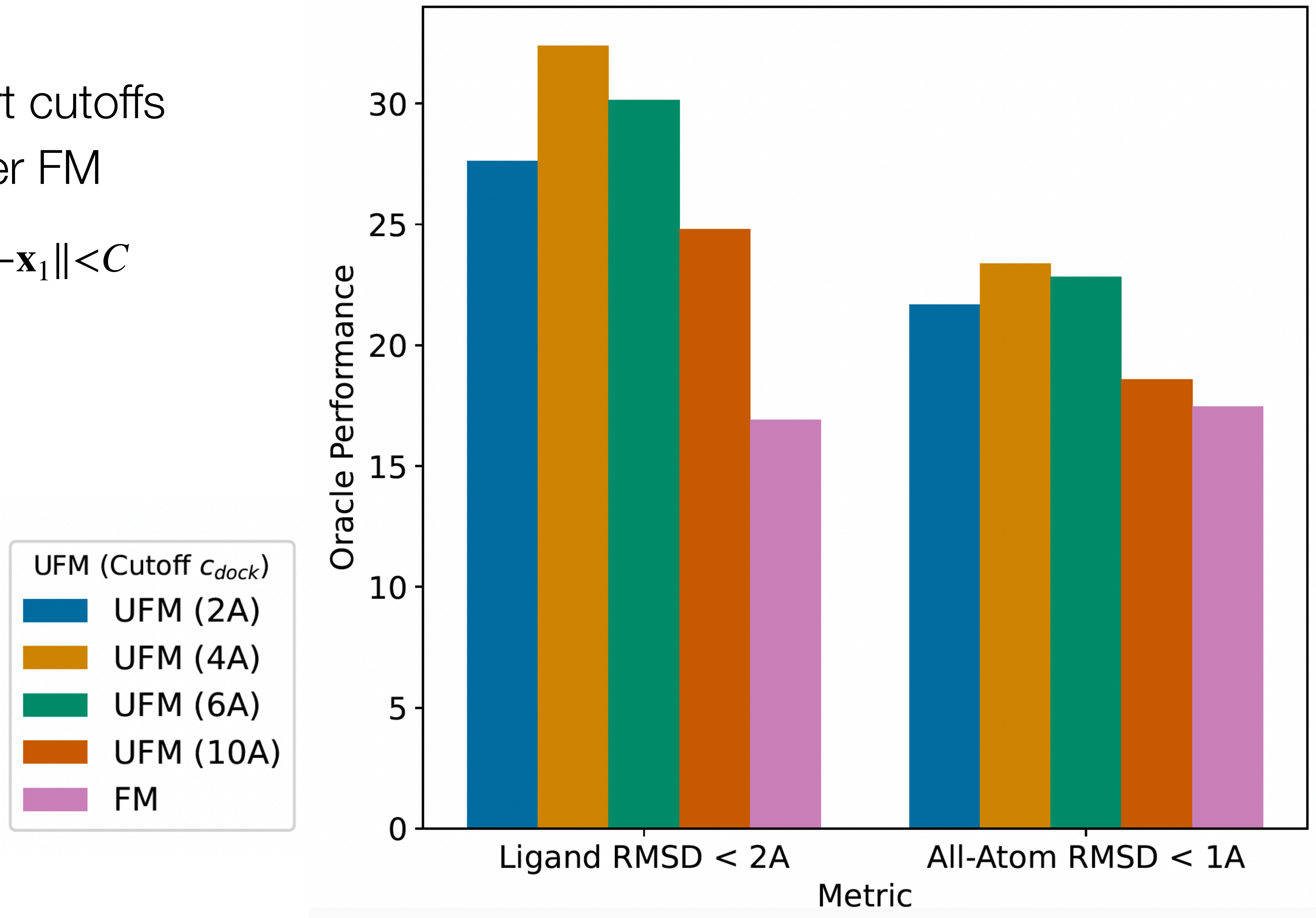
Then, given q, the UFM objective boils down to Flow Matching:

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

Unbalanced FM vs FM

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

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



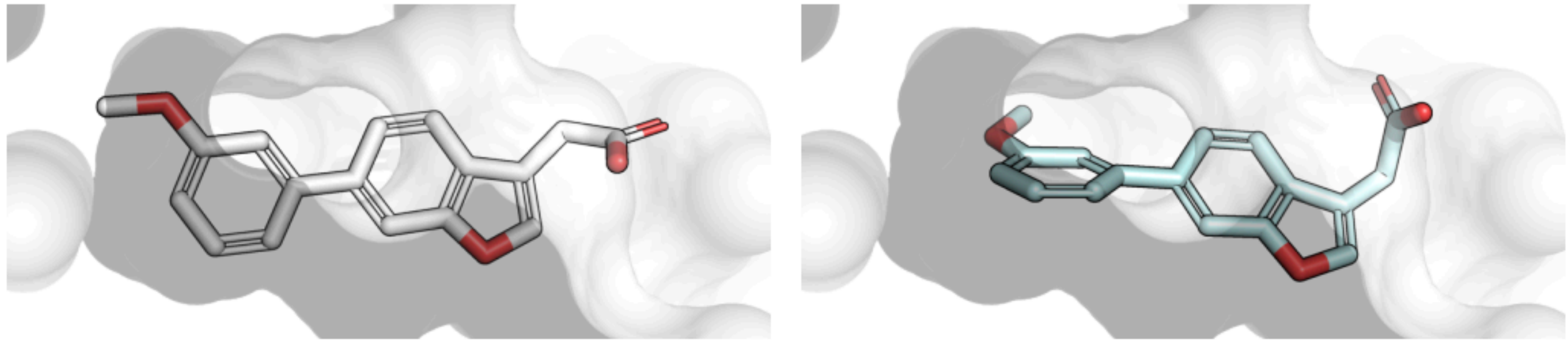
Pose relaxation

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.

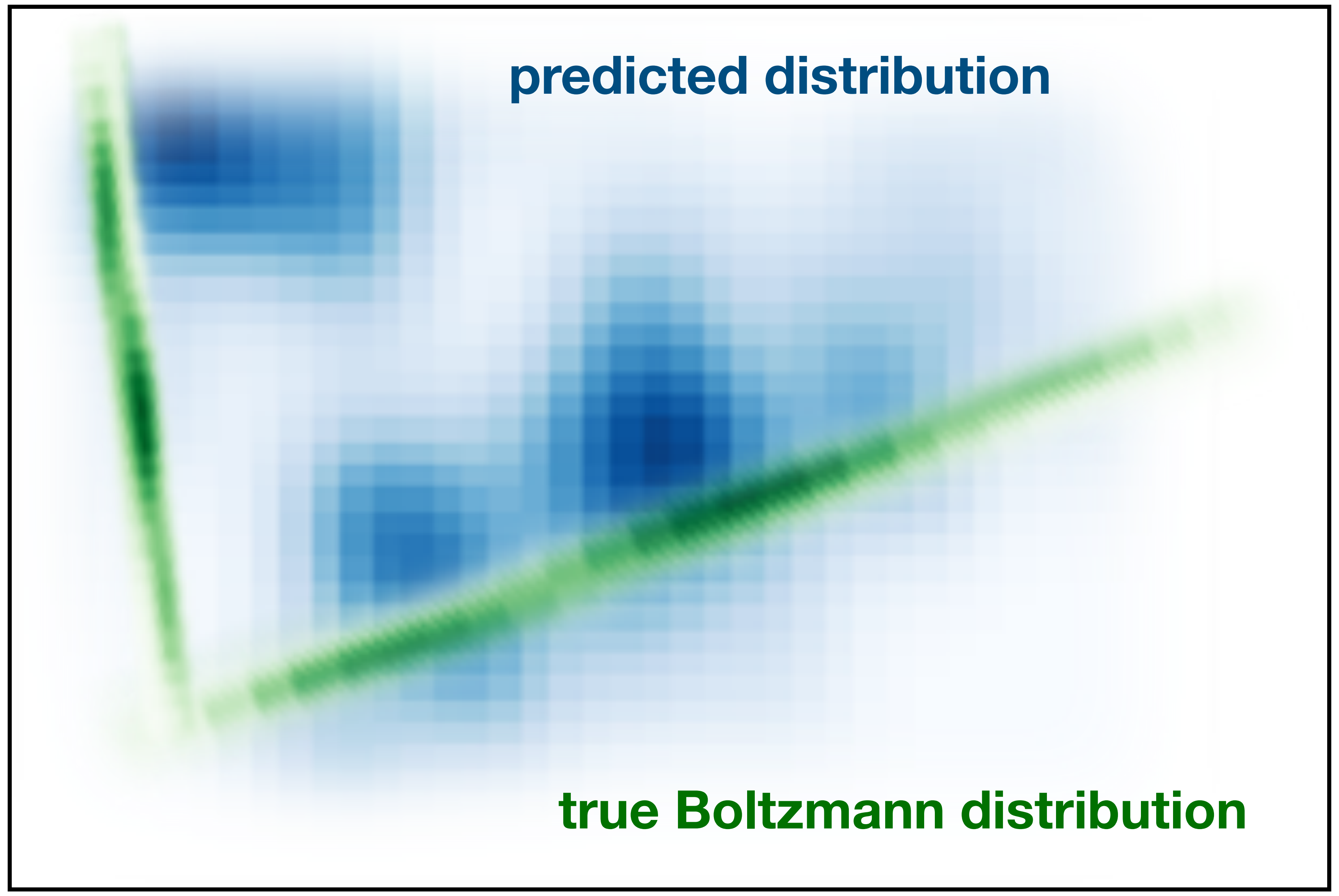
Pose relaxation: refine the structural conformation to find a more energetically favorable

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‡



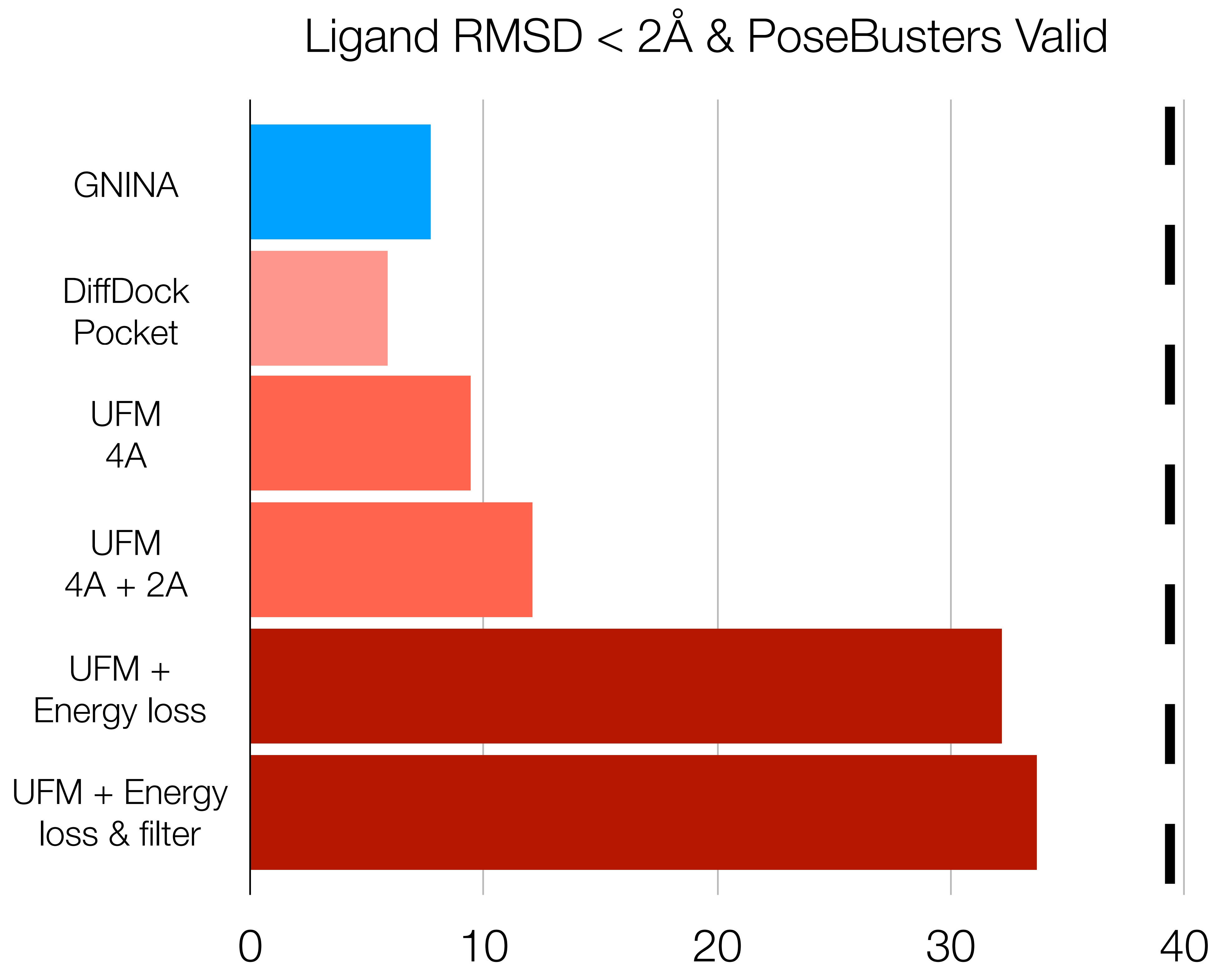
(h) Clash with protein. DiffDock prediction for ligand XQ1 of protein-ligand complex 7L7C. RMSD 1.6 Å.



Pose relaxation with Unbalanced FM

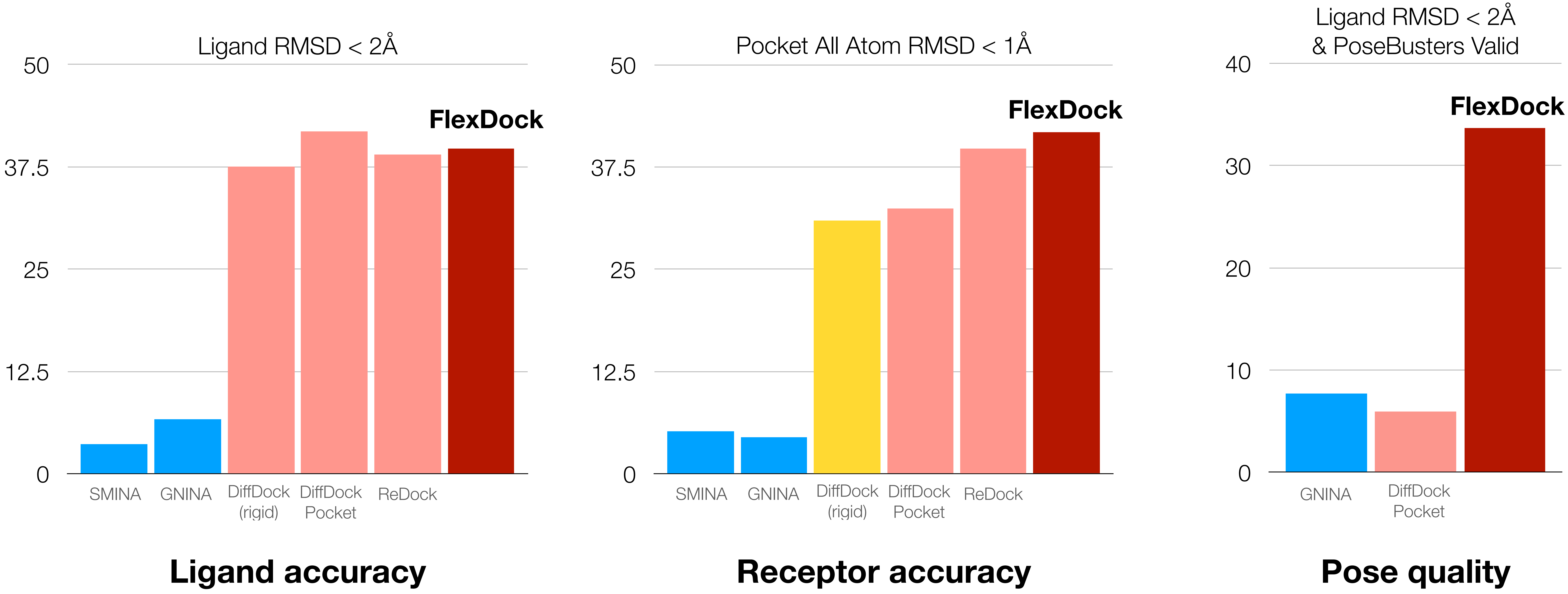
Applying “vanilla” Unbalanced FM but with a smaller cutoff

To incentivize the model to preserve physicality also in very narrow degrees of freedom we add an energy loss



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

Pocket-based Flexible Docking



ETH zürich

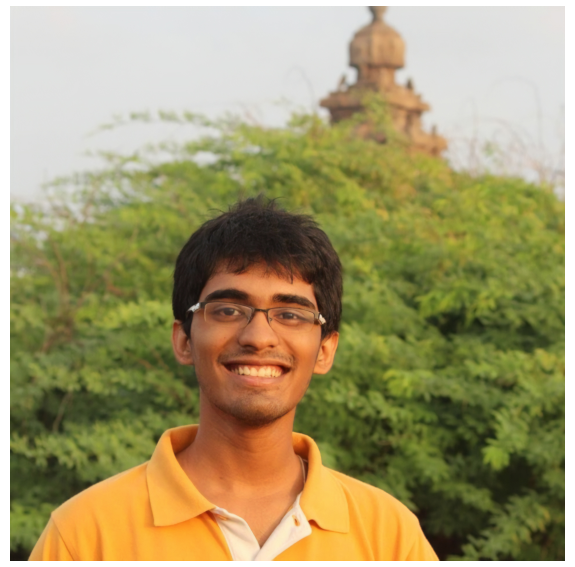
JAMEEL CLINIC

MIT CSAIL

FlexDock

Composing Unbalanced Flows for Flexible Docking and Relaxation

*Gabriele Corso** *Vignesh Ram Somnath** *Noah Getz** *Regina Barzilay* *Tommi Jaakkola* *Andreas Krause*



All links available in our GitHub: github.com/vsomnath/flexdock