Can machines know what they don't know?

Fanny Yang, Department of Computer Science, ETH Zurich



Statistical Machine Learning group

AI & Synthetic Biology, Paris, November 26

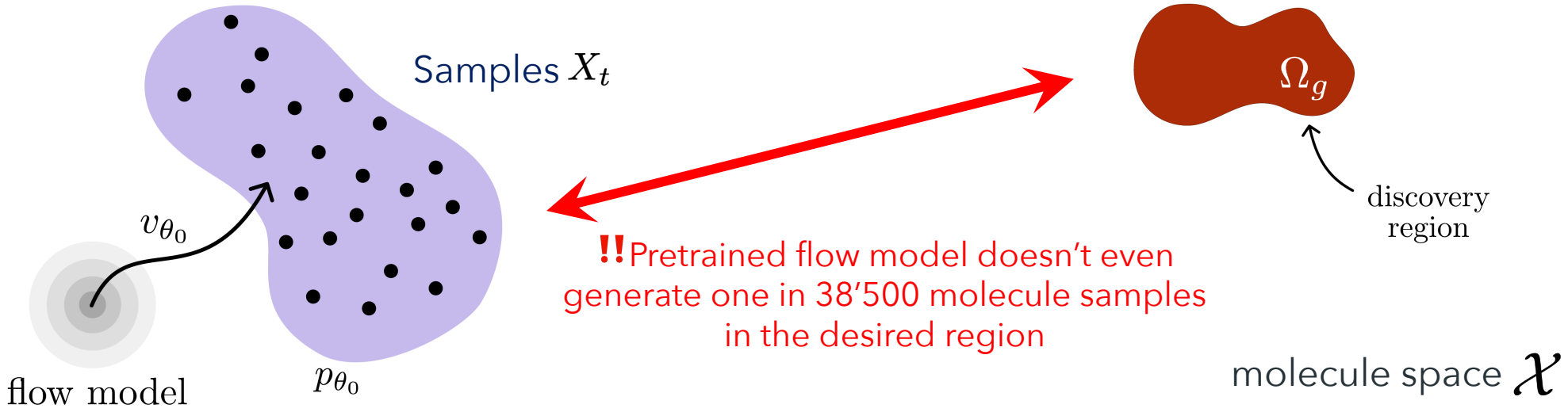
Motivating example: LLM-to-Flow Distillation for discovery

Natural-language discovery goal

g = 'neutral, connected molecules with a six-membered aza-aromatic ring, bearing NH_2 next to a pyridine-like N, and a distinct saturated ring containing a free secondary amine'

Desired discovery region

$$\Omega_g := \{x \in \mathcal{X} : f_g(x) = 1\}$$



"LLM-to-Flow Distillation: Teaching Natural Language Goals to Scientific Generators", De Santi, Di Gennaro, Gnazzo, Tang, Chatterjee, **Y.**, Ceylan, Bronstein, Krause; arxiv '26

Motivating example: LLM-to-Flow Distillation for discovery

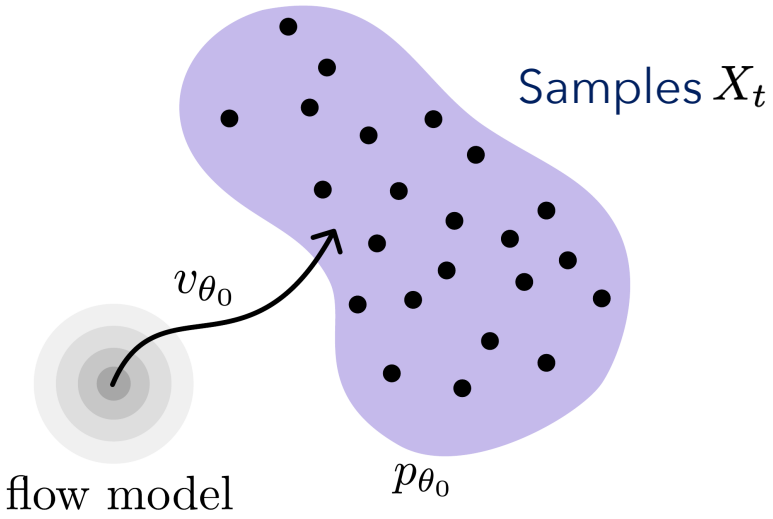
Natural-language discovery goal

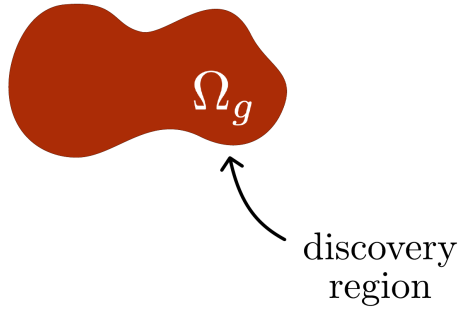
g = 'neutral, connected molecules with a six-membered aza-aromatic ring, bearing NH_2 next to a pyridine-like N, and a distinct saturated ring containing a free secondary amine'

Desired discovery region

$$\Omega_g := \{x \in \mathcal{X} : f_g(x) = 1\}$$

$$g_t \leftarrow \textbf{LLMSetSubgoal}\ (g, g_{t-1}, \mathcal{O}(X_t))$$





molecule space $\mathcal{X}$

Motivating example: LLM-to-Flow Distillation for discovery

Natural-language discovery goal

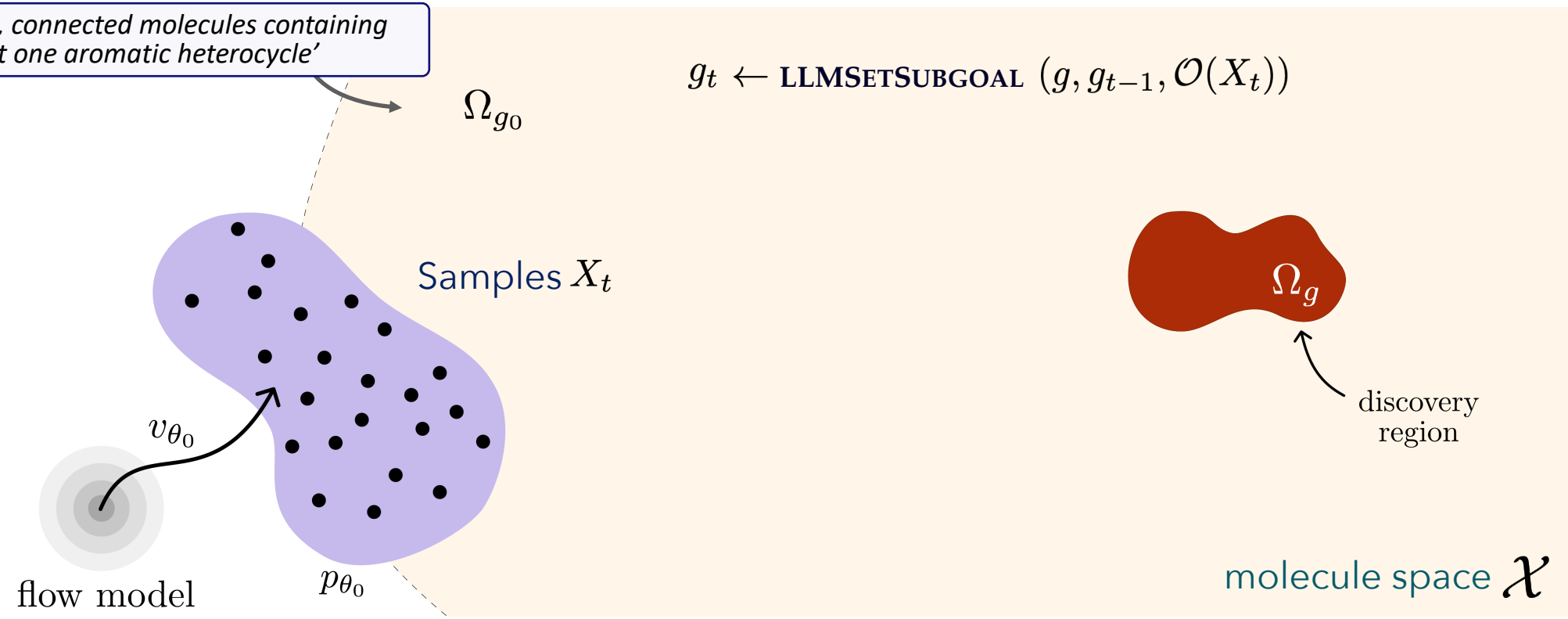
g = 'neutral, connected molecules with a six-membered aza-aromatic ring, bearing NH_2 next to a pyridine-like N, and a distinct saturated ring containing a free secondary amine'

Desired discovery region

$$\Omega_g := \{x \in \mathcal{X} : f_g(x) = 1\}$$

g_0 = 'neutral, connected molecules containing at least one aromatic heterocycle'

$$g_t \leftarrow \textsc{LLMSetSubgoal}\ (g, g_{t-1}, \mathcal{O}(X_t))$$



Motivating example: LLM-to-Flow Distillation for discovery

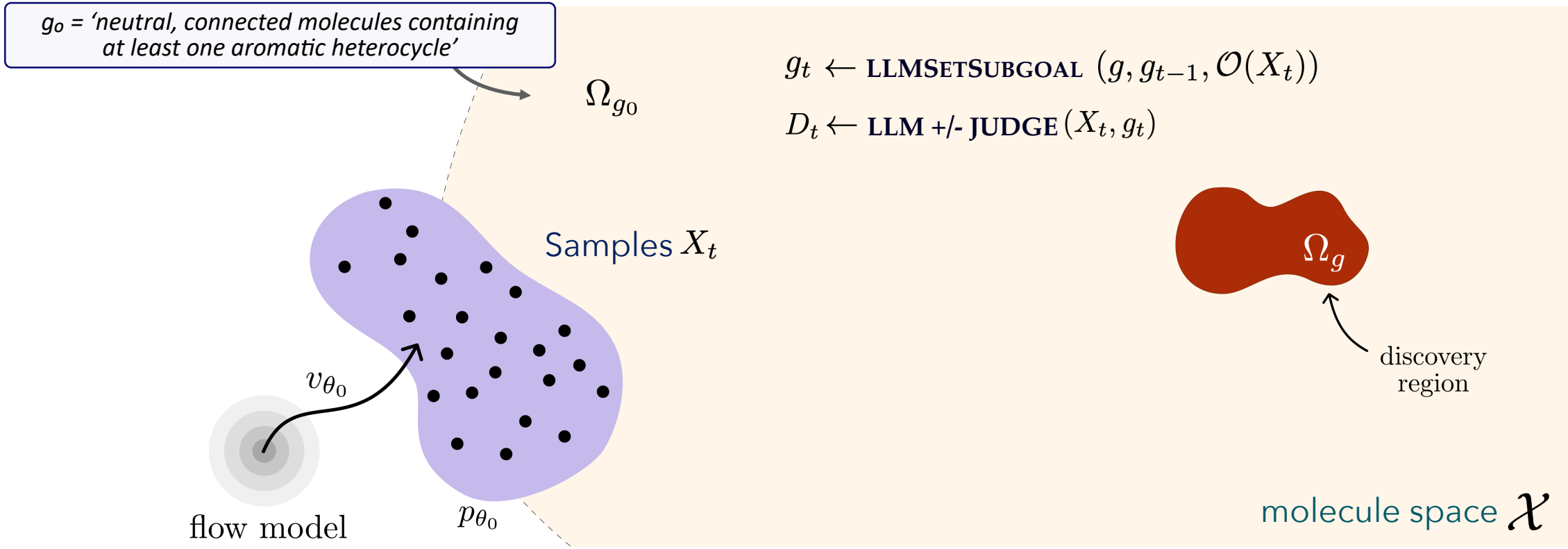
Natural-language discovery goal

g = 'neutral, connected molecules with a six-membered aza-aromatic ring, bearing NH_2 next to a pyridine-like N, and a distinct saturated ring containing a free secondary amine'

Desired discovery region

$$\Omega_g := \{x \in \mathcal{X} : f_g(x) = 1\}$$

g_0 = 'neutral, connected molecules containing at least one aromatic heterocycle'



Motivating example: LLM-to-Flow Distillation for discovery

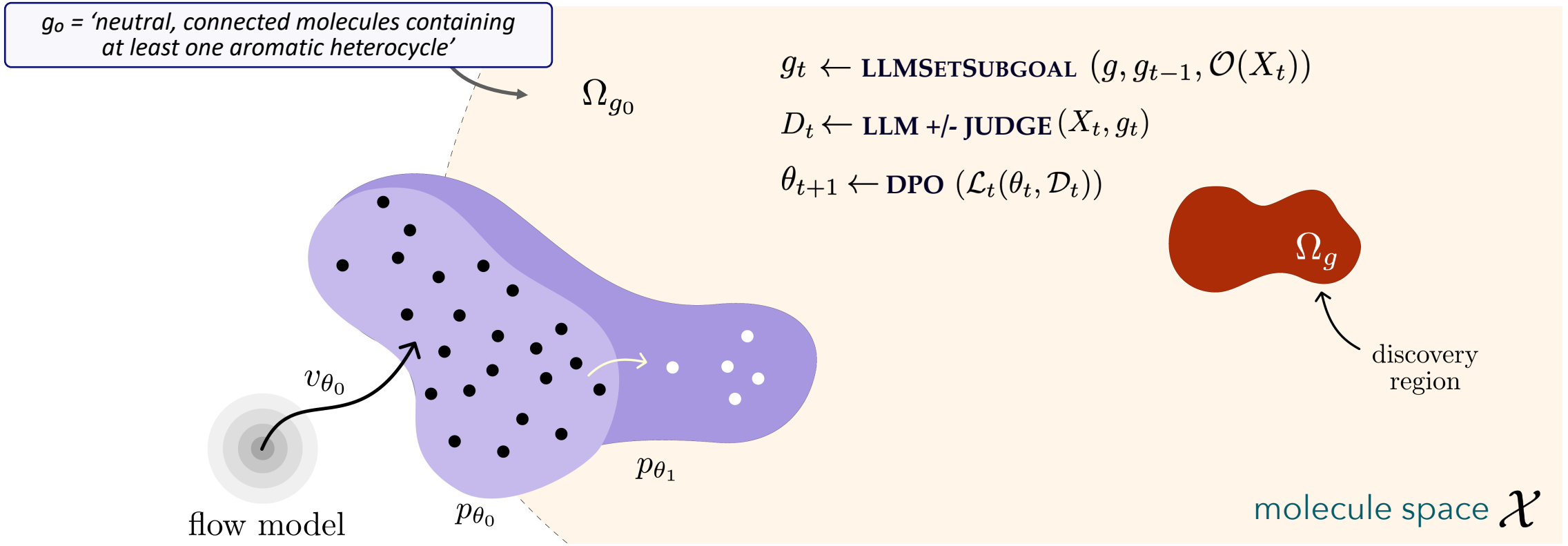
Natural-language discovery goal

g = 'neutral, connected molecules with a six-membered aza-aromatic ring, bearing NH_2 next to a pyridine-like N, and a distinct saturated ring containing a free secondary amine'

Desired discovery region

$$\Omega_g := \{x \in \mathcal{X} : f_g(x) = 1\}$$

g_0 = 'neutral, connected molecules containing at least one aromatic heterocycle'



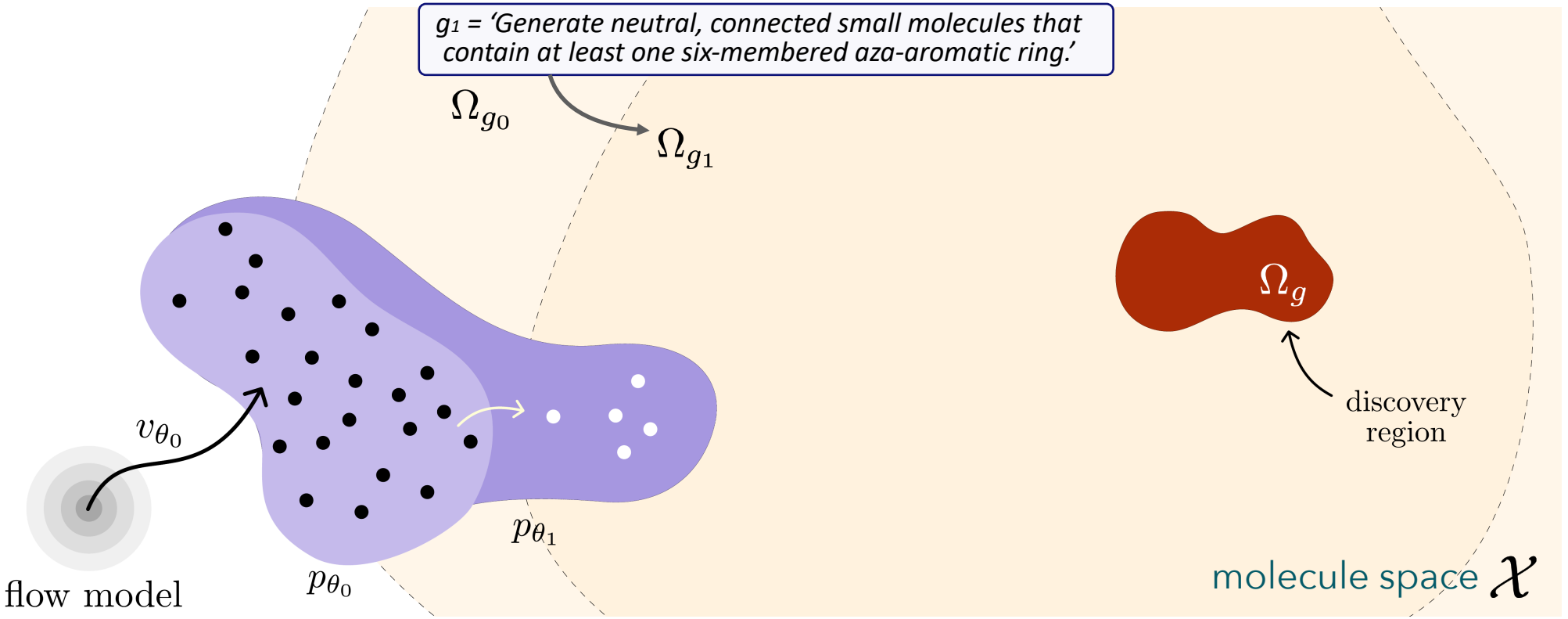
Motivating example: LLM-to-Flow Distillation for discovery

Natural-language discovery goal

g = 'neutral, connected molecules with a six-membered aza-aromatic ring, bearing NH_2 next to a pyridine-like N, and a distinct saturated ring containing a free secondary amine'

Desired discovery region

$$\Omega_g := \{x \in \mathcal{X} : f_g(x) = 1\}$$



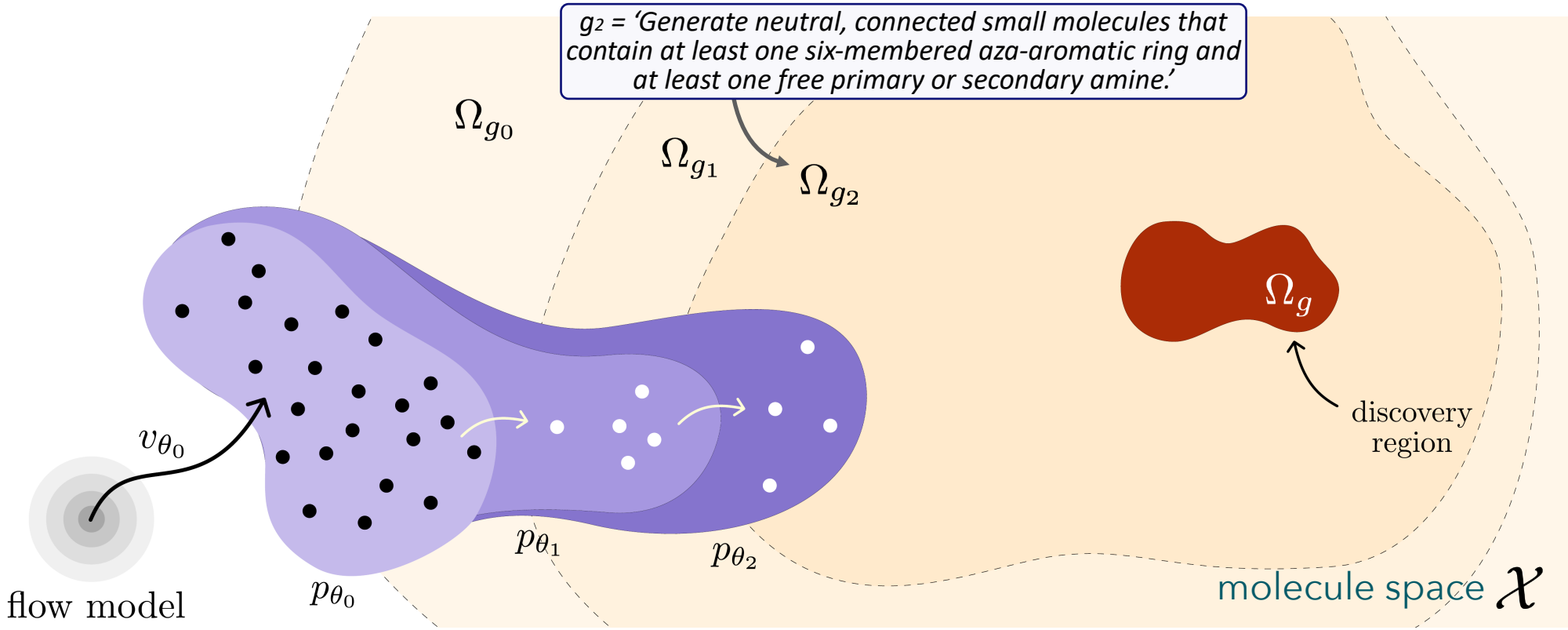
Motivating example: LLM-to-Flow Distillation for discovery

Natural-language discovery goal

g = 'neutral, connected molecules with a six-membered aza-aromatic ring, bearing NH_2 next to a pyridine-like N, and a distinct saturated ring containing a free secondary amine'

Desired discovery region

$$\Omega_g := \{x \in \mathcal{X} : f_g(x) = 1\}$$



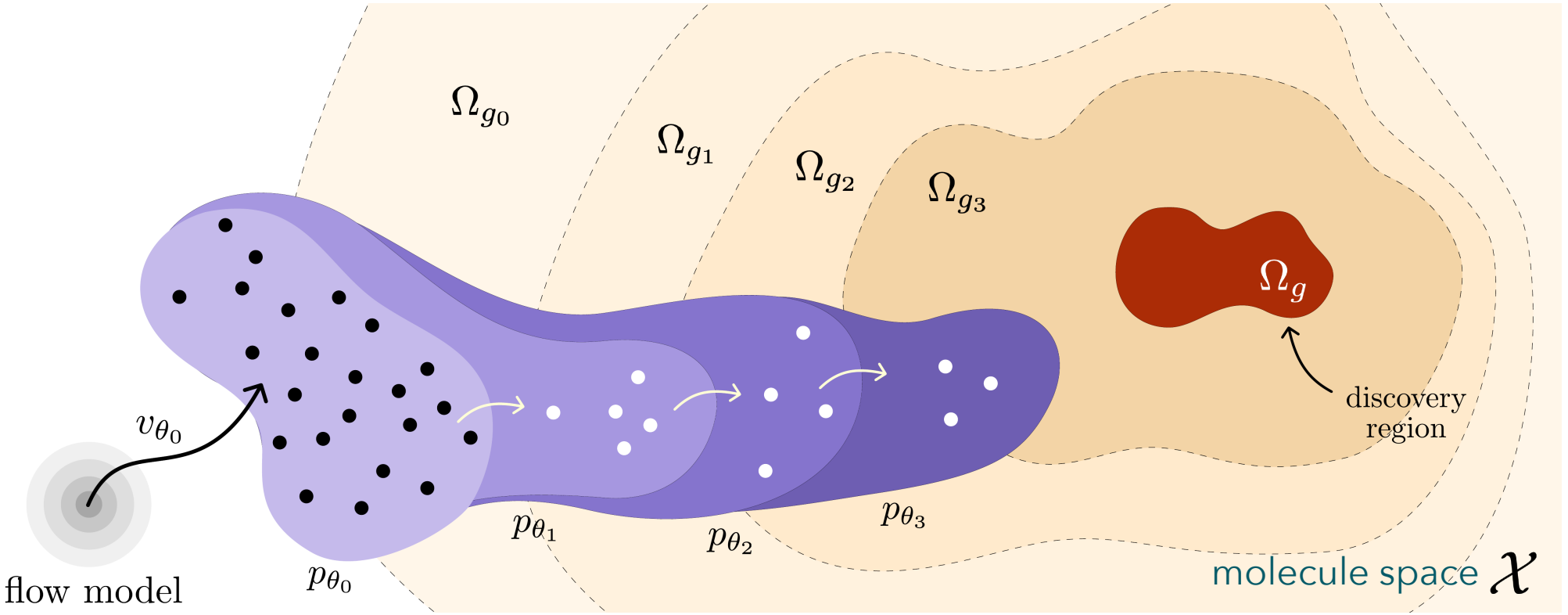
Motivating example: LLM-to-Flow Distillation for discovery

Natural-language discovery goal

g = 'neutral, connected molecules with a six-membered aza-aromatic ring, bearing NH_2 next to a pyridine-like N, and a distinct saturated ring containing a free secondary amine'

Desired discovery region

$$\Omega_g := \{x \in \mathcal{X} : f_g(x) = 1\}$$



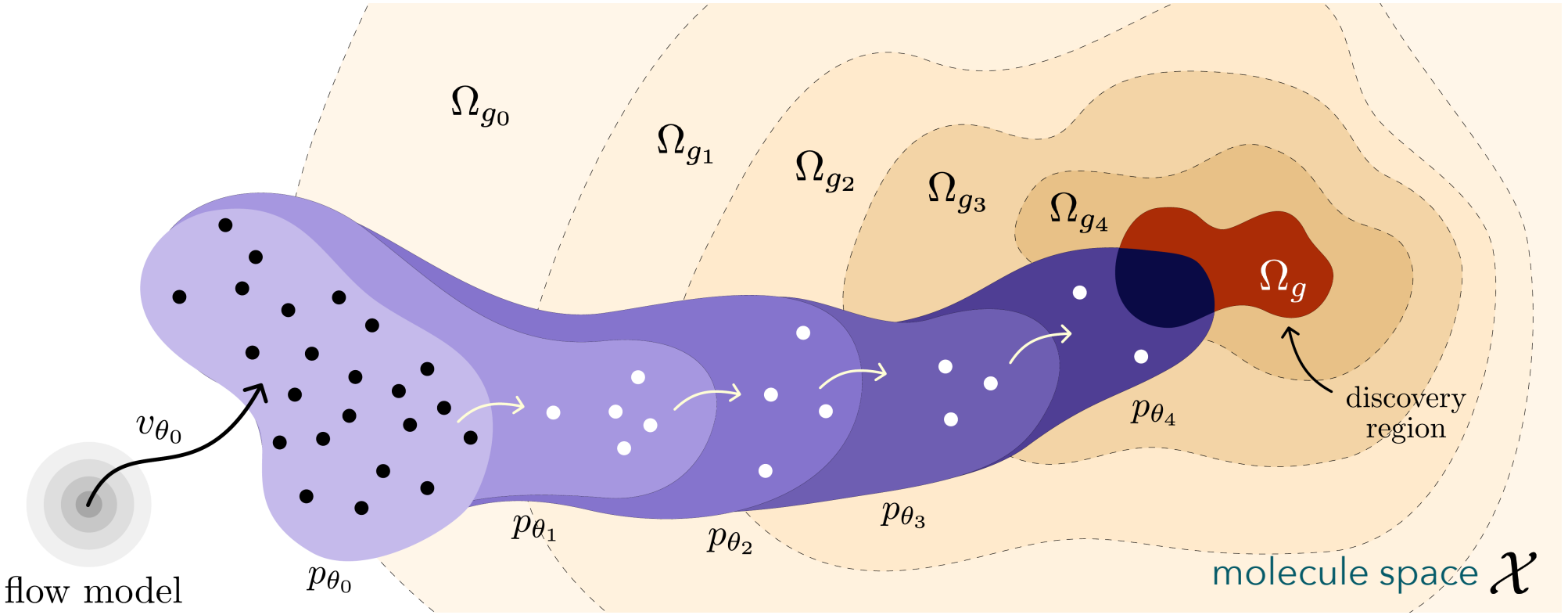
Motivating example: LLM-to-Flow Distillation for discovery

Natural-language discovery goal

g = 'neutral, connected molecules with a six-membered aza-aromatic ring, bearing NH_2 next to a pyridine-like N, and a distinct saturated ring containing a free secondary amine'

Desired discovery region

$$\Omega_g := \{x \in \mathcal{X} : f_g(x) = 1\}$$



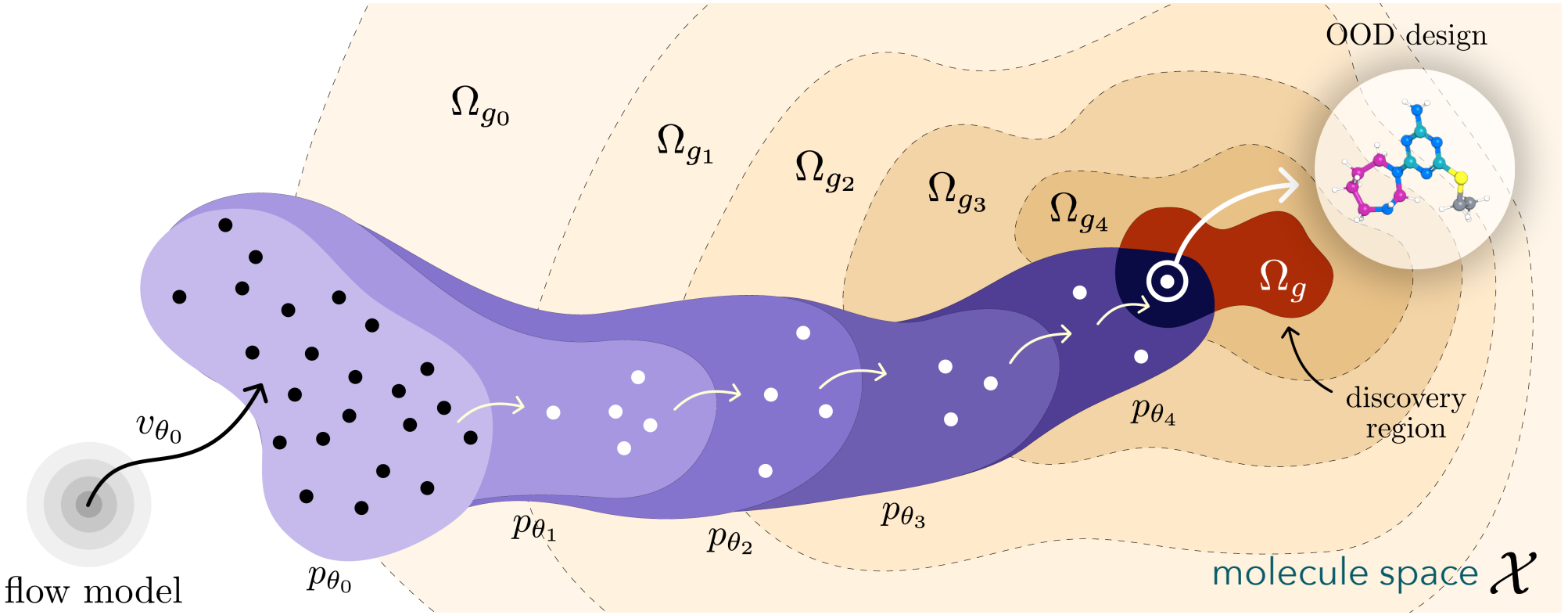
Motivating example: LLM-to-Flow Distillation for discovery

Natural-language discovery goal

g = 'neutral, connected molecules with a six-membered aza-aromatic ring, bearing NH_2 next to a pyridine-like N, and a distinct saturated ring containing a free secondary amine'

Desired discovery region

$$\Omega_g := \{x \in \mathcal{X} : f_g(x) = 1\}$$



Motivating example: LLM-to-Flow Distillation for discovery

Natural-language discovery goal

g = 'neutral, connected molecules with a six-membered aza-aromatic ring, bearing NH_2 next to a pyridine-like N, and a distinct saturated ring containing a free secondary amine'

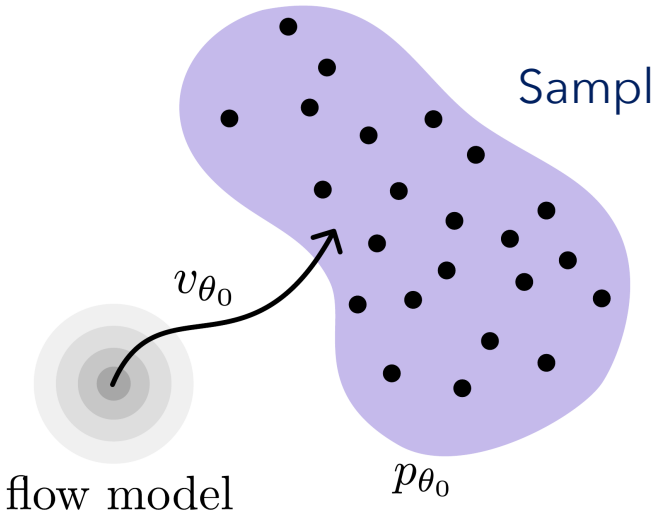
Desired discovery region

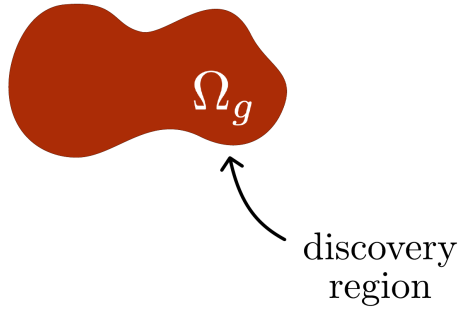
$$\Omega_g := \{x \in \mathcal{X} : f_g(x) = 1\}$$

$$g_t \leftarrow \textbf{LLMSETSUBGOAL}\ (g, g_{t-1}, \mathcal{O}(X_t))$$

$$D_t \leftarrow \textbf{LLM +/- JUDGE}\,(X_t, g_t)$$

Question: Can an LLM actually solve these tasks reliably?



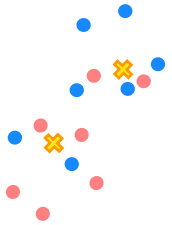


molecule space $\mathcal{X}$

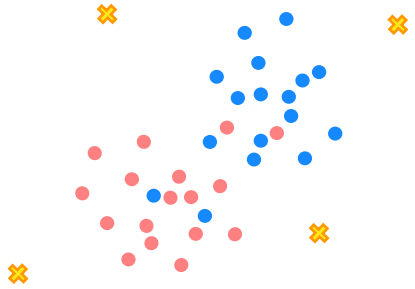
More details can be found in https://llm-to-flow-distillation.github.io/

Different reasons behind "not knowing"

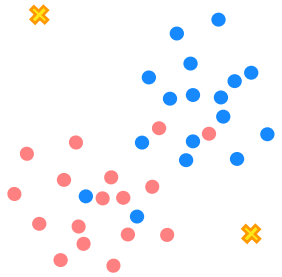
Uncertain test points labeled points seen during training



Noisy data and
too few samples



Anomalous input x
w/o any assumptions
on function class



Anomalous input x
undecidable even for a
given structure of function

may be unnecessarily
conservative

What to do when truly uncertain



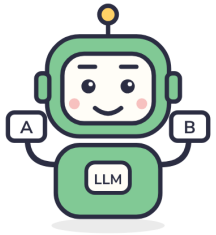
Will tomorrow's *grève* mess up my commute?

Give multiple possible answers that contains the right one/ are all correct w/o extra info

It depends on how you travel — here are the likely cases:

- **A** — **Metro or RER**
  Probably yes: reduced service, packed trains
- **B** — **Bus, bike or on foot**
  Probably not, but avoid the march route

Which one fits your trip?



Part I

For LLMs: Ask for clarification

Could be — it depends on your route. Quick question first:

Which metro or RER lines do you take, and **around what time** do you travel?

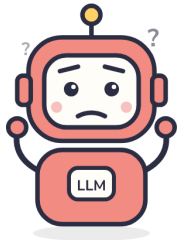
Our focus today



Admit it! (Abstain or answer w/ uncertainty score)

I don't know. I don't have live updates on the strike, so I can't tell you reliably — and I'd rather not guess.

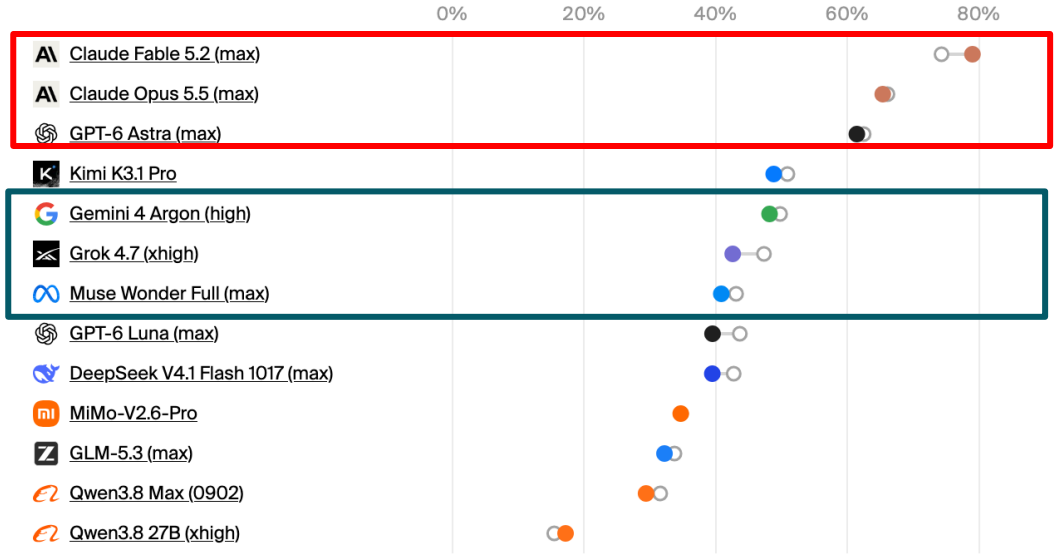
The RATP app will have tomorrow's traffic forecast.



Part II

Status quo: Abstention for LLMs (hallucination rate)

Ranking wrt accuracy



Ranking wrt $\frac{\text{\# incorrect}}{\text{\# incorrect + \# partial corr. / abstained}}$



GPT-5 generated questions & answers (human-checked) in diff. domains, LLM judge checks similarity w/ GT

Despite improvements, best frontier models still don't know when they don't know

Status quo: Diverse and true prediction sets for LLMs

#	System	$utility_{10}$
Raw models	One model, sampled ten times.	
1	claude-opus-5 2026-09-10	2.90
2	claude-fable-5-1 2026-09-10	2.85
3	gpt-6-astra 2026-09-10	2.67
4	gpt-5.6-sol 2026-09-10	2.62
5	gpt-5.6-terra 2026-09-10	2.62
6	gpt-5.6-luna 2026-09-10	2.53
7	gemma-2-27b-it 2025-03-27	2.00

specialized calls

#	System	$utility_{10}$
Inference-time algorithms	A scaffold around a base model,	
1	novamine-ai (Panton, Novamine AI, 2026) runs on claude-opus-5 inference-time system 2026-08-19	6.65
2	novamine-ai-trill (Panton, Novamine AI, 2026) runs on claude-fable-5 inference-time system 2026-08-19	6.25
3	claude-fable-5-1 in-context regeneration 2026-09-10	5.32
4	claude-opus-5 in-context regeneration 2026-09-10	4.94
5	gpt-6-astra in-context regeneration 2026-09-10	4.45

Prompts with ≥ 10 correct answers (tell a joke, list a country capital); $utility_{10}$: ~ different useful outputs out of 10 samples

Unsolved: Frontier models (even specialized) can't generate diverse correct outputs!

PART I: Prediction sets
(conformal prediction framework)

Traditional approaches to output prediction sets instead of points

Goal: Given some labeled data (x_i, y_i) and a test input x_{n+1}, output a set of labels including the true one y_{n+1}

	Ass. on estimator	Guaranteed coverage
• Confidence (Bayesian) prediction intervals	consistent (known noise family)	asymptotically valid as $n \to \infty$ (finite sample in special cases)
• Bootstrap/Jackknife	smooth/regular	asymptotically valid as $n \to \infty$
• CV+/Jackknife+	symmetric	for finite samples $n < \infty$
• Conformal prediction	none	(only marginally)

Trend towards

- fewer assumptions on models & noise
- finite-sample guarantees

Next up: Primer on conformal prediction and variants, usage on LLMs

(Split) Conformal prediction for black-box coverage – in one slide

Example: regression
(for trained model $\hat{\mu}$)

$$C_t(x) = \{Y : |\hat{\mu}(x) - Y| \leq t\}$$

$$t(x, y) = |\hat{\mu}(x) - y|$$

General framework: We're given a model…

- and a way to construct confidence set $C_t(x)$ increasing with t
- define for any point (x, y) score $t(x, y) = \inf\{t : y \in C_t(x)\}$
  smallest $C_t(x)$ *(smallest t)* so that it includes y

(Split) Conformal prediction for black-box coverage – in one slide

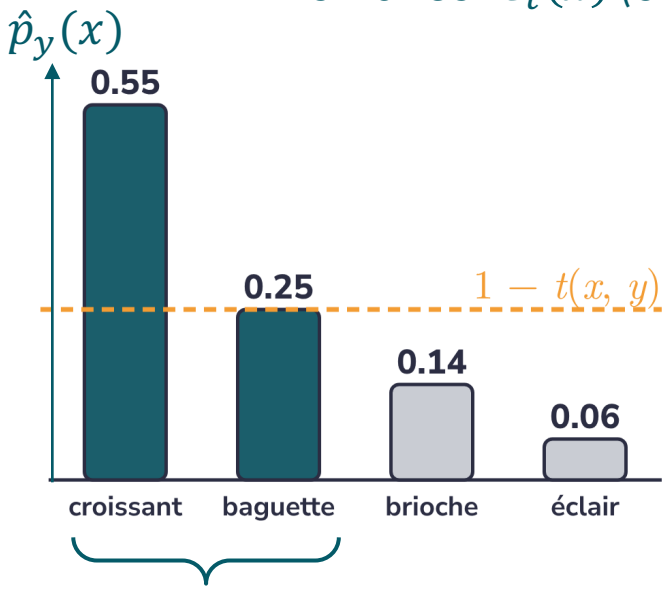
Example: classification
(for trained model $\hat{p}$)

$$C_t(x) = \{Y : \hat{p}_Y(x) \geq 1 - t\}$$

$$t(x, y) = 1 - \hat{p}_y(x)$$



input x
label: baguette



General framework: We're given a model…

- and a way to construct confidence set $C_t(x)$ increasing with t
- define for any point (x, y) score $t(x, y) = \inf\{t : y \in C_t(x)\}$
  smallest $C_t(x)$ *(smallest t)* so that it includes y

smallest $C_t(x)$ including y
is {croissant, baguette}
for $t(x, y) = 1 - 0.25 = 0.75$

(Split) Conformal prediction for black-box coverage – in one slide

Example: classification
(for trained model $\hat{p}$)

$$C_t(x) = \{Y : \hat{p}_Y(x) \geq 1 - t\}$$

$$t(x, y) = 1 - \hat{p}_y(x)$$

General framework: We're given a model…

- and a way to construct confidence set $C_t(x)$ increasing with t
- define for any point (x, y) score $t(x, y) = \inf\{t : y \in C_t(x)\}$ smallest $C_t(x)$ *(smallest t)* so that it includes y

Conformal paradigm tell us which t to choose!

1. Split to train and calibration set $D = \{(x_i, y_i)\}$ with n points



2. Train some model, compute scores $\{t(x_i, y_i)\}$ on calibration set
3. Compute $\hat{q}_{1-\alpha}$ that is $1 - \alpha$ quantile of $\{t(x_i, y_i)\}$
4. For new point x_{n+1} return $C_{\hat{q}_{1-\alpha}}(x_{n+1})$

No matter how good the model is!



Guarantees under exchangeability of $(X_i, Y_i) \in D$ and (X_{n+1}, Y_{n+1}):

$$P\left(Y_{n+1} \in C_{\hat{q}_{1-\alpha}}(X_{n+1})\right) \geq 1 - \alpha \quad \Leftrightarrow \quad P(t(X_{n+1}, Y_{n+1}) \leq q_{1-\alpha}(\{t(X_i, Y_i)\}) \geq 1 - \alpha$$

Distribution shifts and extrapolation

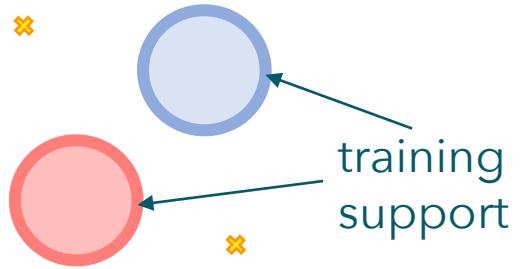
Conformal prediction under covariate shift $P_{test}(x) \neq P_{train}(x)$

1. Split to train and calibration set $D = \{(x_i, y_i)\}$



2. Train some model, compute scores $\{t(x_i, y_i)\}$ on calibration set
3. Estimate weights $w(x) = \frac{p_{test}(x)}{p_{train}(x)}$ and normalize $w_i = \frac{w(x_i)}{\sum_{i=1}^{n+1} w(x_i)}$
4. Compute $\tilde{q}_{1-\alpha}$ that is $1-\alpha$ quantile of $\{t(x_i, y_i)$ weighted by $w_i\}$
5. For new point x_{n+1} return $C_{\tilde{q}_{1-\alpha}}(x_{n+1})$

For unseen support: $x \notin supp_{train}$



- no calibration set for which exchangeability holds
- reweighting would give ∞

Unsolved: conformal cannot be used for OOD detection and abstention!

Extension: Conformal risk control

General goal: Given way to construct confidence set $C_t(x)$ (e.g. $= \{Y : \hat{p}_Y(x) \geq 1 - t\}$)

Find t s.t. $\mathbb{E}[\ell(C_t(x_{n+1}), y_{n+1})] \leq \alpha$ (over calibration data and test point)

for $\ell(C_t(x_{n+1}), y_{n+1}) = \mathbb{I}\{y_{n+1} \notin C_t(x_{n+1})\}$ is just conformal prediction

For example: Assume $Y \subset [K]$ (multiple labels out of K are correct)

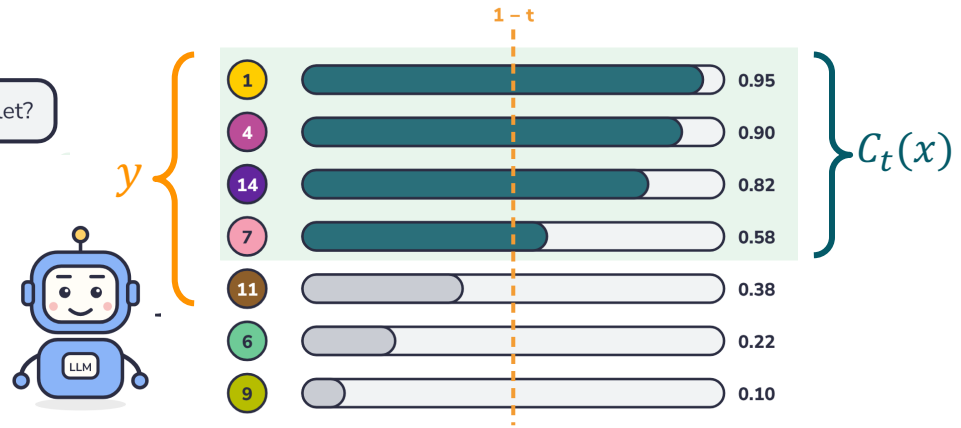
$\ell(C_t(x), y) = 1 - \frac{|y \cap C_t(x)|}{|y|} \triangleq$ false negative rate (how many true labels were missed)

input x:



Which metro lines stop at Châtelet?

true $y = \{1,4,7,11,14\}$



$C_t(x)$ → $\ell(C_t(x), y) = 1 - \frac{4}{5} = \frac{1}{5}$

Extension: Conformal risk control

General goal: Given way to construct confidence set $C_t(x)$ (e.g. $= \{Y : \hat{p}_Y(x) \geq 1 - t\}$)

Find t s.t. $\mathbb{E}[\ell(C_t(x_{n+1}), y_{n+1})] \leq \alpha$ (over calibration data and test point)

for $\ell(C_t(x_{n+1}), y_{n+1}) = \mathbb{I}\{y_{n+1} \notin C_t(x_{n+1})\}$ is just conformal prediction

For example: Assume $Y \subset [K]$ (multiple labels out of K are correct)

$\ell(C_t(x), y) = 1 - \frac{|y \cap C_t(x)|}{|y|} \triangleq$ false negative rate (how many true labels were missed)

Choose $\hat{t} = \inf\left\{t : \frac{1}{n}\sum_{i=1}^{n} \ell(C_t(x_i), y_i) \leq \alpha\right\}$ and output $C_{\hat{t}}(x_{n+1})$

Unsolved: hard to achieve "useful" diversity for sampling text instead of numbers!

Extension: Conformal prediction to combat LLM hallucinations

preprocess sampled $\hat{y}$

- Separate $\hat{y}$ into individual claims $c_1(\hat{y}), \ldots, c_K(\hat{y})$ (e.g. using LLM)
- Score each claim $c_i(\hat{y})$ for correctness → $s(c_i(\hat{y}))$ (e.g. LLM as judge)



Extension: Conformal prediction to combat LLM hallucinations

preprocess sampled $\hat{y}$

- Separate $\hat{y}$ into individual claims $c_1(\hat{y}), \ldots, c_K(\hat{y})$ (e.g. using LLM)
- Score each claim $c_i(\hat{y})$ for correctness → $s(c_i(\hat{y}))$ (e.g. LLM as judge)

calibrated rule for test time

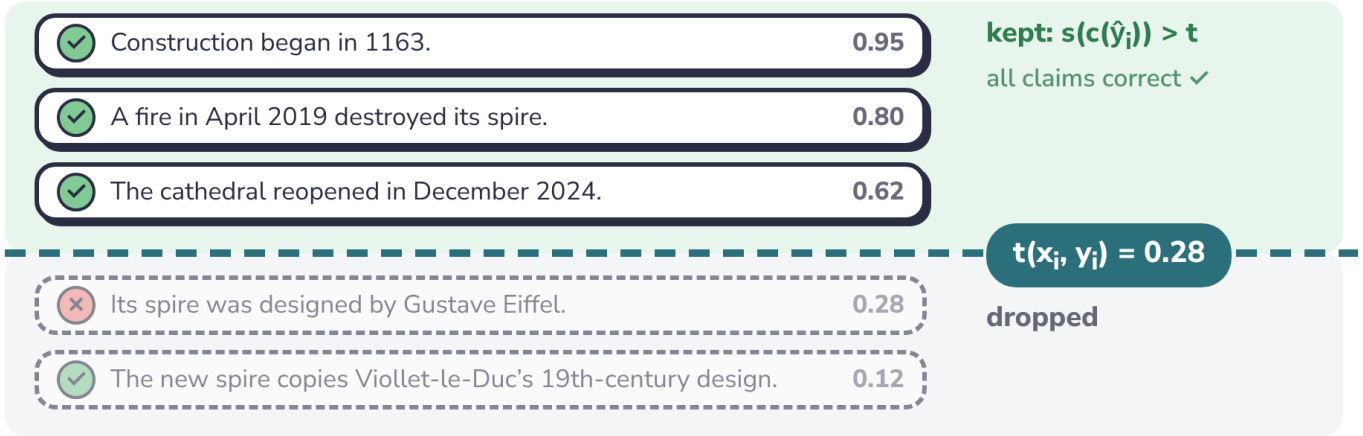
- For all calibration samples (x_i, y_i) compute scores $t(x_i, y_i)$

  s.t. all c_j with $s\left(c_j(\hat{y}_i)\right) \geq t$ are correct (i.e. implied by y_i)

For one calibration example:

Prompt x_i: "Summarise the history of Notre-Dame de Paris."

Reference answer y_i (ground truth)
Construction began in 1163 on the Île de la Cité. The 19th-century spire by Eugène Viollet-le-Duc collapsed in the April 2019 fire; it was rebuilt to the same design, and the cathedral reopened in December 2024.



Extension: Conformal prediction to combat LLM hallucinations

preprocess sampled $\hat{y}$

- Separate $\hat{y}$ into individual claims $c_1(\hat{y}), \ldots, c_K(\hat{y})$ (e.g. using LLM)
- Score each claim $c_i(\hat{y})$ for correctness → $s(c_i(\hat{y}))$ (e.g. LLM as judge)

calibrated rule for test time

- For all calibration samples (x_i, y_i) compute scores $t(x_i, y_i)$

  s.t. all c_j with $s\left(c_j(\hat{y}_i)\right) \geq t$ are correct (i.e. implied by y_i)
- Compute quantile $q_{1-\alpha}$ of $\{t(x_i, y_i)\}$

During test time

- For a new test input x_{n+1}, sample a response $\hat{y}$ from the data, split to claims and

  output a merger of all claims $c_i(\hat{y})$ with score $s\left(c_i(\hat{y})\right) \geq q_{1-\alpha}$



Expensive: Calibration to be done for each model anew, LLM judge might be bad

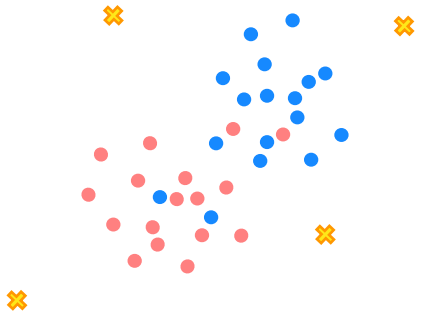
PART II: Abstention from extrapolation

What are points in unseen regions we cannot extrapolate on?

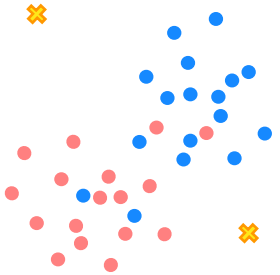
Anomalous inputs x or featurized z without any assumptions on optimal prediction function

Anomalous inputs x or featurized z even given structural assumptions on optimal prediction function

Uncertain test points



sometimes too conservative



1. Estimate marginal P_x / P_z or conditional $P_{x|y}$ / $P_{z|y}$ and define "typicality" based on these
2. Flag any with small typicality

1. Determine which points x can't we extrapolate, for given structural ass. & training data
2. Only abstain on those!

Next: Some examples of structural assumptions of functions where we can extrapolate on anomalous x

Extrapolation for simple models (e.g. linear)

Assume we know the Bayes optimal classifier is linear

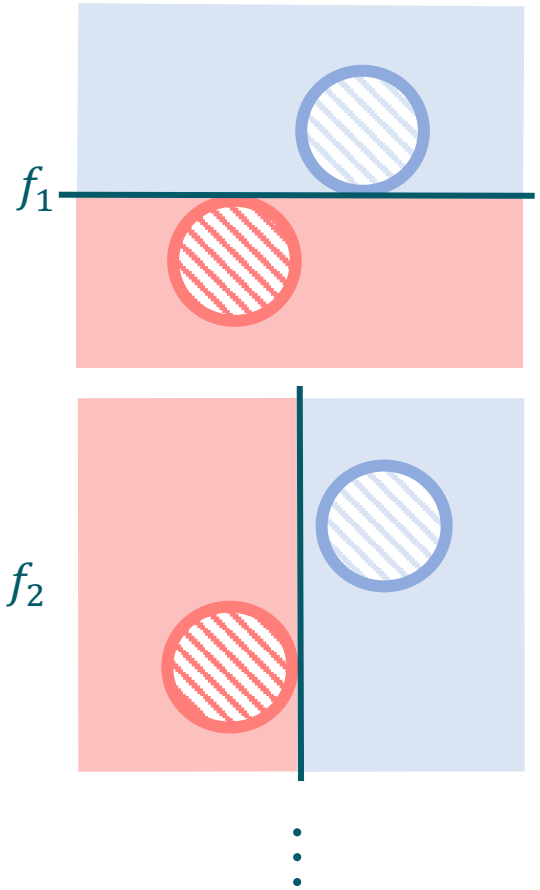
Linear classifiers

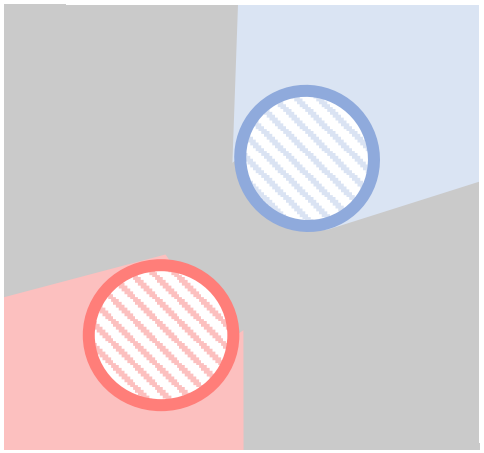
that correctly fit

training distributions



with predictions





 : input region where all linear models good on training data predict the same

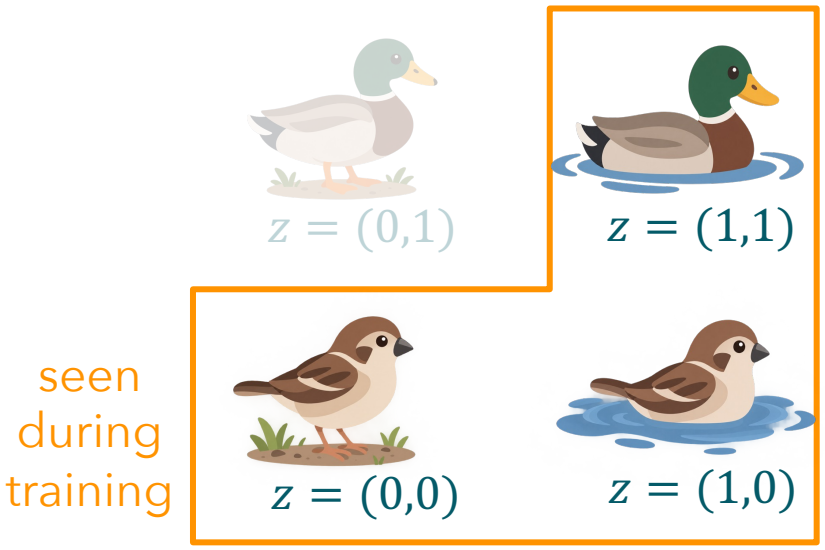
 : input region where some of these models disagree

Can't extrapolate, should abstain on

How to identify?

Extrapolation via compositionality

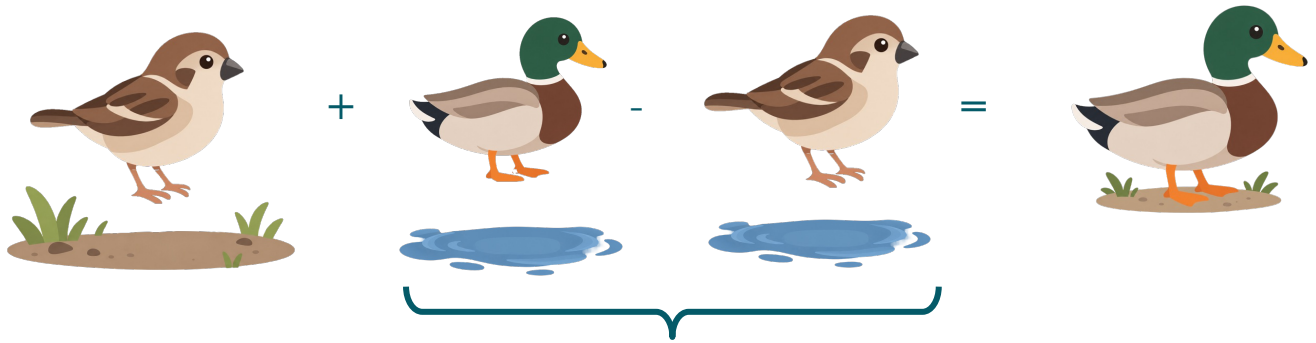
One modeling assumption allowing extrapolation to unseen combinations of factors: additive compositionality



z = (birdtype, background)

Can we "predict" how $z = (0,1)$ looks given training data?

$z = (0,1) = (0,0) + (1,1) - (1,0)$



background cancels

Some other types of extrapolation via compositionality: Beyond additive compositions (Wiedemer et al. NeurIPS '23)
Chained functions (Ramesh et al. ICML '24), modular tasks (Schug et al ICML'24), language skills (Lake & Baroni ICML'18, Zhao et al NeurIPS '24)

e.g. "Compositional Risk Minimization", Mahajan, Pezeshki, Arnal, Mitliagkis, Ahuja, Vincent @ICML '25

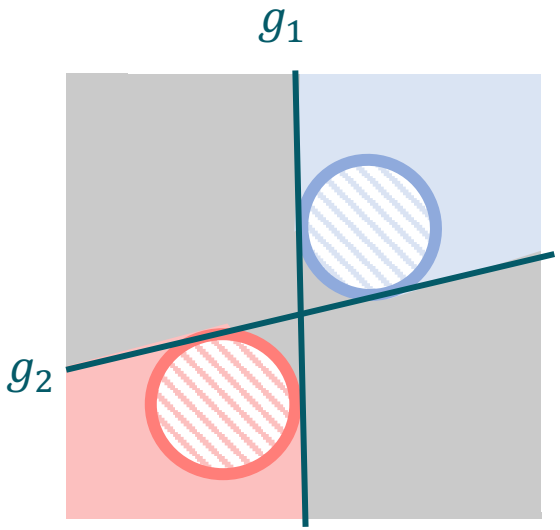
Using ensemble-based disagreement (e.g. for linear)

Assume we know the Bayes optimal classifier is linear

Fit an ensemble of models $\{\hat{p}_i(y|x)\}$ (e.g. $\{g_1, g_2\}$) that

- have low validation accuracy on training dist.
- predict differently outside of training distribution

→ abstain on samples where $\hat{p}_i(y|x)$ very different



: input region where all linear models good on training data predict the same

: input region where some of these models disagree

Ensembles can be obtained via

- vanilla ensembles[1]
- Bayesian posterior over models[2]

} typically not diverse enough!

- artificial labels[3] (see next slide)

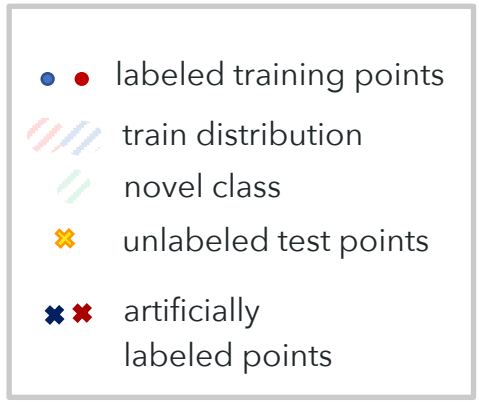
Can't extrapolate, should abstain on

How to identify?

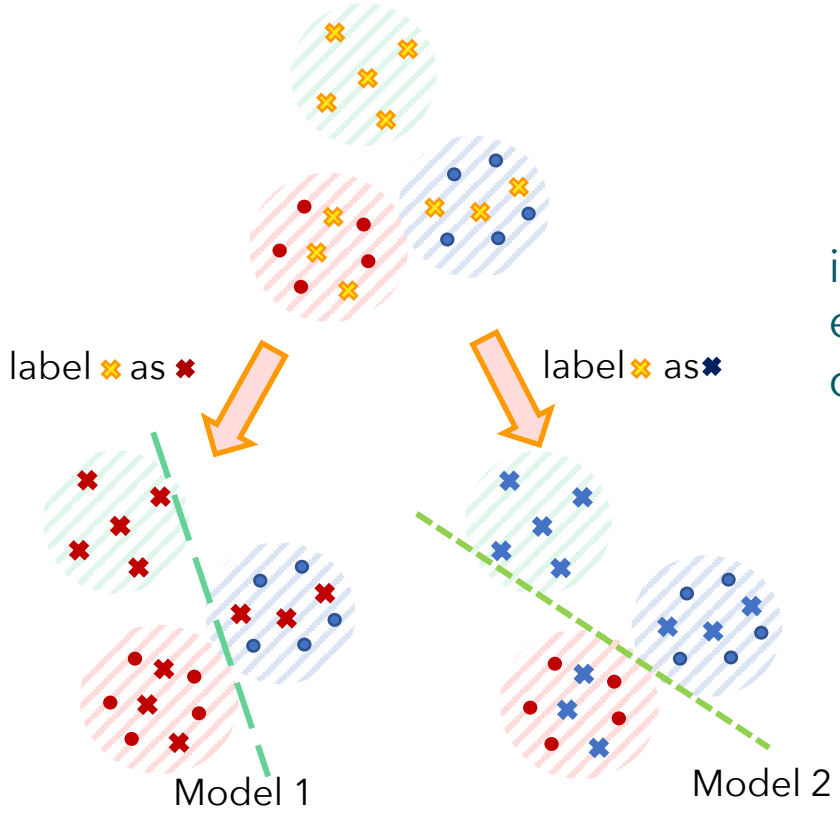
[1]Gal-Gahramani @ICML'16, Kristiadi et al. @ICML'20; [2]Lakshminarayanan et al @NeurIPS'17, [3]see also Bibas et al @NeurIPS'21

Creating diverse ensembles with artificial labels

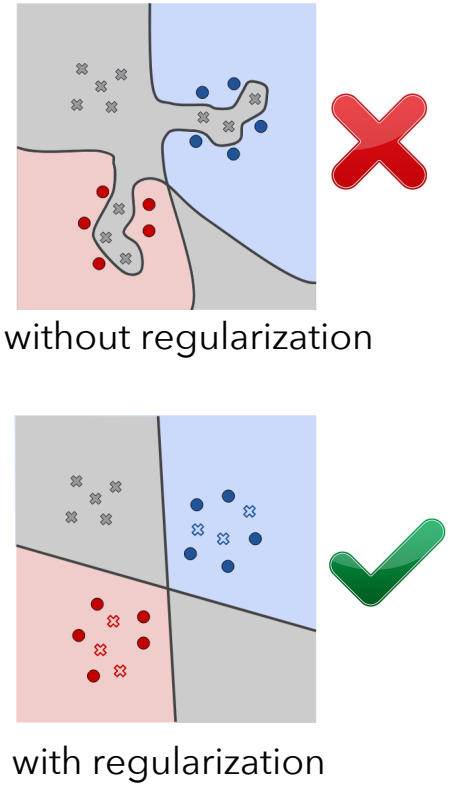
Given a dataset that has both OOD and extrapolatable inputs:



Goal: obtain ensembles that fit training dist. well in function class but disagree maximally on rest

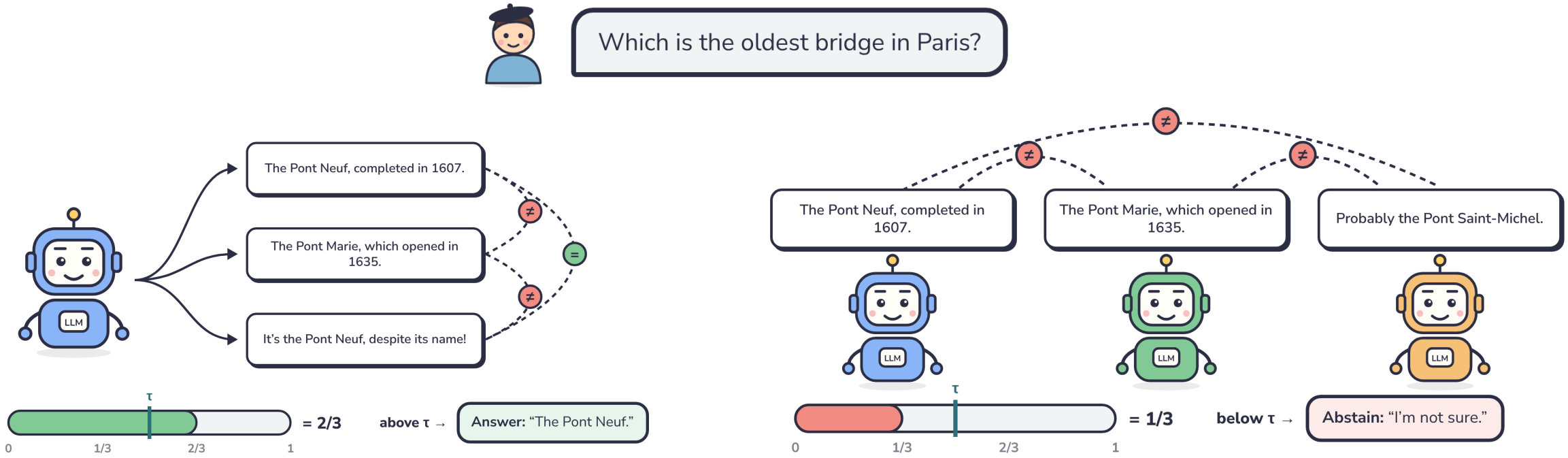


identify where ensembles disagrees



Ensemble with regularized disagreement: They have to do well on the validation set of training dist.!

Disagreement-based Abstention for LLMs



- Sample multiple responses from one (left) or different models (right)
- Determine pairwise disagreements and max # responses that agree
- Compare with a threshold τ (e.g. via conformal) → if too small, abstain

Unsolved: Cannot be extended when multiple responses are correct

Right, e.g. "SAC³: Reliable Hallucination Detection in Black-Box Language Models via Semantic-aware Cross-check Consistency", Zhang et al. @EMNLP Findings '23;
Left, e.g. "Mitigating LLM Hallucinations via Conformal Abstention", Abbasi-Yadkori, Kuzborskij, Stutz, György, Fisch, Doucet, Beloshapka, Weng, Yang, Szepesvári, Cemgil, Tomasev; arxiv '24

Back to LLM-to-Flow Distillation (LFD) for discovery

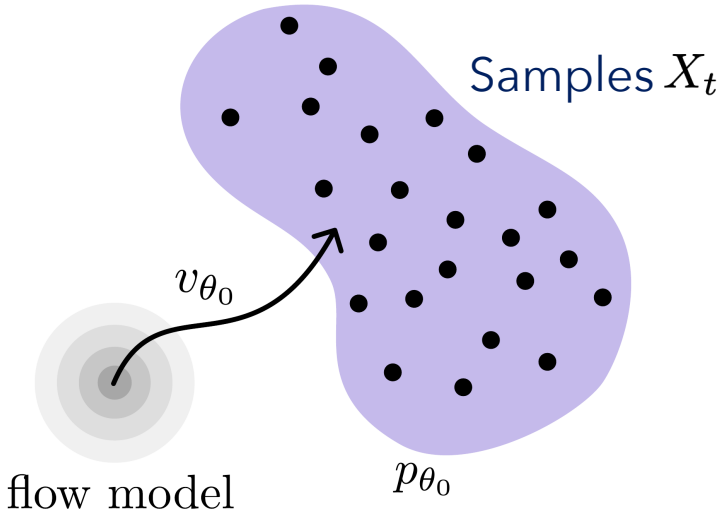
Natural-language discovery goal

g = 'neutral, connected molecules with a six-membered aza-aromatic ring, bearing NH_2 next to a pyridine-like N, and a distinct saturated ring containing a free secondary amine'



Desired discovery region

$$\Omega_g := \{x \in \mathcal{X} : f_g(x) = 1\}$$



$$g_t \leftarrow \textbf{LLMSETSUBGOAL}\ (g, g_{t-1}, \mathcal{O}(X_t))$$

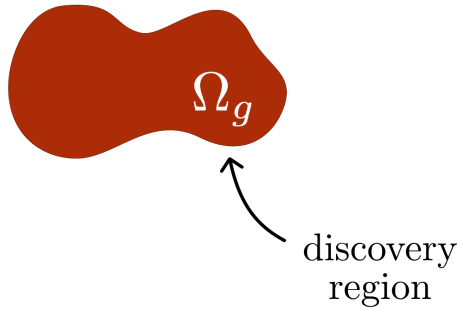
$$D_t \leftarrow \textbf{LLM +/- JUDGE}\,(X_t, g_t)$$

Question: Can an LLM actually solve these tasks reliably?

Current LLMs cannot

- produce diverse possible subgoals g_t instead of abstaining in that case
- reliably flag when they can't judge which samples in X_t align with g_t

Especially hard when no perfect verifiers exist.



molecule space $\mathcal{X}$

$\Longrightarrow$ Still lots to be done for reliable scientific discovery with generative AI…

Papers mentioned in the talk

Motivating example and LLM uncertainty benchmarks

- "LLM-to-Flow Distillation: Teaching Natural Language Goals to Scientific Generators", De Santi, Di Gennaro, Gnazzo, Tang, Chatterjee, **Y.**, Ceylan, Bronstein, Krause; arxiv preprint '26
- "AA-Omniscience: Evaluating Cross-Domain Knowledge Reliability in Large Language Models", Jackson, Keating, Cameron, Hill-Smith; arxiv '25
- "NoveltyBench: Evaluating Language Models for Humanlike Diversity", Zhang, Diddee, Holm, Liu, Liu, Samuel, Wang, Ippolito @COLM '25

Conformal prediction

- "A gentle introduction to Conformal Prediction and Distribution-Free Uncertainty quantification", Angelopoulos, Bates; arxiv preprint '21
- "Conformal prediction under covariate shift", Tibshirani, Barber, Candès, Ramdas @NeurIPS '19
- "Conformal risk control", Angelopoulos, Bates, Fisch, Lei, Schuster @ICLR '24
- "Language Models with Conformal Factuality Guarantees", Mohri, Hashimoto @ICML '24

Extrapolation, Outlier detection and abstention

- "Compositional risk minimization", Mahajan, Pezeshki, Arnal, Mitliagkis, Ahuja, Vincent @ICML '25
- "Semi-supervised novelty detection using ensembles with regularized disagreement", Tifrea, Stavarache, **Y.** @UAI '22





Statistical Machine Learning Group

sml.inf.ethz.ch