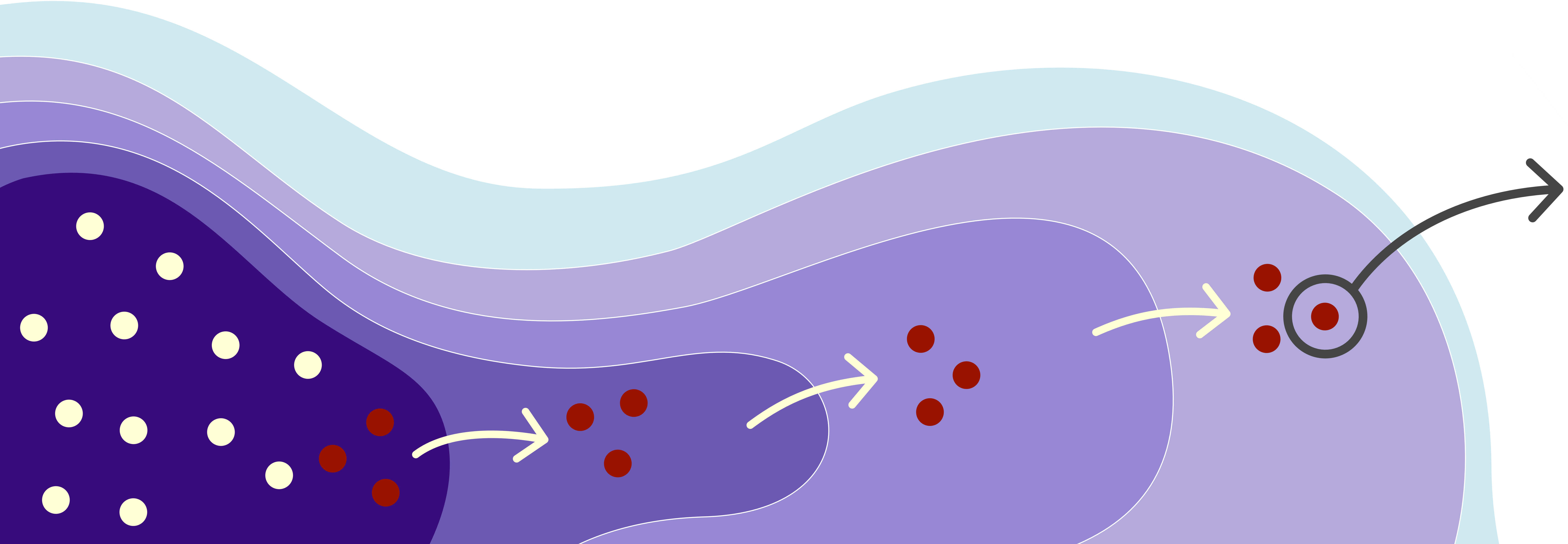
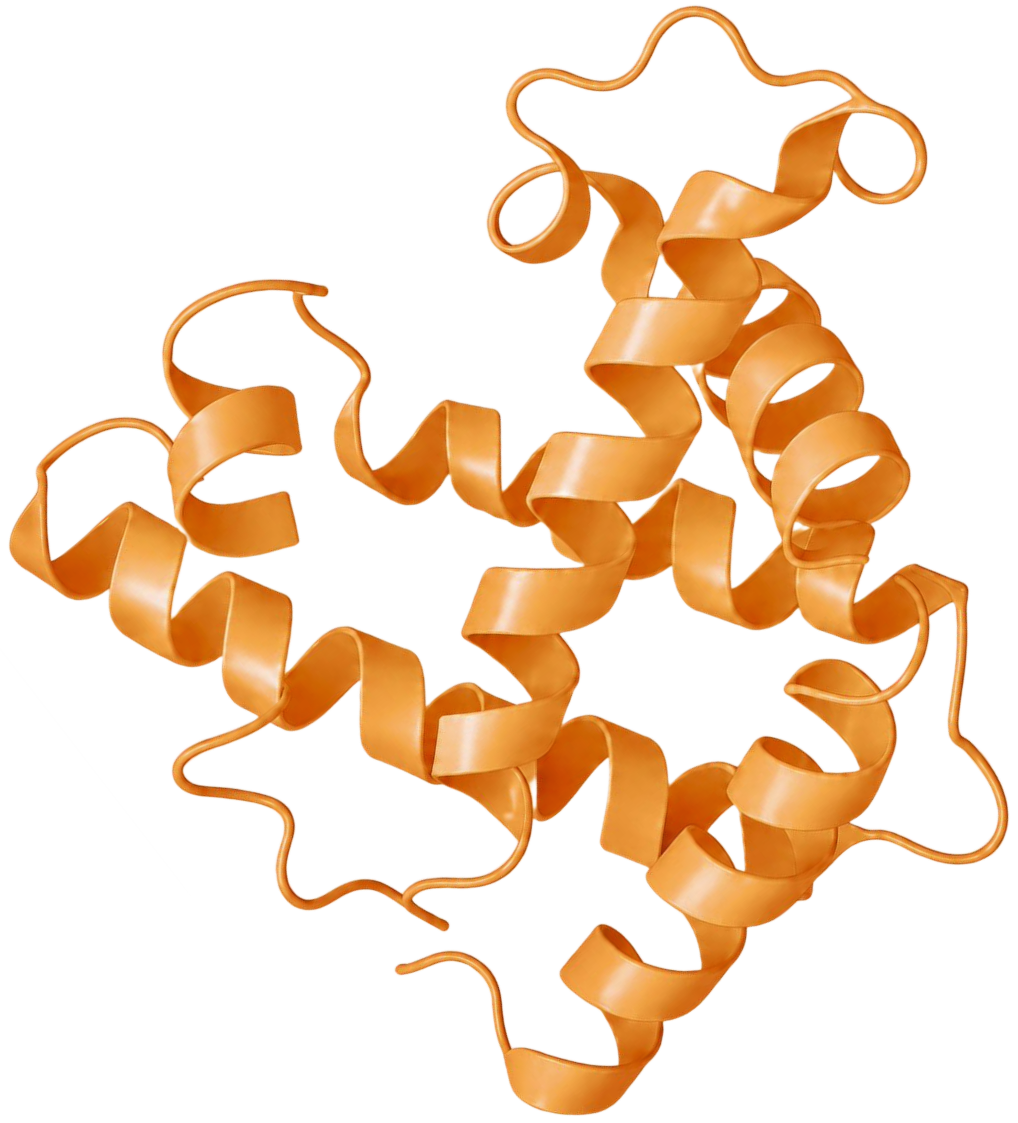
Active Generative Models for Out-of-Distribution Discovery: from Theory to Molecules

Riccardo De Santi

Doctoral Fellow at ETH AI Center

@r_de_santi rdesanti@ethz.ch





UNIST, July 13th, 2026

Team

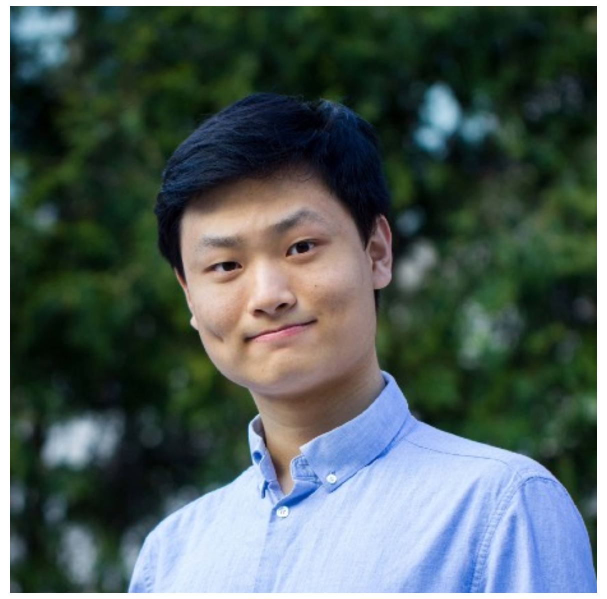

Riccardo De Santi


Bruce Lee


Cristian P. Jensen


Kimon Protopapas


Sophia Tong


Chenghao Liu


Pranam Chatterjee

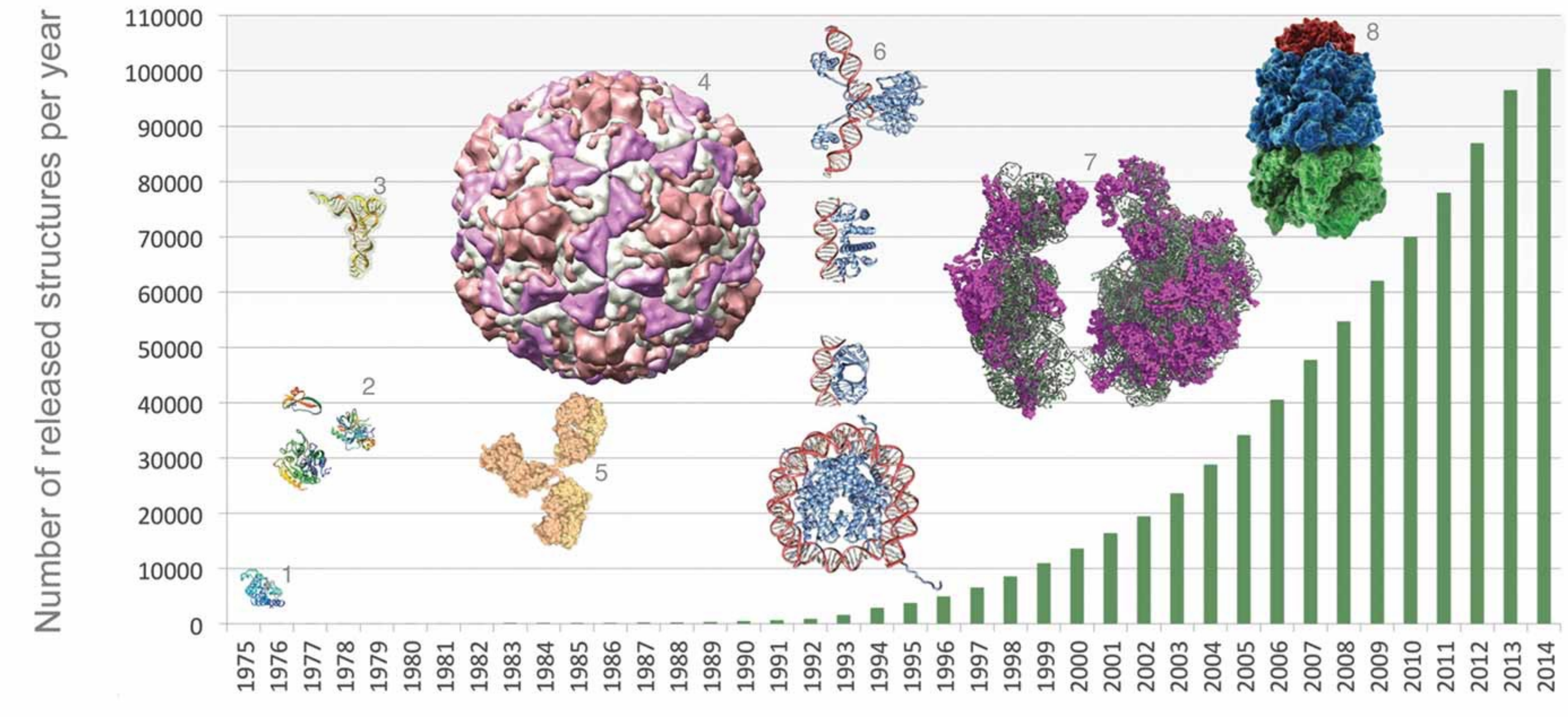

Yisong Yue


Andreas Krause

Generative Models

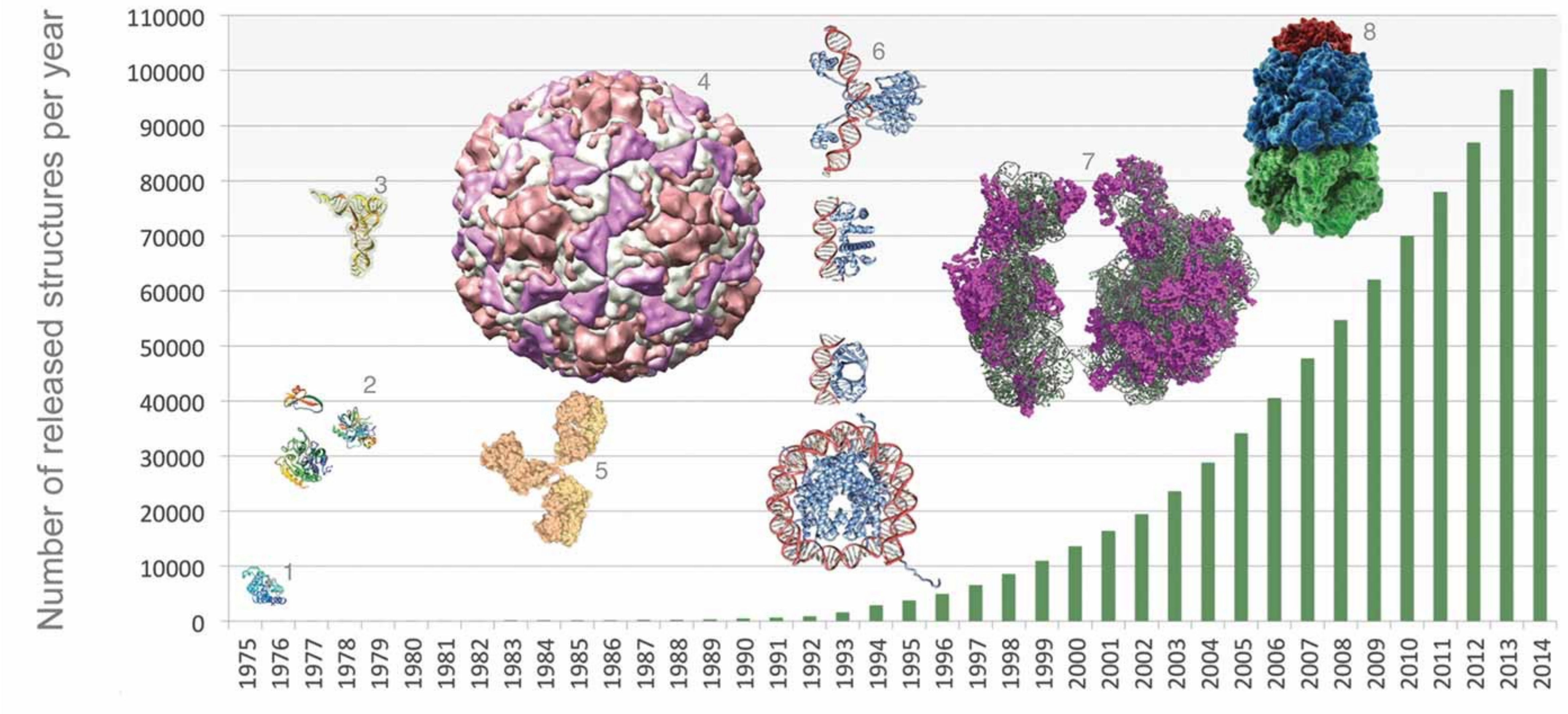
[Steven Wommack]

Generative Models for Science



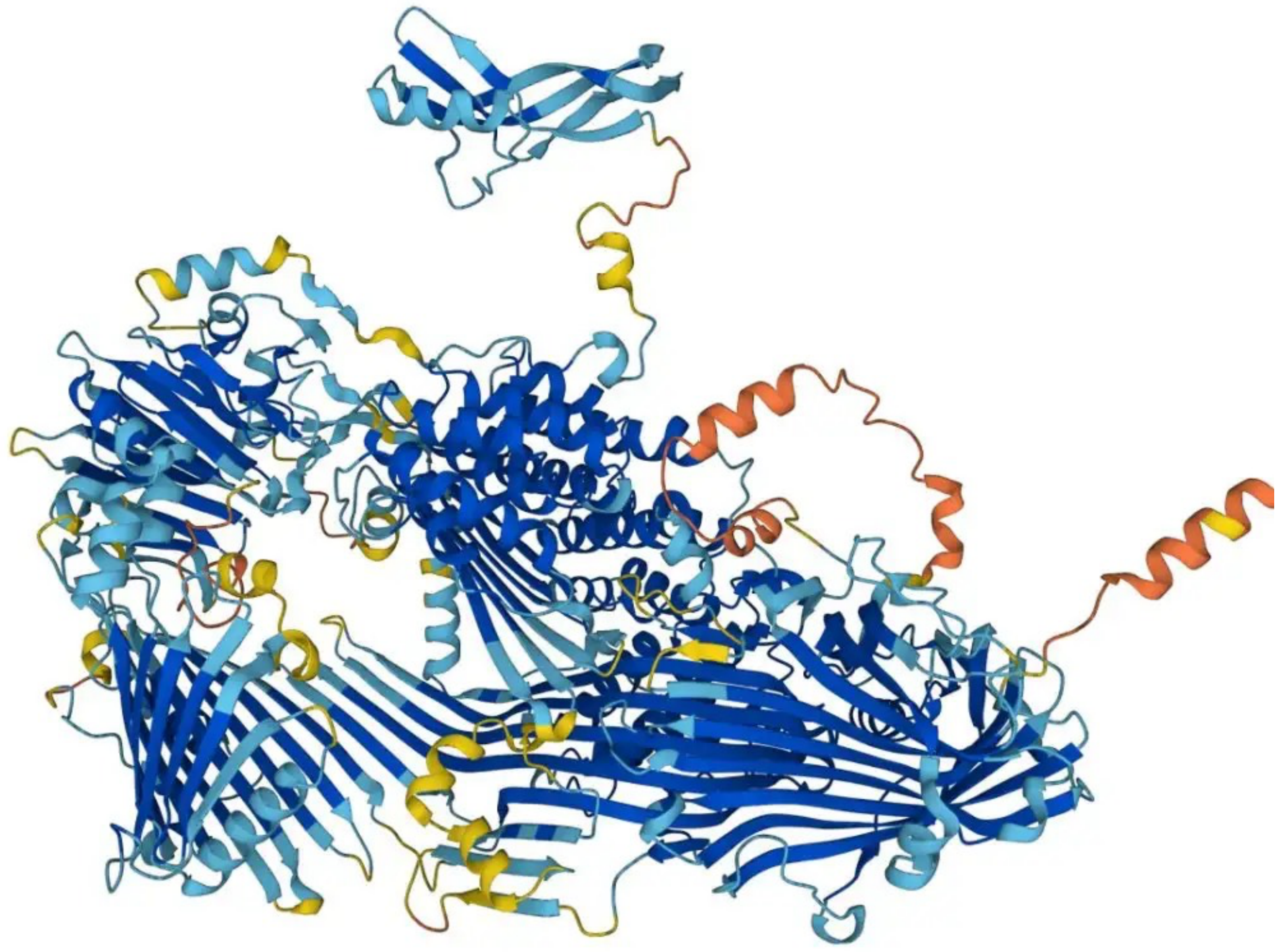
Protein Data Databank **(PDB)**

Generative Models for Science

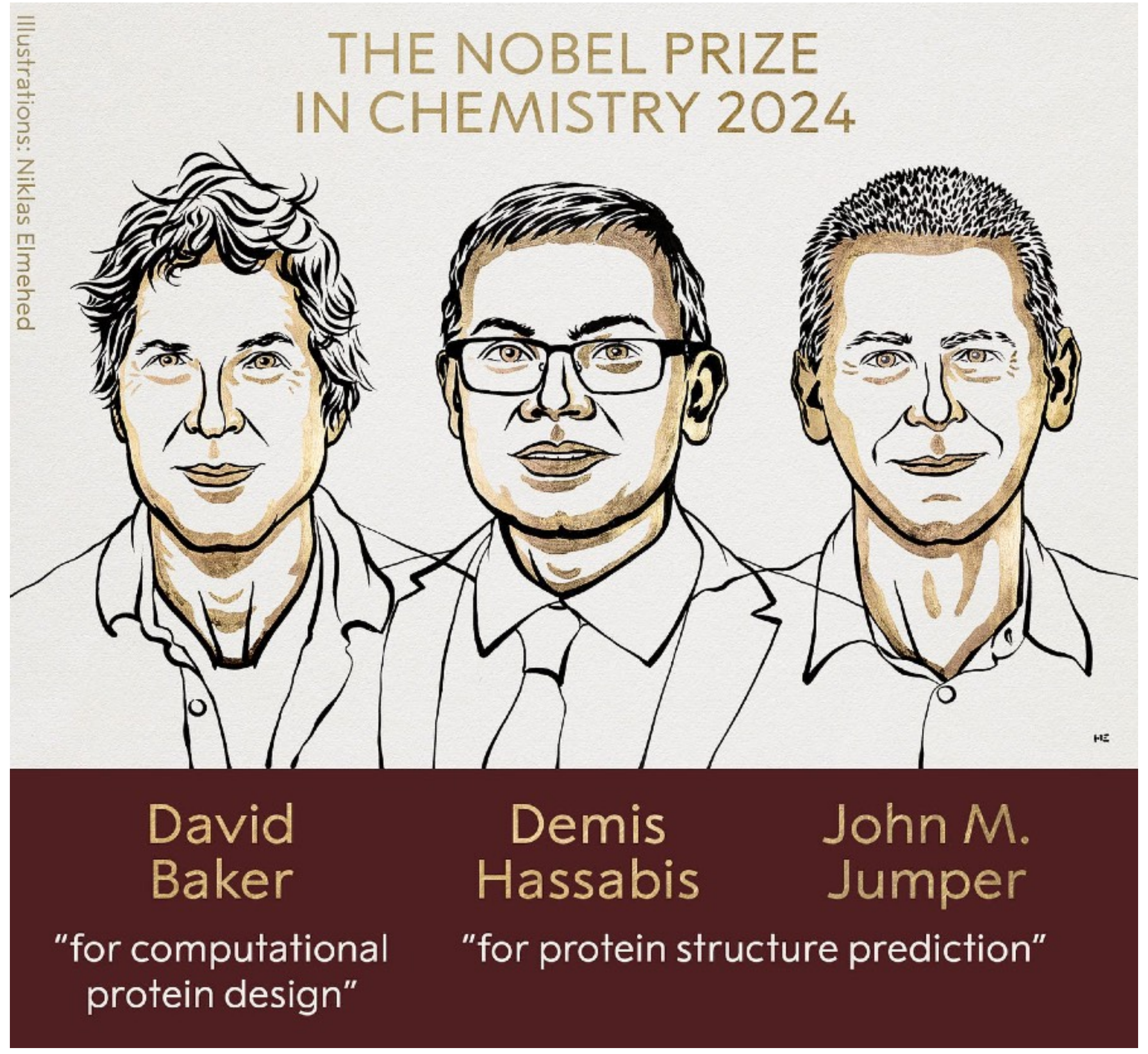


Protein Data Databank **(PDB)**

Protein structure prediction and design



2024 Chemistry Nobel Prize



Generative ~~Modeling~~ Discovery?

What is Discovery?

black-box function (reward)

$$\underset{x \in \Omega}{argmax}\; f(x)$$

valid design space (e.g., drugs, materials)

(or **action space**)

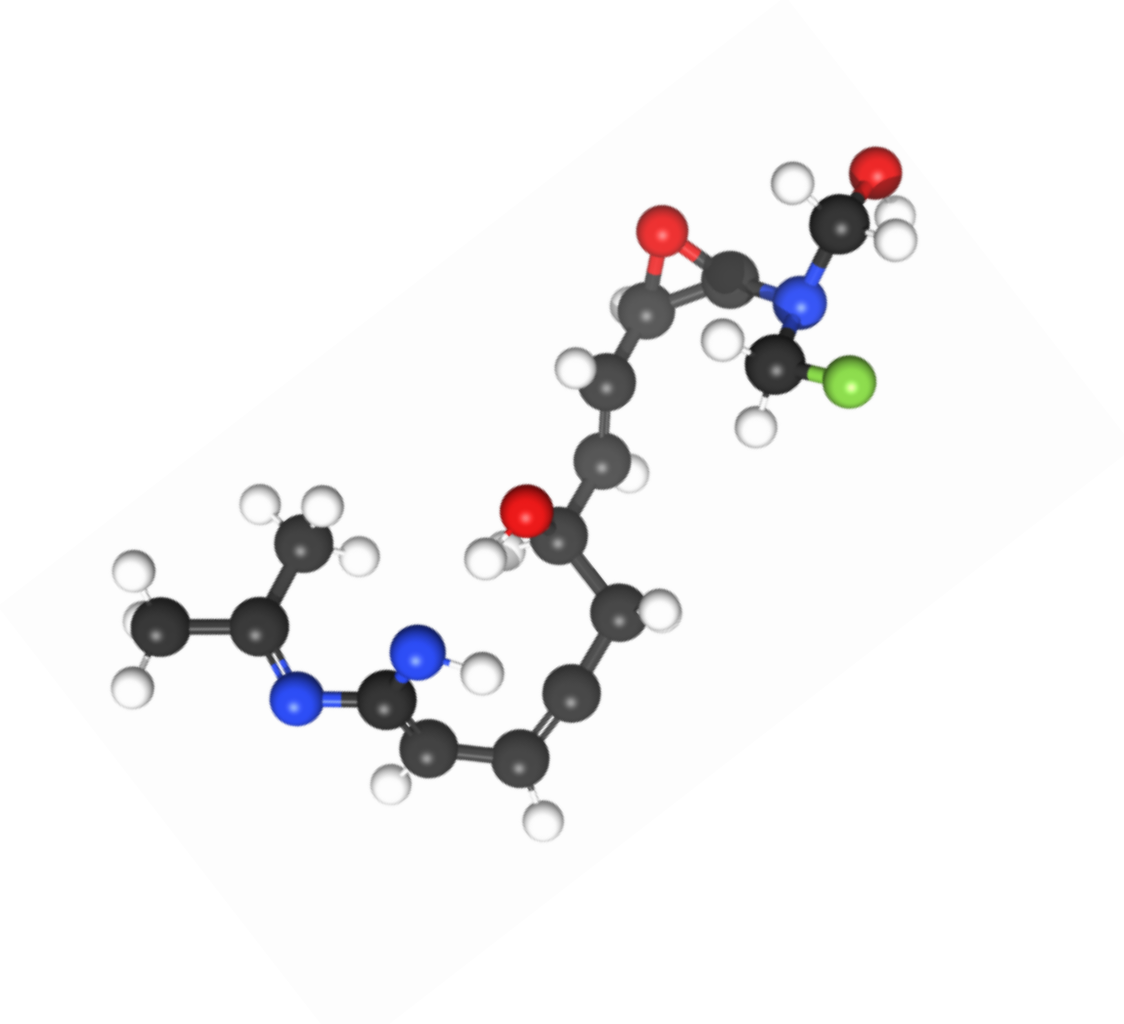
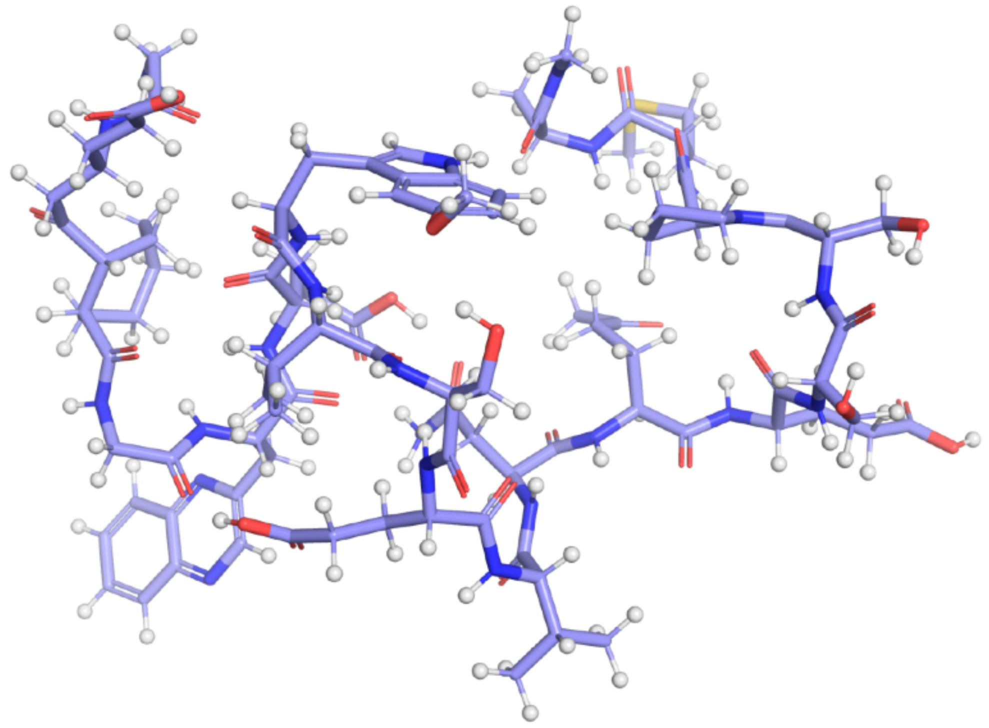
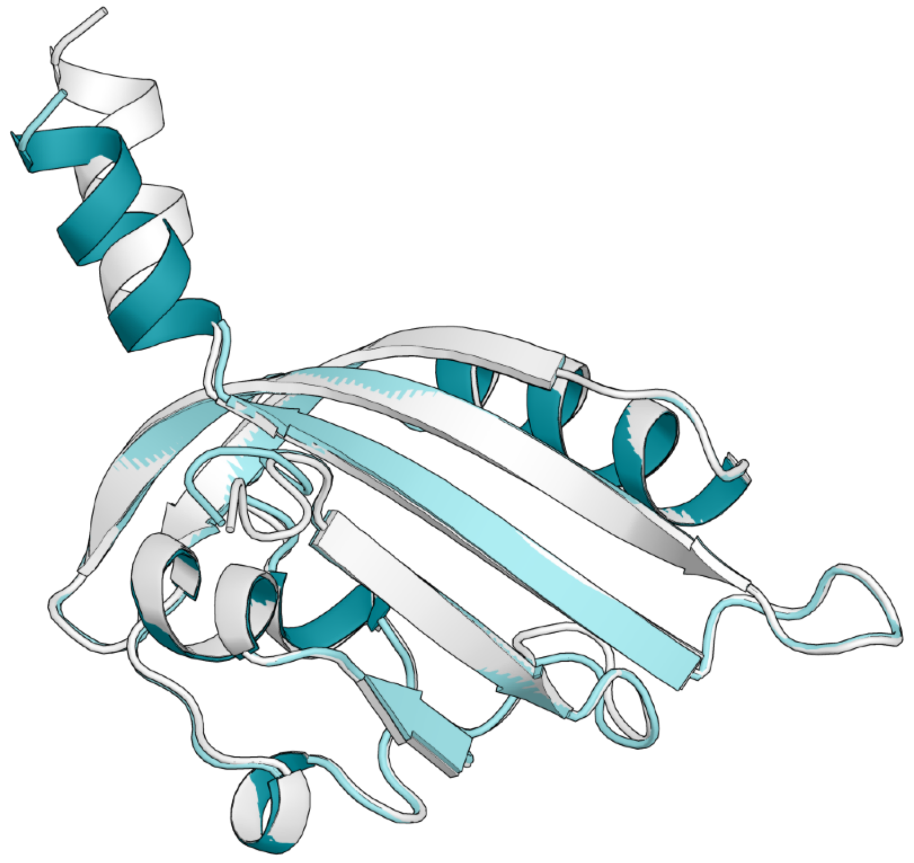
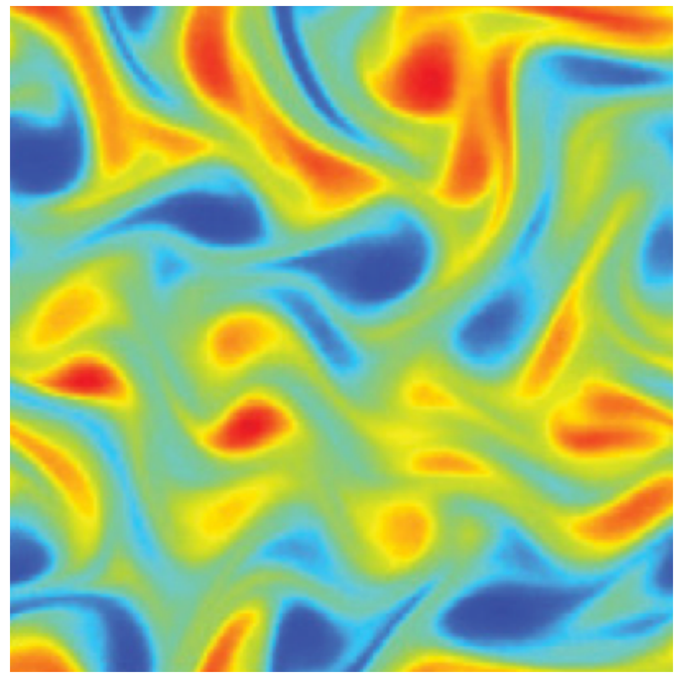
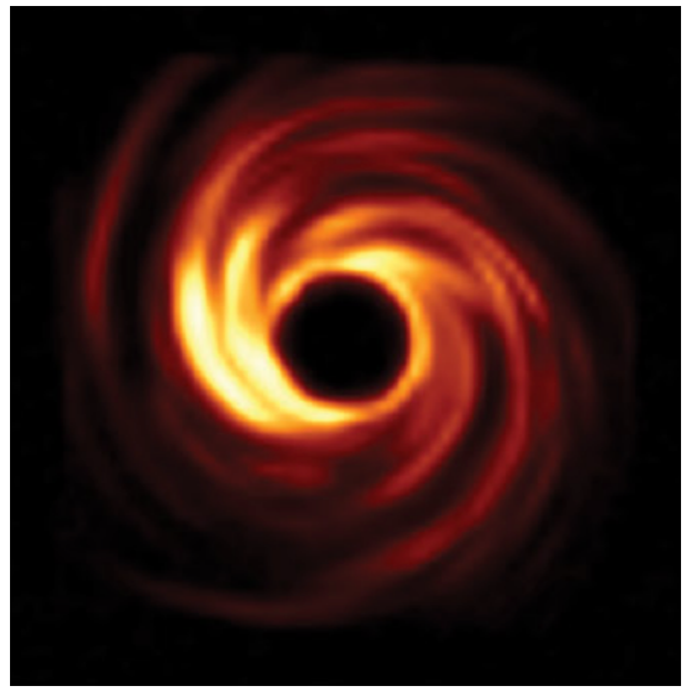
What is Discovery?

black-box function (reward)

$$\underset{x \in \Omega}{argmax}\ f(x)$$

valid design space (e.g., drugs, materials)

(or **action space**)

[Zheng et al., 2025]

What is Discovery?

black-box function (reward)

$$\underset{x \in \Omega}{argmax}\ f(x)$$

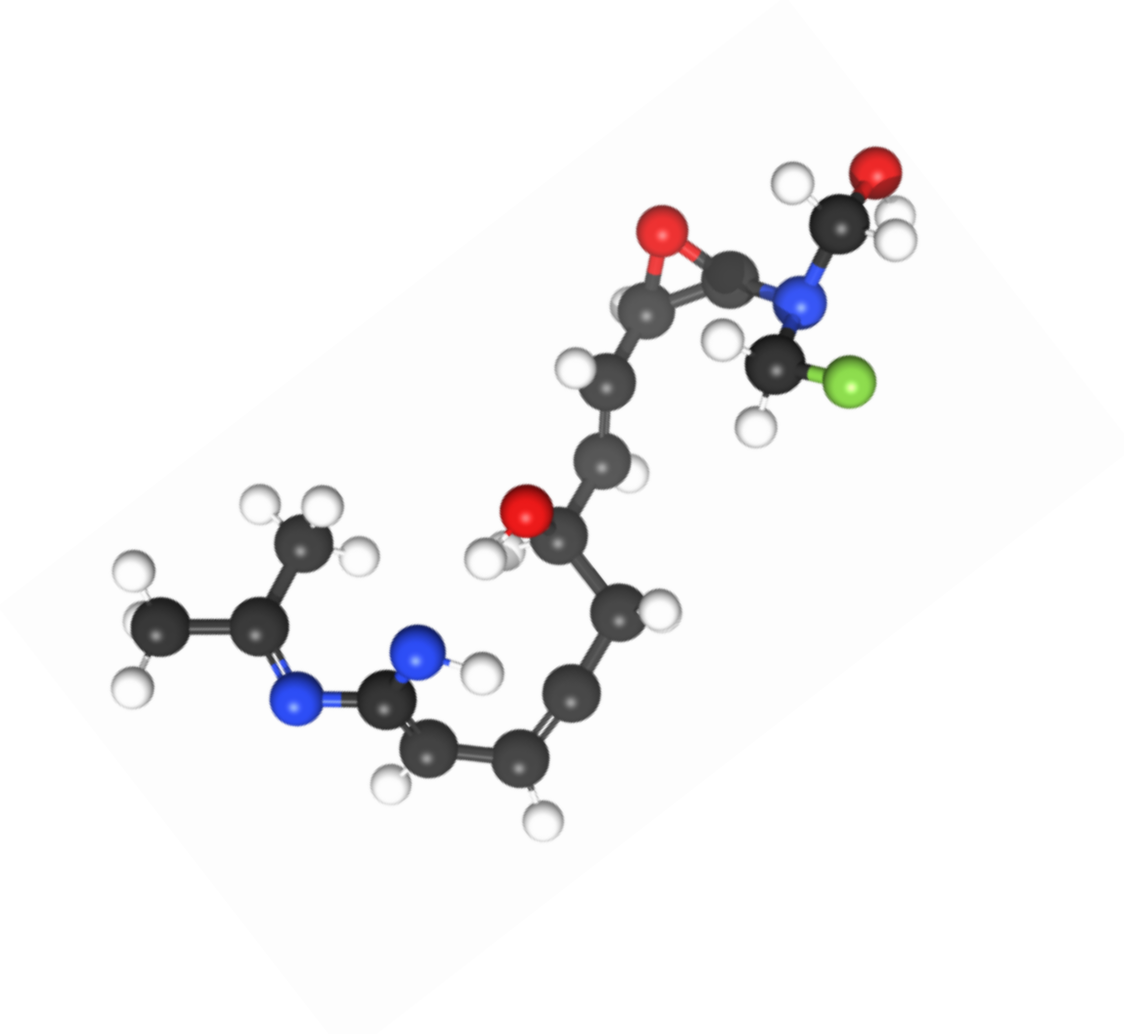
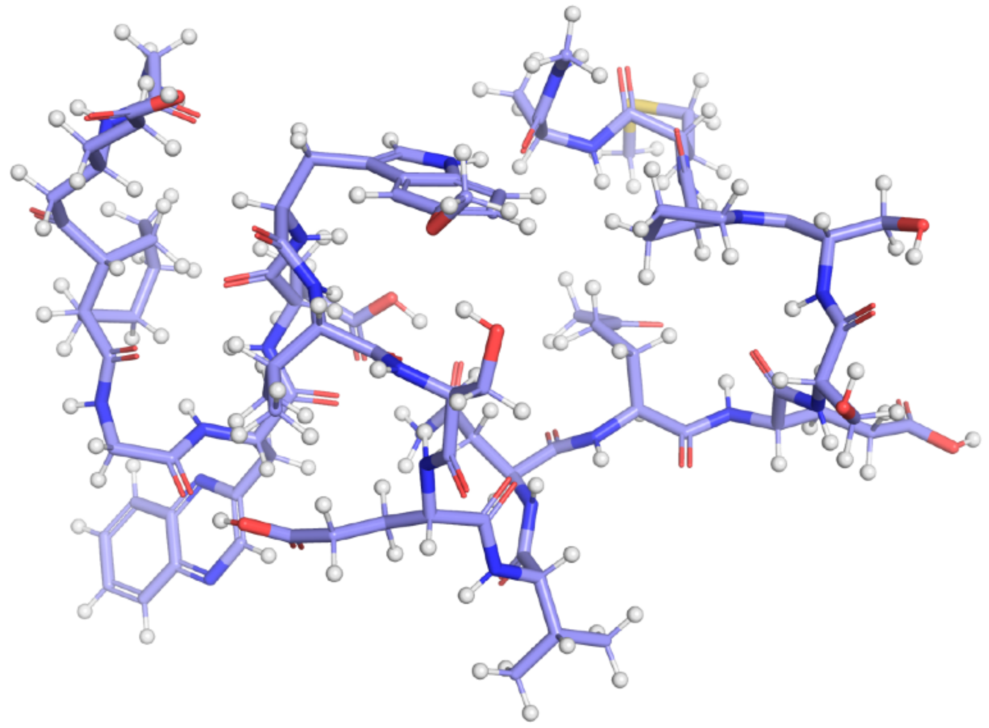
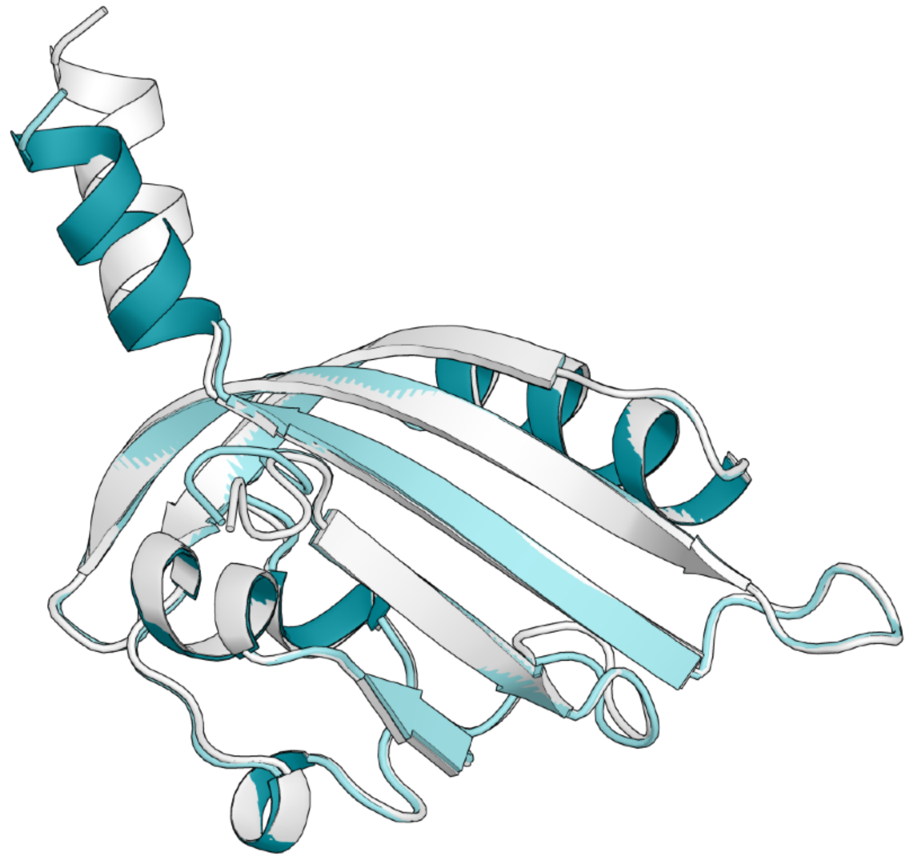
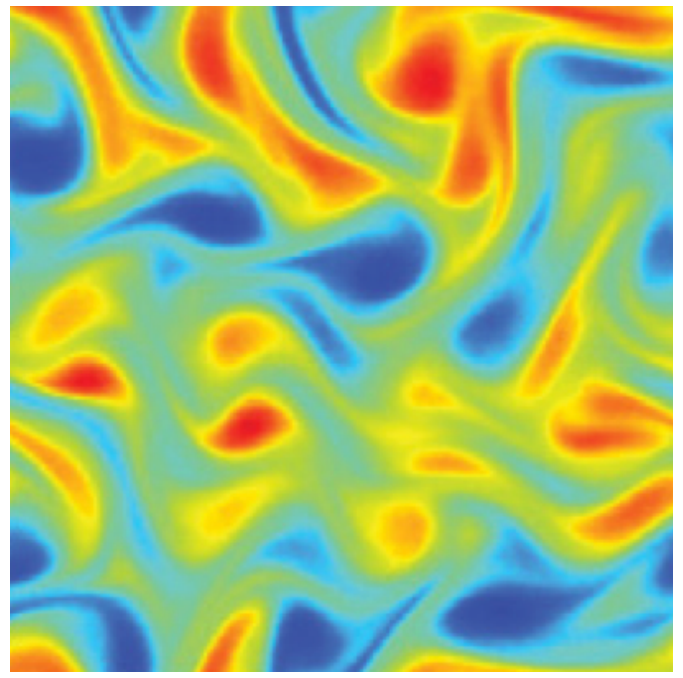
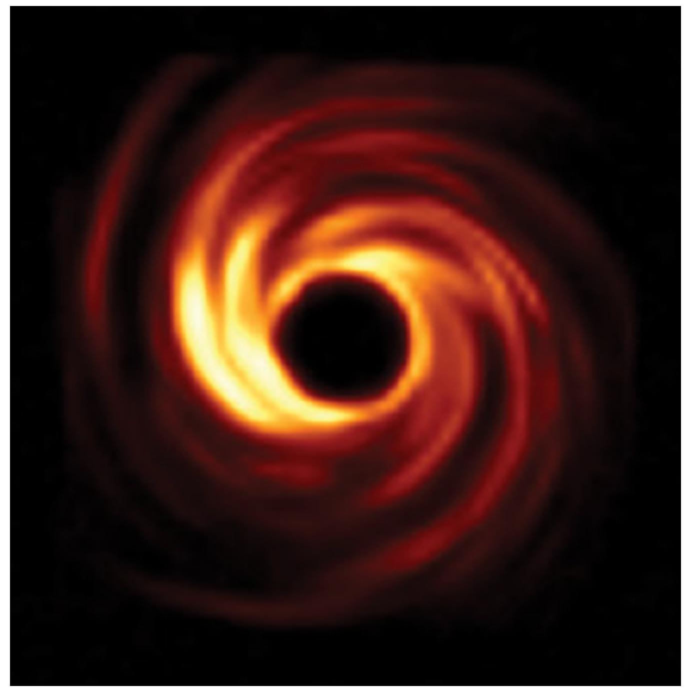
valid design space (e.g., drugs, materials)

(or **action space**)

$\approx 10^{60}$ drug discovery

$\approx 10^{80}$ material design

unknown or **hard to represent** and **search**

[Zheng et al., 2025]

What is Discovery?

black-box function (reward)

$$\underset{x \in \Omega}{argmax} \; f(x)$$

valid design space (e.g., drugs, materials)

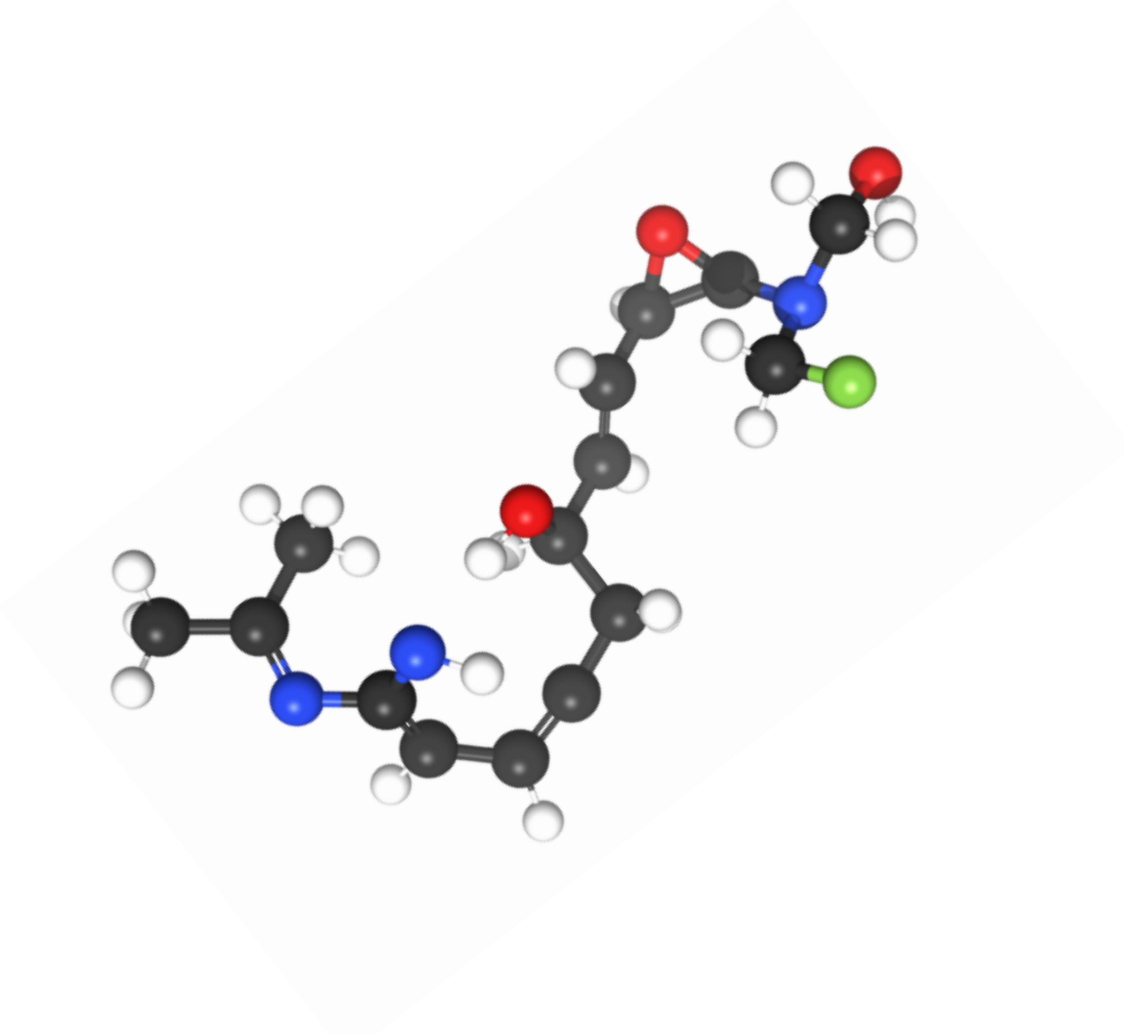
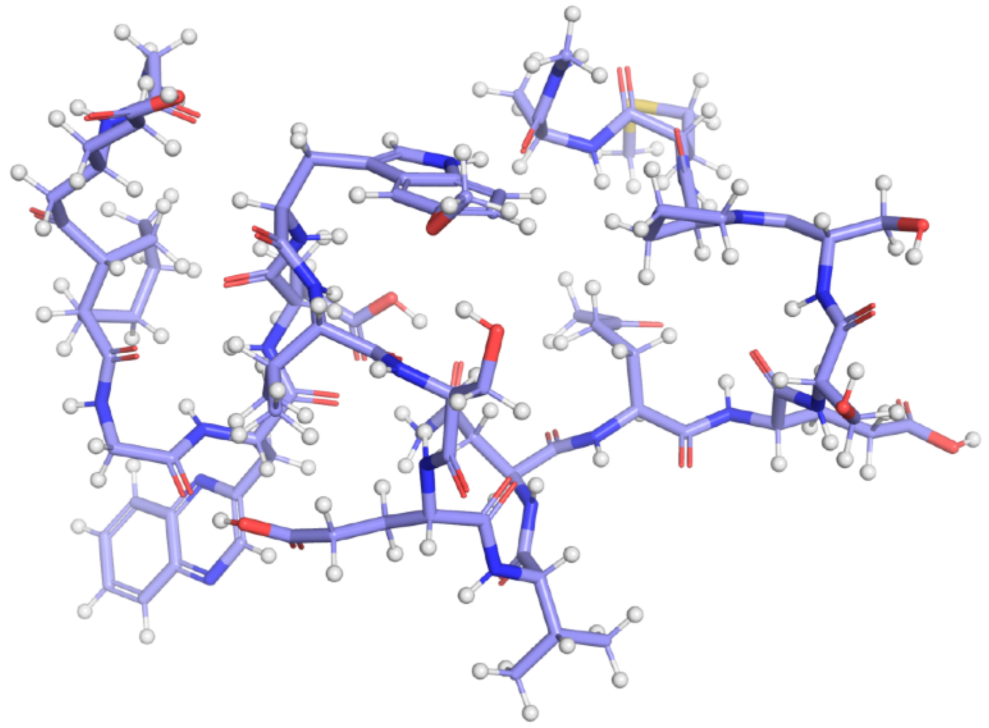
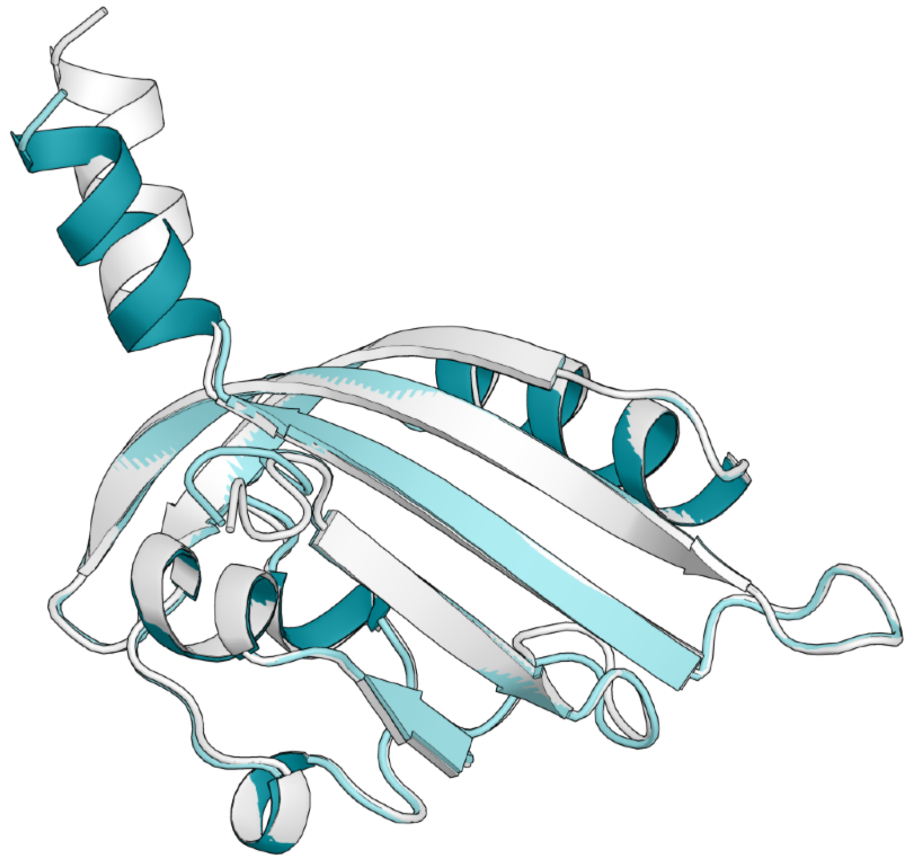
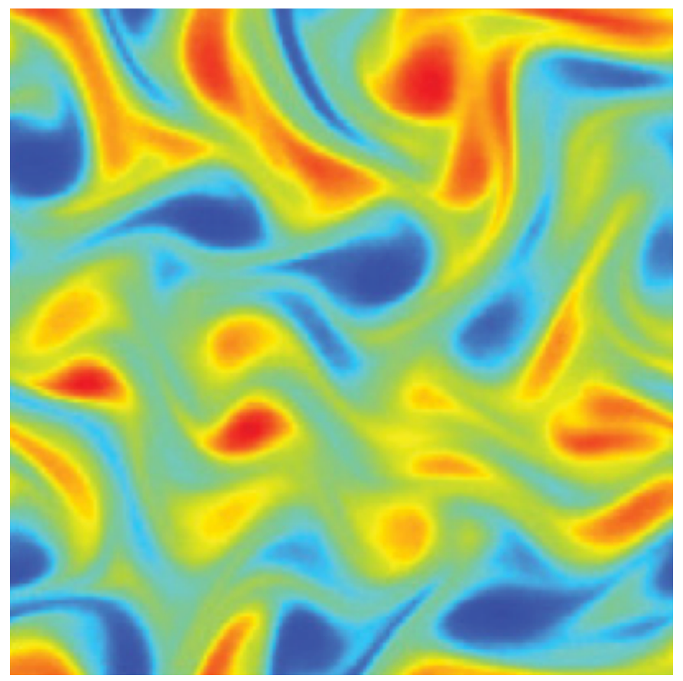
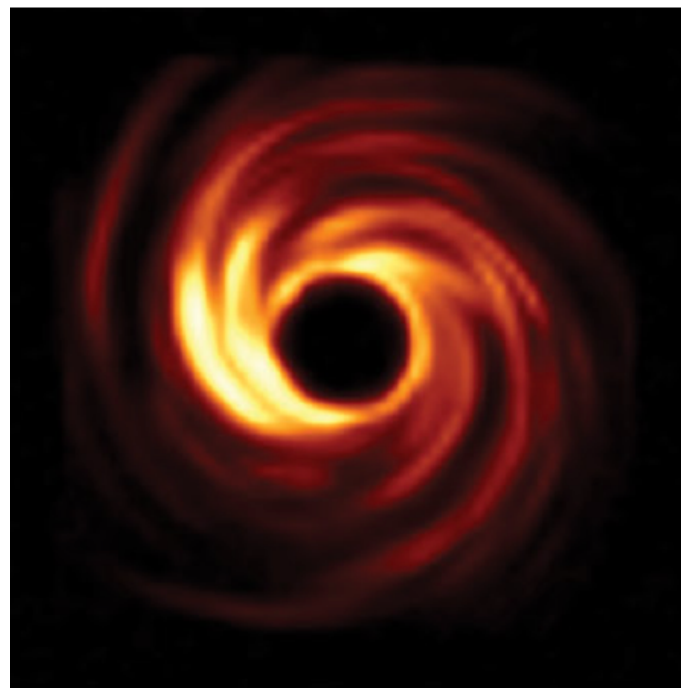
(or **action space**)

$\approx 10^{60}$ drug discovery

$\approx 10^{80}$ material design

unknown or **hard to represent** and **search**

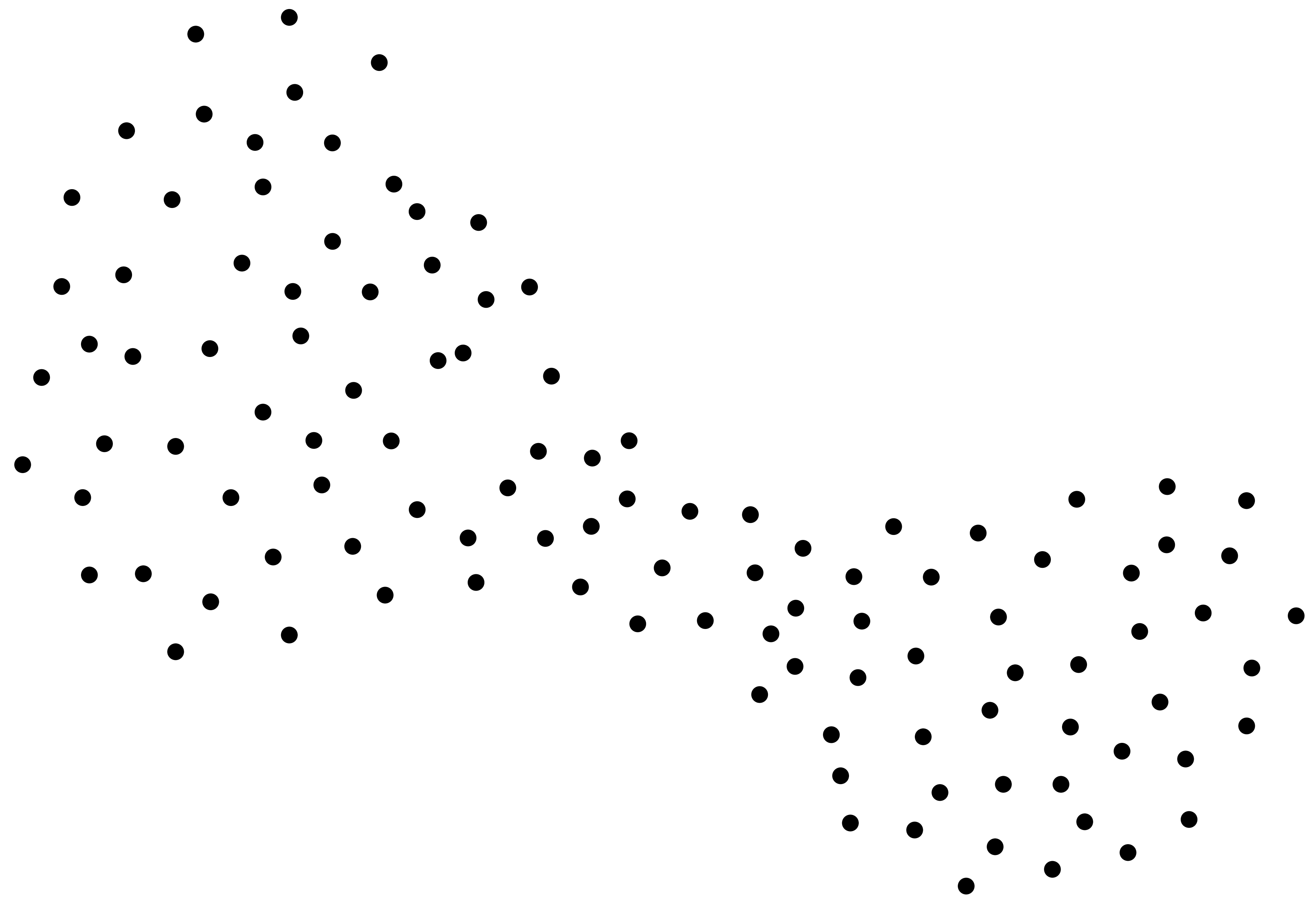
→ Generative Discovery

[Zheng et al., 2025]

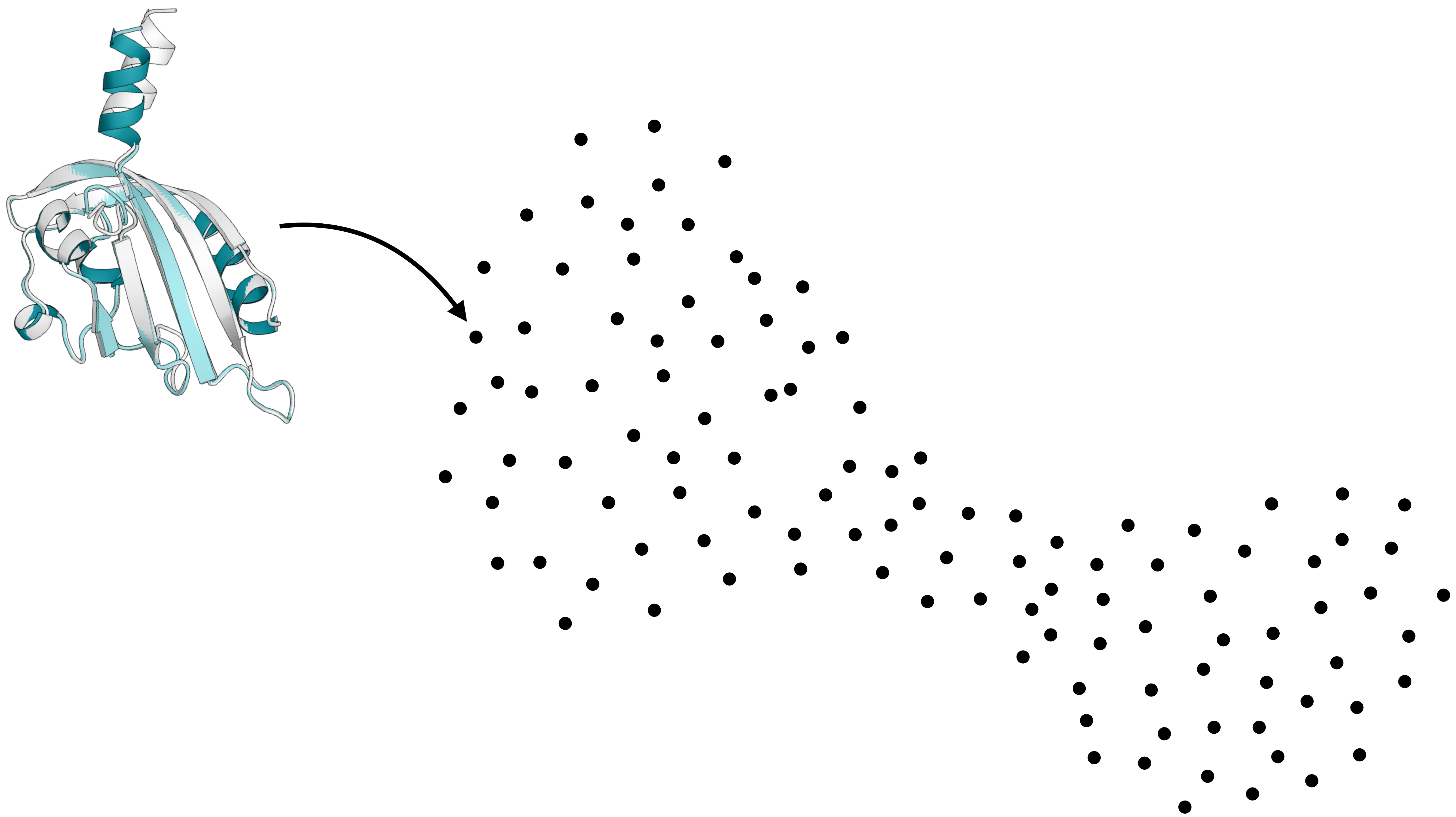
Generative Discovery Paradigm

Data $D = \{x_i\}_{i=1}^n$



Generative Discovery Paradigm

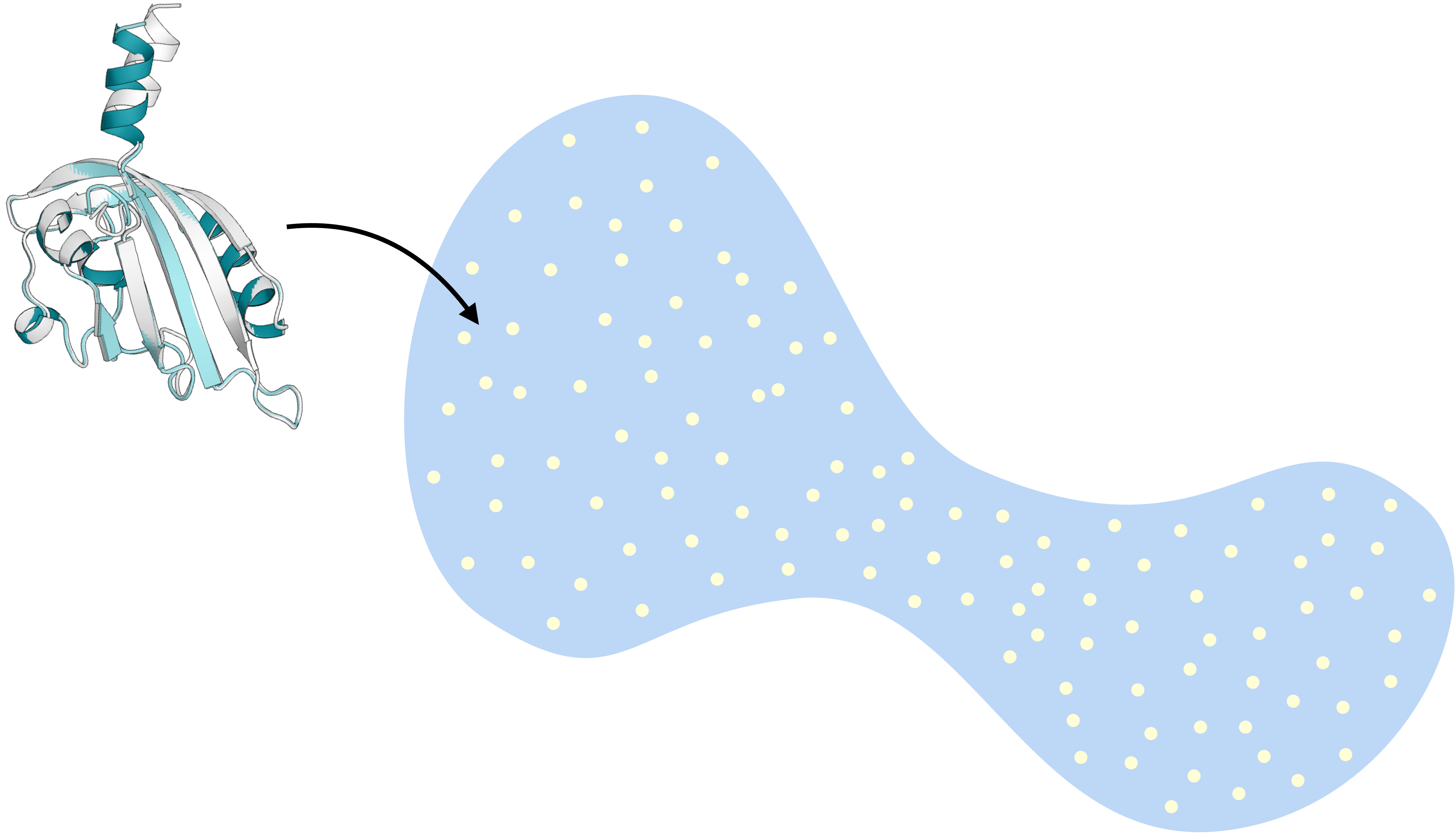
Data $D = \{x_i\}_{i=1}^n$



Generative Discovery Paradigm

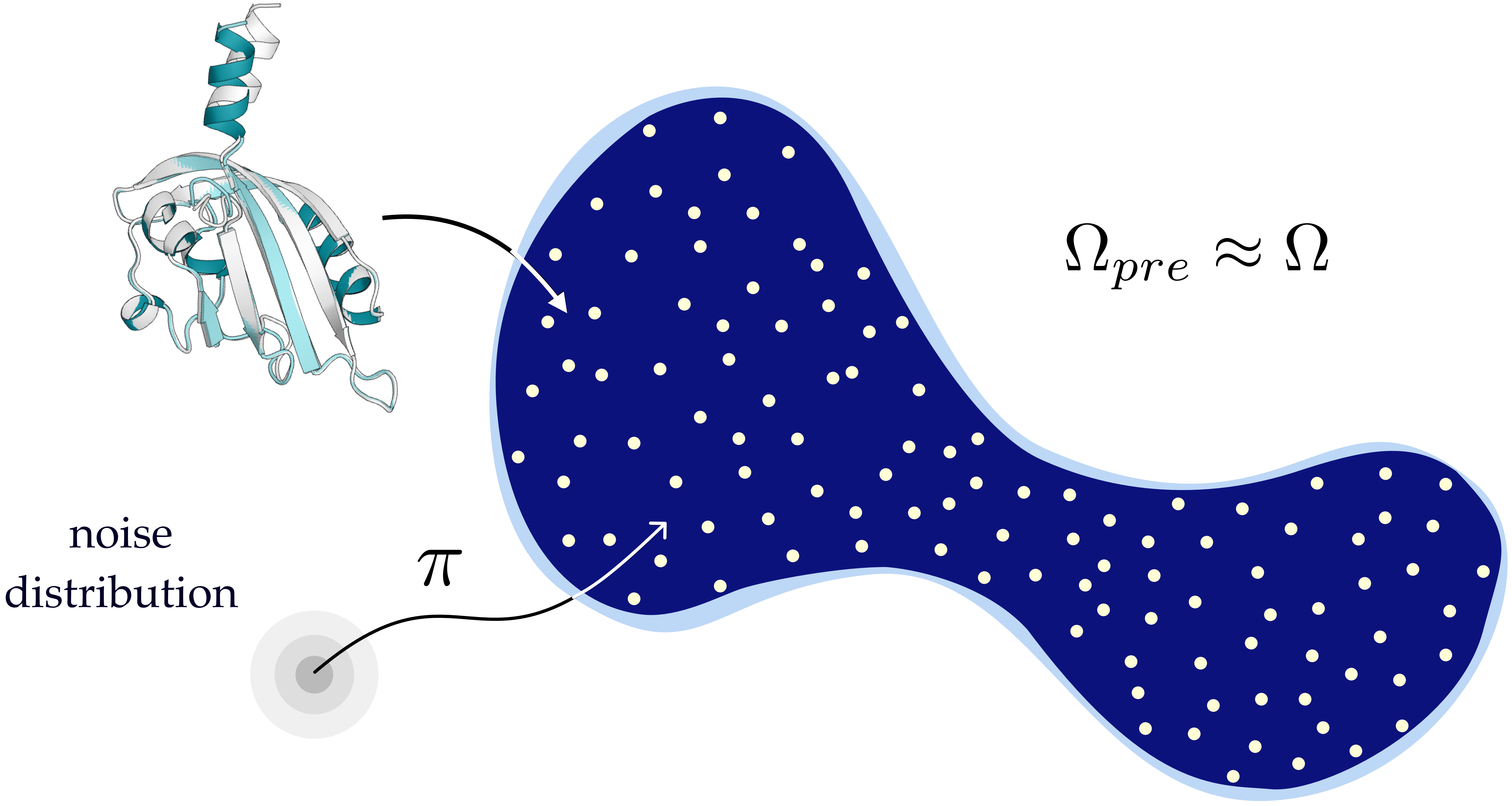
valid design space Ω (e.g., protein structures)

Data $D = \{x_i\}_{i=1}^n$



Generative Discovery Paradigm

valid design space Ω — generable set of pre-trained gen. model Ω_{pre}



$\Omega_{pre} \approx \Omega$

Data $D = \{x_i\}_{i=1}^n$

$\downarrow$

Pre-trained generative model π

The Dream: Discovery via Generative Optimization

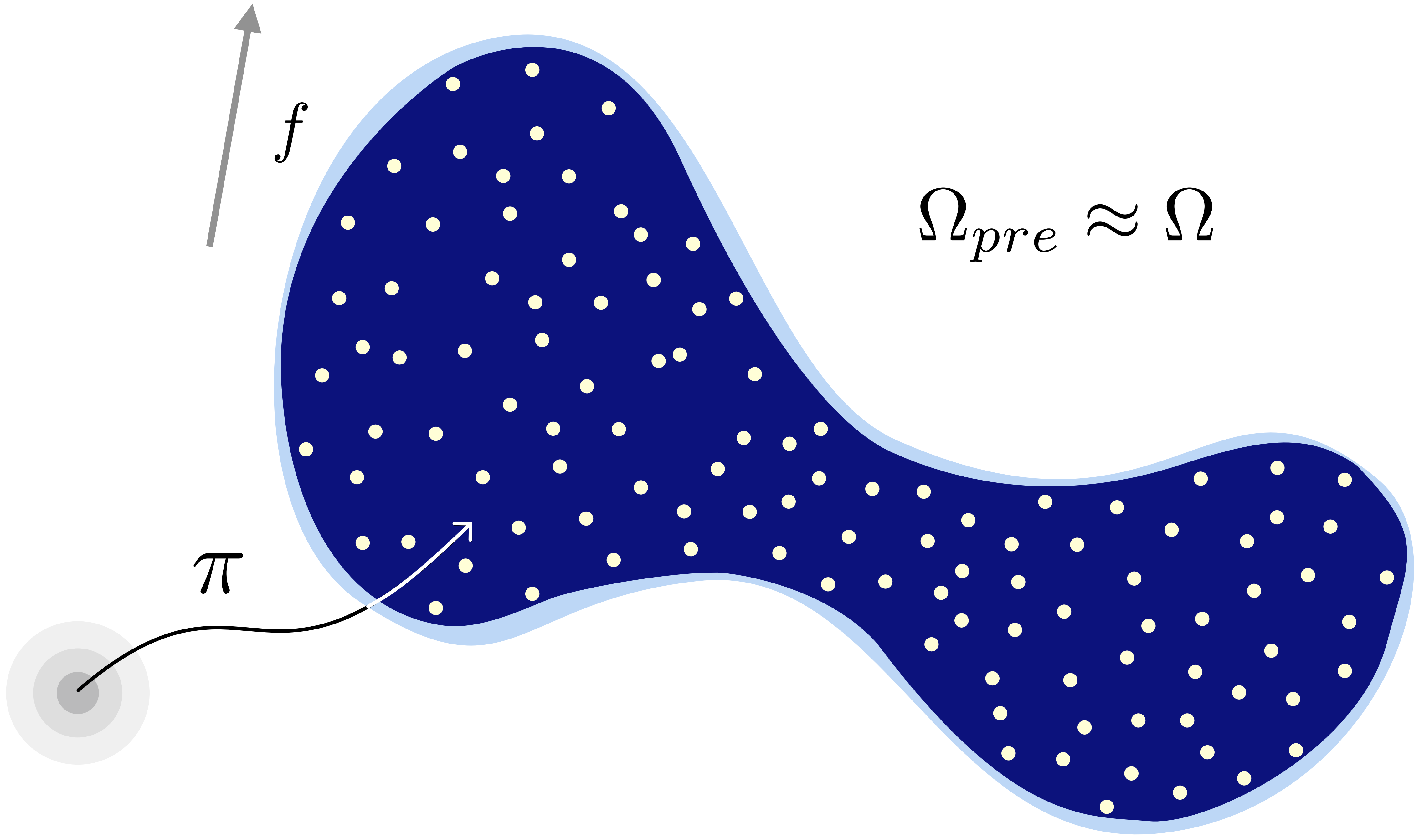
valid design space Ω

generable set of pre-trained gen. model Ω_{pre}

Data $D = \{x_i\}_{i=1}^{n}$

$\downarrow$

Pre-trained generative model π

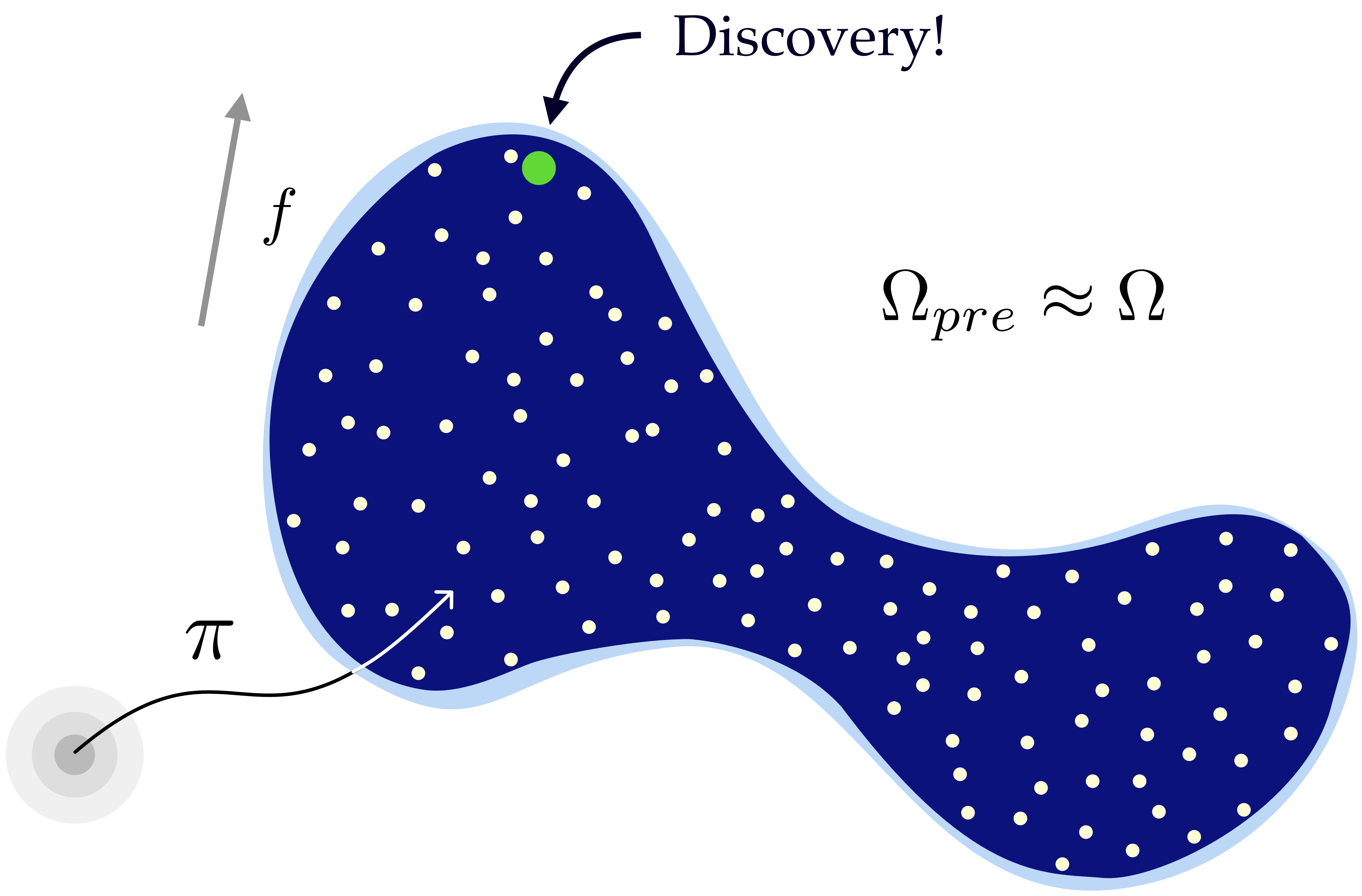


Discovery problem

$$\underset{x \in \Omega}{argmax}\ f(x)$$

The Dream: Discovery via Generative Optimization

valid design space Ω — generable set of pre-trained gen. model Ω_{pre}



Data $D = \{x_i\}_{i=1}^n$

↓

Pre-trained generative model π

↓

Adapt π to optimize f

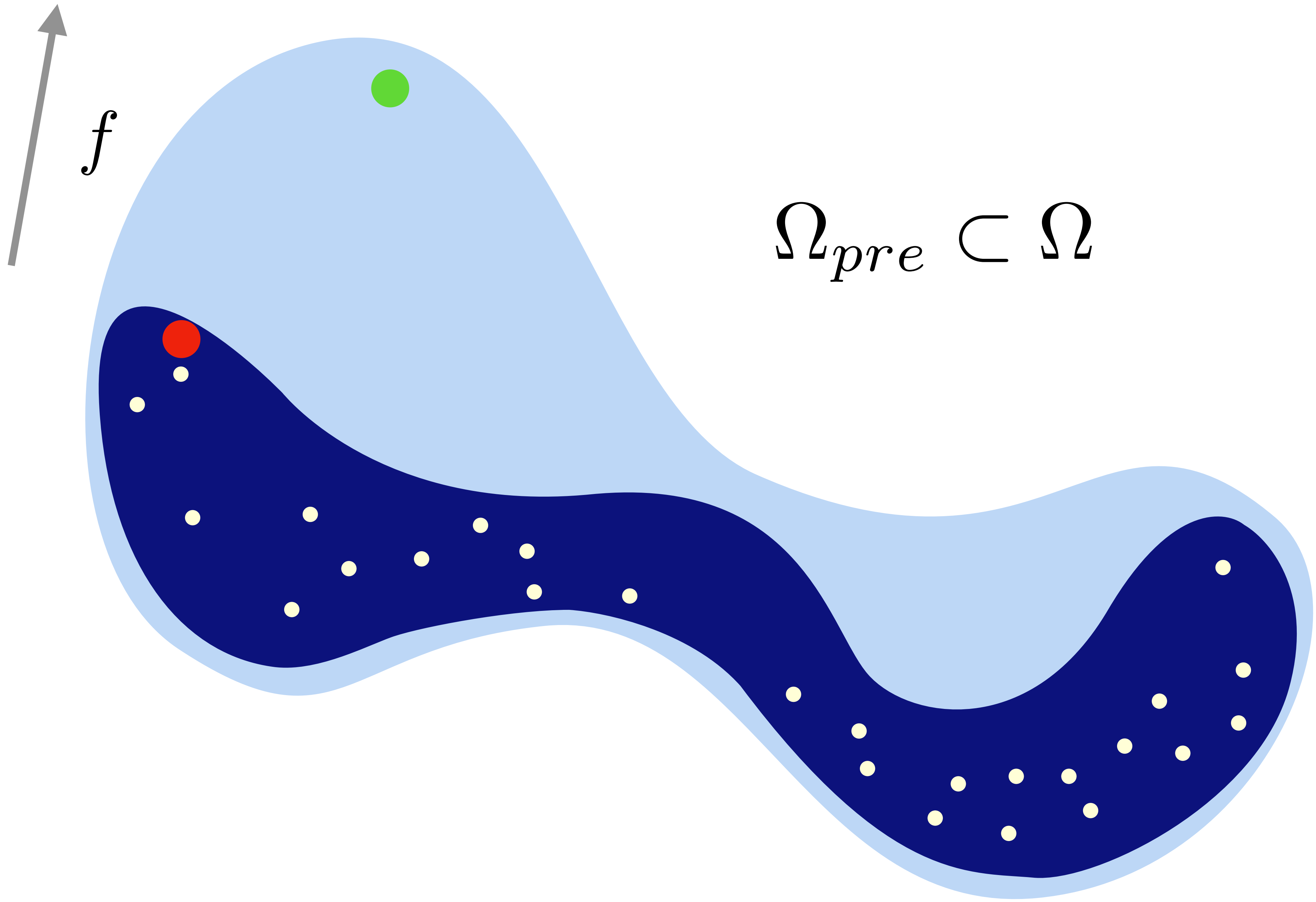
Discovery problem

$$\underset{x \in \Omega}{argmax}\ f(x)$$

The Reality: Generative Optimization Limitation

valid design space Ω

generable set of pre-trained gen. model Ω_{pre}



$\Omega_{pre} \subset \Omega$

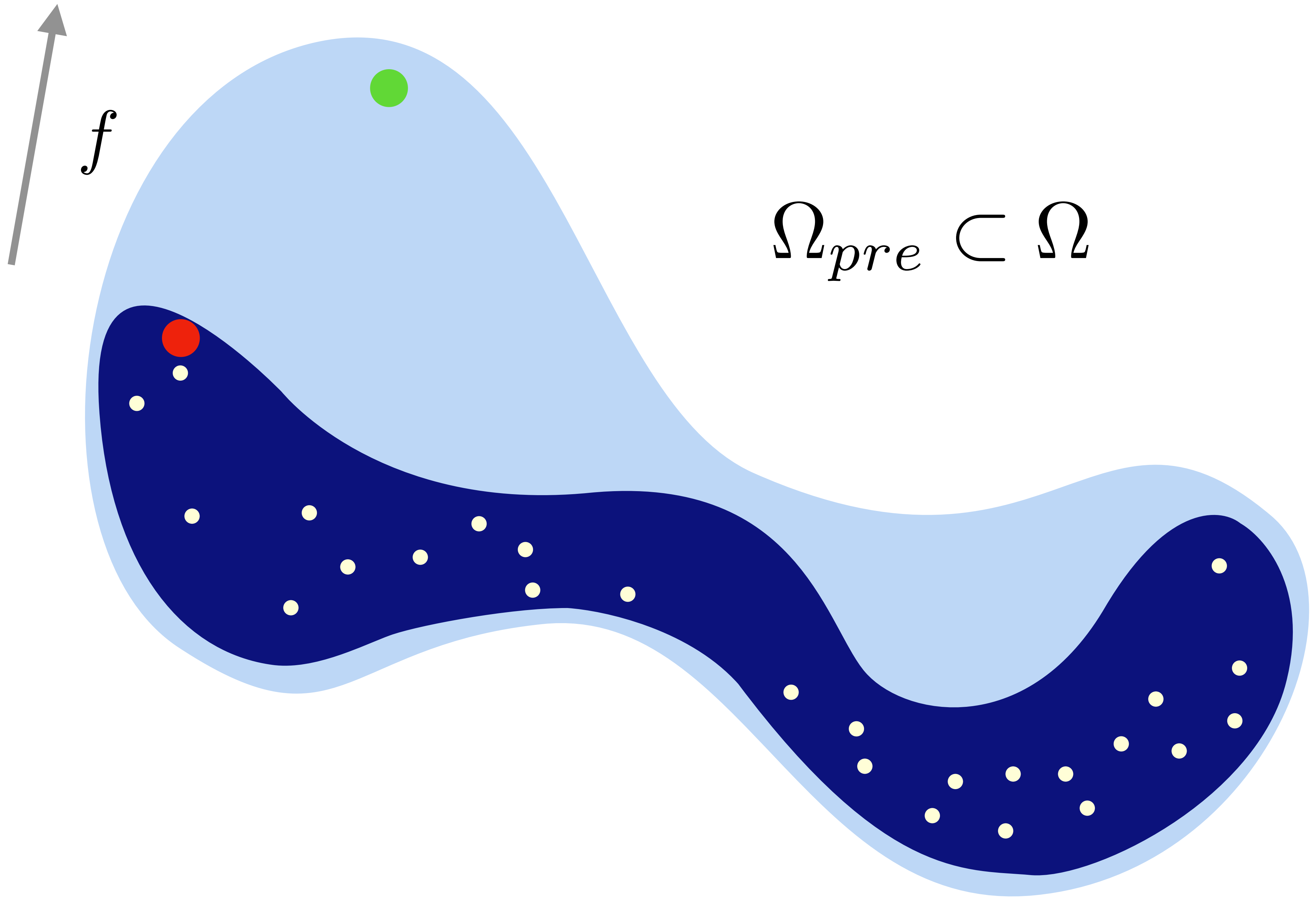
Generative models generalize locally over training data.

Discovery problem

$$\underset{x \in \Omega}{argmax}\ f(x)$$

The Reality: Generative Optimization Limitation

valid design space Ω — generable set of pre-trained gen. model Ω_{pre}



$\Omega_{pre} \subset \Omega$

Generative models generalize locally over training data.

↓

Especially in natural sciences:

E.g.,
- Chemistry
- Biology
- Physics

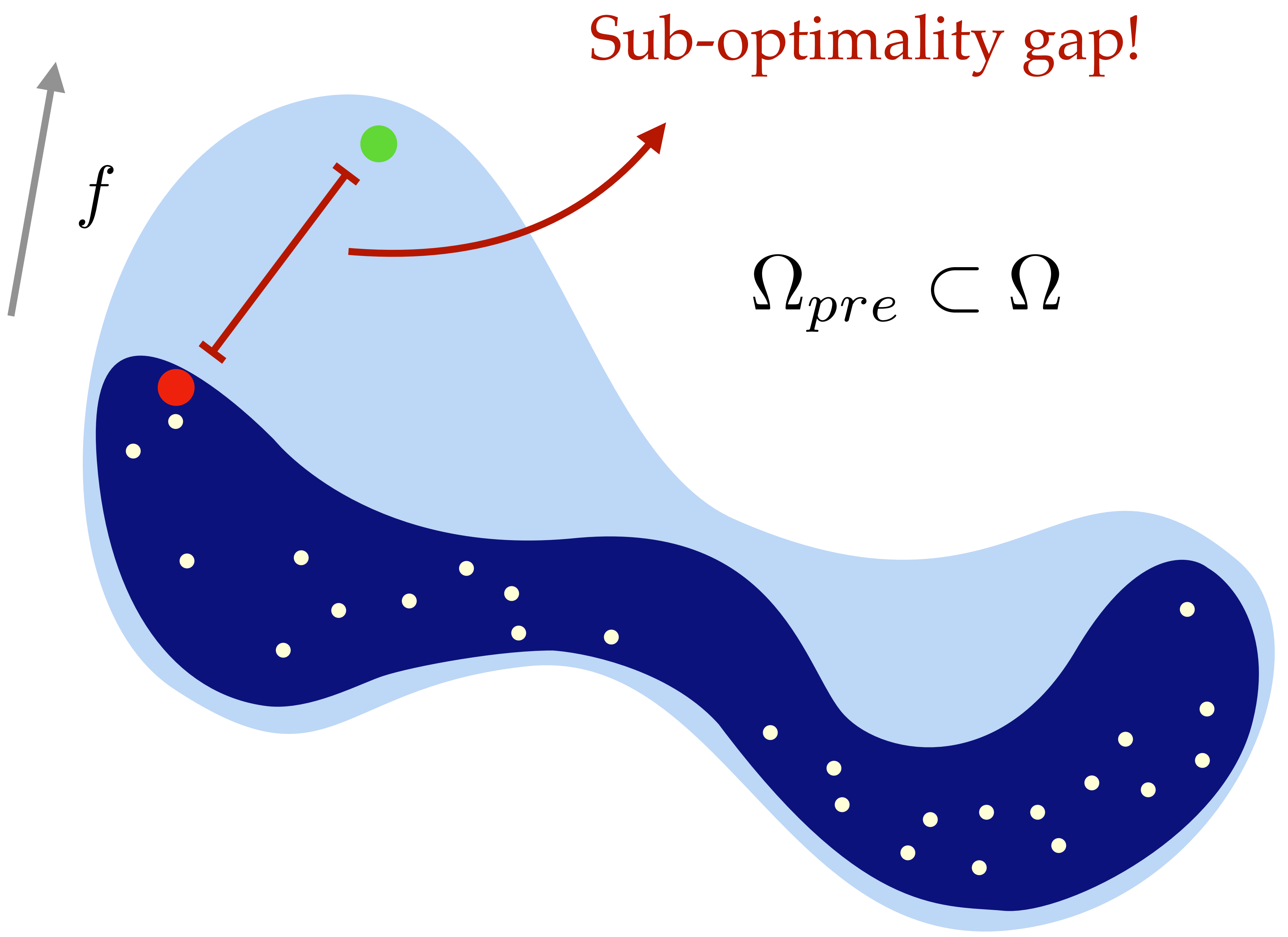
Discovery problem

$$\underset{x \in \Omega}{argmax}\ f(x)$$

The Reality: Generative Optimization Limitation

valid design space Ω

generable set of pre-trained gen. model Ω_{pre}

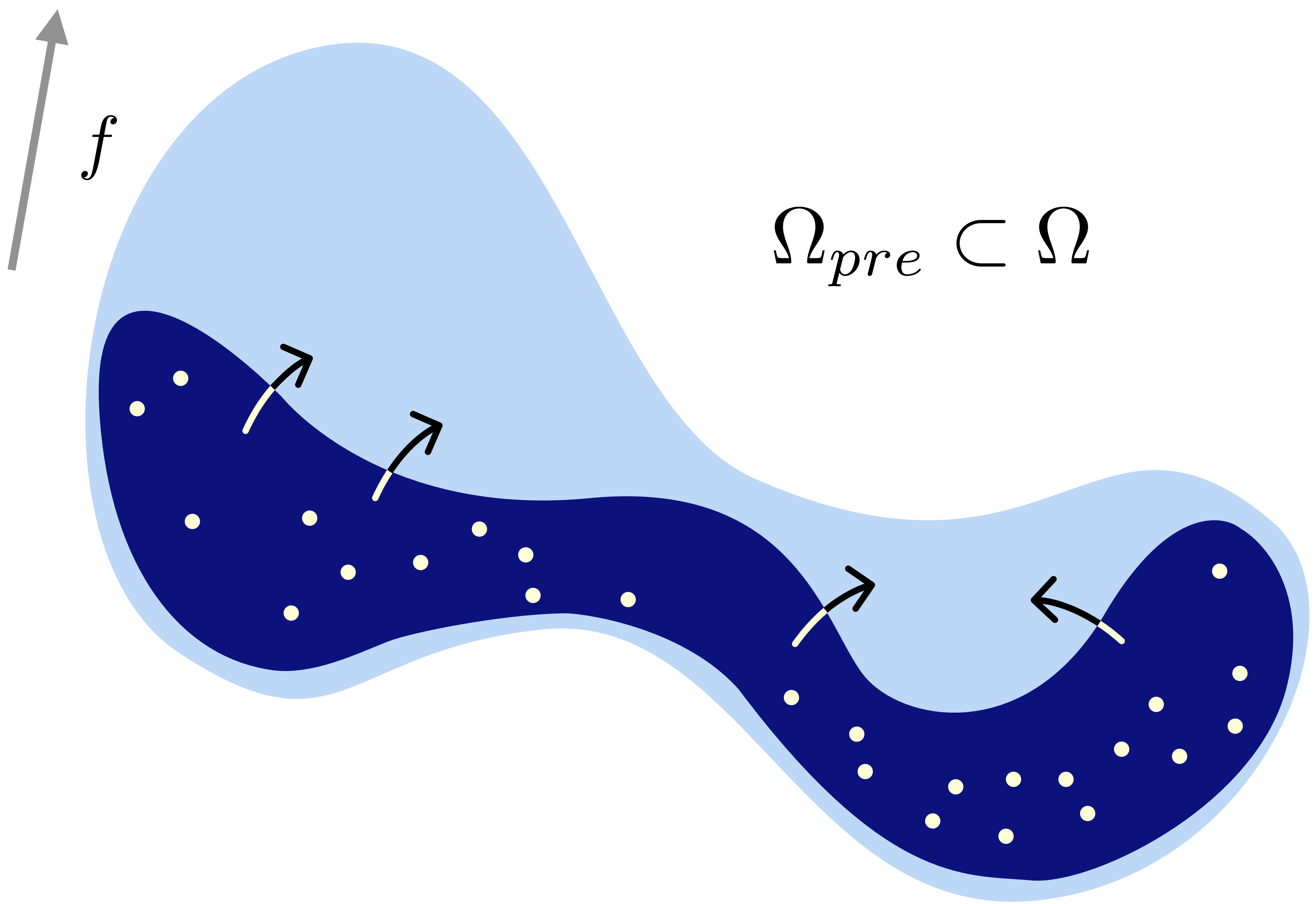


Discovery problem

$$\underset{x \in \Omega}{argmax}\ f(x)$$

Generative discovery entails a new exploration problem

$$\underset{x \in \Omega}{argmax}\ f(x)$$



Generative discovery entails a new exploration problem

$$\underset{x \in \Omega}{argmax}\ f(x)$$

Generative discovery entails a new exploration problem

$$\underset{x \in \Omega}{argmax}\ f(x)$$

classic optimization
(e.g., convex optimization)

f known

Ω known

Generative discovery entails a new exploration problem

$$\underset{x \in \Omega}{argmax} \; f(x)$$

classic optimization (e.g., convex optimization)		**decision-making** (e.g., bandits, BO, RL)	
f	known	f	learned
Ω	known	Ω	known

Generative discovery entails a new exploration problem

$$\underset{x \in \Omega}{argmax}\ f(x)$$

classic optimization (e.g., convex optimization)		**decision-making** (e.g., bandits, BO, RL)		**generative discovery** (e.g., drug design)	
f	known	f	learned	f	learned
Ω	known	Ω	known	Ω	learned

Generative discovery entails a new exploration problem

$$\underset{x \in \Omega}{argmax}\ f(x)$$

classic optimization (e.g., convex optimization)		**decision-making** (e.g., bandits, BO, RL)		**generative discovery** (e.g., drug design)	
f	known	f	learned	f	learned
Ω	known	Ω	known	Ω	learned

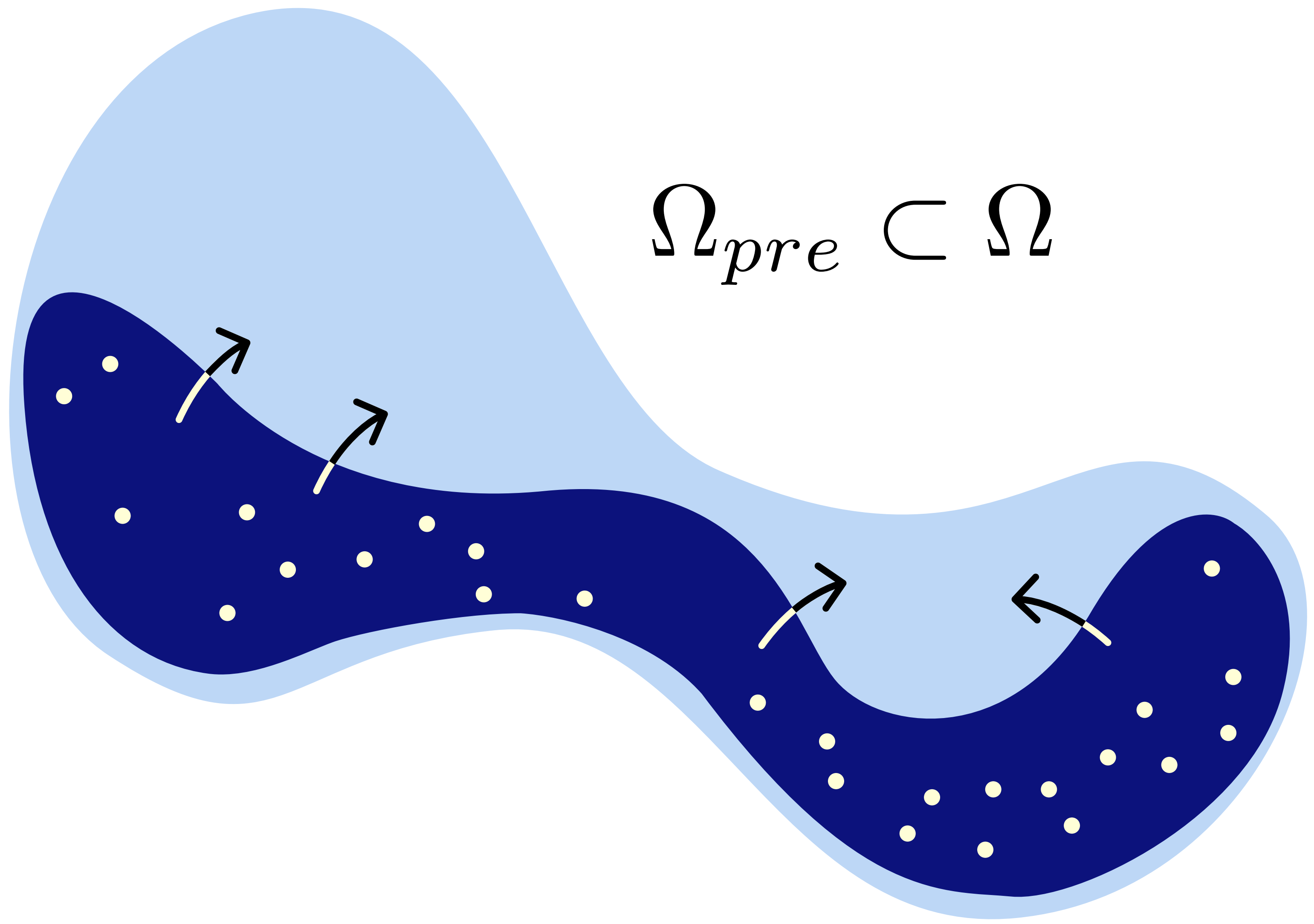
Takeaway

Discovery requires the machine to **expand its action space** (i.e., **learn what actions are possible**)

How? Out-of-Distribution Flow Modeling

valid design space Ω

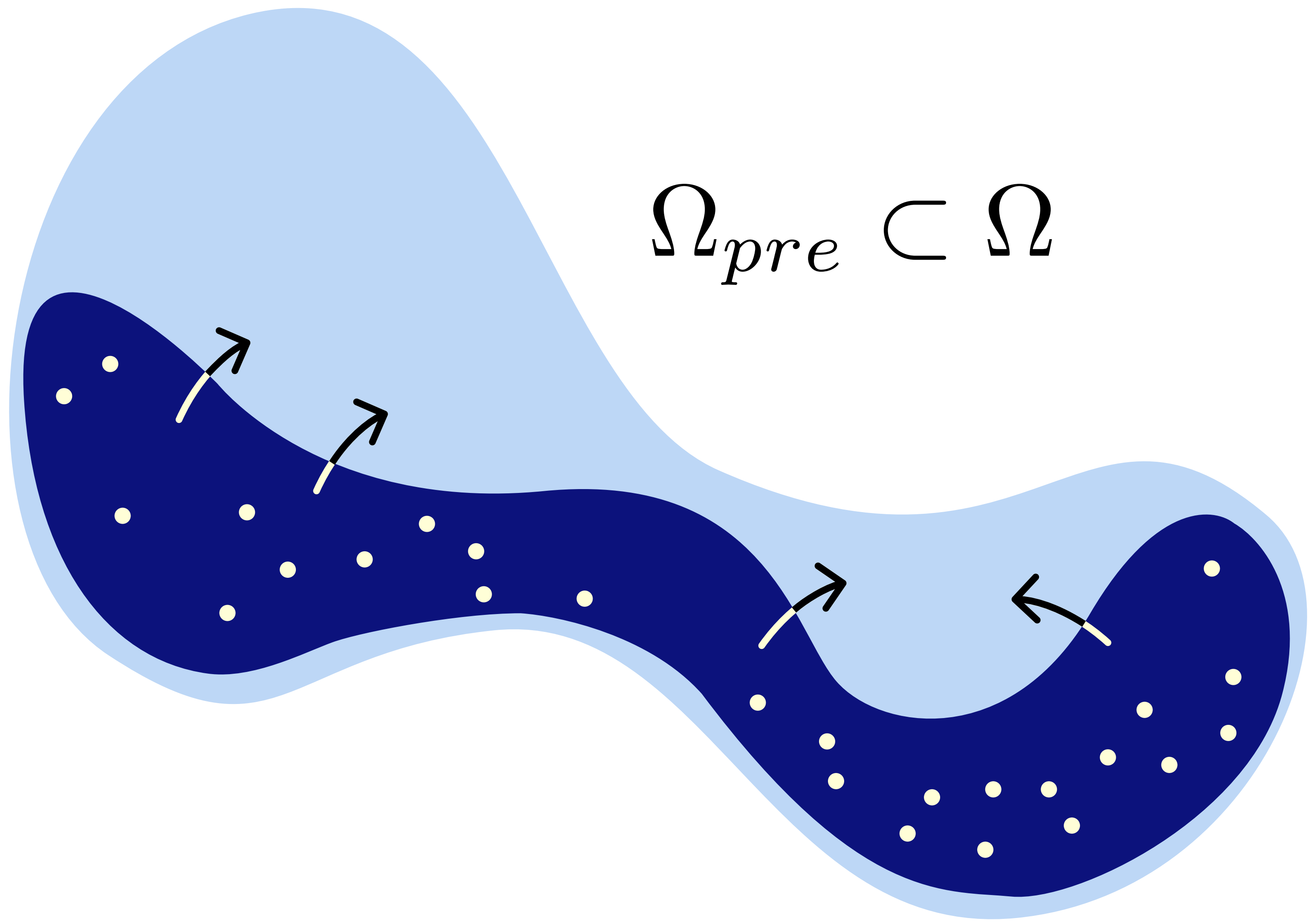
generable set of pre-trained gen. model Ω_{pre}



Discovery requires the machine to **expand its action space** (i.e., **learn what actions are possible**)

How? Out-of-Distribution Flow Modeling

valid design space Ω generable set of pre-trained gen. model Ω_{pre}

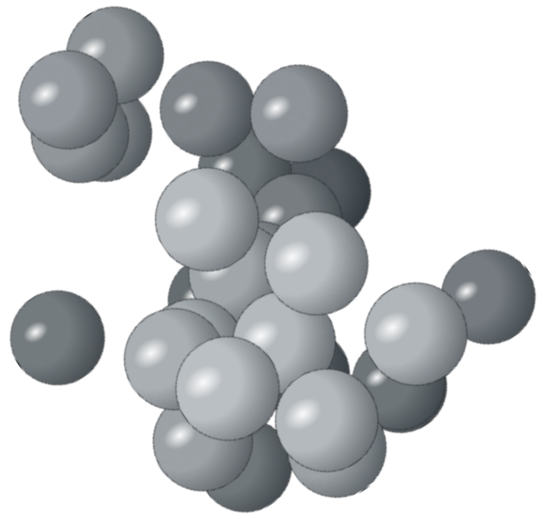


$\Omega_{pre} \subset \Omega$

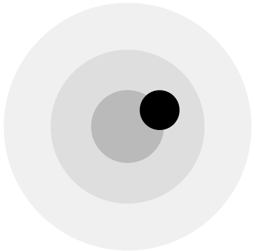
1. Formalize OOD generative modeling
2. How? Active Flow Expansion (ActFlow)
3. Guarantees for OOD generative modeling
4. Results on chemical and biological data

Discovery requires the machine to **expand its action space** (i.e., **learn what actions are possible**)

Diffusion and Flow Modeling: Generation



noise
distribution

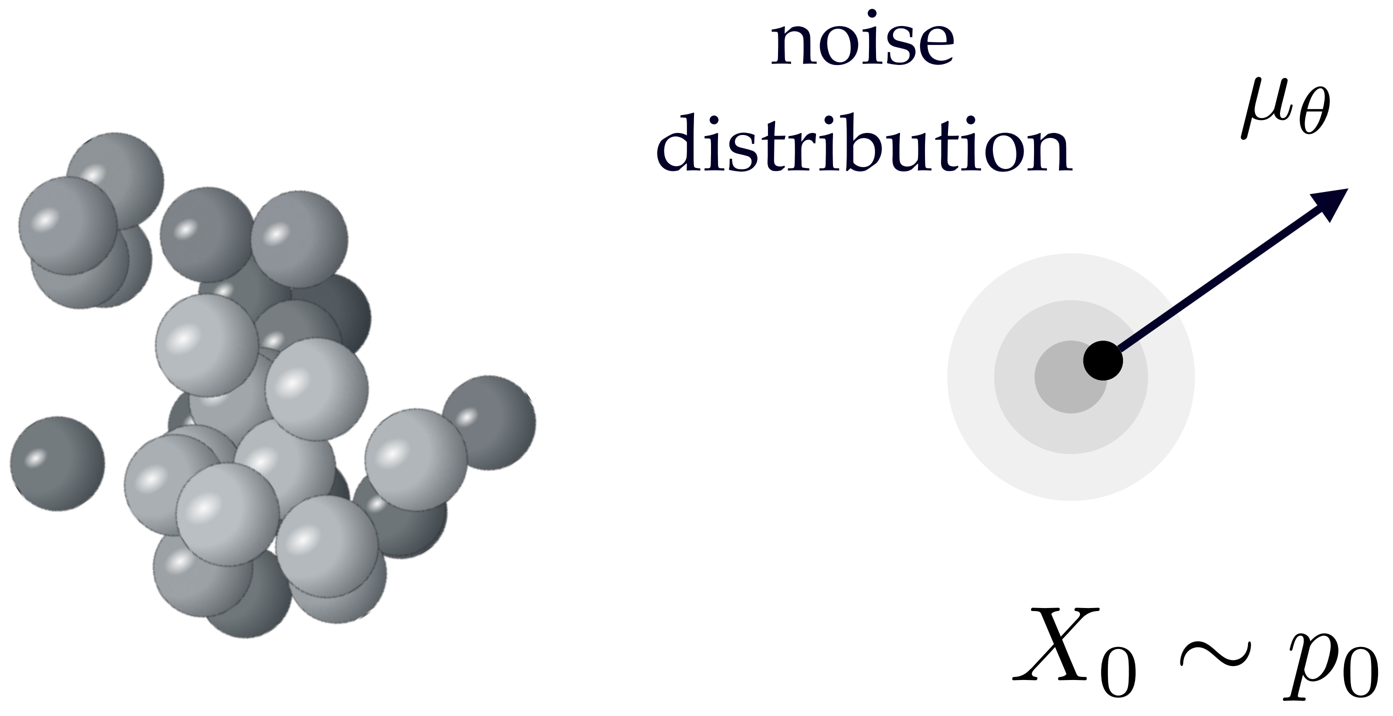


$X_0 \sim p_0$

Generation via learned vector field

Sample noise $X_0 \sim p_0$ $\longrightarrow$ Follow learned vector field μ_θ

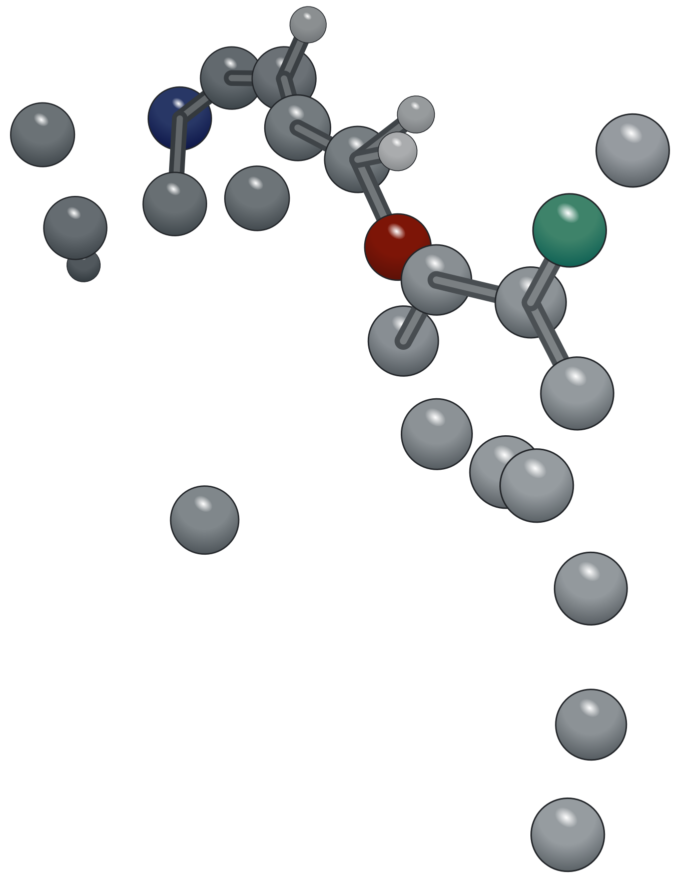
Diffusion and Flow Modeling: Generation

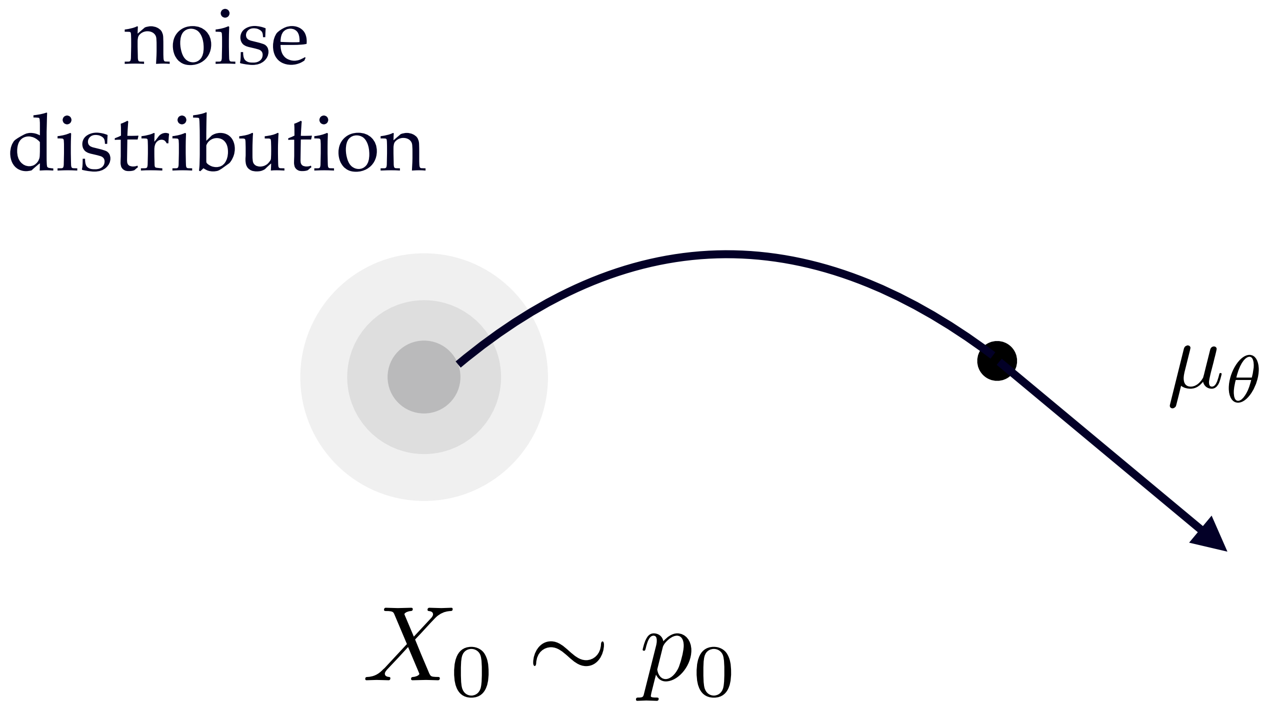


Generation via learned vector field

Sample noise $X_0 \sim p_0$ $\longrightarrow$ Follow learned vector field μ_θ

Diffusion and Flow Modeling: Generation

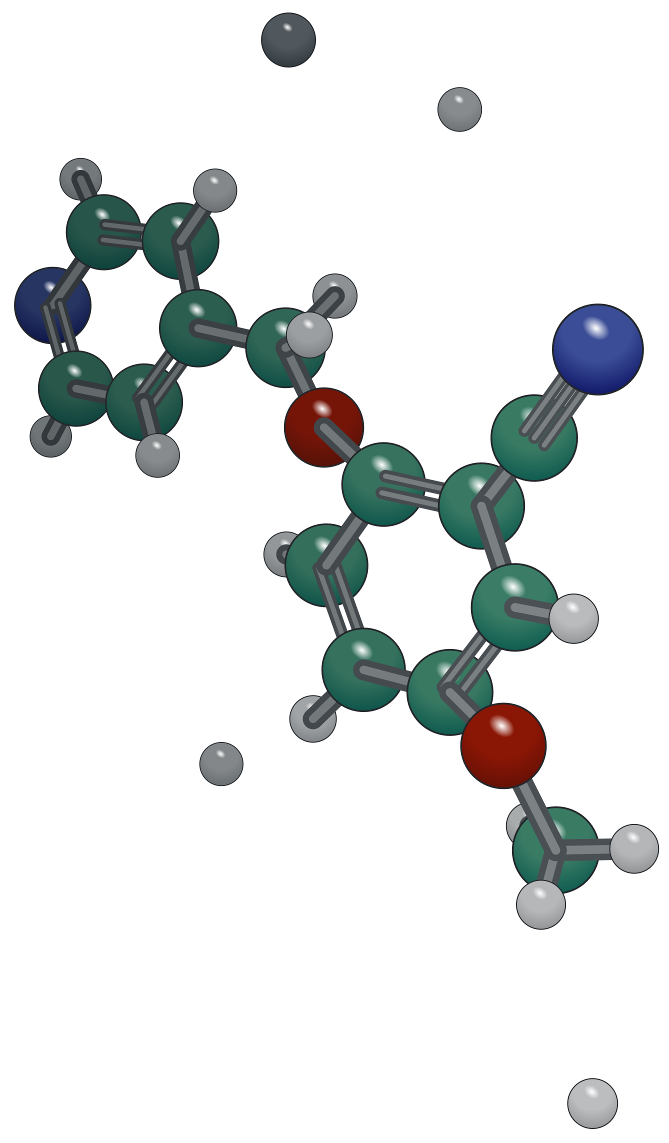


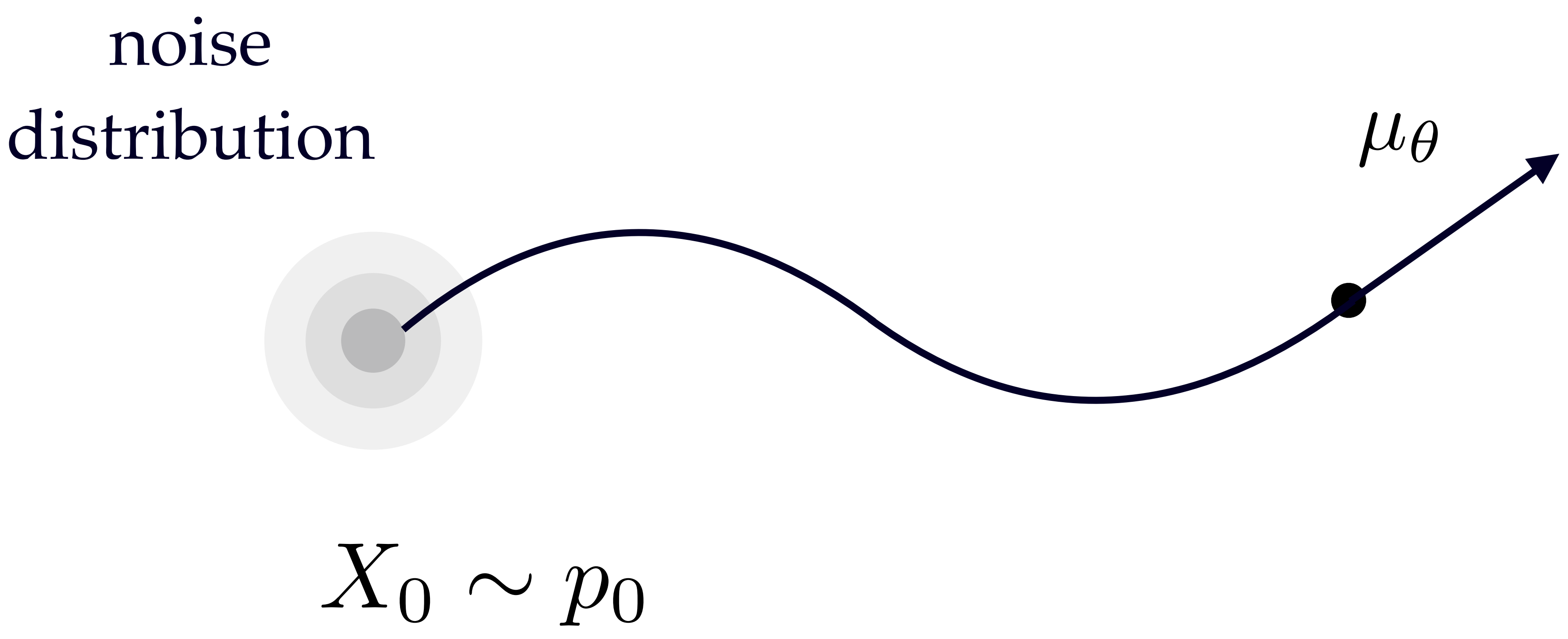


Generation via learned vector field

Sample noise $X_0 \sim p_0$ $\longrightarrow$ Follow learned vector field μ_θ

Diffusion and Flow Modeling: Generation

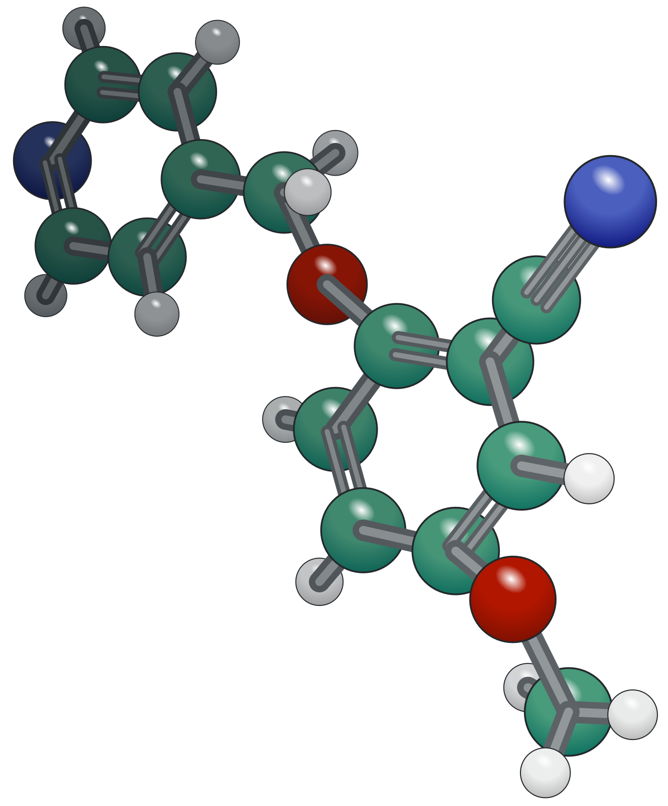


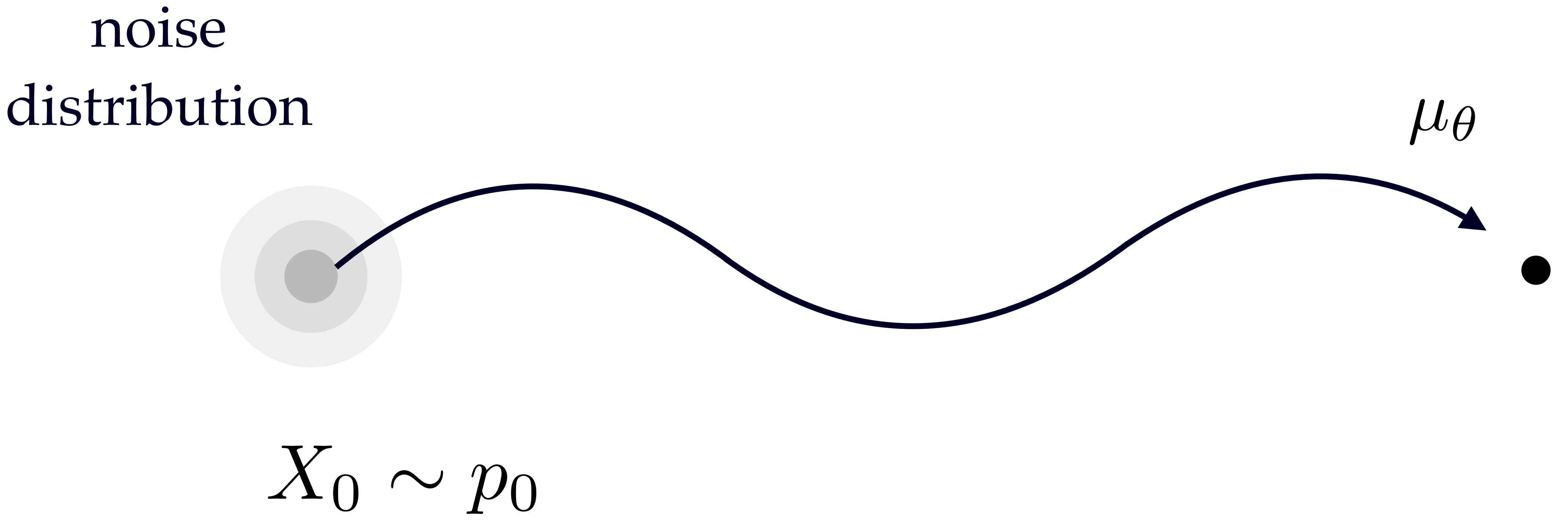


Generation via learned vector field

Sample noise $X_0 \sim p_0$ $\longrightarrow$ Follow learned vector field μ_θ

Diffusion and Flow Modeling: Generation

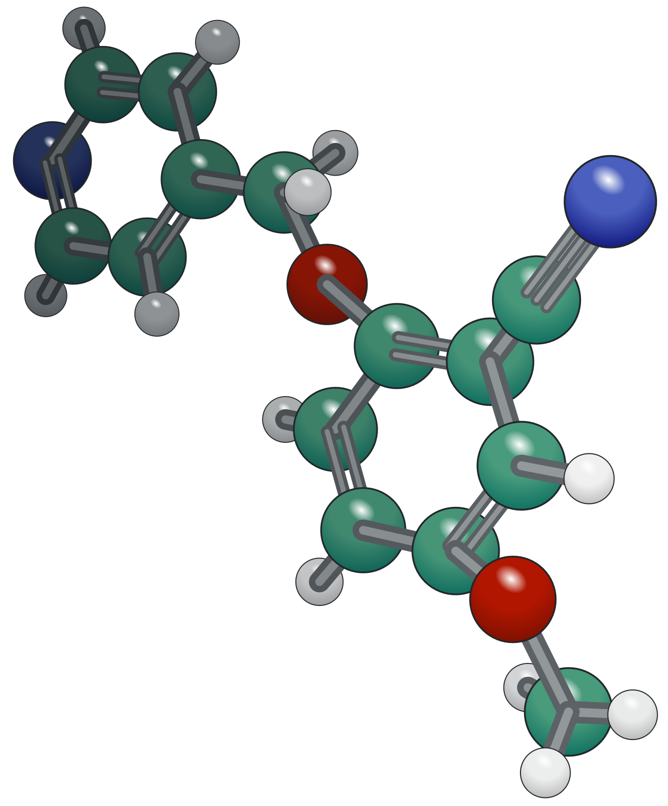


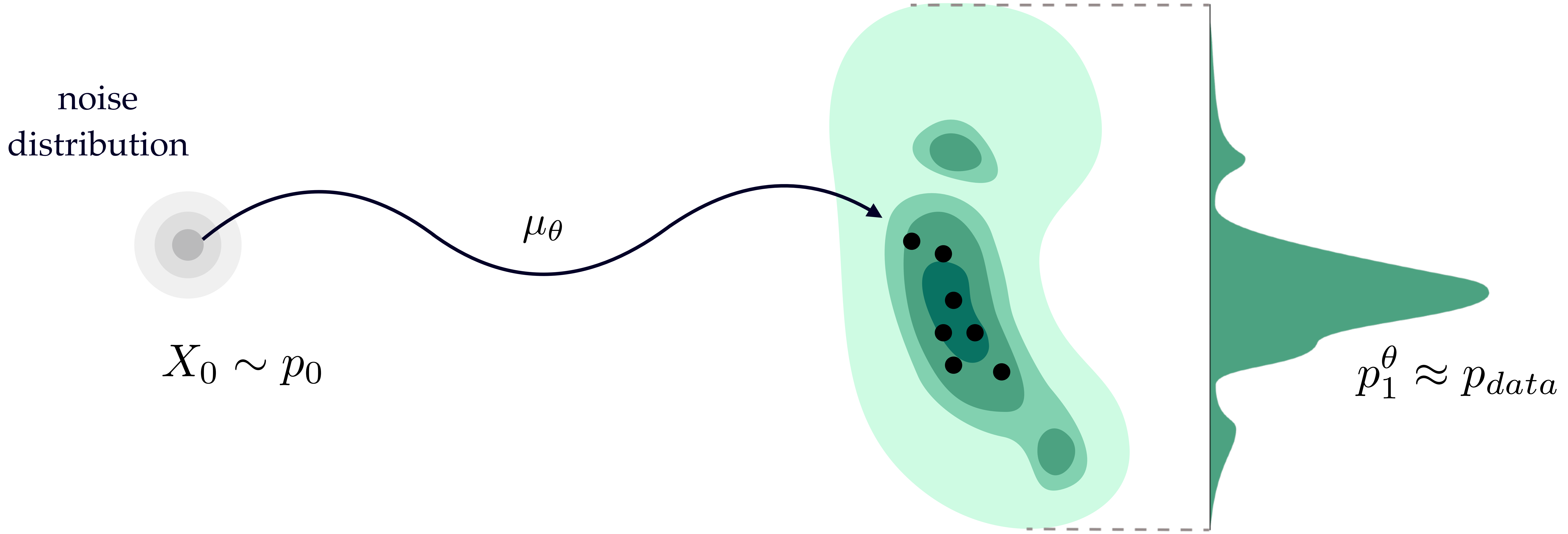


Generation via learned vector field

Sample noise $X_0 \sim p_0$ $\longrightarrow$ Follow learned vector field μ_θ

Diffusion and Flow Modeling: Generation

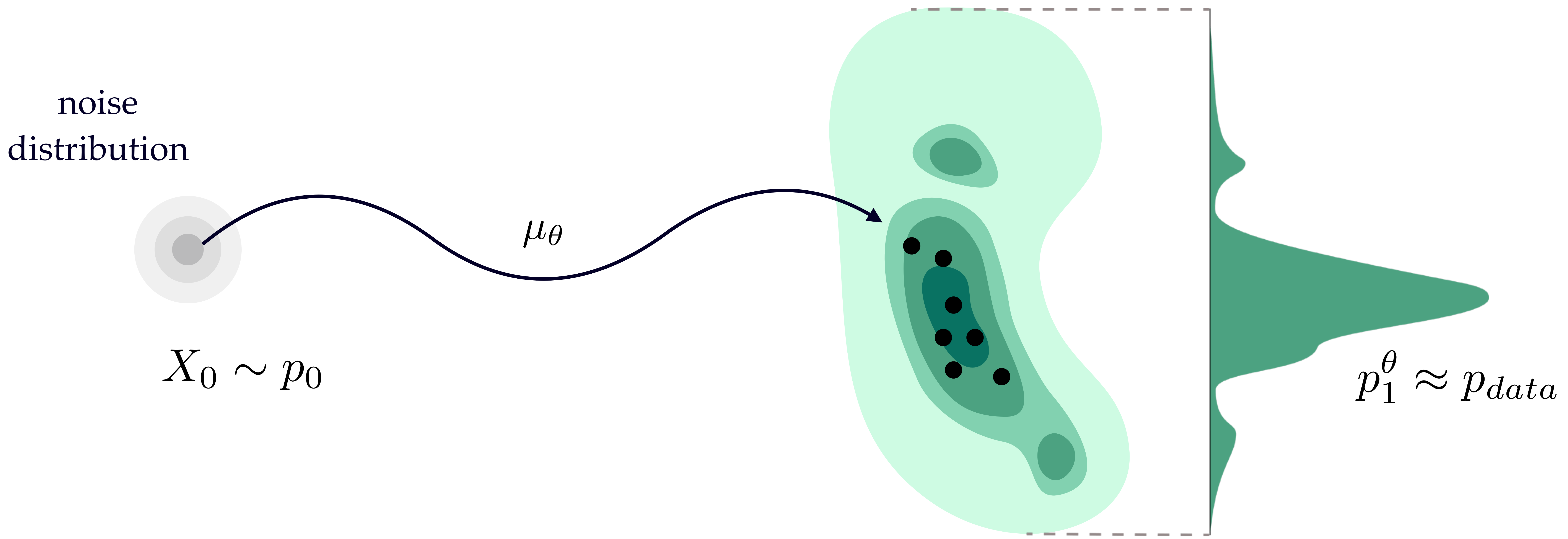




Generation via learned vector field

Sample noise $X_0 \sim p_0$ $\longrightarrow$ Follow learned vector field μ_θ

Diffusion and Flow Modeling: Pre-Training

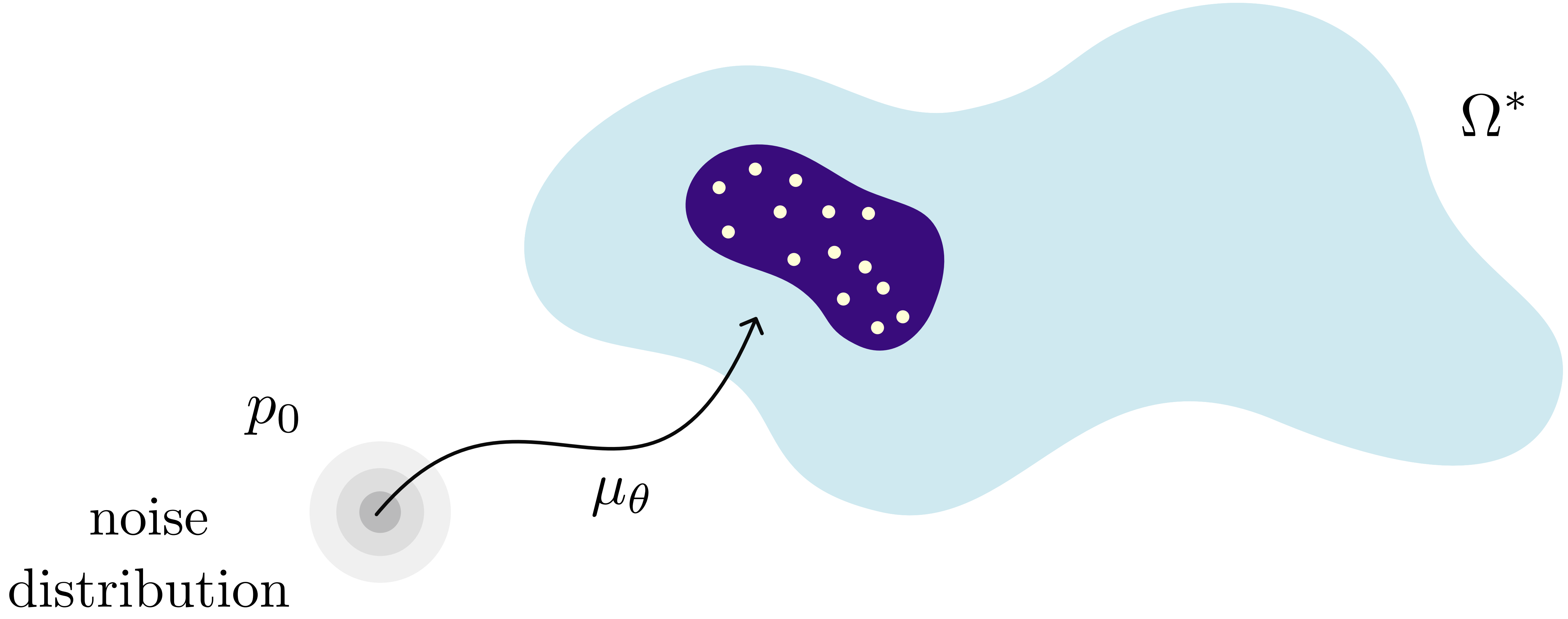


Distribution Matching: Standard Flow Pre-training

learn θ s.t. $p_1^\theta \approx p_{data}$

Implement via e.g., Flow/Score Matching

Flow Pre-training for Discovery: Key Limitation



Distribution Matching: Standard Flow Pre-training

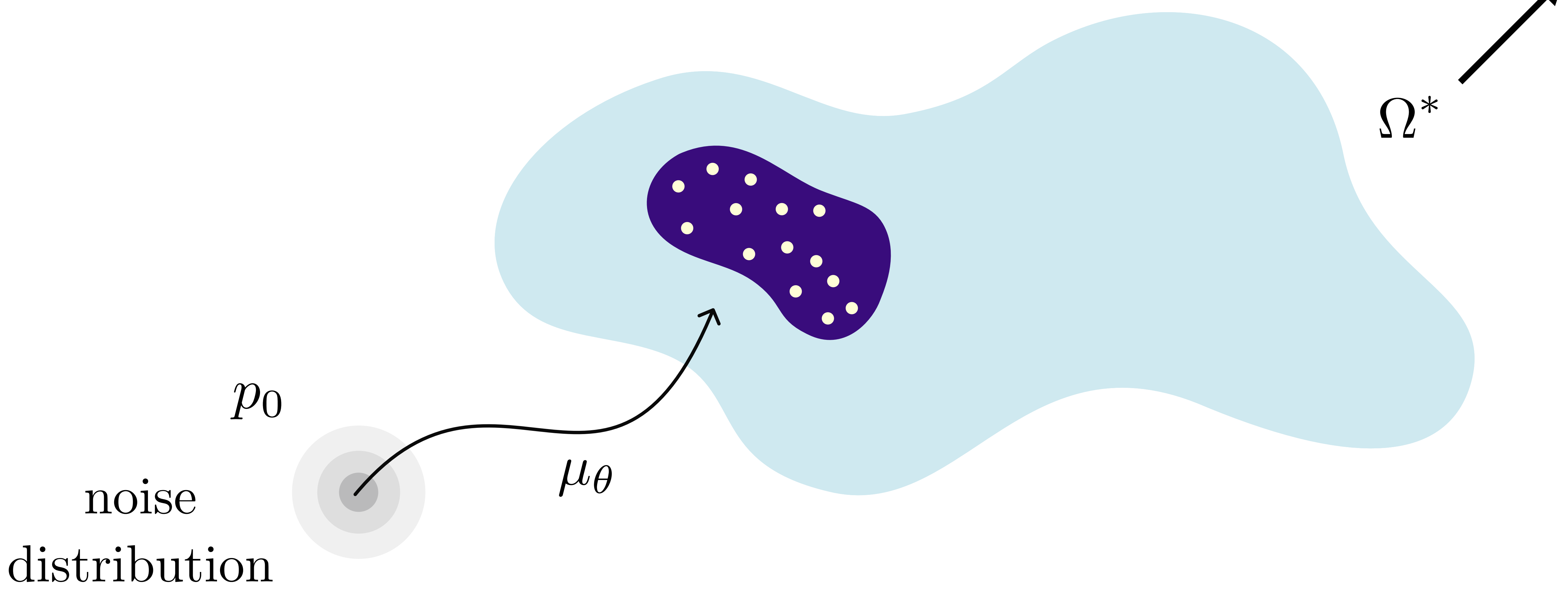
learn θ s.t. $p_1^\theta \approx p_{data}$

Flow Pre-training for Discovery: Key Limitation

Valid design space

$$\Omega^* := \{x \in \mathcal{X} : v(x) = 1\}$$

Unknown laws of nature (e.g., chemistry, biology)



Distribution Matching: Standard Flow Pre-training

learn θ s.t. $p_1^\theta \approx p_{data}$

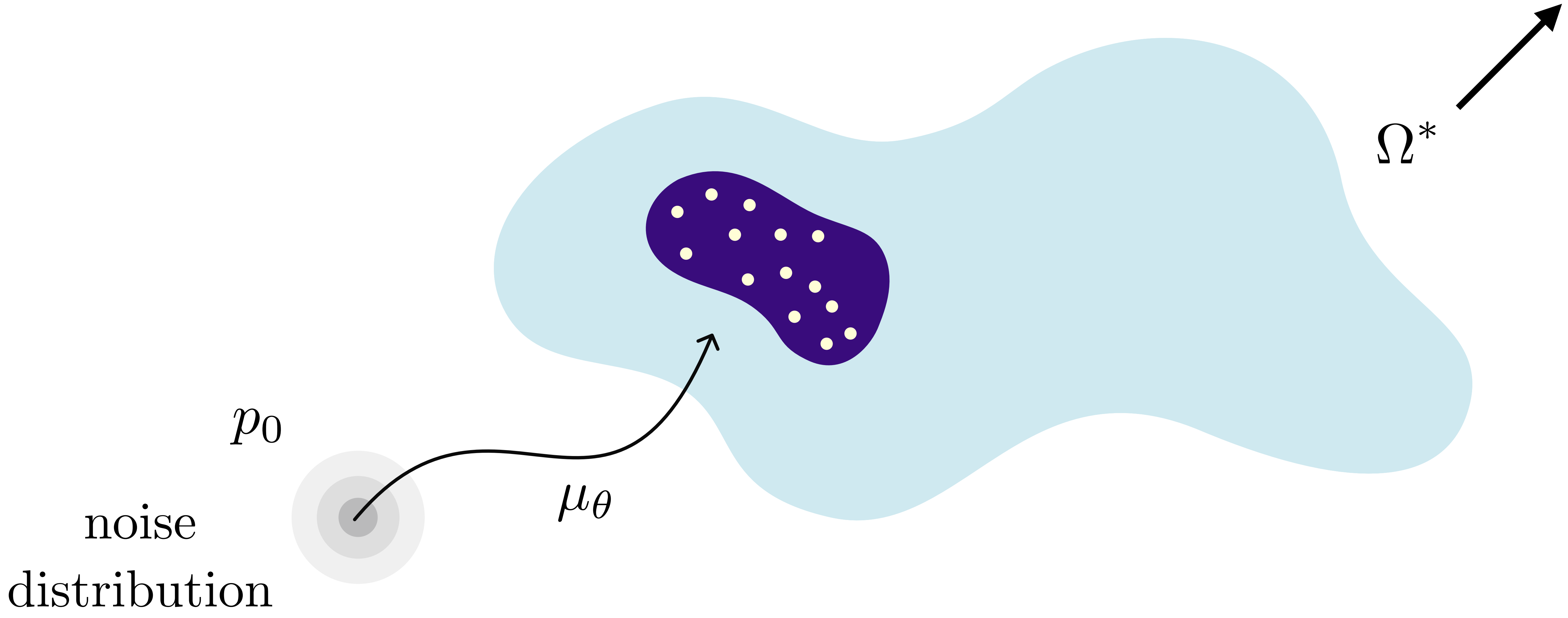
Flow Pre-training for Discovery: Key Limitation

How to formalize limited coverage?

Valid design space

$$\Omega^* := \{x \in \mathcal{X} : v(x) = 1\}$$

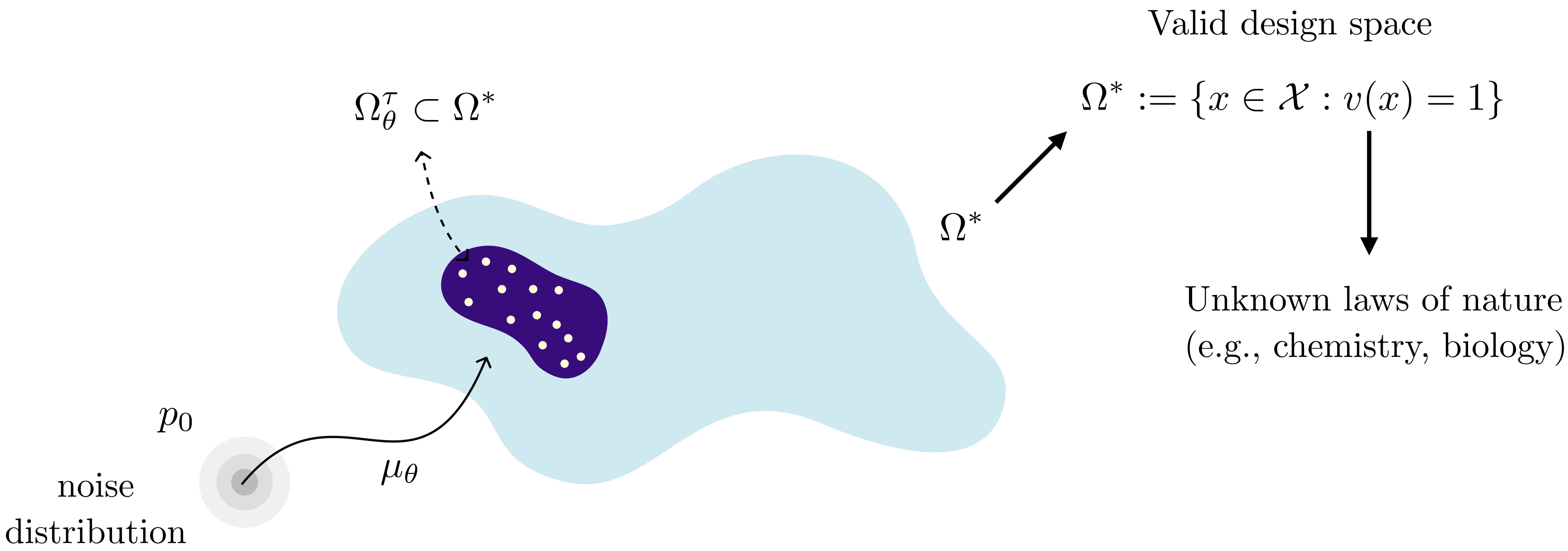
Unknown laws of nature
(e.g., chemistry, biology)



Distribution Matching: Standard Flow Pre-training

learn θ s.t. $p_1^\theta \approx p_{data}$

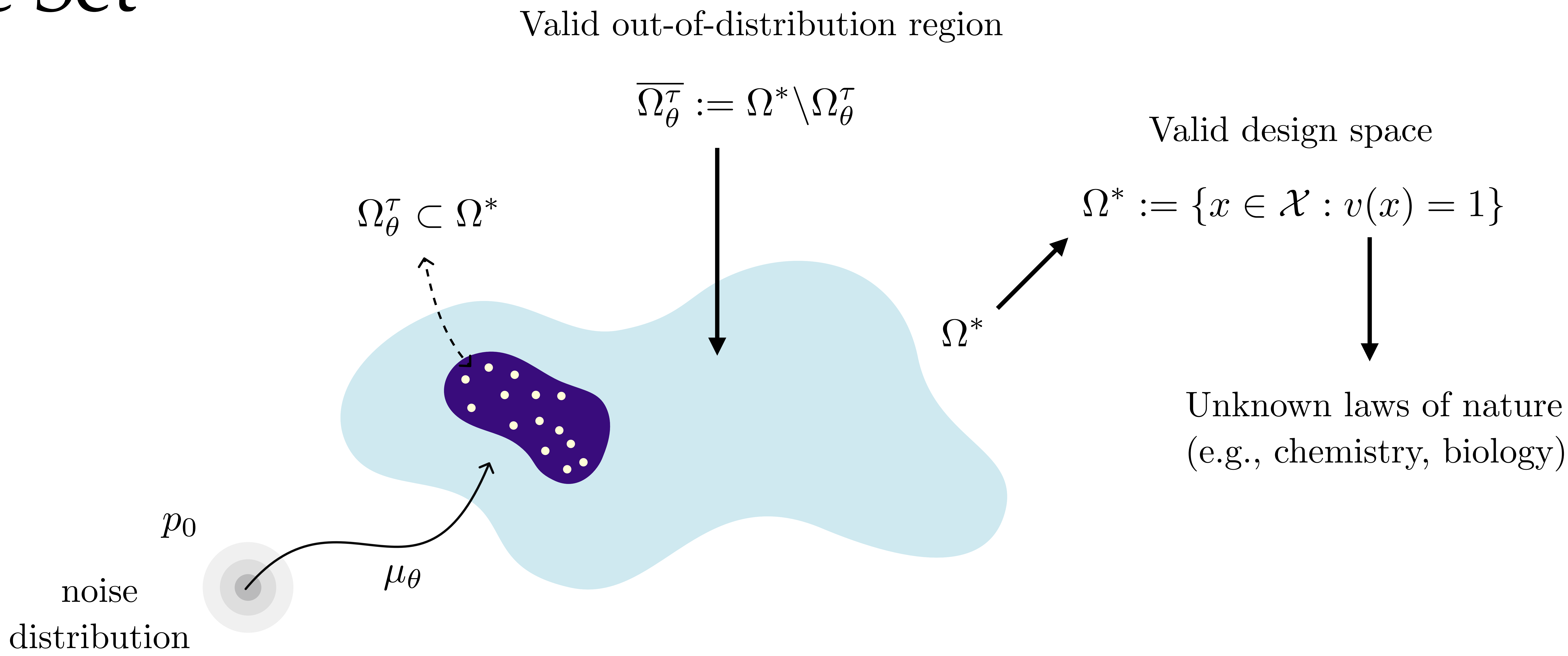
Generable Set



Definition (Generable Set). Let μ_θ inducing density p_1^θ. For $\tau > 0$, we define its τ-level generable set as:

$$\Omega_\theta^\tau := \{x \in \mathcal{X} : p_1^\theta(x) \geq \tau\}$$

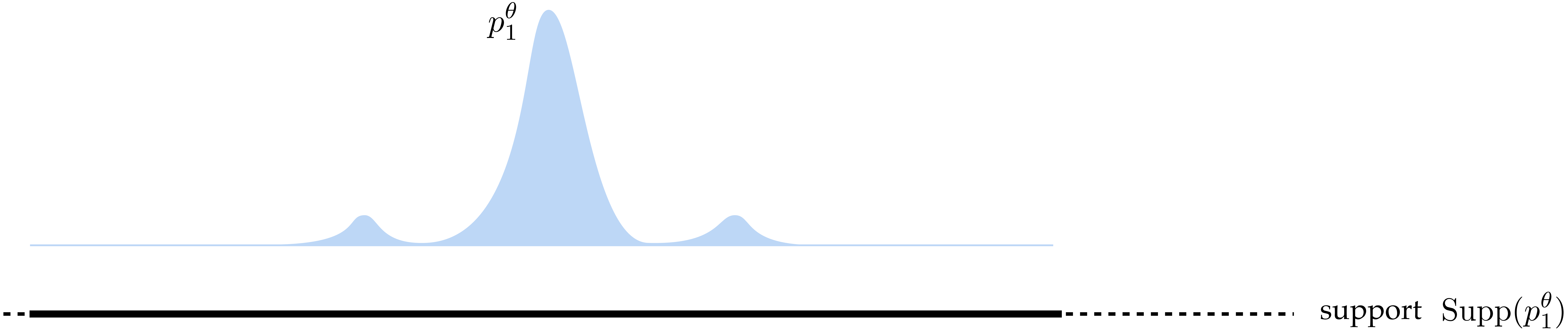
Generable Set



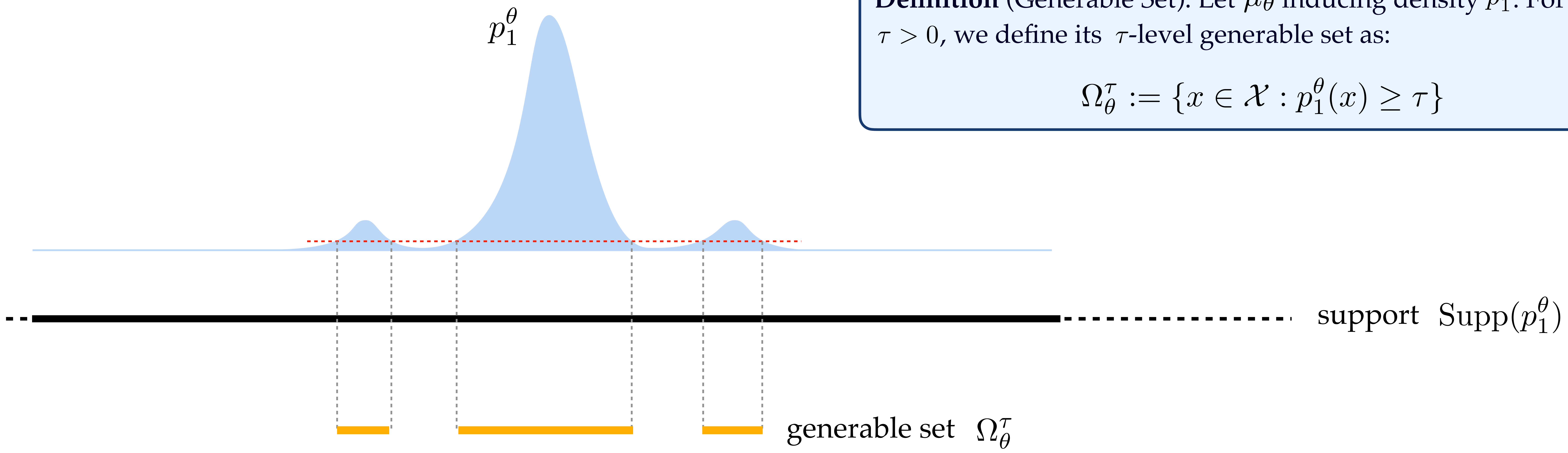
Definition (Generable Set). Let μ_θ inducing density p_1^θ. For $\tau > 0$, we define its τ-level generable set as:

$$\Omega_\theta^\tau := \{x \in \mathcal{X} : p_1^\theta(x) \geq \tau\}$$

Generable Set vs Support



Generable Set vs Support



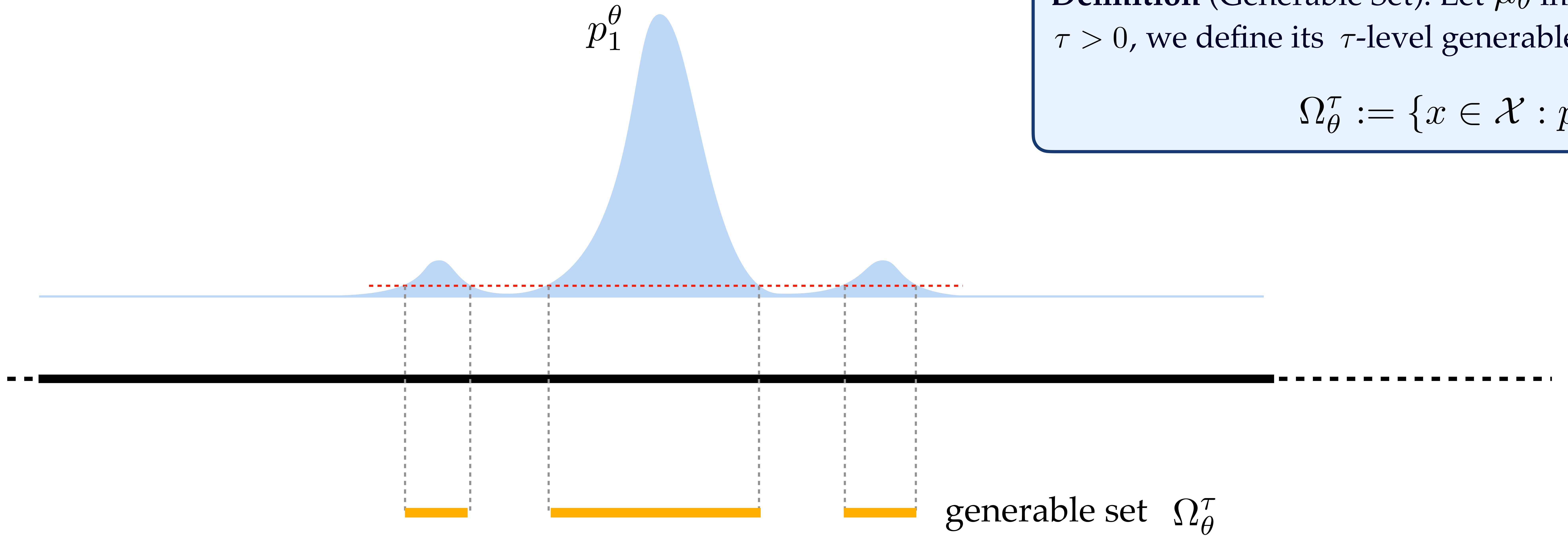
Definition (Generable Set). Let μ_θ inducing density p_1^θ. For $\tau > 0$, we define its τ-level generable set as:

$$\Omega_\theta^\tau := \{x \in \mathcal{X} : p_1^\theta(x) \geq \tau\}$$

Generable Set vs Support

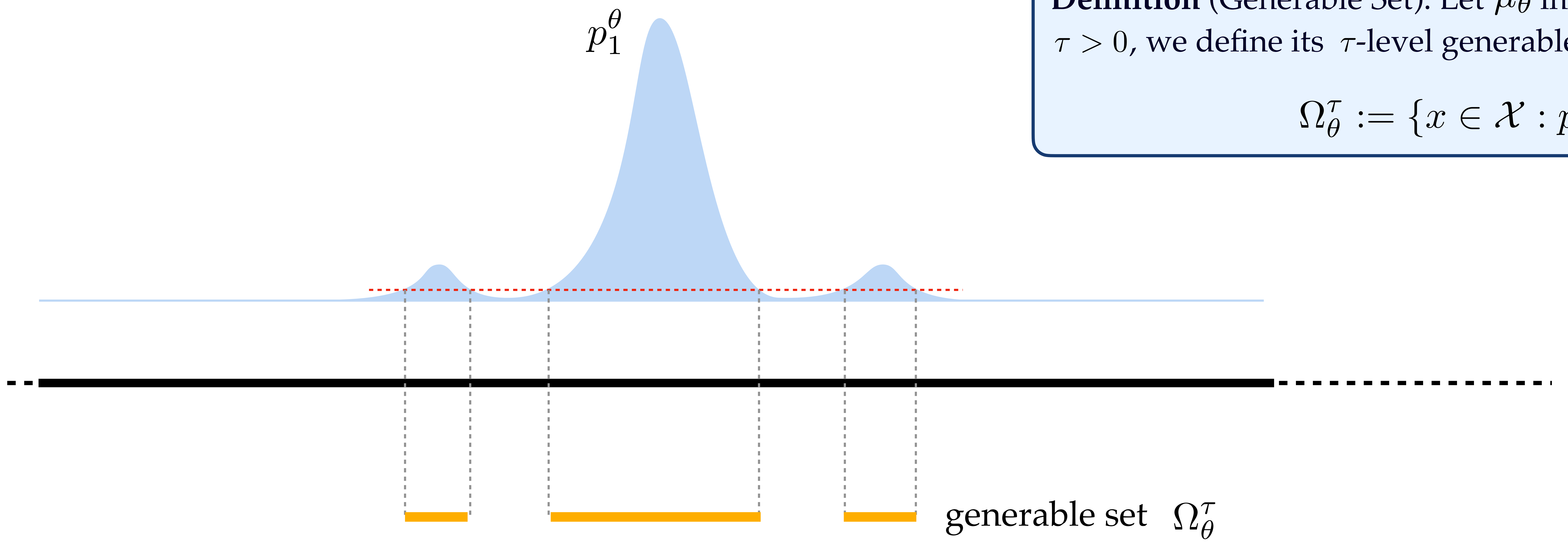
Definition (Generable Set). Let μ_θ inducing density p_1^θ. For $\tau > 0$, we define its τ-level generable set as:

$$\Omega_\theta^\tau := \{x \in \mathcal{X} : p_1^\theta(x) \geq \tau\}$$



$$\Omega_\theta^\tau \xrightarrow{\tau \to 0} \mathrm{Supp}(p_1^\theta)$$

Generable Set vs Support

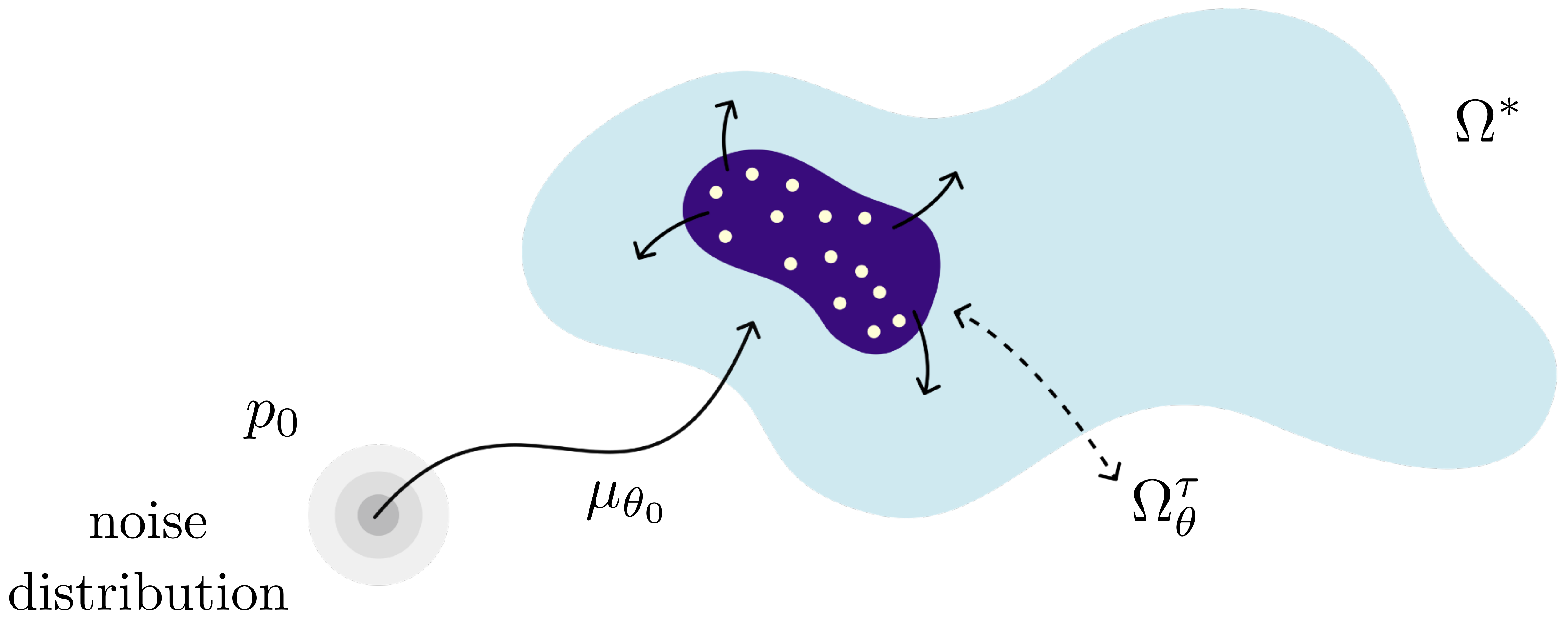


Definition (Generable Set). Let μ_θ inducing density p_1^θ. For $\tau > 0$, we define its τ-level generable set as:

$$\Omega_\theta^\tau := \{x \in \mathcal{X} : p_1^\theta(x) \geq \tau\}$$

generable set
- region likely to be covered by a finite draw
- region where the generative model is reliable (e.g., sufficient validity signal)

Generable Set Expansion for OOD Flow Modeling

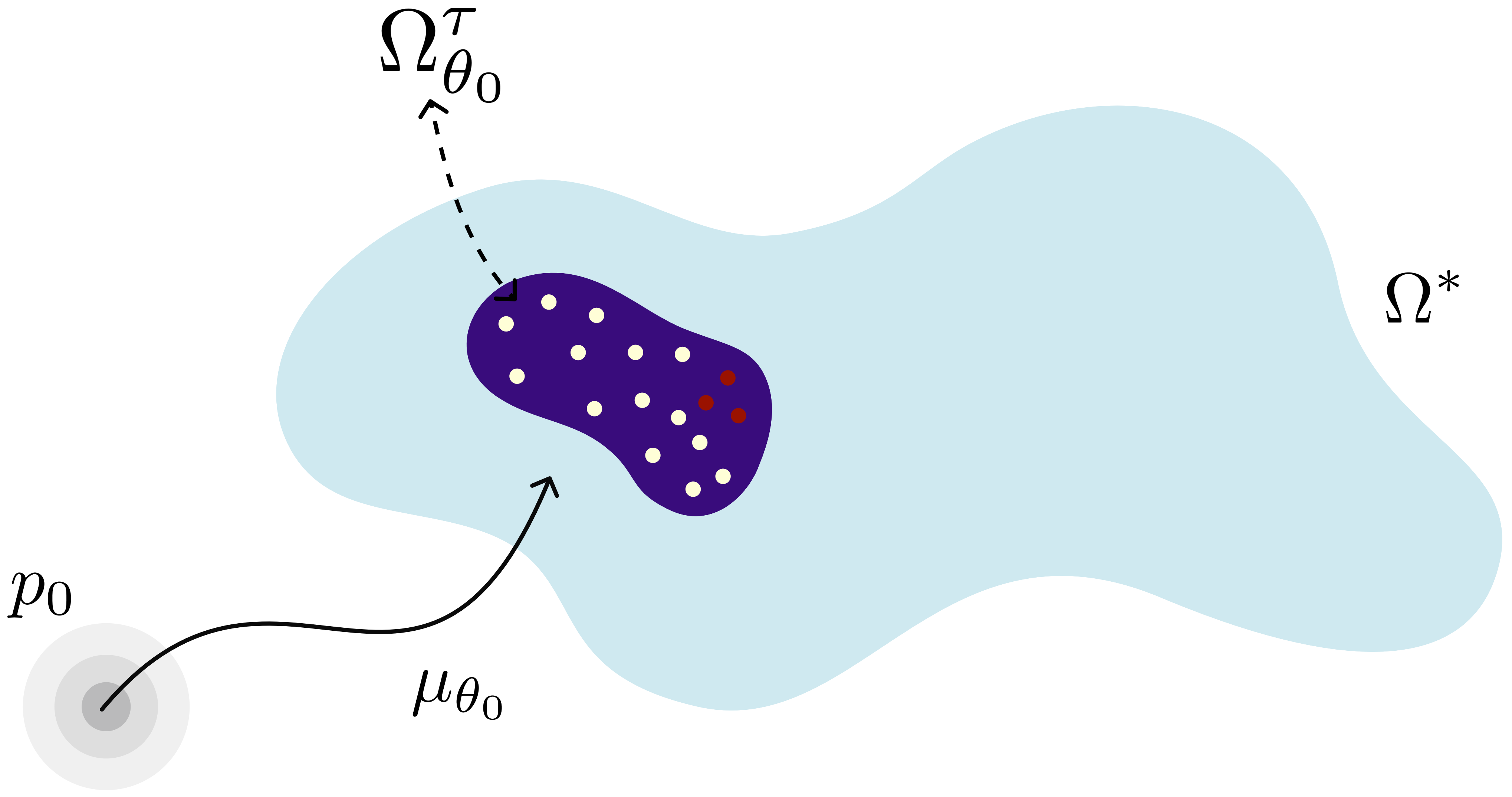


Requires acquiring information about Ω^* beyond the data

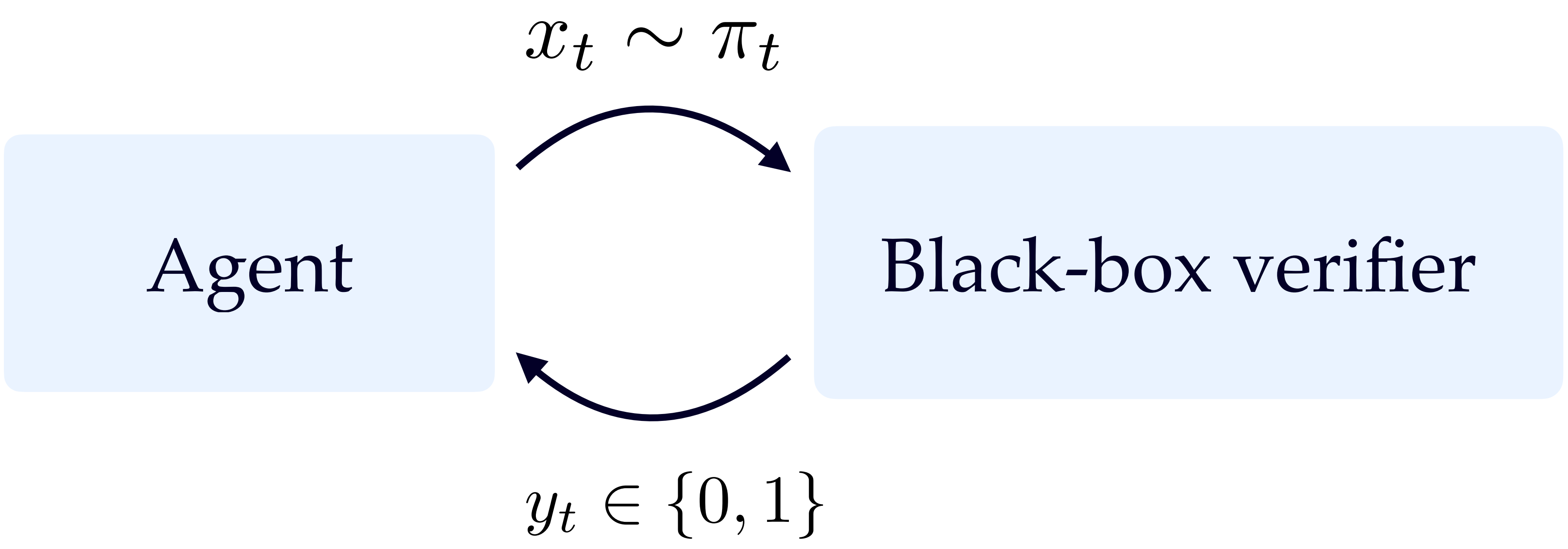
Generable Set Expansion: Out-of-Distribution Flow Modeling

learn θ s.t. $\Omega_\theta^\tau \approx \Omega^*$

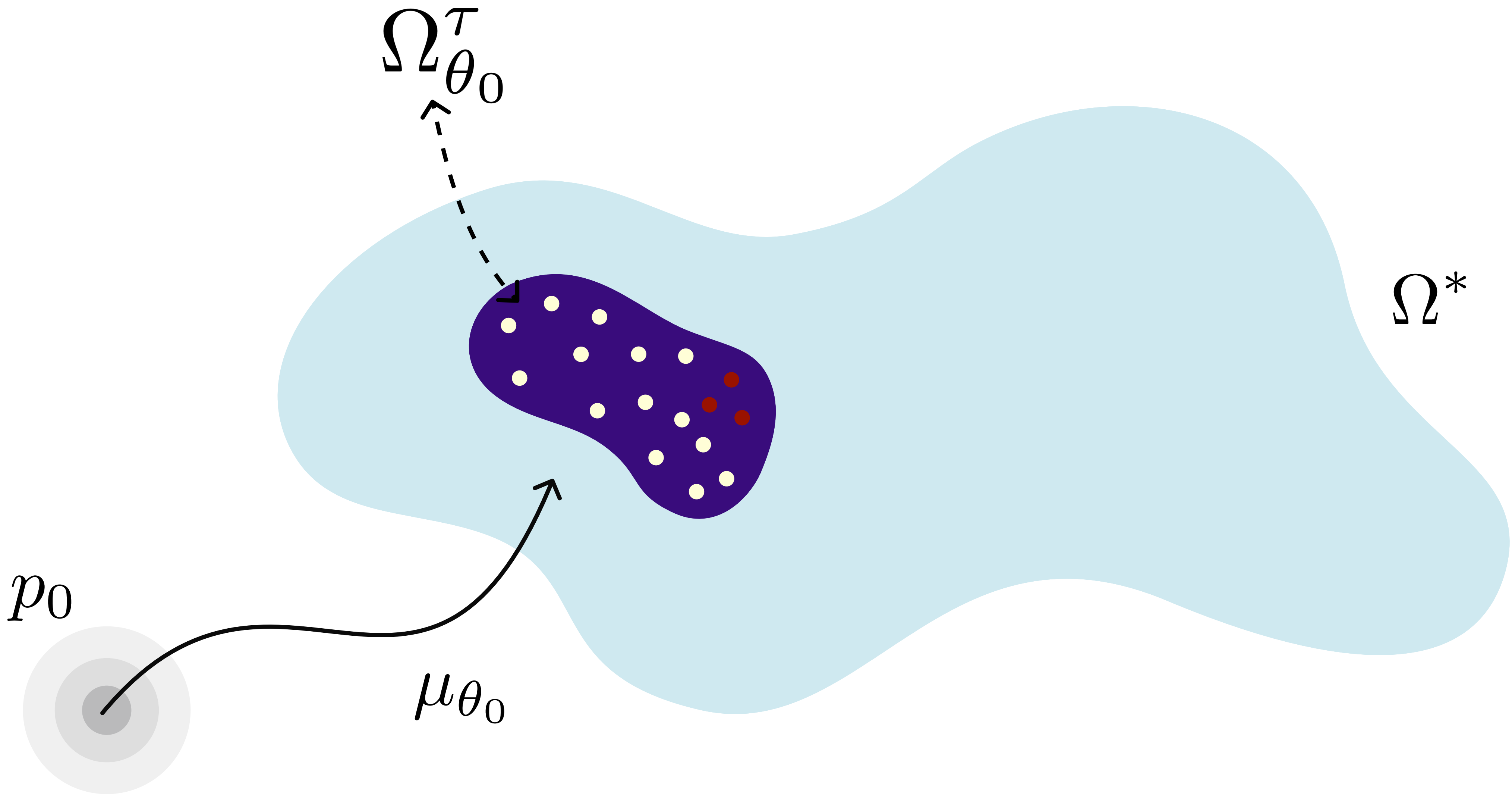
Generable Set Expansion for OOD Flow Modeling



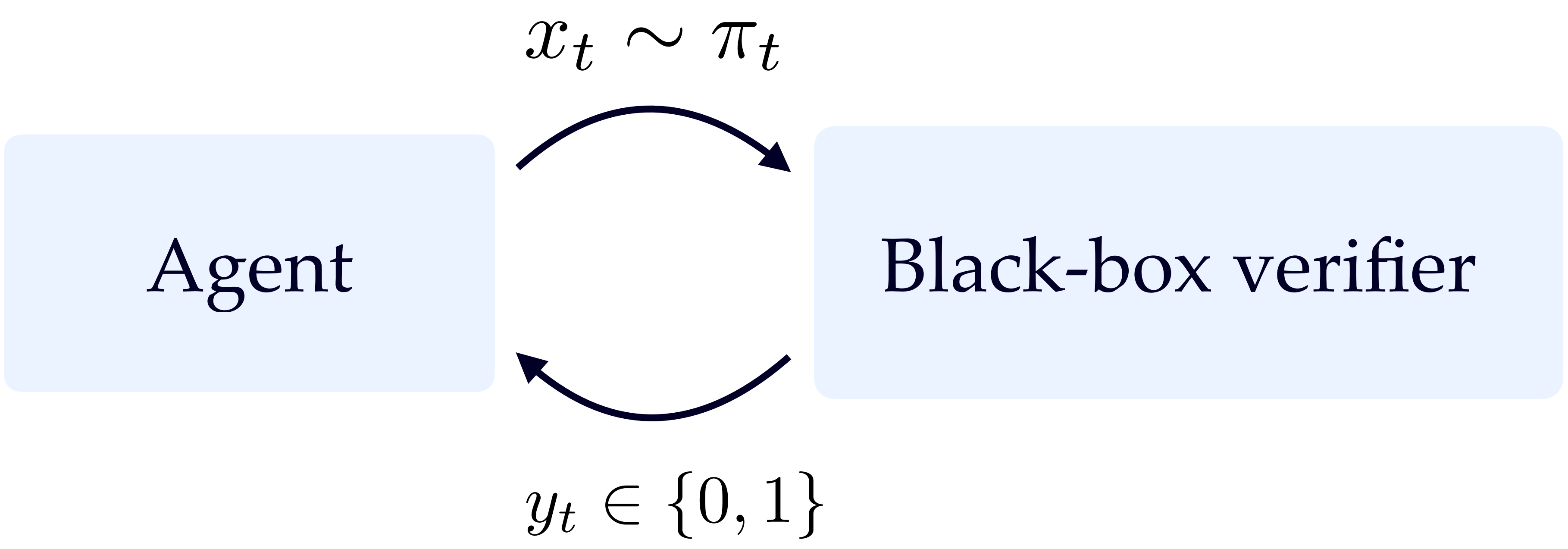
Black-box noisy binary verifier feedback



Generable Set Expansion for OOD Flow Modeling



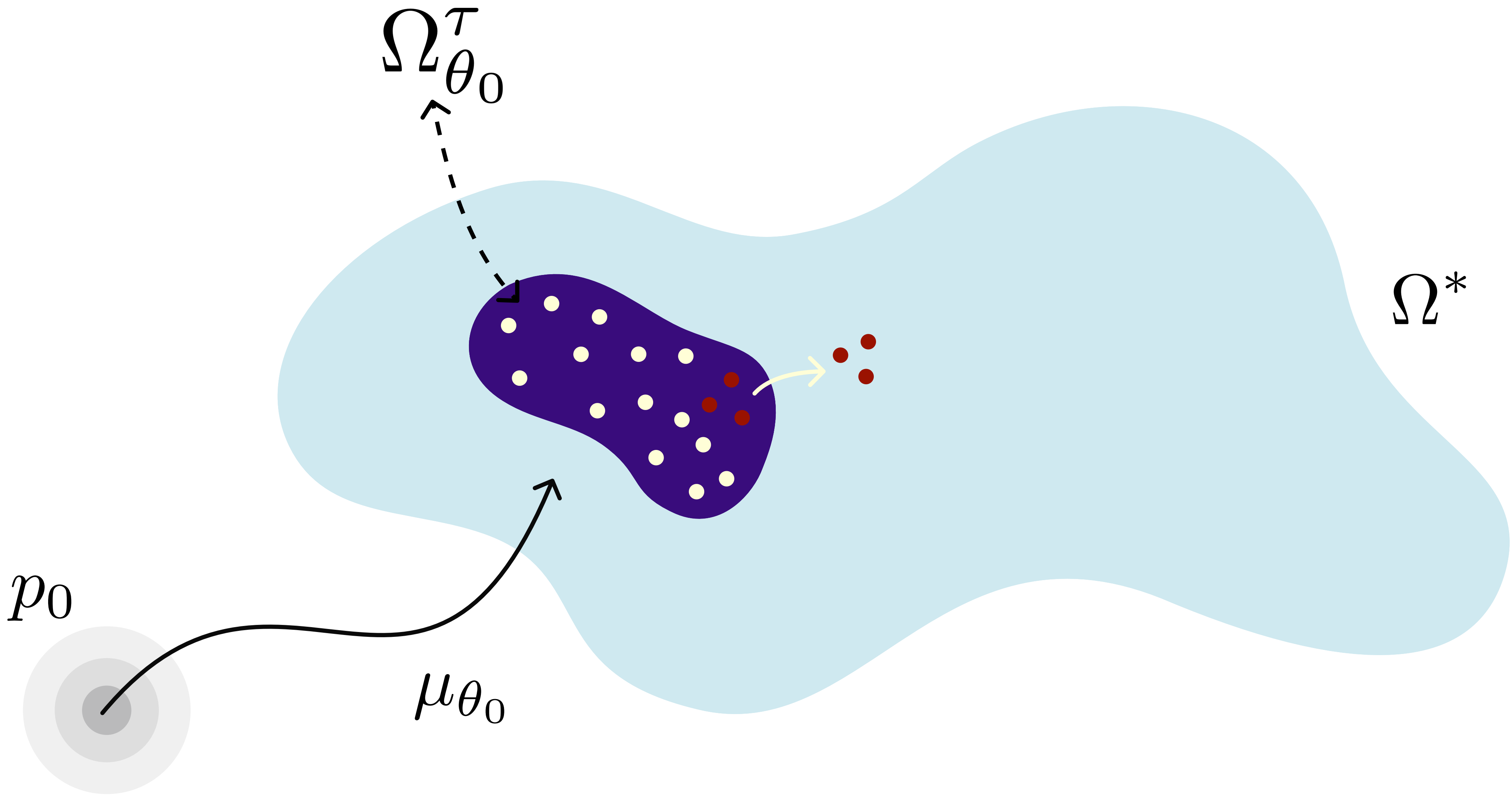
Black-box noisy binary verifier feedback



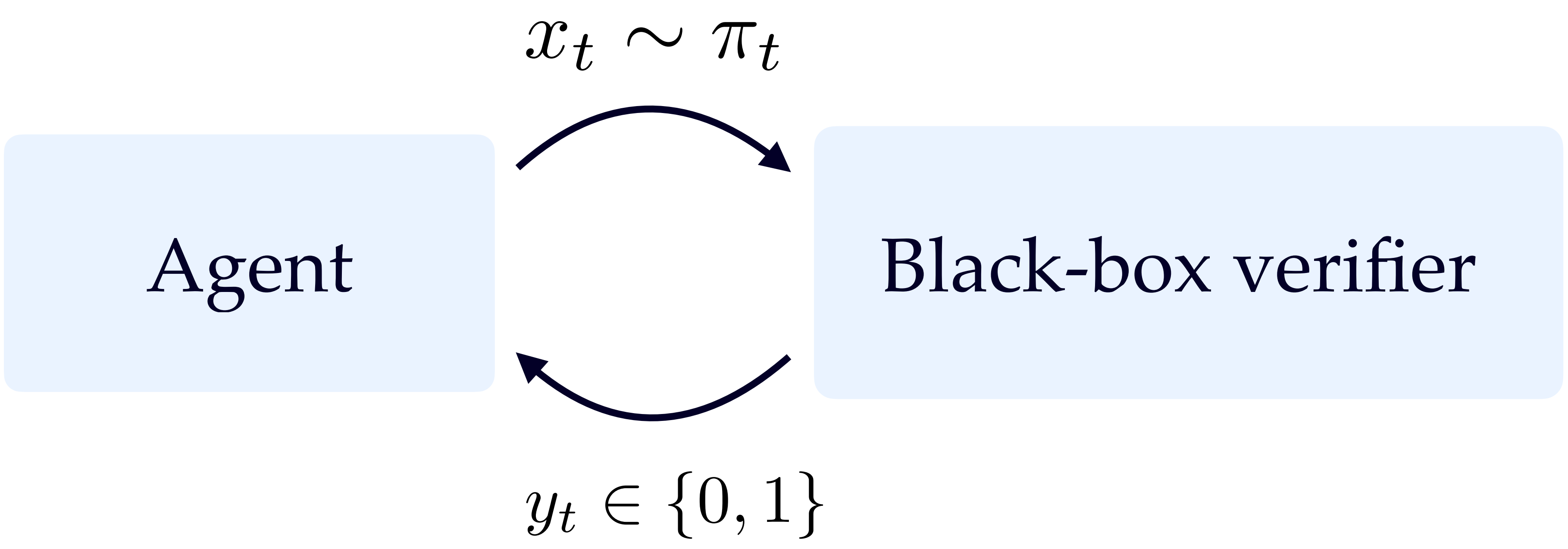
$$\theta^{\text{pre}} = \theta_0$$

$$\Omega_{\theta_0}^{\tau}$$

Generable Set Expansion for OOD Flow Modeling



Black-box noisy binary verifier feedback

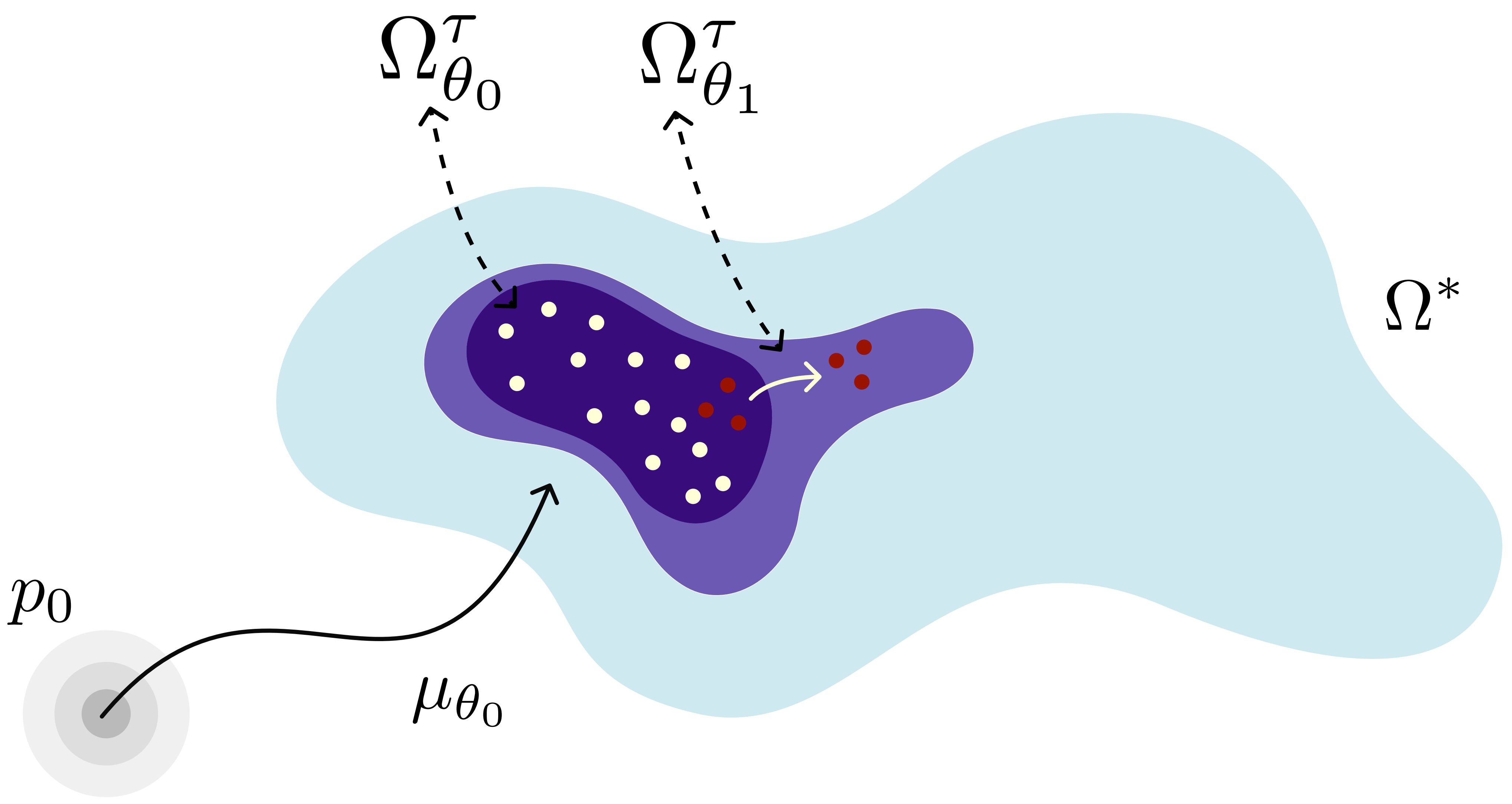


$$\theta^{\text{pre}} = \theta_0$$

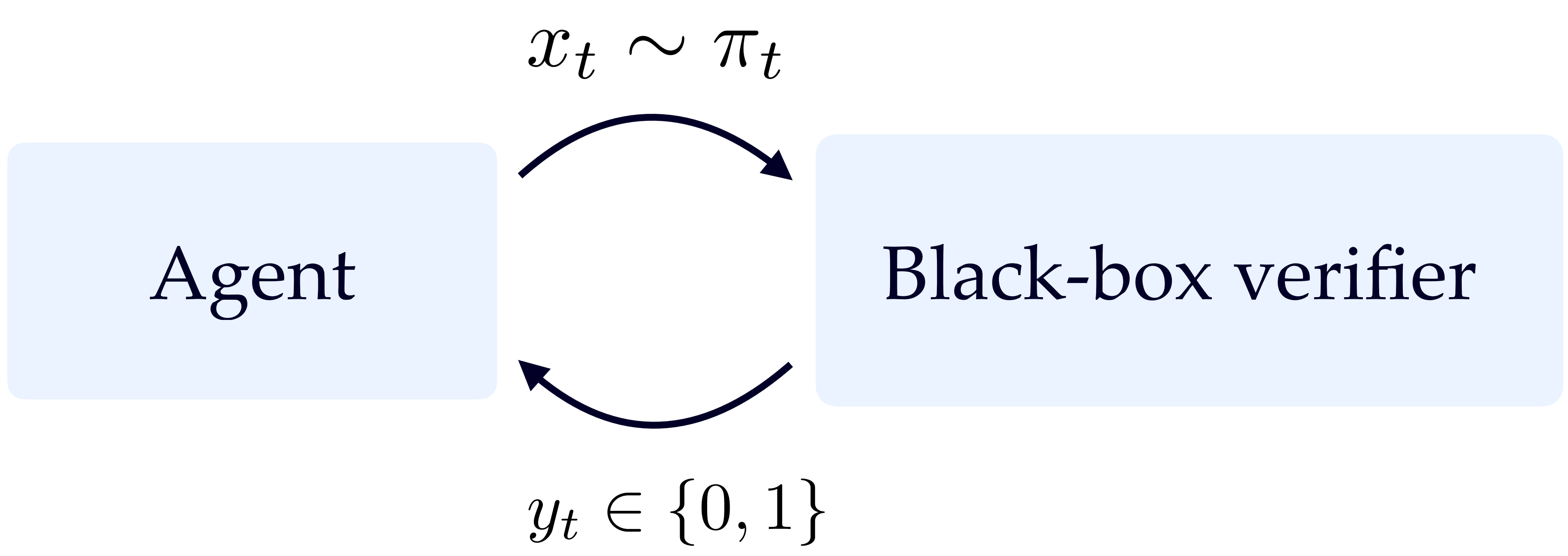
$$\Omega^{\tau}_{\theta_0}$$

Continual expansion via adaptation on valid **frontier samples**

Generable Set Expansion for OOD Flow Modeling



Black-box noisy binary verifier feedback

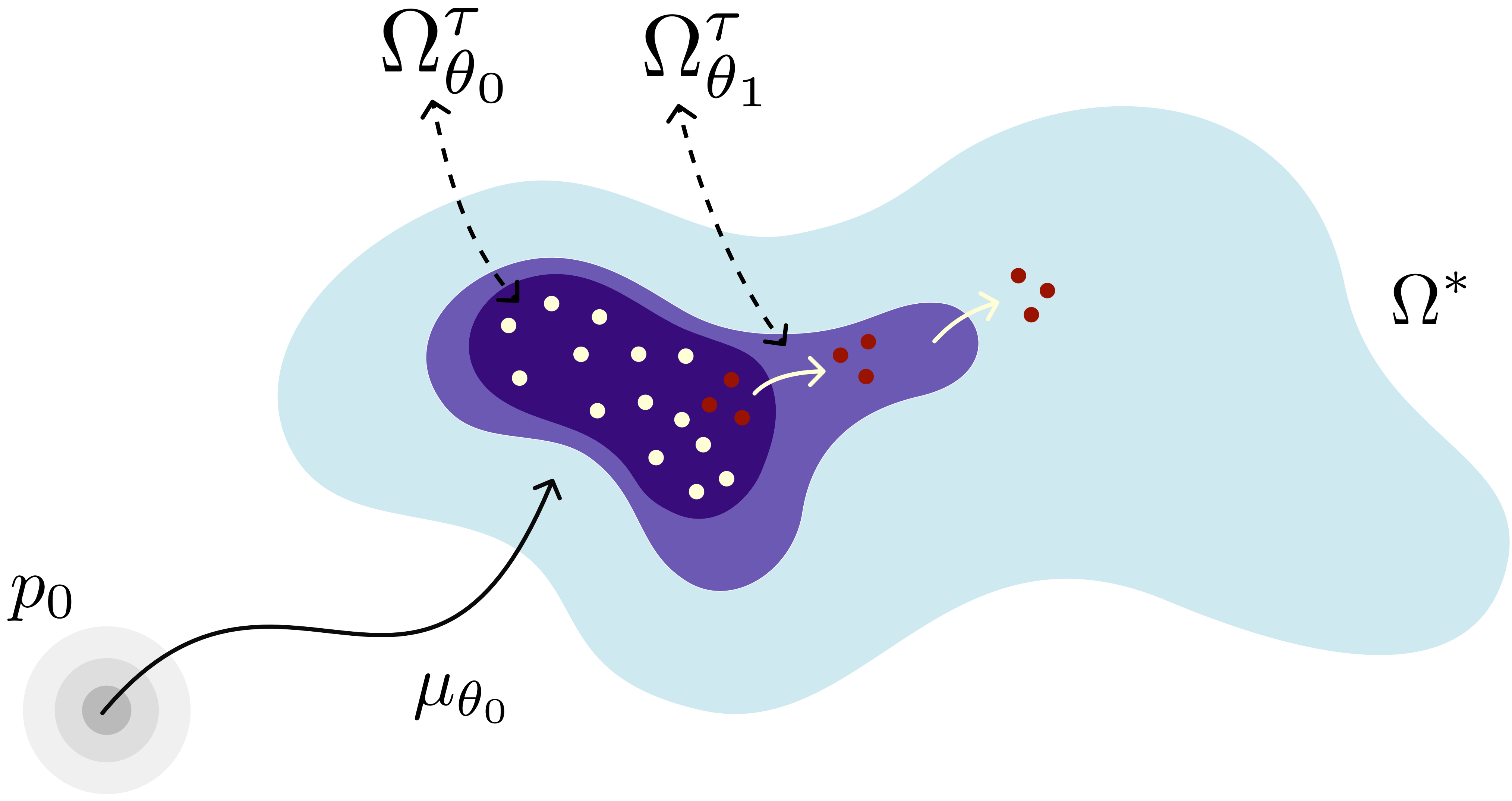


$$\theta^{\text{pre}} = \theta_0 \to \theta_1$$

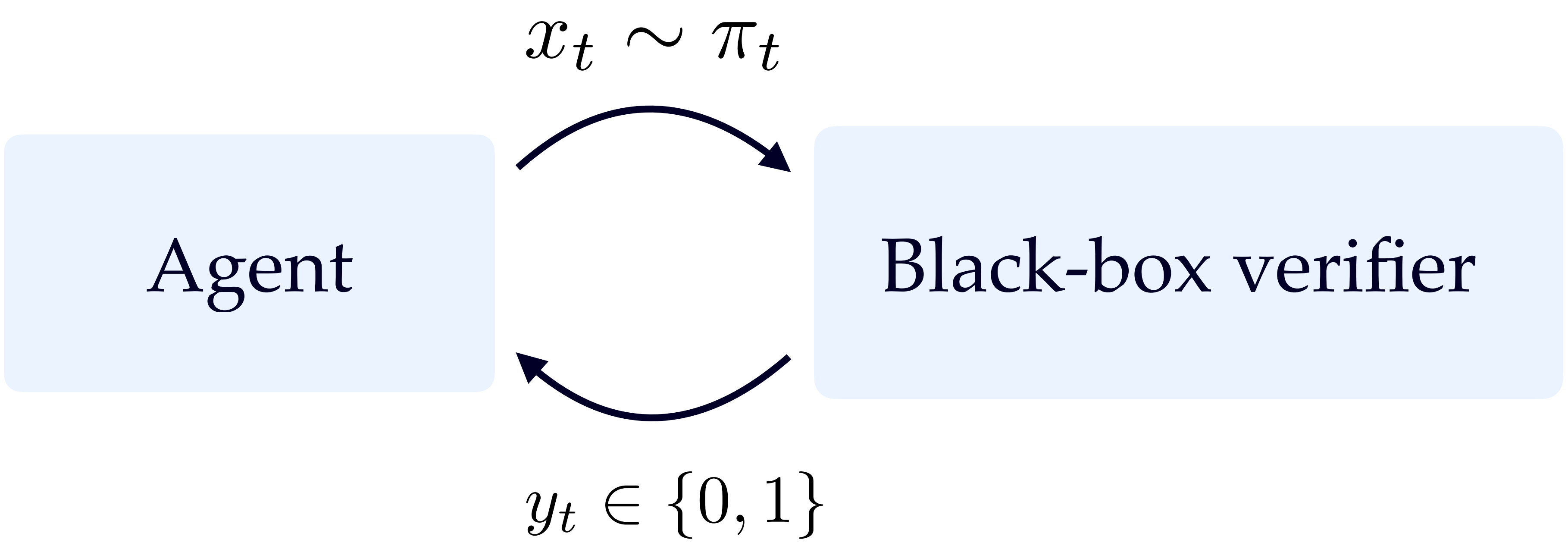
$$\Omega^{\tau}_{\theta_0} \subseteq \Omega^{\tau}_{\theta_1}$$

Continual expansion via adaptation on valid **frontier samples**

Generable Set Expansion for OOD Flow Modeling



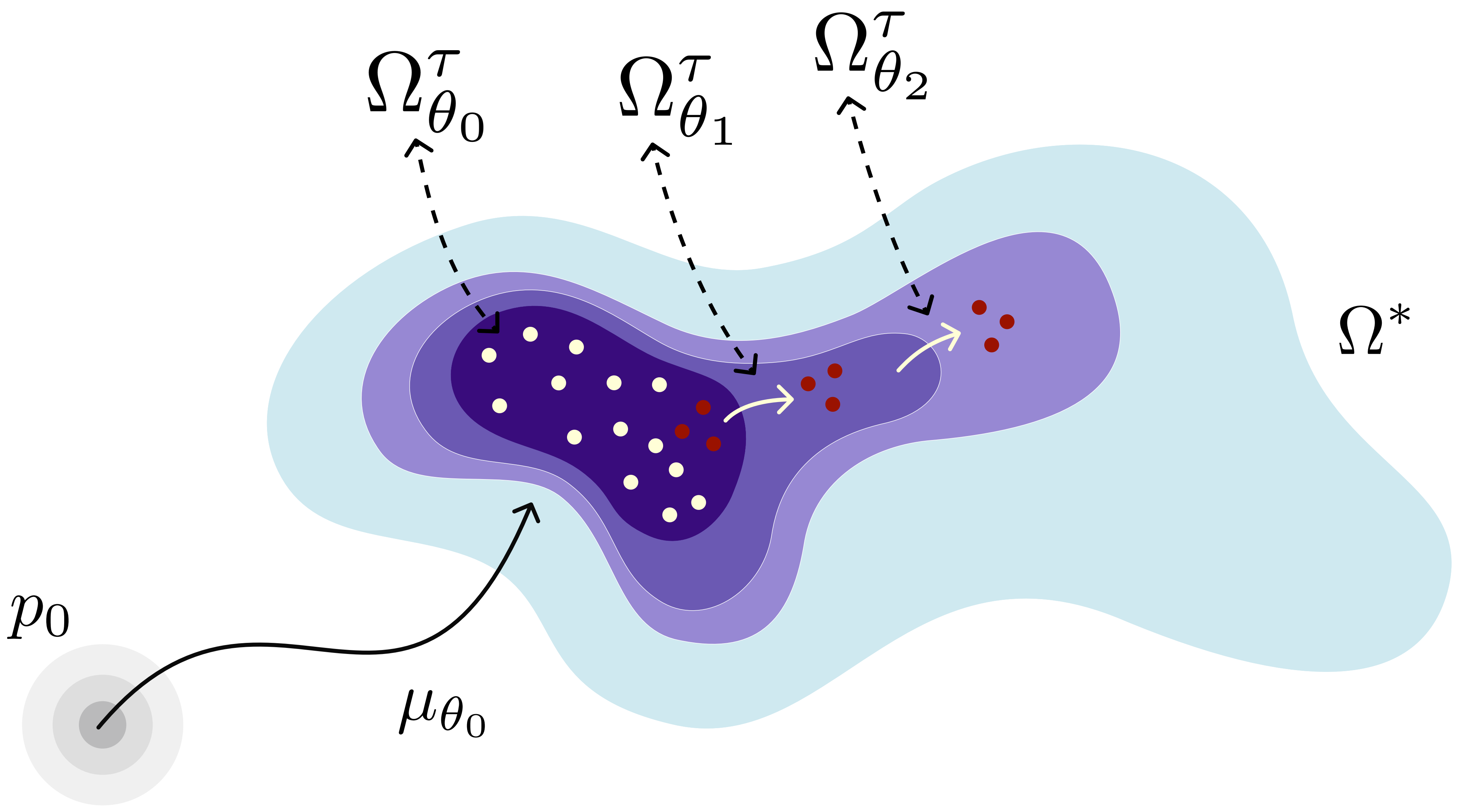
Black-box noisy binary verifier feedback



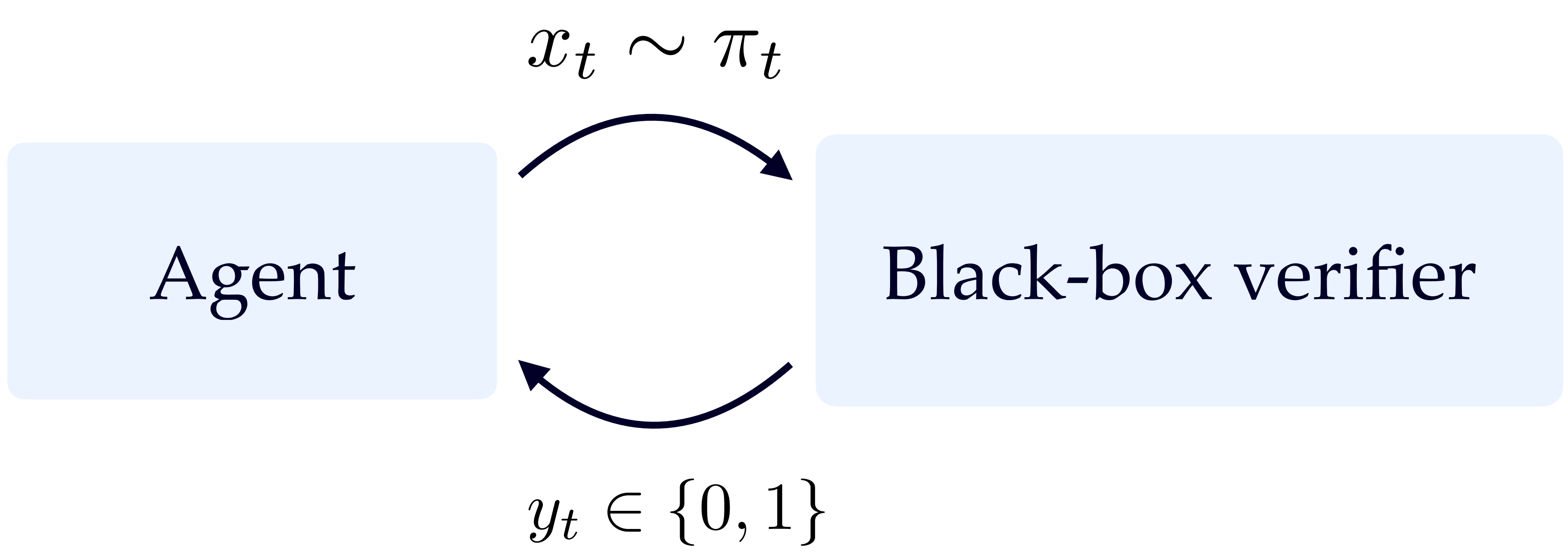
$$\theta^{\text{pre}} = \theta_0 \rightarrow \theta_1 \rightarrow \qquad \Omega^{\tau}_{\theta_0} \subseteq \Omega^{\tau}_{\theta_1} \subseteq$$

Continual expansion via adaptation on valid **frontier samples**

Generable Set Expansion for OOD Flow Modeling



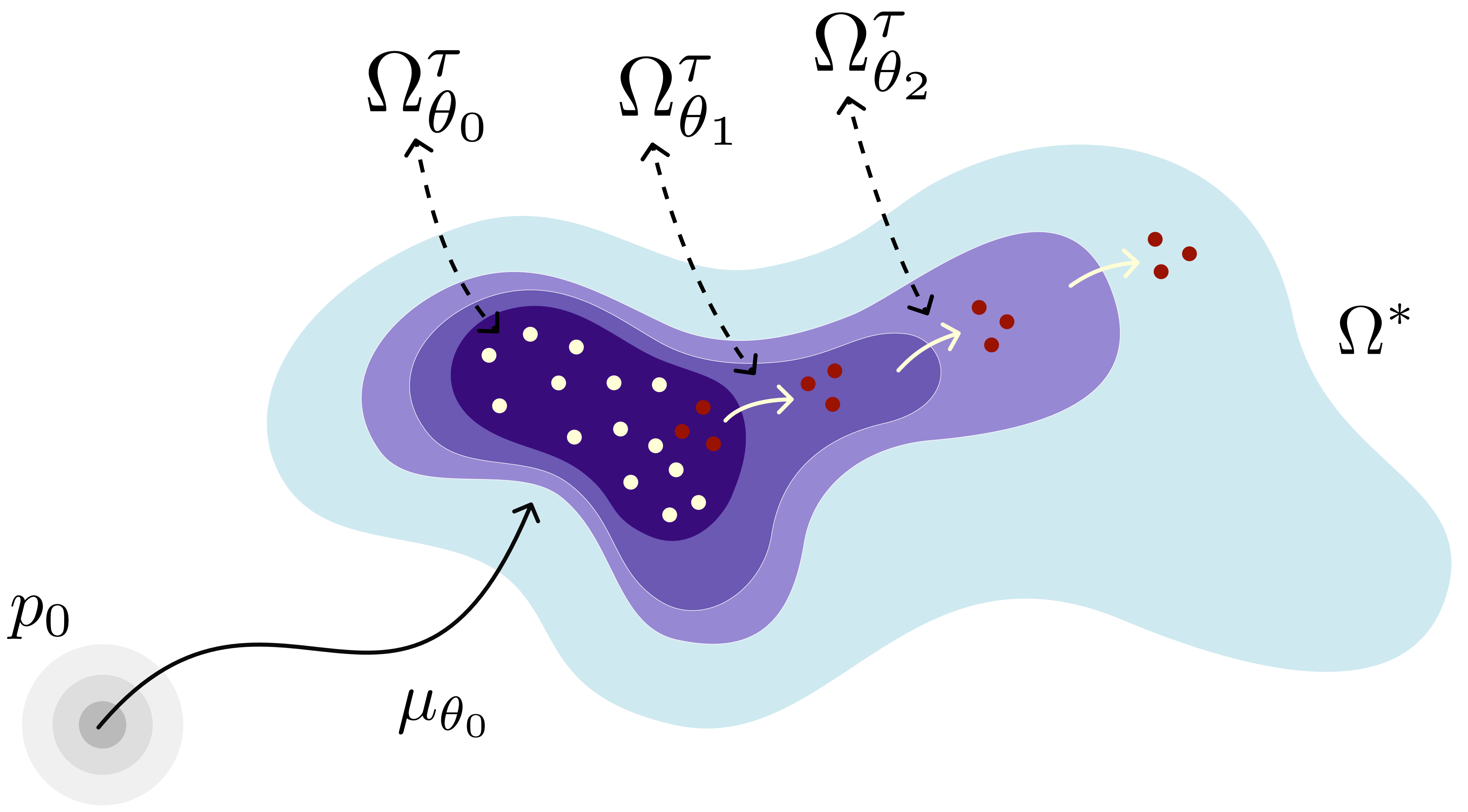
Black-box noisy binary verifier feedback



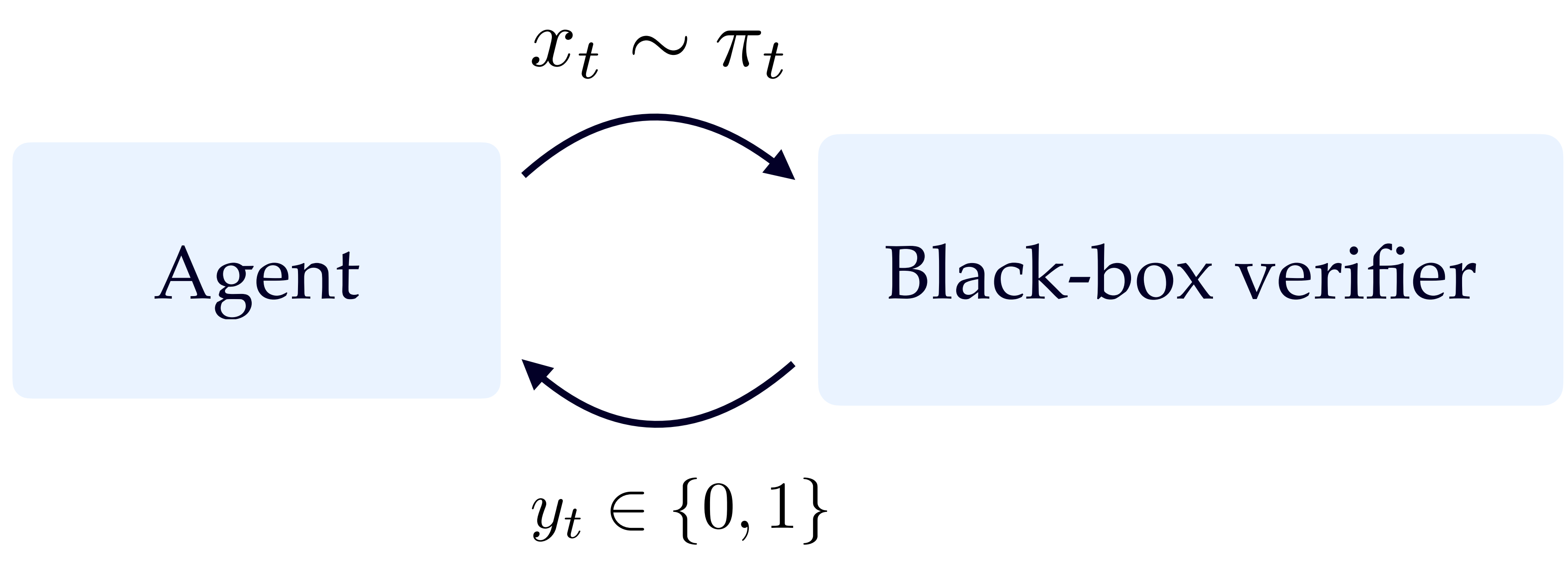
$$\theta^{\text{pre}} = \theta_0 \to \theta_1 \to \cdots \qquad \Omega^\tau_{\theta_0} \subseteq \Omega^\tau_{\theta_1} \subseteq \cdots$$

Continual expansion via adaptation on valid **frontier samples**

Generable Set Expansion for OOD Flow Modeling



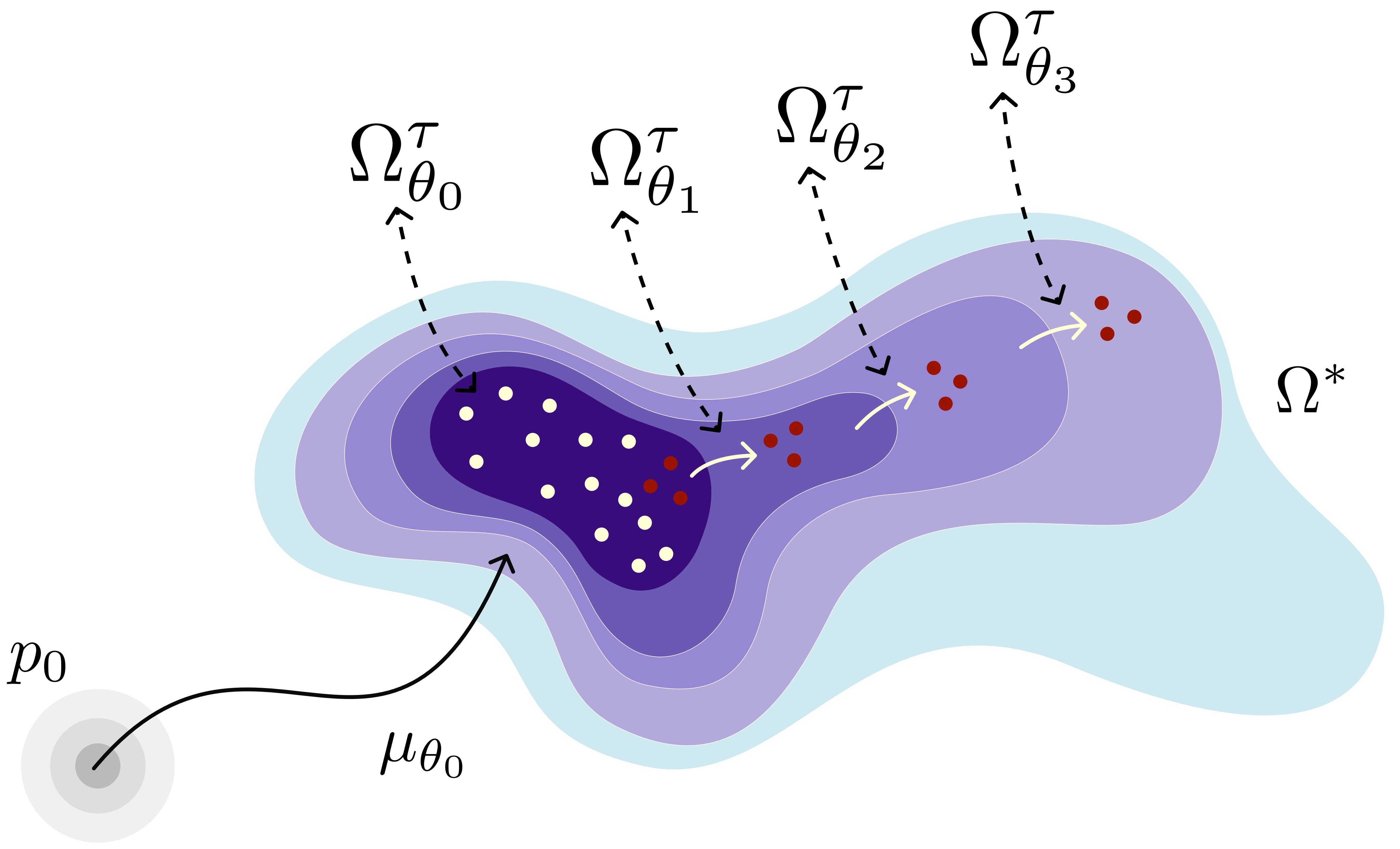
Black-box noisy binary verifier feedback



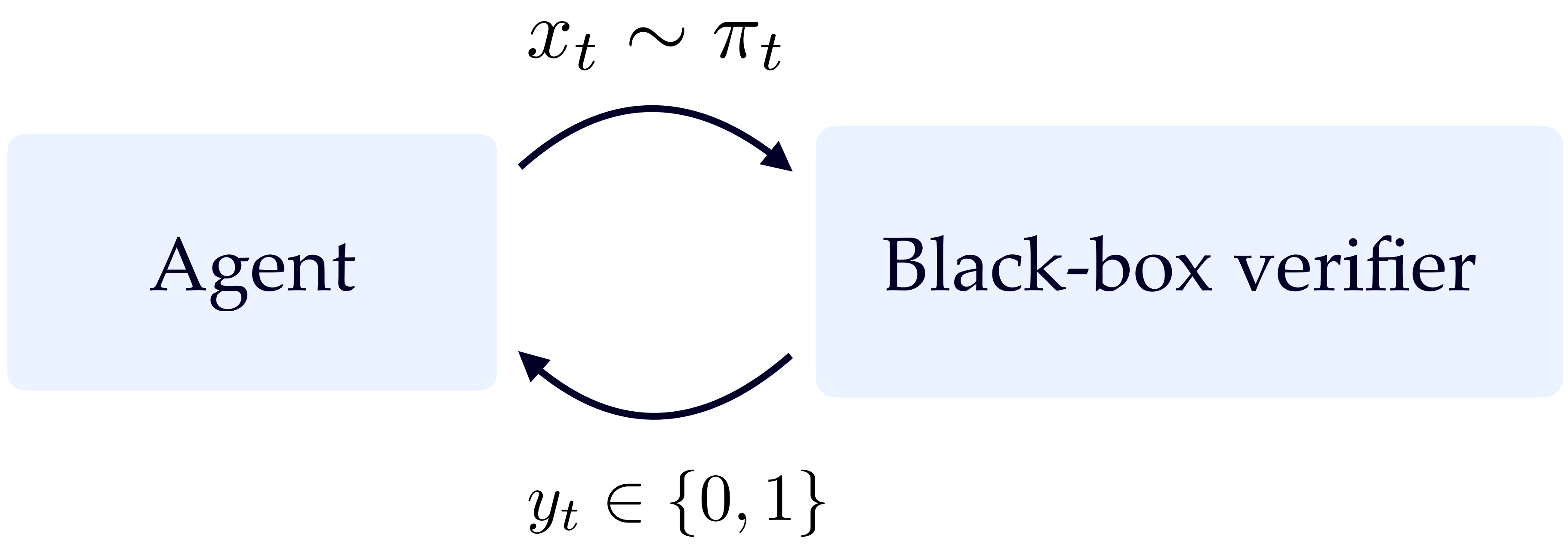
$$\theta^{\text{pre}} = \theta_0 \to \theta_1 \to \cdots \to \qquad \Omega^{\tau}_{\theta_0} \subseteq \Omega^{\tau}_{\theta_1} \subseteq \cdots \subseteq$$

Continual expansion via adaptation on valid **frontier samples**

Generable Set Expansion for OOD Flow Modeling



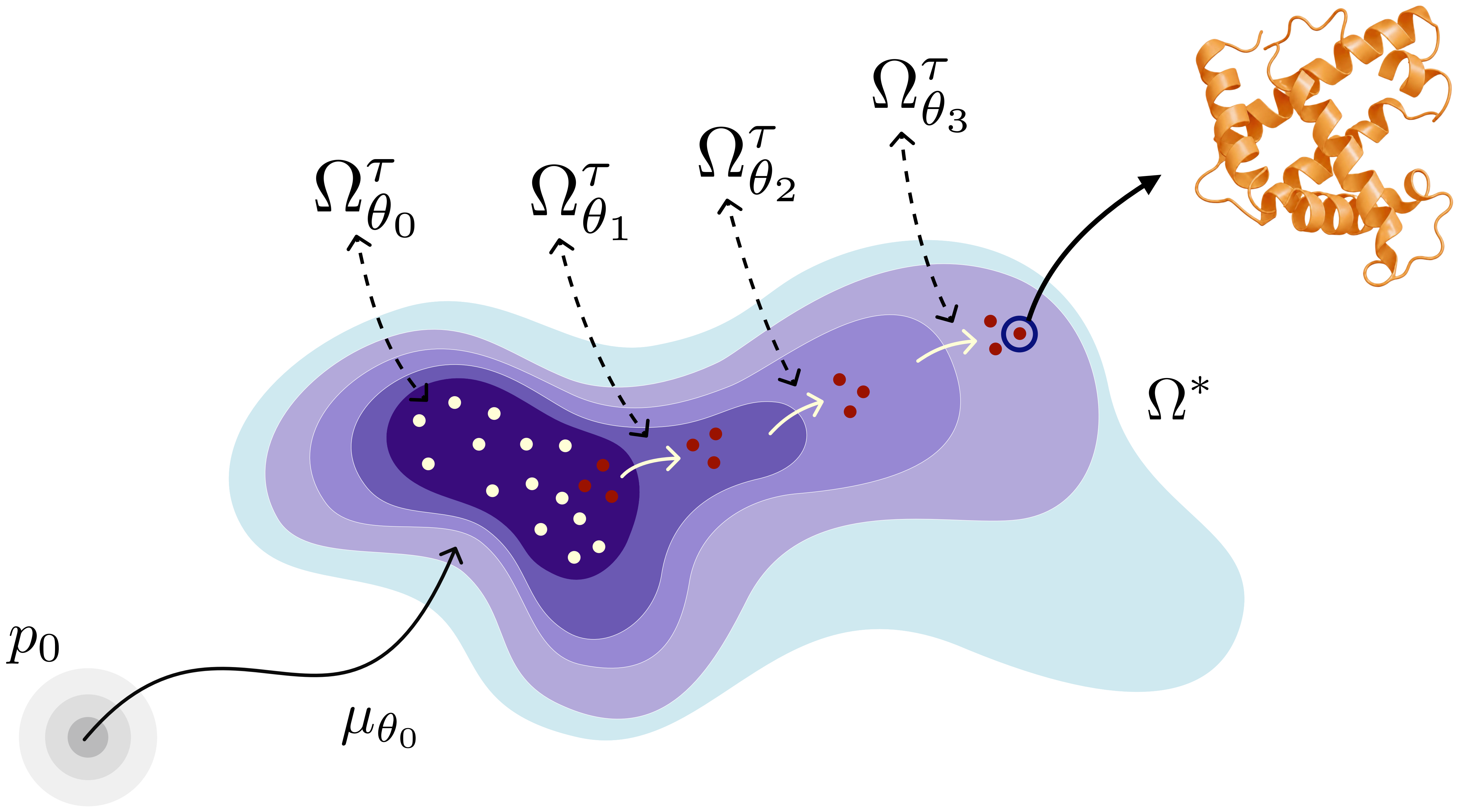
Black-box noisy binary verifier feedback



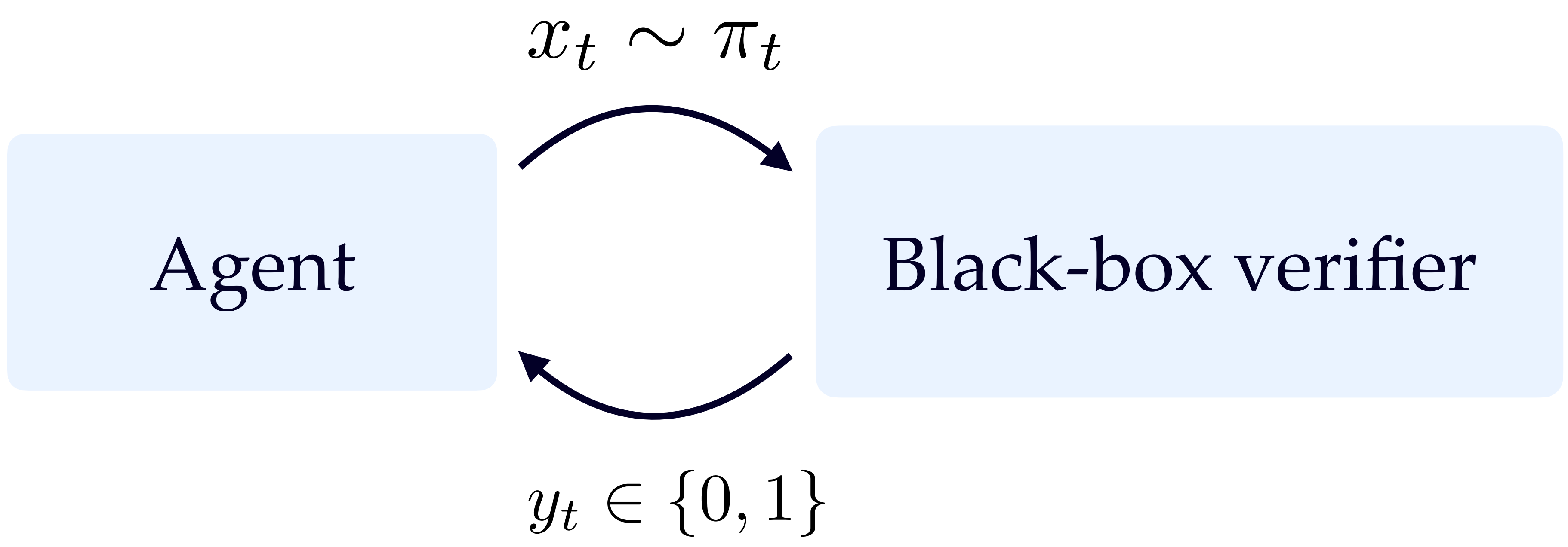
$$\theta^{\text{pre}} = \theta_0 \to \theta_1 \to \cdots \to \theta_T \qquad \Omega^\tau_{\theta_0} \subseteq \Omega^\tau_{\theta_1} \subseteq \cdots \subseteq \Omega^\tau_{\theta_T}$$

Continual expansion via adaptation on valid **frontier samples**

Generable Set Expansion for OOD Flow Modeling



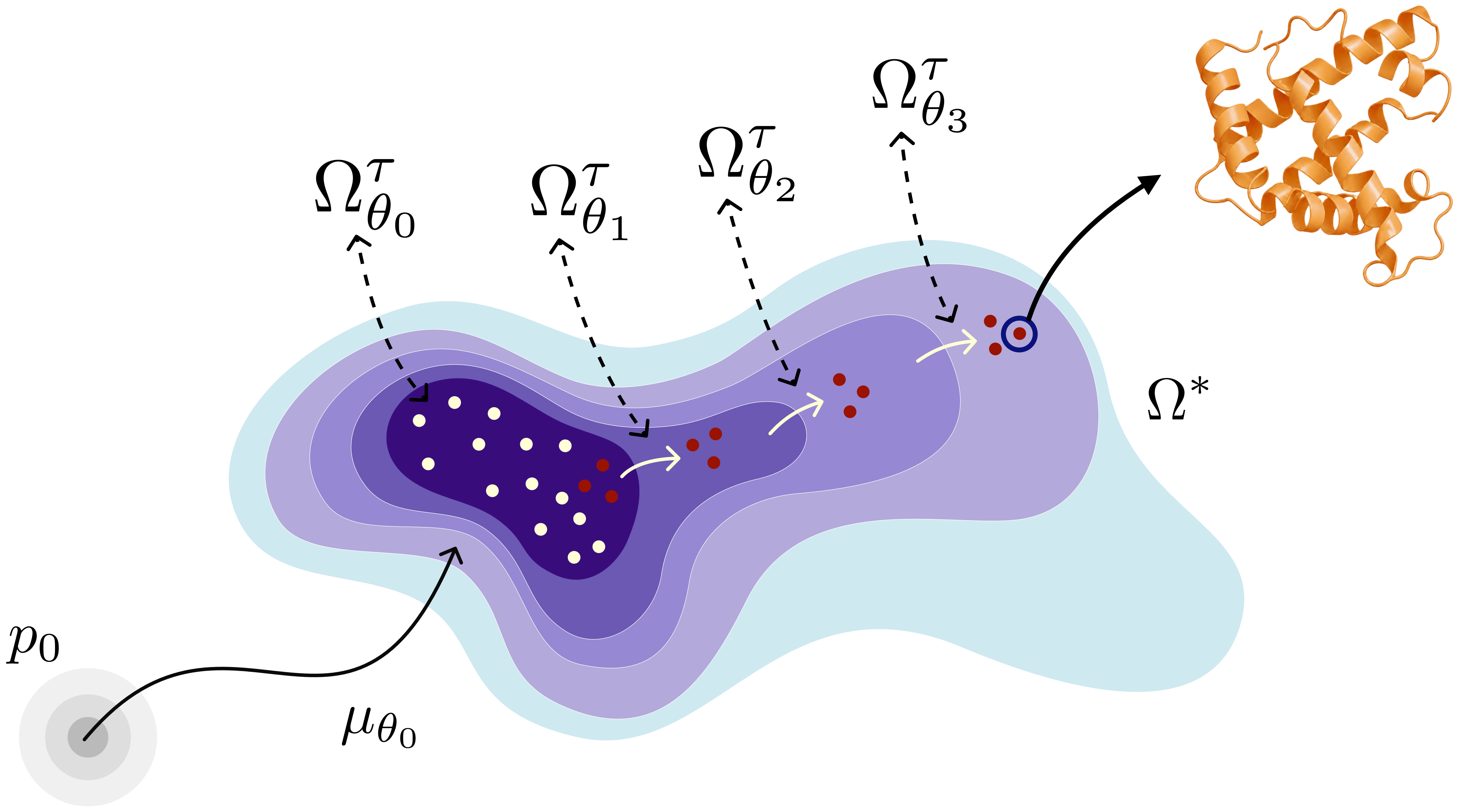
Black-box noisy binary verifier feedback



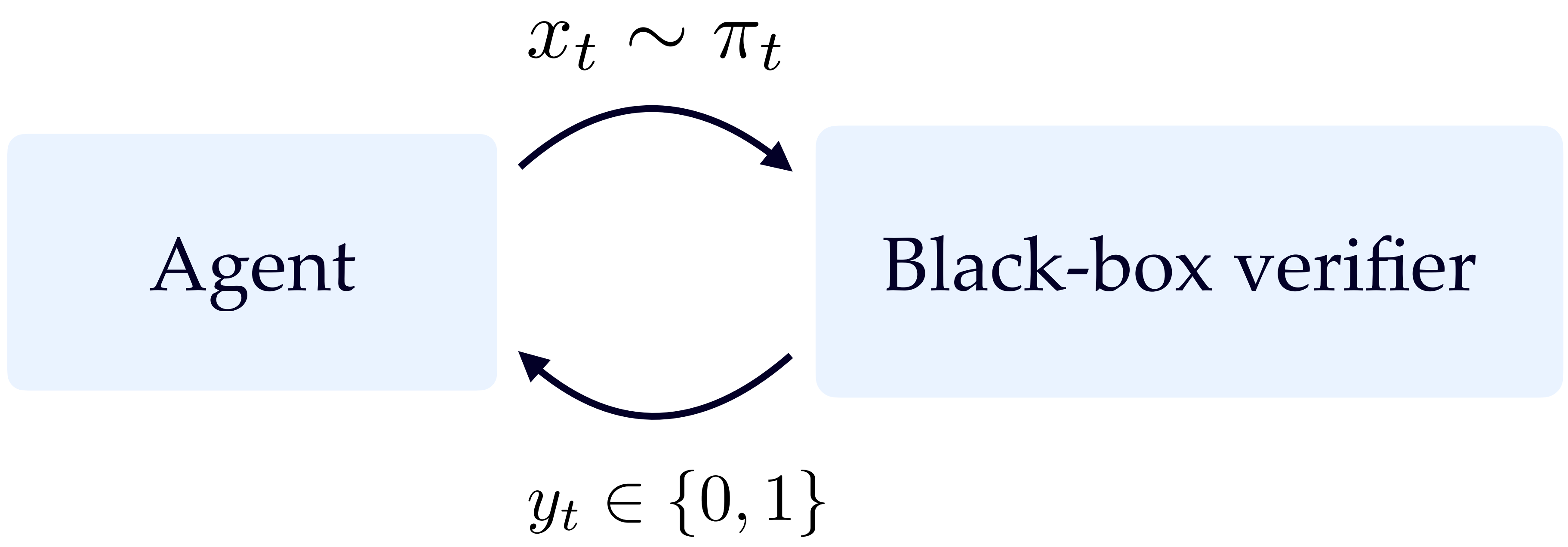
$$\theta^{\text{pre}} = \theta_0 \to \theta_1 \to \cdots \to \theta_T \qquad \Omega^{\tau}_{\theta_0} \subseteq \Omega^{\tau}_{\theta_1} \subseteq \cdots \subseteq \Omega^{\tau}_{\theta_T}$$

Continual expansion via adaptation on valid **frontier samples**

Generable Set Expansion for OOD Flow Modeling



Black-box noisy binary verifier feedback



$$\theta^{\text{pre}} = \theta_0 \to \theta_1 \to \cdots \to \theta_T, \qquad \Omega^{\tau}_{\theta_0} \subseteq \Omega^{\tau}_{\theta_1} \subseteq \cdots \subseteq \Omega^{\tau}_{\theta_T} \overset{?}{\approx} \Omega^{\star}$$

Continual expansion via adaptation on valid **frontier samples**

How to self-generate frontier samples? Generative Uncertainty Sampling

inference-time distribution

$$x_{t+1} \sim \tilde{p}_t \in \arg\max_{q}$$

next data point

generates data

How to self-generate frontier samples? Generative Uncertainty Sampling

verifier uncertainty estimate on $\mathcal{Z}_s^t$

inference-time distribution

$$x_{t+1} \sim \tilde{p}_t \in \arg\max_q \mathbb{E}_{x \sim q}\left[\sigma_t\left(\phi_s^t(x)\right)\right]$$

next data point

noised flow representation

$$\mathcal{Z}_s^t = \phi_s^t(\mathcal{X})$$

generates data with **high verifier (i.e., laws of nature) uncertainty**

How to self-generate frontier samples? Generative Uncertainty Sampling

verifier uncertainty estimate on $\mathcal{Z}_s^t$

inference-time distribution

$$x_{t+1} \sim \tilde{p}_t \in \arg\max_q \mathbb{E}_{x \sim q}\left[\sigma_t\left(\phi_s^t(x)\right)\right] - \beta \, \mathrm{KL}\left(q \,\|\, p_1^{\theta_t}\right)$$

next data point

noised flow representation
$\mathcal{Z}_s^t = \phi_s^t(\mathcal{X})$

current-model (i.e., prior)
regularization

generates data with **high verifier (i.e., laws of nature) uncertainty**, yet with **sufficient validity prior**

How to self-generate frontier samples? Generative Uncertainty Sampling

verifier uncertainty estimate on $\mathcal{Z}_s^t$

inference-time distribution

$$x_{t+1} \sim \tilde{p}_t \in \arg\max_q \mathbb{E}_{x\sim q}\left[\sigma_t\left(\phi_s^t(x)\right)\right] - \beta\,\mathrm{KL}\left(q \,\|\, p_1^{\theta_t}\right)$$

next data point

noised flow representation
$\mathcal{Z}_s^t = \phi_s^t(\mathcal{X})$

current-model (i.e., prior)
regularization

generates data with **high verifier (i.e., laws of nature) uncertainty**, yet with **sufficient validity prior**

$\longrightarrow$ highly informative **frontier samples**

ActFlow: Active Flow Expansion

$$x_{t+1} \sim \tilde{p}_t \in \arg\max_q \mathbb{E}_{x\sim q}\left[\sigma_t\left(\phi_s^t(x)\right)\right] - \beta\,\mathrm{KL}\left(q \,\|\, p_1^{\theta_t}\right)$$

ActFlow: Active Flow Expansion

Init: Initial flow model μ_{θ_0}, black-box verifier $\tilde{v}$, iterations T, $\mathcal{D}_0 \leftarrow \emptyset$

For $t = 0, 1, \ldots, T-1$:

Update surrogate uncertainty σ_t from $\mathcal{D}_t$

Self-generate:

$$x_{t+1} \sim \tilde{p}_t \in \arg\max_q \mathbb{E}_{x\sim q}\left[\sigma_t\left(\phi_s^t(x)\right)\right] - \beta\,\mathrm{KL}\left(q \,\|\, p_1^{\theta_t}\right)$$

ActFlow: Active Flow Expansion

Init: Initial flow model μ_{θ_0}, black-box verifier $\tilde{v}$, iterations T, $\mathcal{D}_0 \leftarrow \emptyset$

For $t = 0, 1, \ldots, T-1$:

Update surrogate uncertainty σ_t from $\mathcal{D}_t$

Self-generate:

$$x_{t+1} \sim \tilde{p}_t \in \arg\max_q \mathbb{E}_{x\sim q}\left[\sigma_t\left(\phi_s^t(x)\right)\right] - \beta\, \mathrm{KL}\left(q \,\|\, p_1^{\theta_t}\right)$$

Query verifier $y_{t+1} \leftarrow \tilde{v}(x_{t+1})$

$\mathcal{D}_{t+1} \leftarrow \mathcal{D}_t \cup \{(x_{t+1}, y_{t+1})\}$

$\theta_{t+1} \leftarrow \textsc{UpdateFlow}(\theta_t, \mathcal{D}_{t+1})$

Return μ_{θ_T}

Theoretical Guarantees for Out-of-Distribution Flow Modeling

Distribution matching

$$p_1^\theta \approx p_{data}$$

$\longrightarrow$

Design space coverage

$$\Omega_\theta^\tau \approx \Omega^*$$

Generative modeling beyond data distribution matching

Theoretical Guarantees for Out-of-Distribution Flow Modeling

Theorem (Informal, generable representation set covers reachable set).
Fix $\epsilon > 0$ and an integer $H \geq 1$. Then, with

$$T^\star \gtrsim \left(\frac{\alpha \gamma_{HT^*}^{\Omega_\star}}{\epsilon} \right)^2$$

By running ActFlow, it holds with probability at least $1 - \delta$ that after T^* verified samples:

$$R_\epsilon^H(S_0) \subseteq \Omega_{T^\star}^\tau$$

Theoretical Guarantees for Out-of-Distribution Flow Modeling

Theorem (Informal, generable representation set covers reachable set).
Fix $\epsilon > 0$ and an integer $H \geq 1$. Then, with

$$T^\star \gtrsim \left(\frac{\alpha \gamma_{HT^*}^{\Omega_\star}}{\epsilon} \right)^2$$

By running ActFlow, it holds with probability at least $1 - \delta$ that after T^* verified samples:

$$R_\epsilon^H(S_0) \subseteq \Omega_{T^\star}^{\tau} \longrightarrow \text{generable representation set}$$

Theoretical Guarantees for Out-of-Distribution Flow Modeling

Theorem (Informal, generable representation set covers reachable set).
Fix $\epsilon > 0$ and an integer $H \geq 1$. Then, with

$$T^\star \gtrsim \left(\frac{\alpha \gamma_{HT^*}^{\Omega_\star}}{\epsilon} \right)^2$$

By running ActFlow, it holds with probability at least $1 - \delta$ that after T^* verified samples:

$$R_\epsilon^H(S_0) \subseteq \Omega_{T^\star}^\tau$$

generable representation set

reachable set given **H-fold** recursive application of **one-step reachability operator** over flow representation:

$$R_\epsilon(S) := \{z \in \mathcal{Z} : \exists z' \in S \text{ s.t. } s(g(z')) - L_s L_g d(z, z') - \epsilon \geq h\}$$

Theoretical Guarantees for Out-of-Distribution Flow Modeling

Theorem (Informal, generable representation set covers reachable set).
Fix $\epsilon > 0$ and an integer $H \geq 1$. Then, with

$$T^\star \gtrsim \left(\frac{\alpha\gamma_{HT^*}^{\Omega_\star}}{\epsilon}\right)^2$$

By running ActFlow, it holds with probability at least $1-\delta$ that after T^* verified samples:

$$R_\epsilon^H(S_0) \subseteq \Omega_{T^\star}^\tau$$

maximum information gain

$$\gamma_t^A \triangleq \max_{z_1,\ldots,z_t \in A} \frac{1}{2} \log\det\Big(I_t + (\lambda\kappa)^{-1} K_t\Big)$$

captures laws of nature complexity

generable representation set

reachable set given **H-fold** recursive application of **one-step reachability operator** over flow representation:

$$R_\epsilon(S) := \{z \in \mathcal{Z} : \exists z' \in S \text{ s.t. } s(g(z')) - L_s L_g d(z, z') - \epsilon \geq h\}$$

Theoretical Guarantees for Out-of-Distribution Flow Modeling

Theorem (Informal, generable representation set covers reachable set).
Fix $\epsilon > 0$ and an integer $H \geq 1$. Then, with

$$T^\star \gtrsim \left(\frac{\alpha \gamma_{HT^*}^{\Omega_\star}}{\epsilon} \right)^2$$

By running ActFlow, it holds with probability at least $1 - \delta$ that after T^* verified samples:

$$R_\epsilon^H(S_0) \subseteq \Omega_{T^\star}^\tau$$

Allows to identify **structural conditions** and **statistical complexity** for:

$$\Omega^\star \subseteq \Omega_{T^\star}^{X,\tau_X}$$

valid design space

generable set on $\mathcal{X}$

Theoretical Guarantees for Out-of-Distribution Flow Modeling

Theorem (Informal, generable representation set covers reachable set).
Fix $\epsilon > 0$ and an integer $H \geq 1$. Then, with

$$T^{\star} \gtrsim \left(\frac{\alpha \gamma_{HT^{*}}^{\Omega_{\star}}}{\epsilon} \right)^{2}$$

By running ActFlow, it holds with probability at least $1 - \delta$ that after T^* verified samples:

$$R_{\epsilon}^{H}(S_0) \subseteq \Omega_{T^{\star}}^{\tau}$$

Takeaway

ActFlow allows to build (generable) **set-theoretic guarantees** for **out-of-distribution flow modeling**.

Distribution matching

$$p_1^{\theta} \approx p_{data}$$

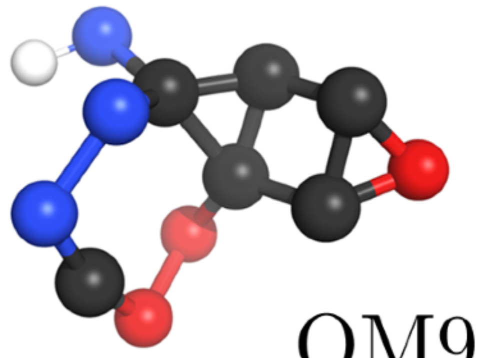
$$\longrightarrow$$

Design space coverage

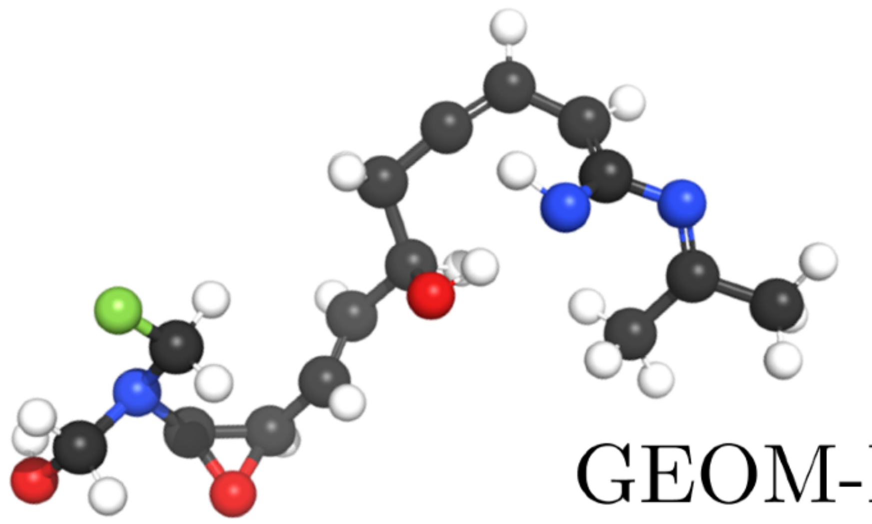
$$\Omega_{\theta}^{\tau} \approx \Omega^{*}$$

Toward a (learning) **theory of** (generative) **discovery**

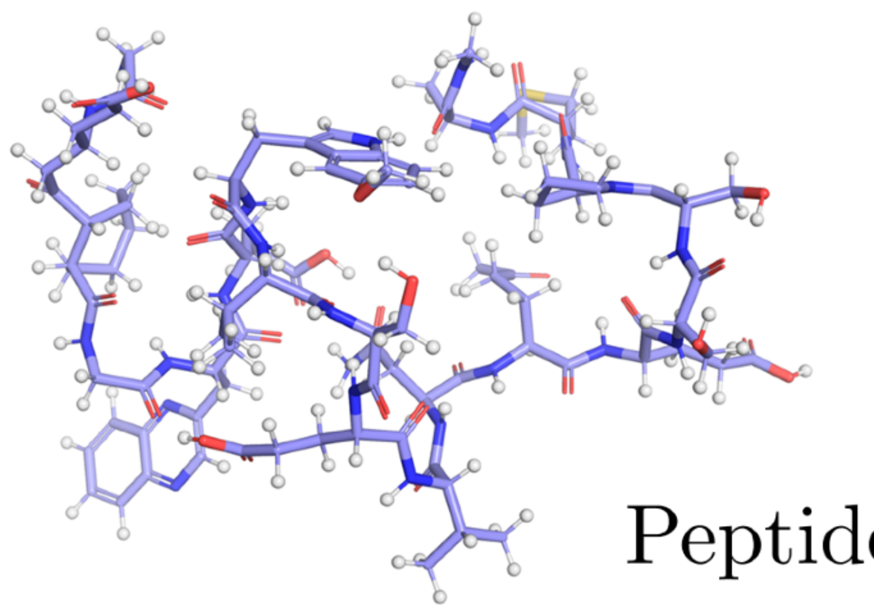
ActFlow for Chemical and Biological Generative Discovery



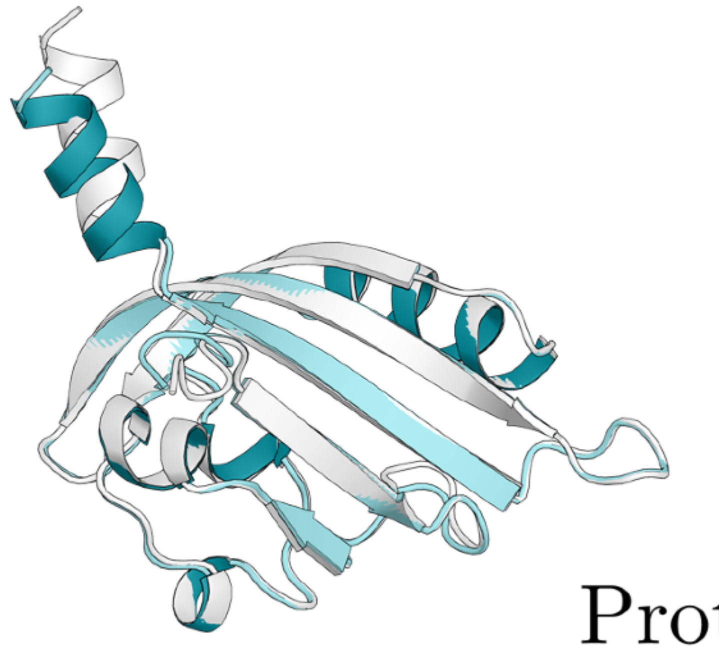
QM9 small
organic molecule



GEOM-Drugs
drug-like molecule

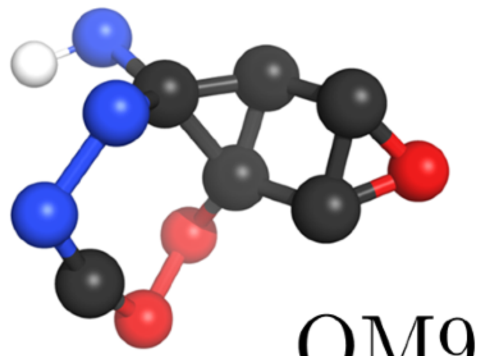


Peptide

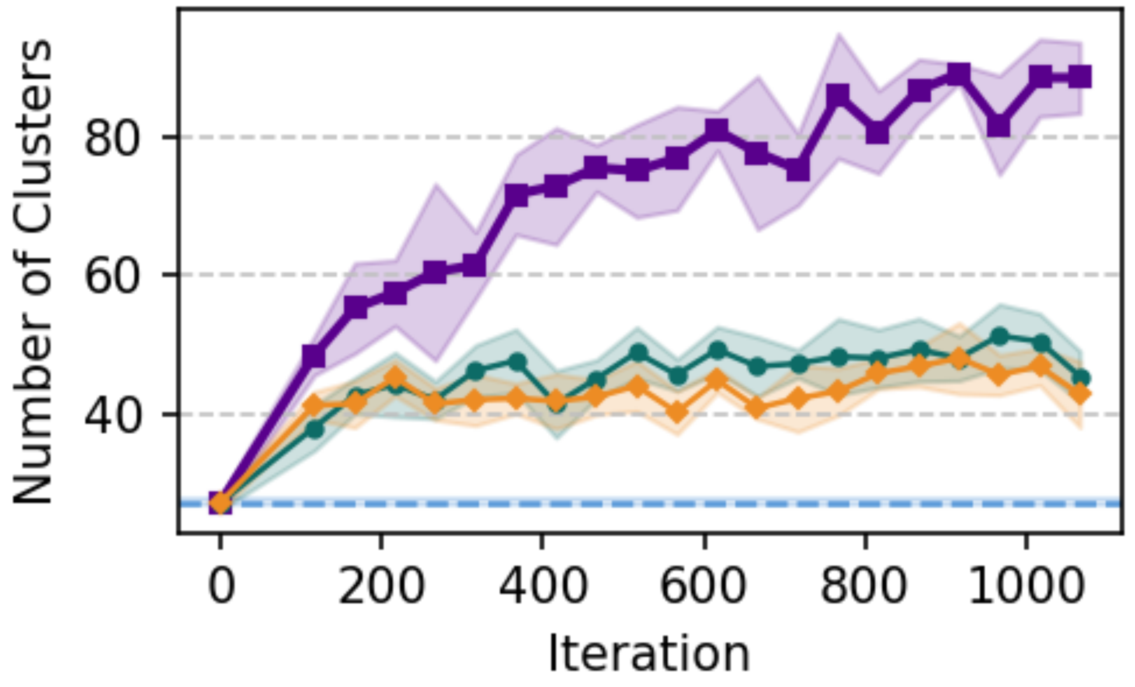


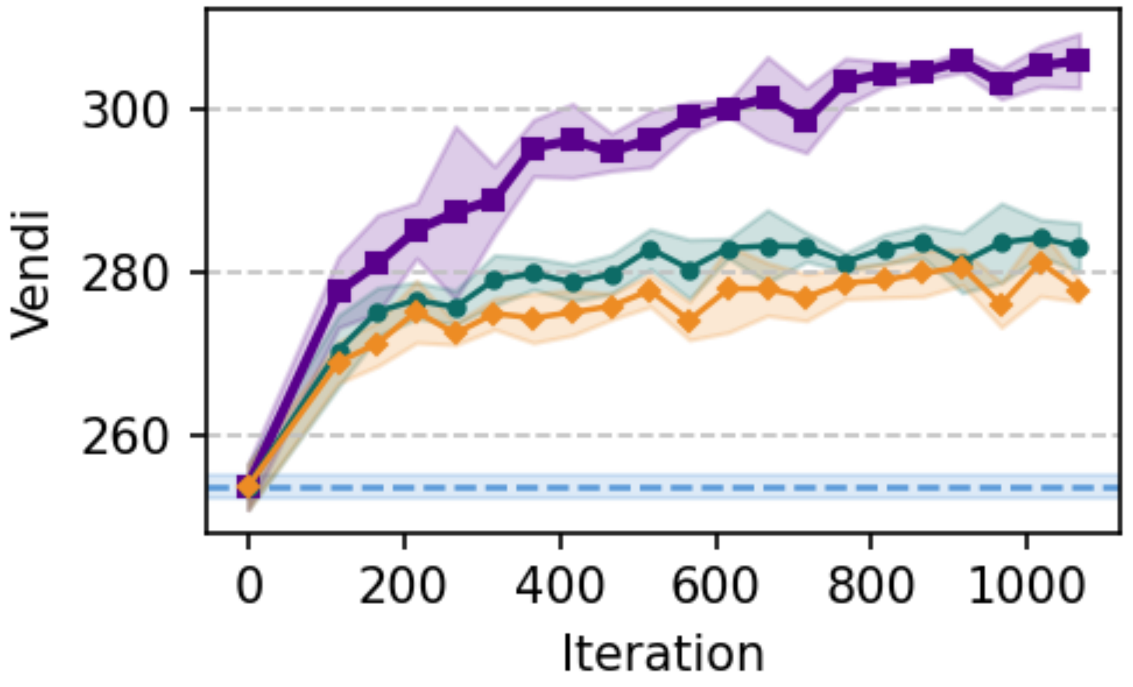
Protein

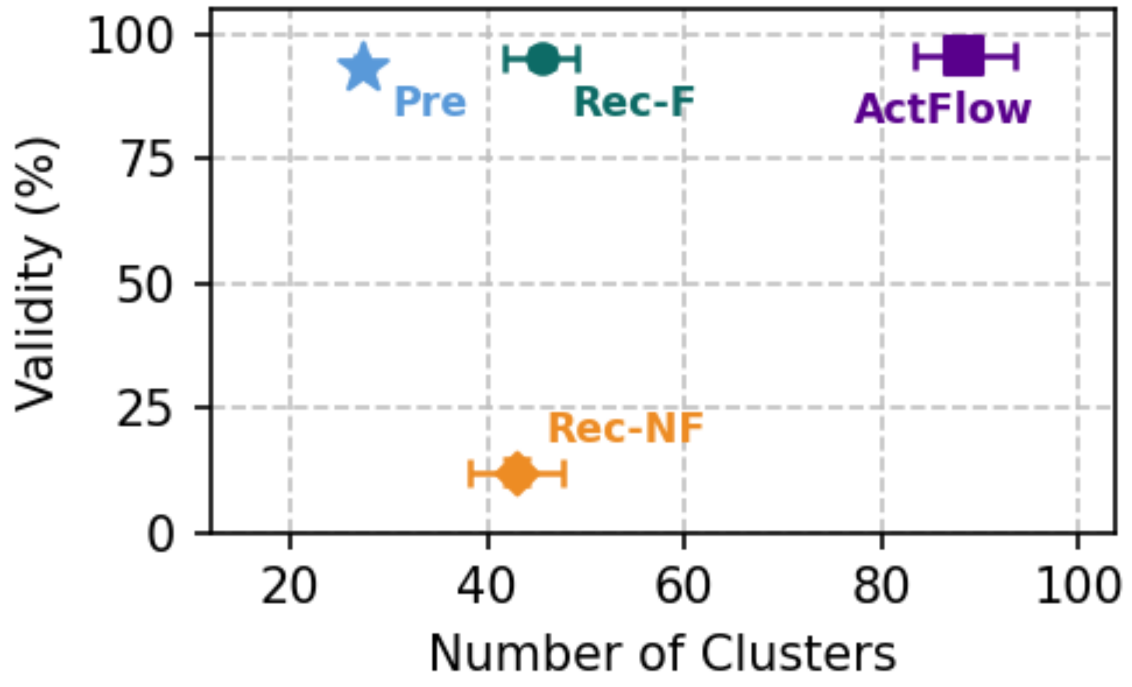
ActFlow for Small Organic and Drug-like Molecules

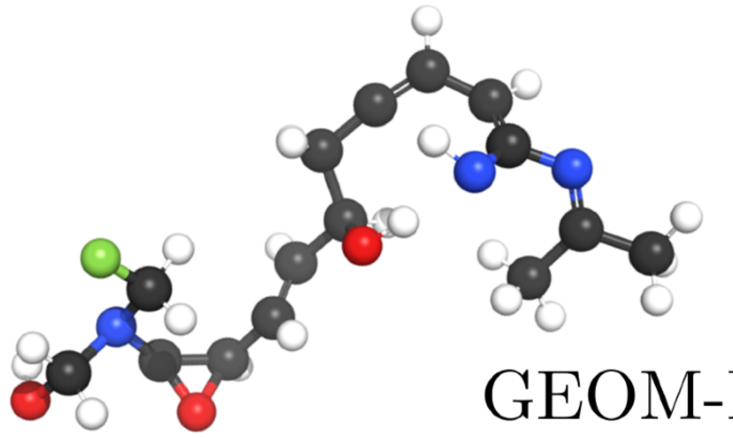


QM9 small organic molecule

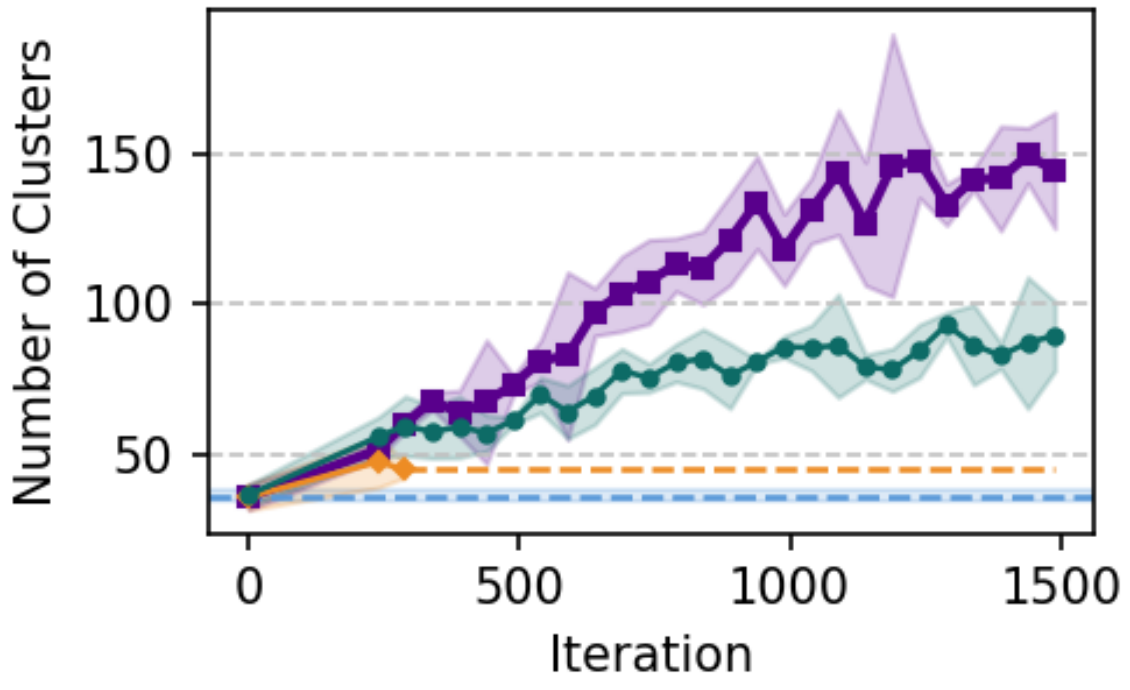


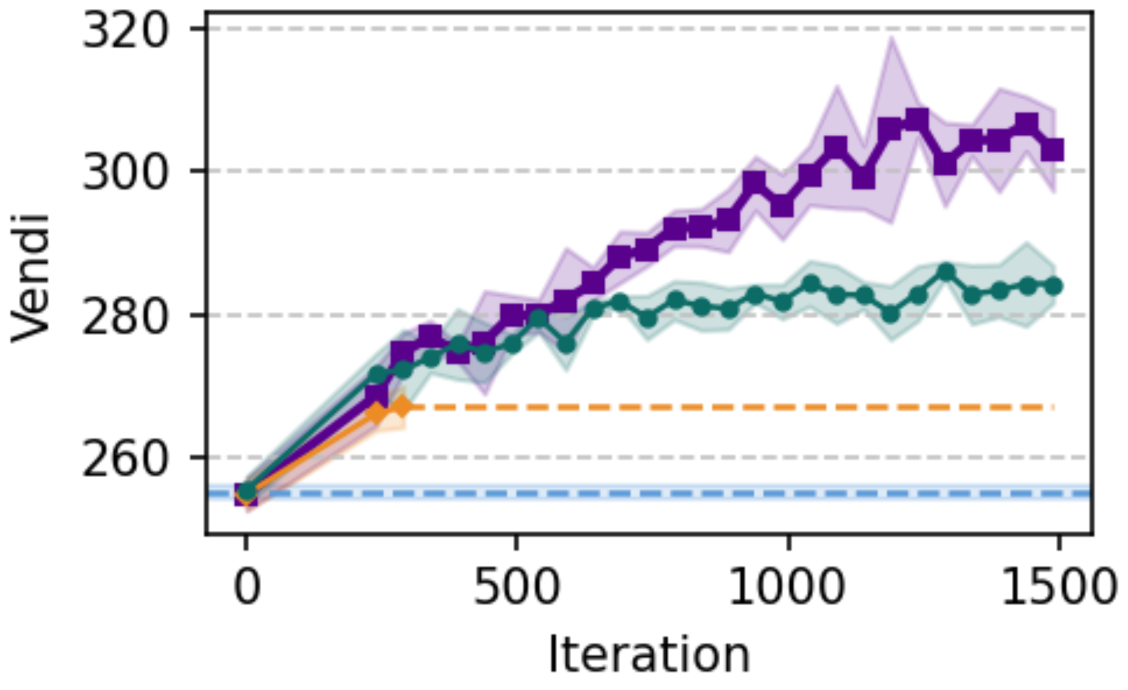


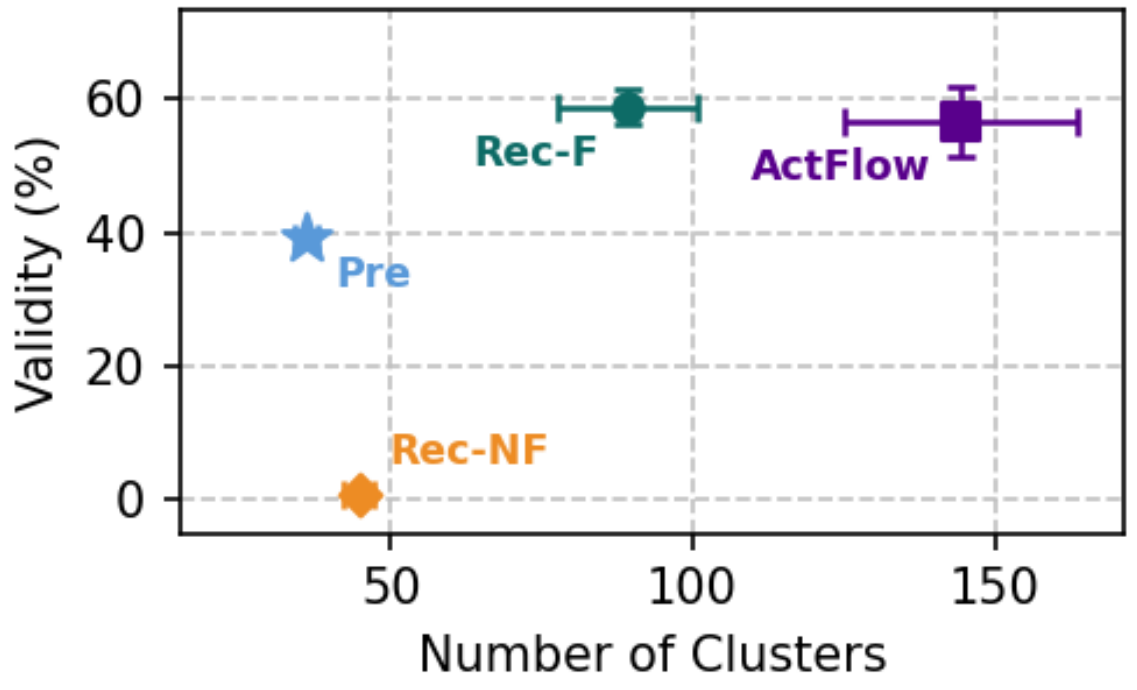




GEOM-Drugs drug-like molecule

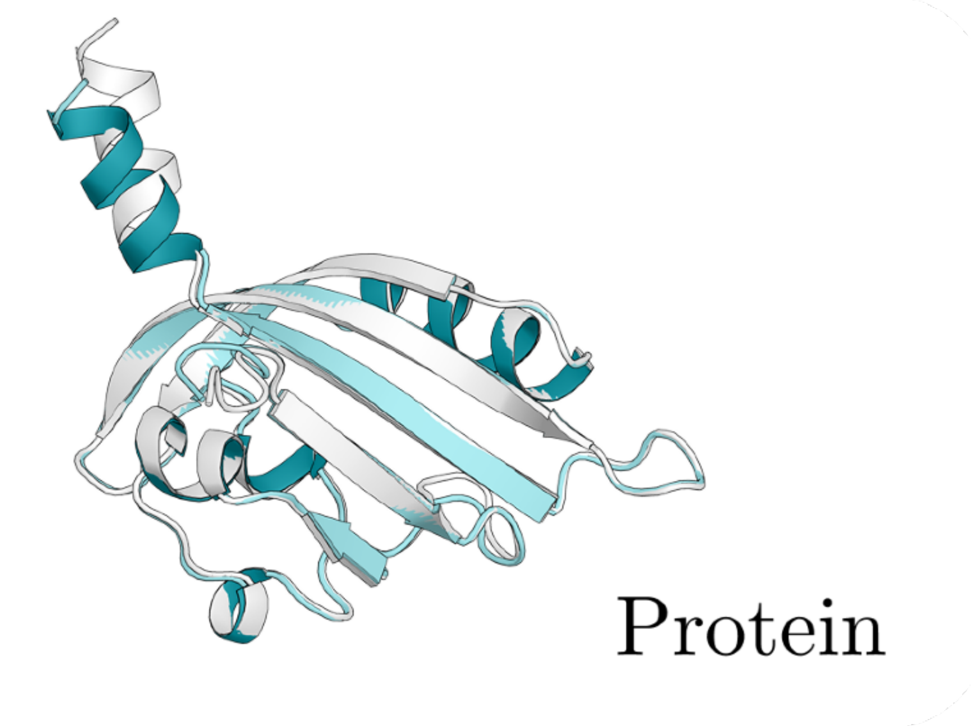


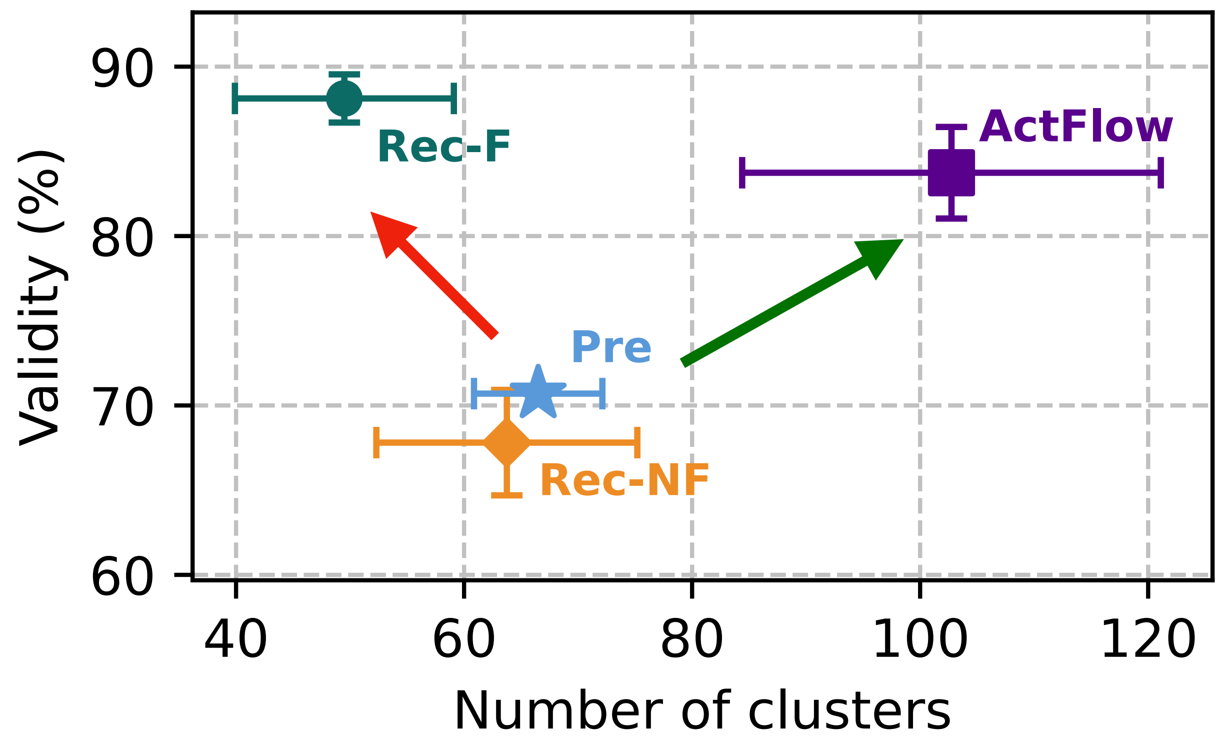




ActFlow for Therapeutic Peptides and Protein Design

	Therapeutic Peptide Design				Protein Sequence Design			
Method	Coverage ↑	Diversity ↑	FID	Validity (%) ↑	Coverage ↑	Diversity ↑	FID	Validity (%) ↑
Pre	61.00 ± 0.00	13.73 ± 0.00	0.00 ± 0.00	40.12 ± 0.00	66.50 ± 5.63	12.87 ± 0.63	0.05 ± 0.01	70.81 ± 1.12
REC-NF	0.00 ± 0.00	0.00 ± 0.00	0.00 ± 0.00	0.00 ± 0.00	63.75 ± 11.45	11.85 ± 0.34	0.27 ± 0.06	67.81 ± 3.11
REC-F	59.67 ± 17.97	13.62 ± 4.34	3.18 ± 1.77	71.16 ± 5.22	49.50 ± 9.60	11.67 ± 0.40	0.25 ± 0.04	88.12 ± 1.42
ACTFLOW	358.33 ± 95.45	58.87 ± 25.98	59.15 ± 44.31	41.59 ± 10.06	102.75 ± 18.36	42.14 ± 10.85	5.45 ± 3.34	83.74 ± 2.70





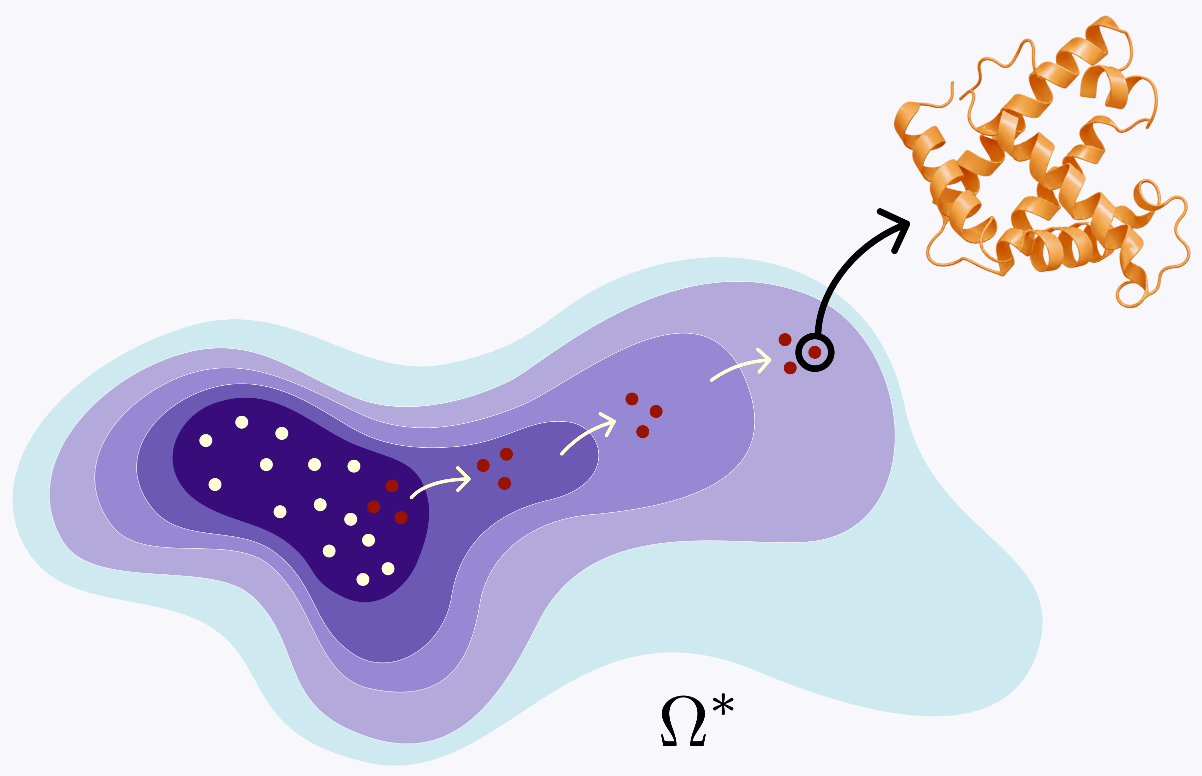
Takeaway

ActFlow allows to **actively expand** pre-trained biochemical flow and diffusion models **over OOD regions**.

Research Program

Foundations of Generative Discovery Beyond the Data

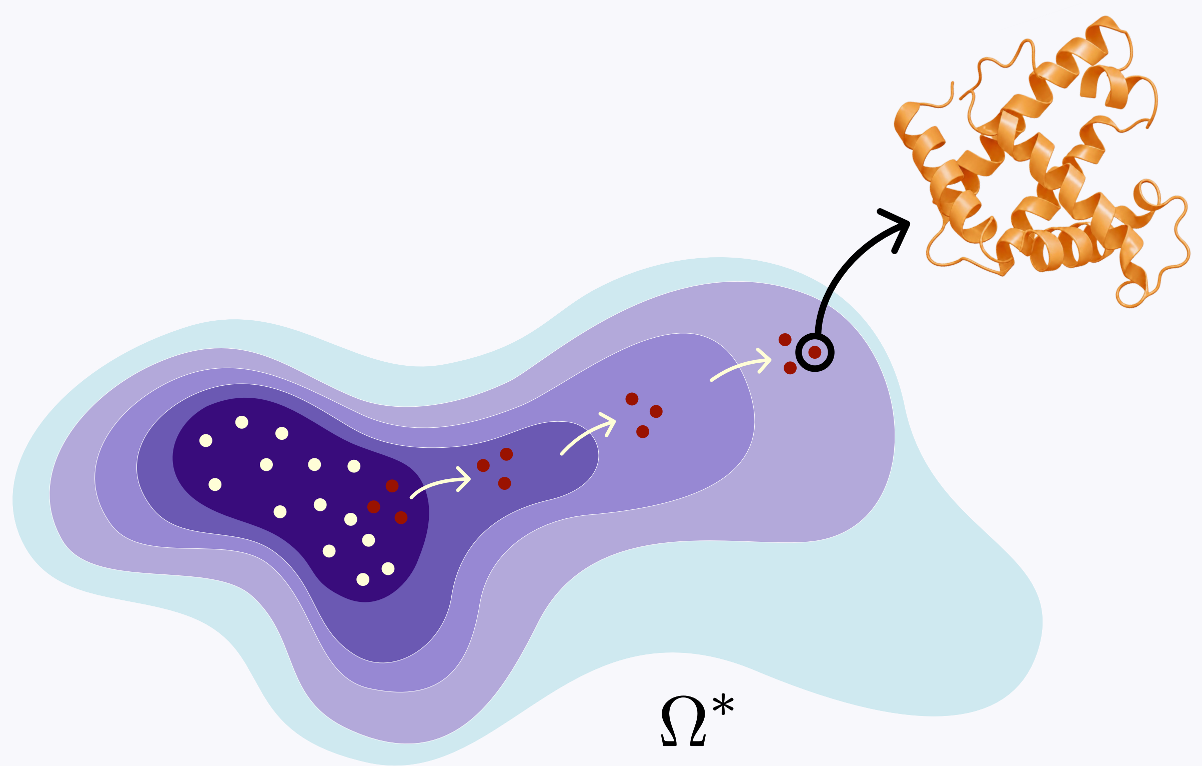
Out-of-Distribution
Flow Modeling



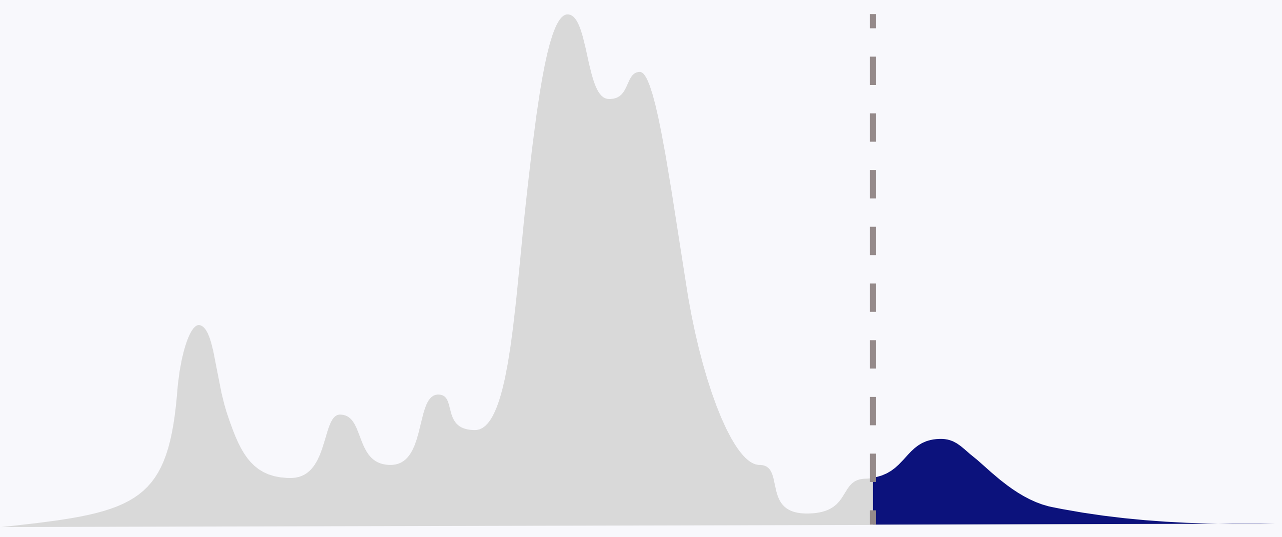
Research Program

Foundations of Generative Discovery Beyond the Data

Out-of-Distribution Flow Modeling

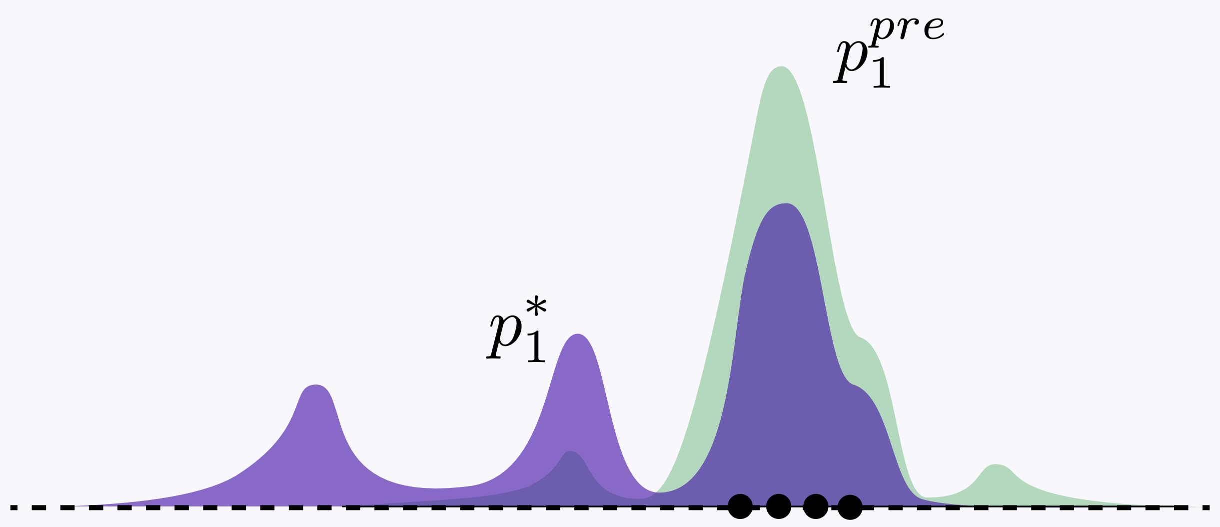


High-value rare structure generation



$\mathrm{CVaR}^{f}_{1-\beta}(p^{\pi}_1)$

Data Debiasing and Hidden Mode Discovery



Thanks!

Collaborators

Kimon Protopapas, Ya-Ping Hsieh, Sven Gutjahr, Luca Schaufelberger, Zifan Wang, Xiaoyu Mo, Malte Franke, Cristian Perez Jensen, Andrey Kharitenko, Zebang Shen, Marin Vlastelica, Guy Schacht, Mojmir Mutny, Ziyad Sheebaelhamd, Bruce Lee, Sophia Tang, Cheng-Hao Liu, Michael M. Zavlanos, Harini Vijayshankar, Dylan Abramson, Basile Wicky, Karl H. Johansson, Kjell Jorner, Florian Dörfler, Pranam Chatterjee, Yisong Yue, Niao He, Andreas Krause.

Funding

NCCR Catalysis

NCCR Automation

ETH AI CENTER

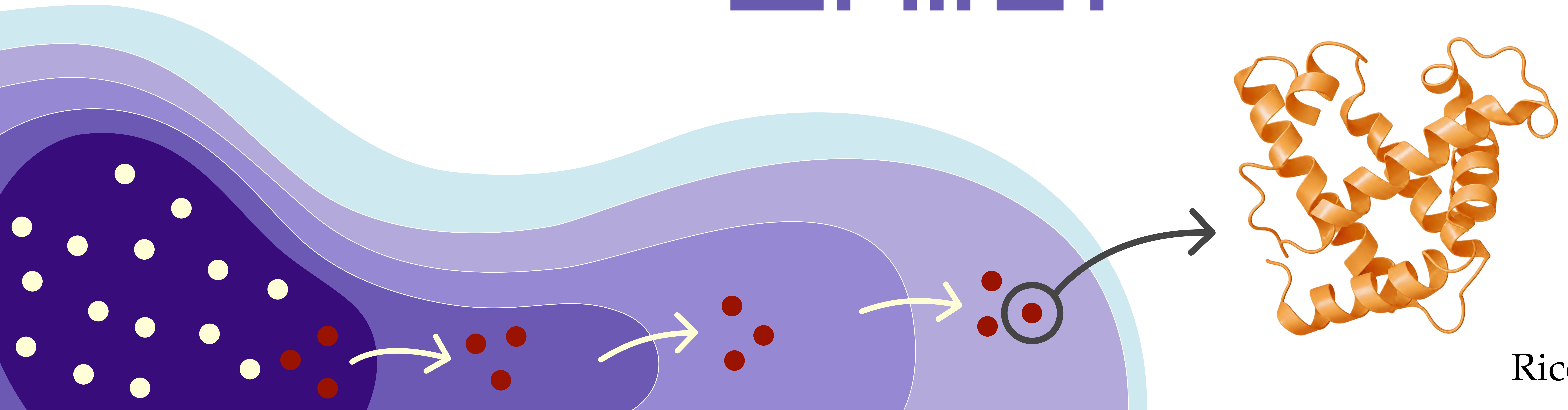
Thanks!

Paper + Code

riccardodesanti.com/research

Connect:

X / twitter: @r_de_santi
email: rdesanti@ethz.ch



Riccardo De Santi