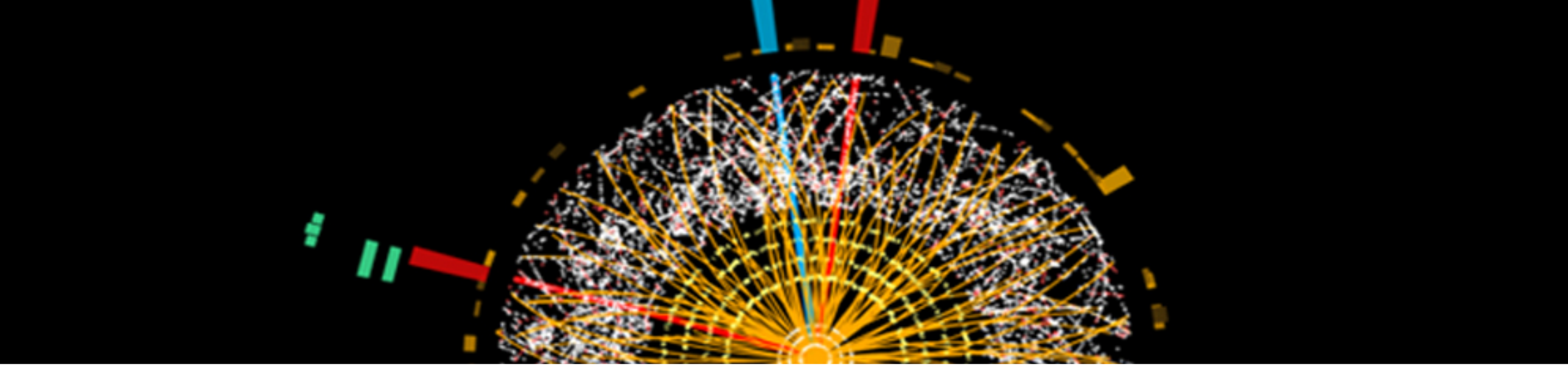


Preference-Optimized Generative Models for Underconstrained Inference in Particle Physics

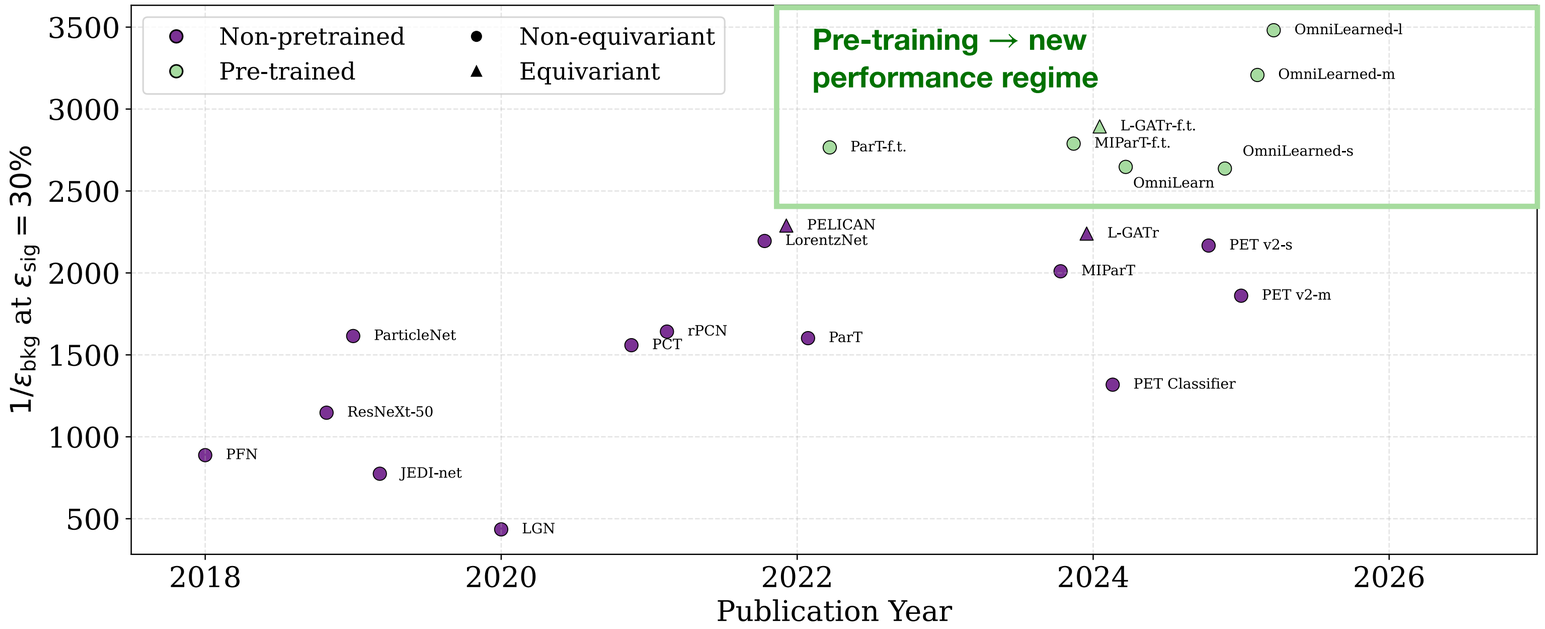
16 Sep 2026, University of Vienna, ML4Jets

Team: Yi-Ren Wu[1], **Ting-Hsiang Hsu[1]**, Yuan-Tang Chou[2], Benjamin Nachman[3], Stathes Paganis[1], Shih-Chieh Hsu[4], Vinicius Massami Mikuni[5], Yulei Zhang[4]

[1] [2] [3] [4] [5]

HEP is Entering an Era of Large-Scale Pre-training

- **Large-scale pre-training** has already demonstrated clear performance gains.
- **Foundation models** now span detector, jet, and event representations, with transfer across diverse downstream tasks.



Detector
TrackingBERT · Panda
FM4NPP

Jets
ParT · MPM · JEPA
OmniJet · OmniLearn · OmniLearned

Event / Analysis
BumbleBee
EveNet

tracking · particle ID
noise detection

tagging · generation
anomaly detection

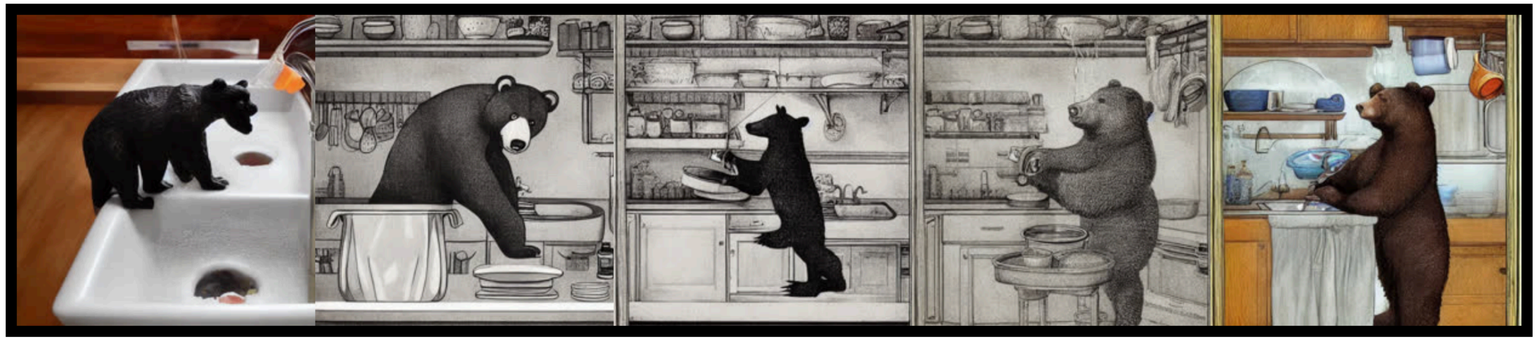
reconstruction · searches
anomaly detection

Pre-training is Only Half the Story - Post-training

Post-training goes beyond traditional ~~task-specific fine-tuning~~:
it shapes pretrained capabilities toward desired behavior through feedback and objectives.

	Pre-training	Reinforce learning / Fine-tuning…	Post-training
Language	Learn language · Context Structure		reasoning · instruction following
Image	Learn visual distribution		Prompt fidelity · human preference

a bear washing dishes →



Black et al., *Training Diffusion Models with Reinforcement Learning*, ICLR 2024

Pre-training is Only Half the Story - Post-training

Post-training goes beyond traditional ~~task-specific fine-tuning~~:
it shapes pretrained capabilities toward desired behavior through feedback and objectives.

	Pre-training	Reinforce learning / Fine-tuning… →	Post-training
Language	Learn language · Context Structure		reasoning · instruction following
Image	Learn visual distribution		Prompt fidelity · human preference

a bear washing dishes →

Black et al., *Training Diffusion Models with Reinforcement Learning*, ICLR 2024

What does “Preference” Mean for Scientific Inference?

Depends on downstream scientific objective…

Physics Fidelity

Analysis Sensitivity

Unbiased Estimation

Etc.…

What does “Preference” Mean for Scientific Inference?

Depends on downstream scientific objective…

Physics Fidelity

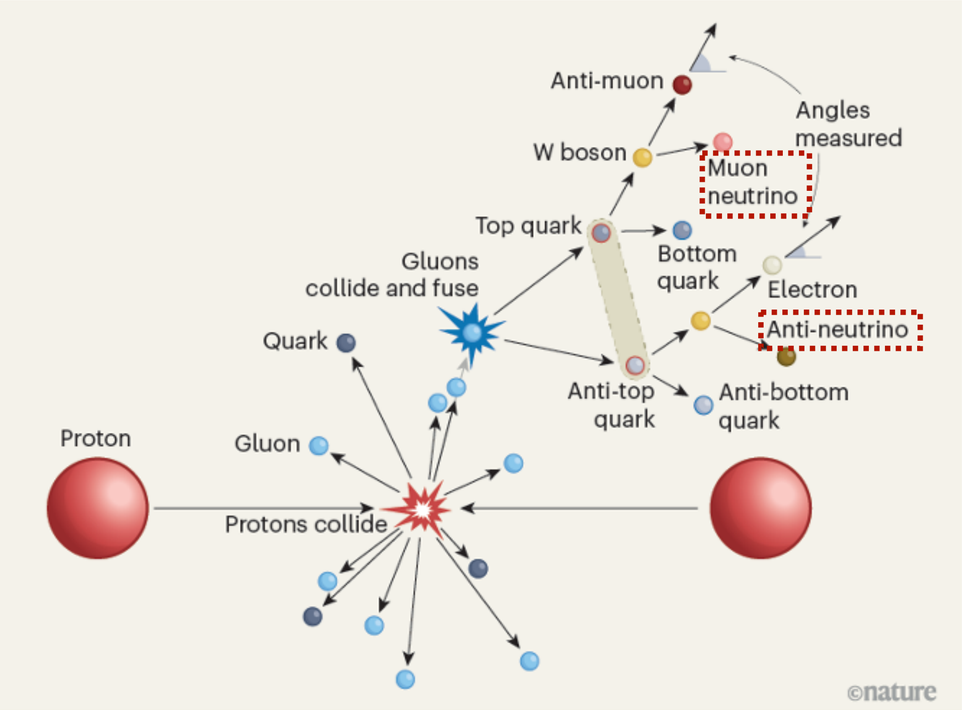
Analysis Sensitivity

Unbiased Estimation

Etc.…

Our Testbed: Underconstrained ν reconstruction

- Essential for **precision measurements** like quantum entanglement.
- No unique solution → generative models.
- Generated model do not guarantee **physically consistent** solutions, especially when the relevant physics constraints are soft rather than exact.



What does “Preference” Mean for Scientific Inference?

Depends on downstream scientific objective…

Physics Fidelity

Analysis Sensitivity

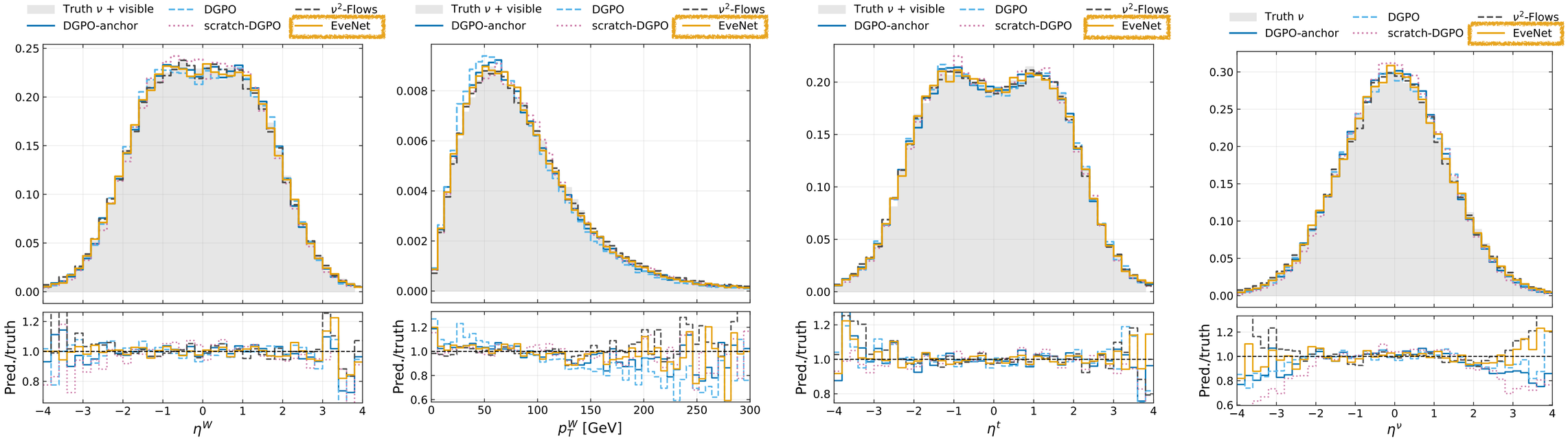
Unbiased Estimation

Etc.…

Our Testbed: Underconstrained ν reconstruction

Foundation Model f.t.

All fit well, but…



What does “Preference” Mean for Scientific Inference?

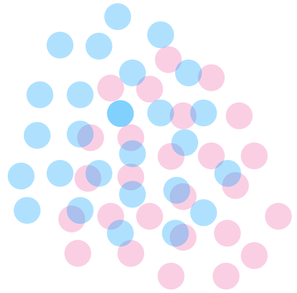
Depends on downstream scientific objective…

Physics Fidelity | Analysis Sensitivity | Unbiased Estimation | Etc.…

Our Testbed: Underconstrained ν reconstruction

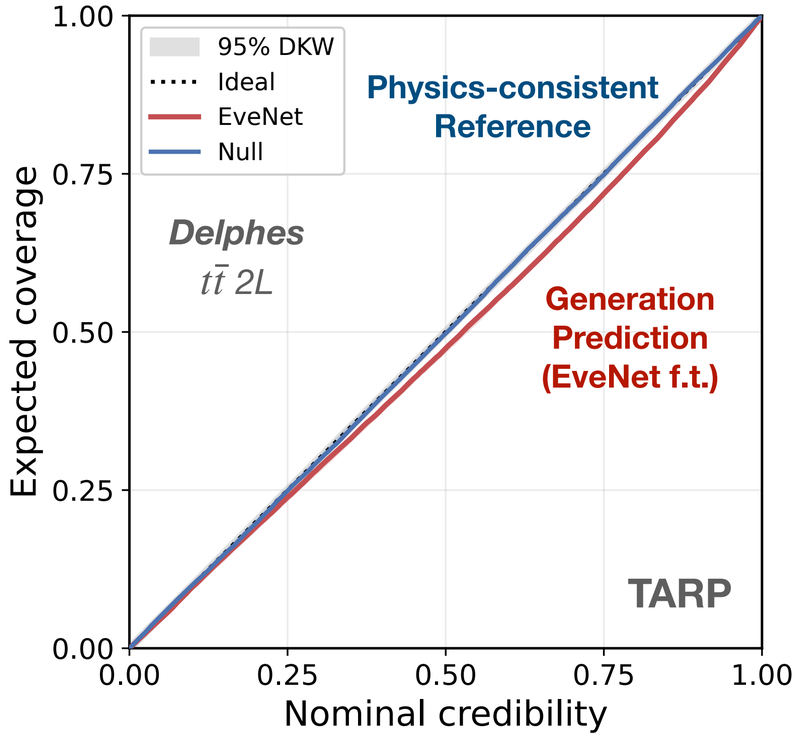
Does the reconstructed posterior reproduce the coverage expected from physics-consistent solutions.

MC: Physics consistent solutions



Generation Prediction

A plausible **generative posterior** is not necessarily physics-consistent. **Deep correlation is hard to learn.**

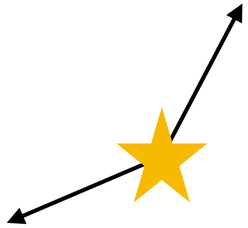


Preference Optimization for Generative Models

Observed Event

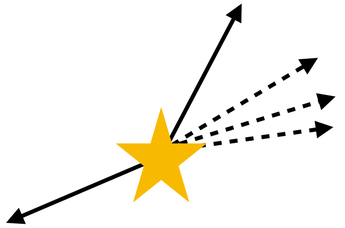
Pretrained Generative Model

Generated ν candidates

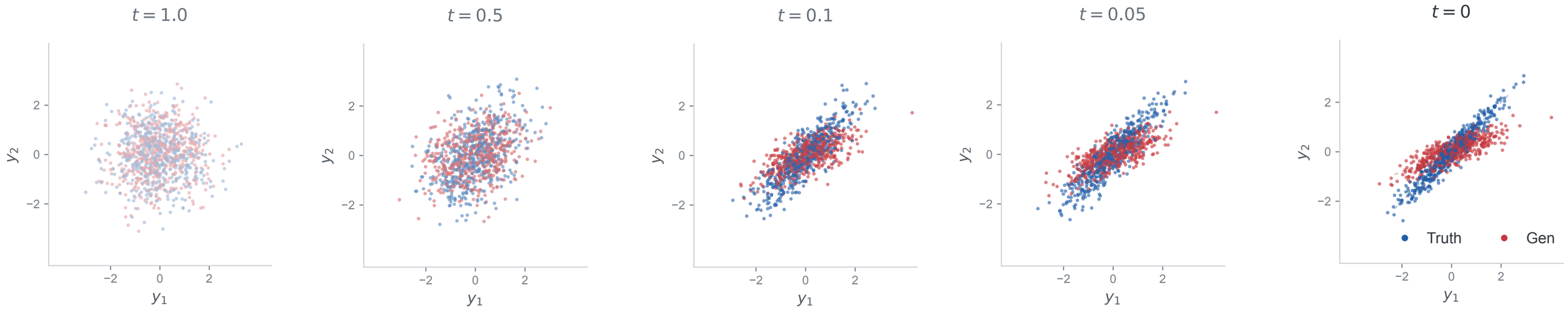


EveNet

Nature Machine Intelligence (accepted), 2026



Reverse diffusion / Sampling



Progressive Noise Injection

Preference Optimization for Generative Models

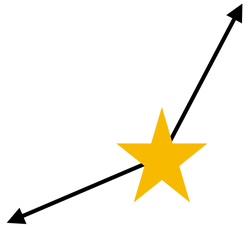
Observed Event

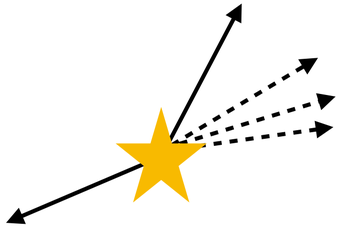
Pretrained Generative Model

Generated ν candidates

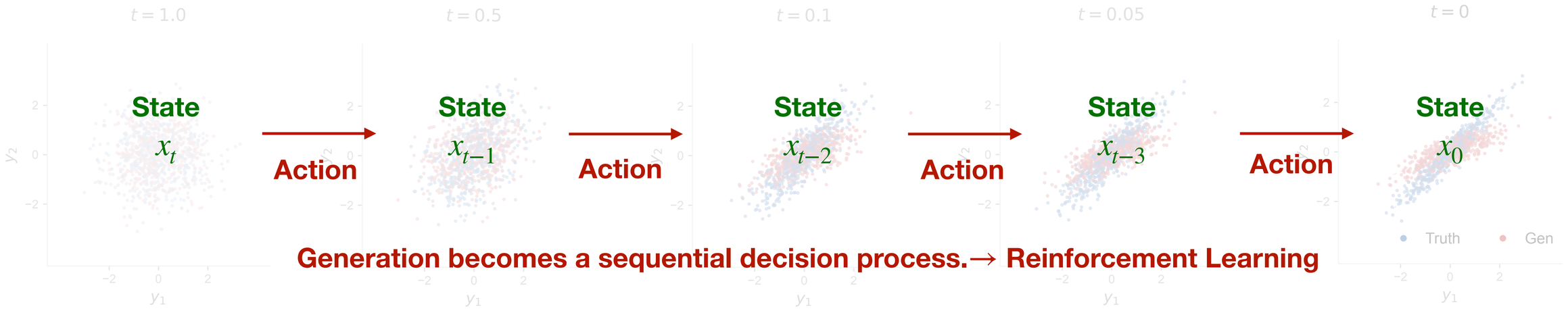
Policy

$$p_\theta\,(x_{t-1}\,|\,x_t\,,\,t,\,c)$$





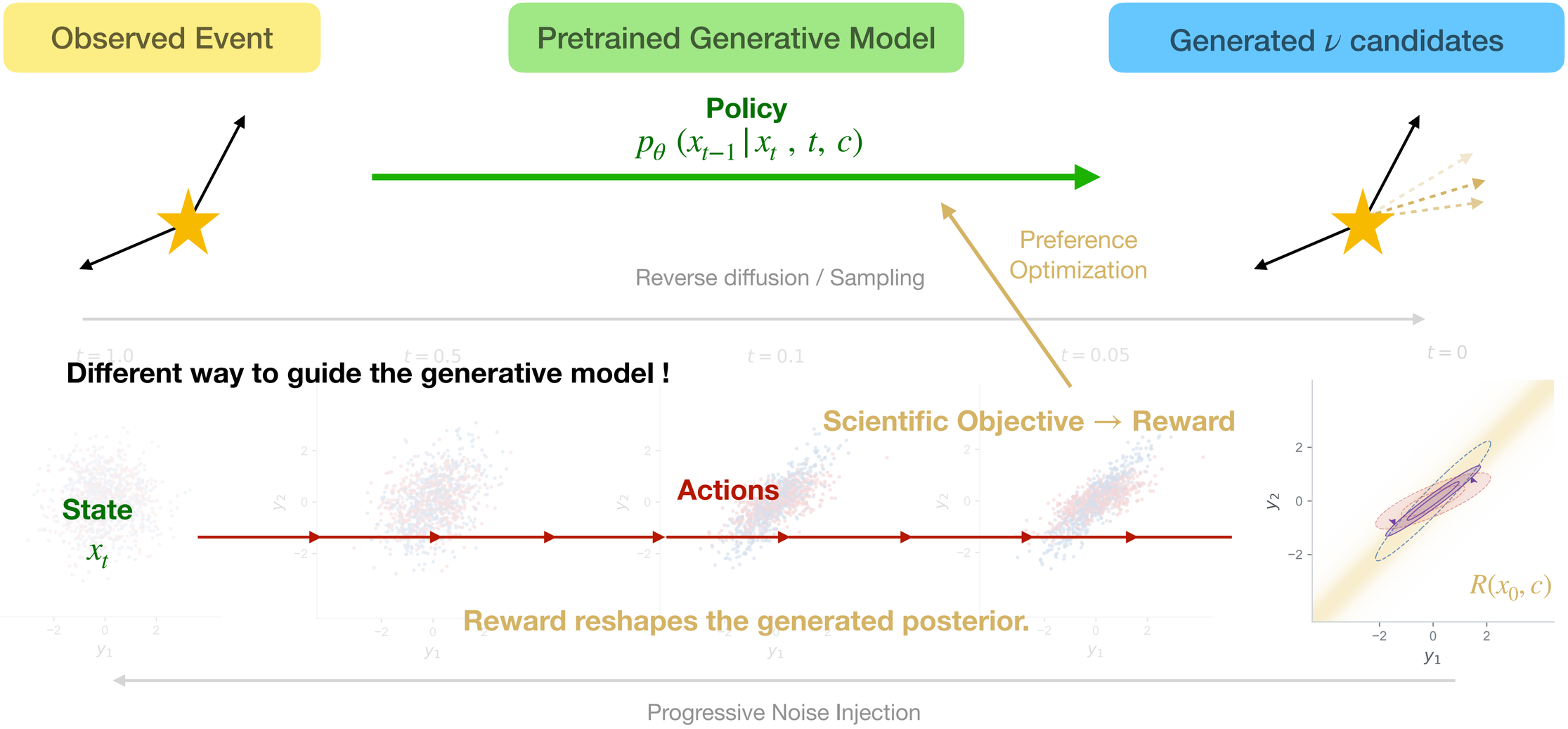
Reverse diffusion / Sampling



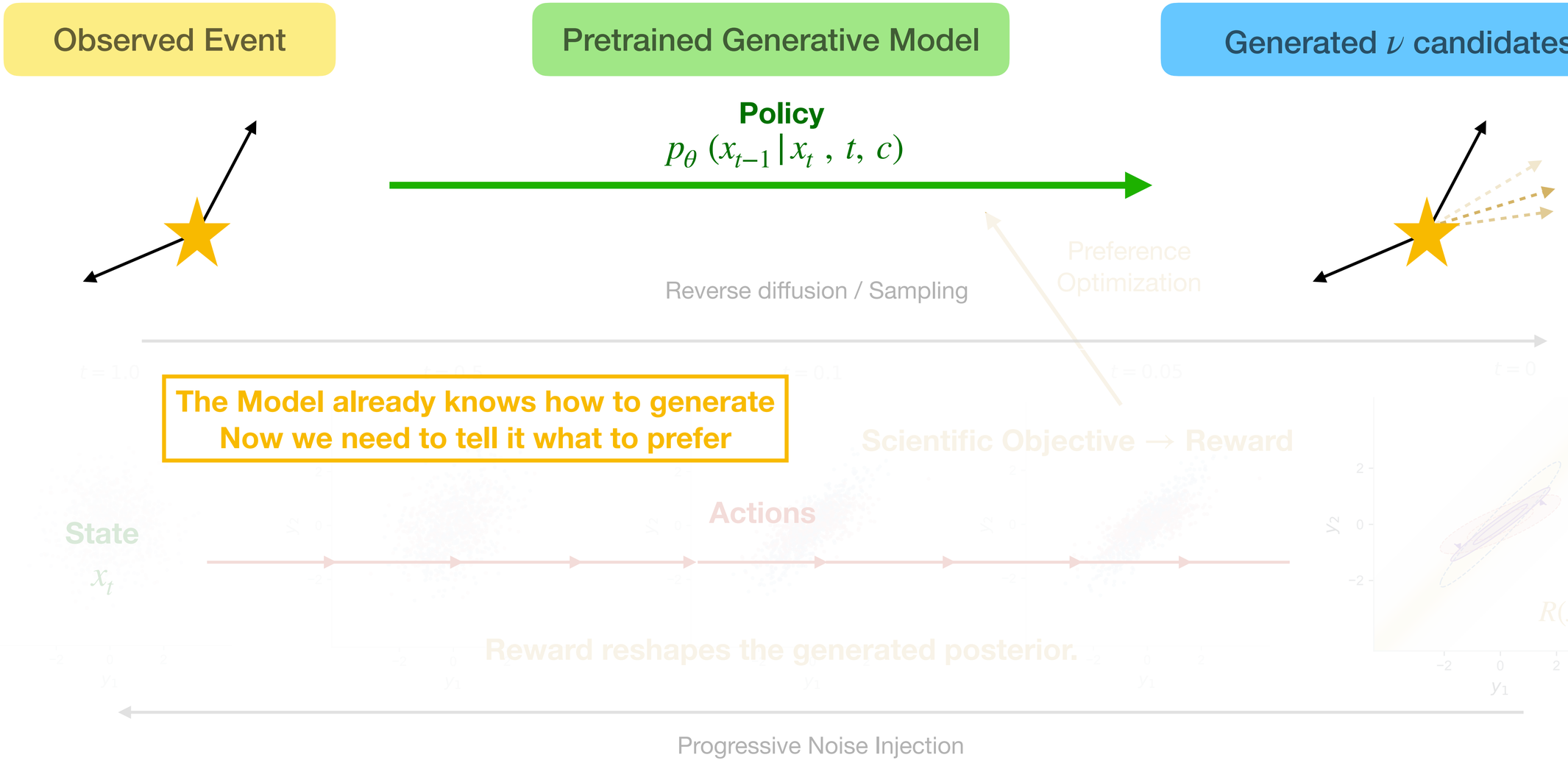
Generation becomes a sequential decision process.→ Reinforcement Learning

Progressive Noise Injection

Preference Optimization for Generative Models



Preference Optimization for Generative Models

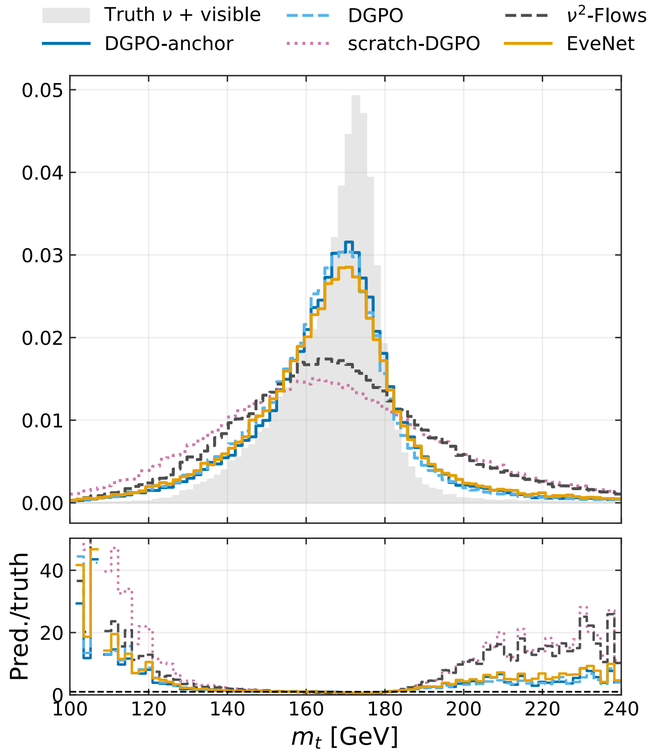


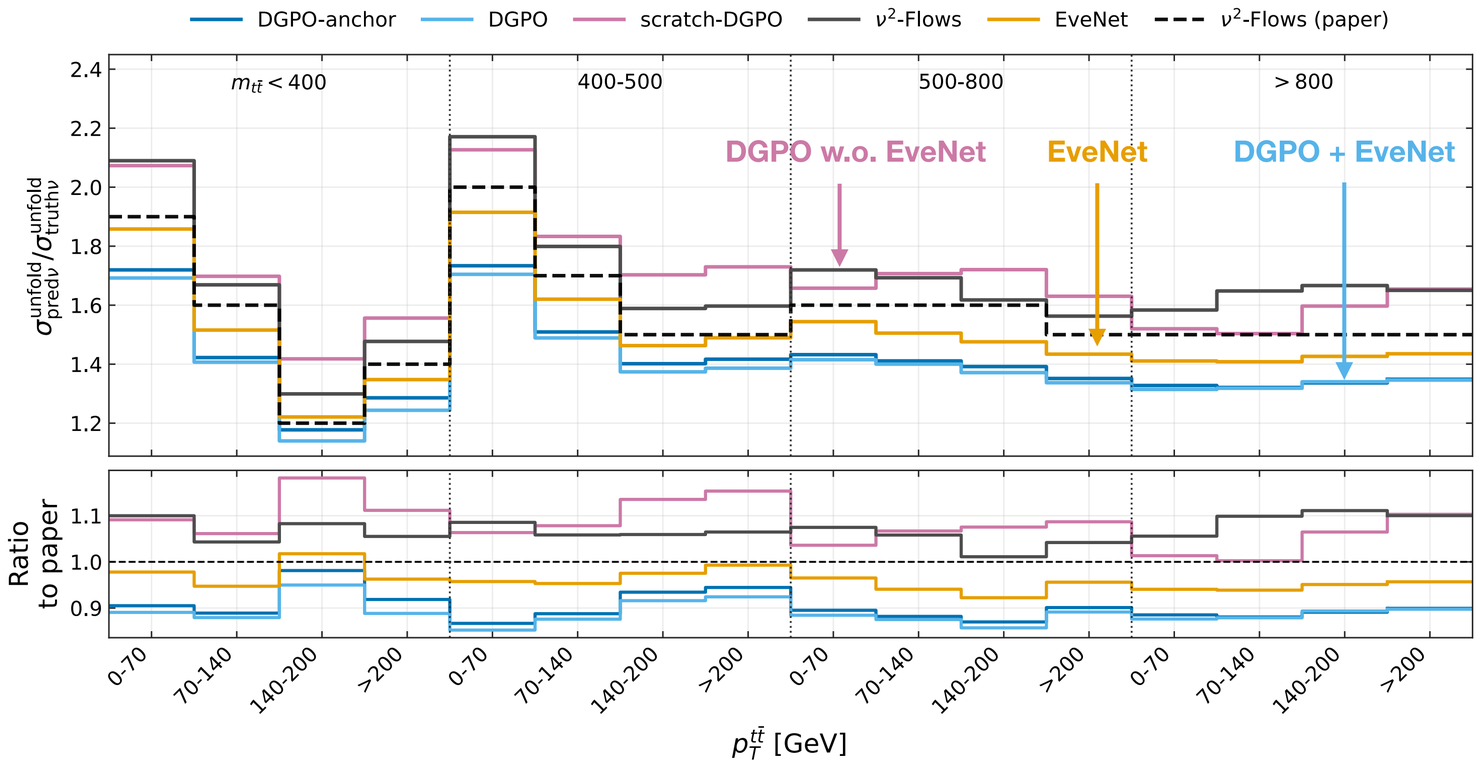
Lesson we learned: Truth-Guided Preferences

- Candidates closer to the **event-wise truth neutrino momenta** receive higher reward.

$$R(\hat{z}, z^*) = -\sum_{j,c} \left(\frac{\hat{p}_{c,j} - p^*_{c,j}}{s_c + \epsilon} \right)^2$$

- Preference optimization on top of EveNet improves reconstruction and downstream statistical precision.

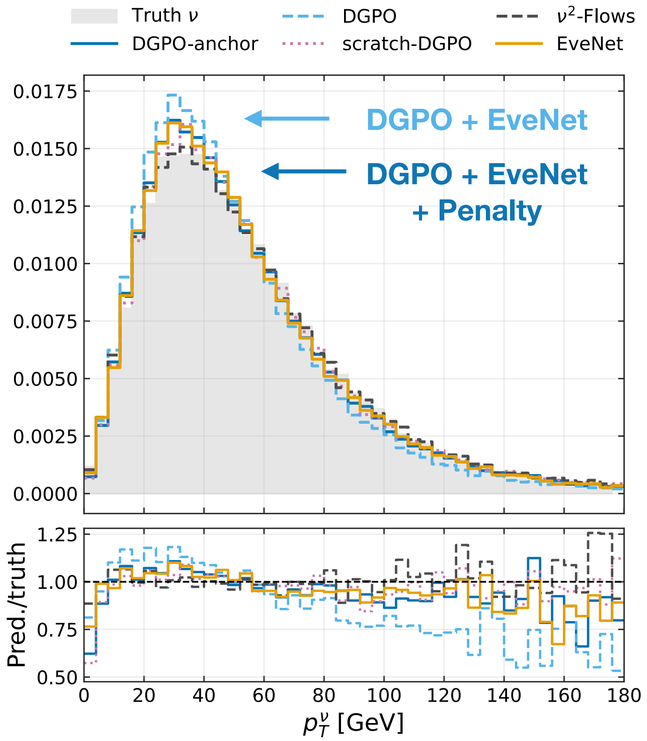


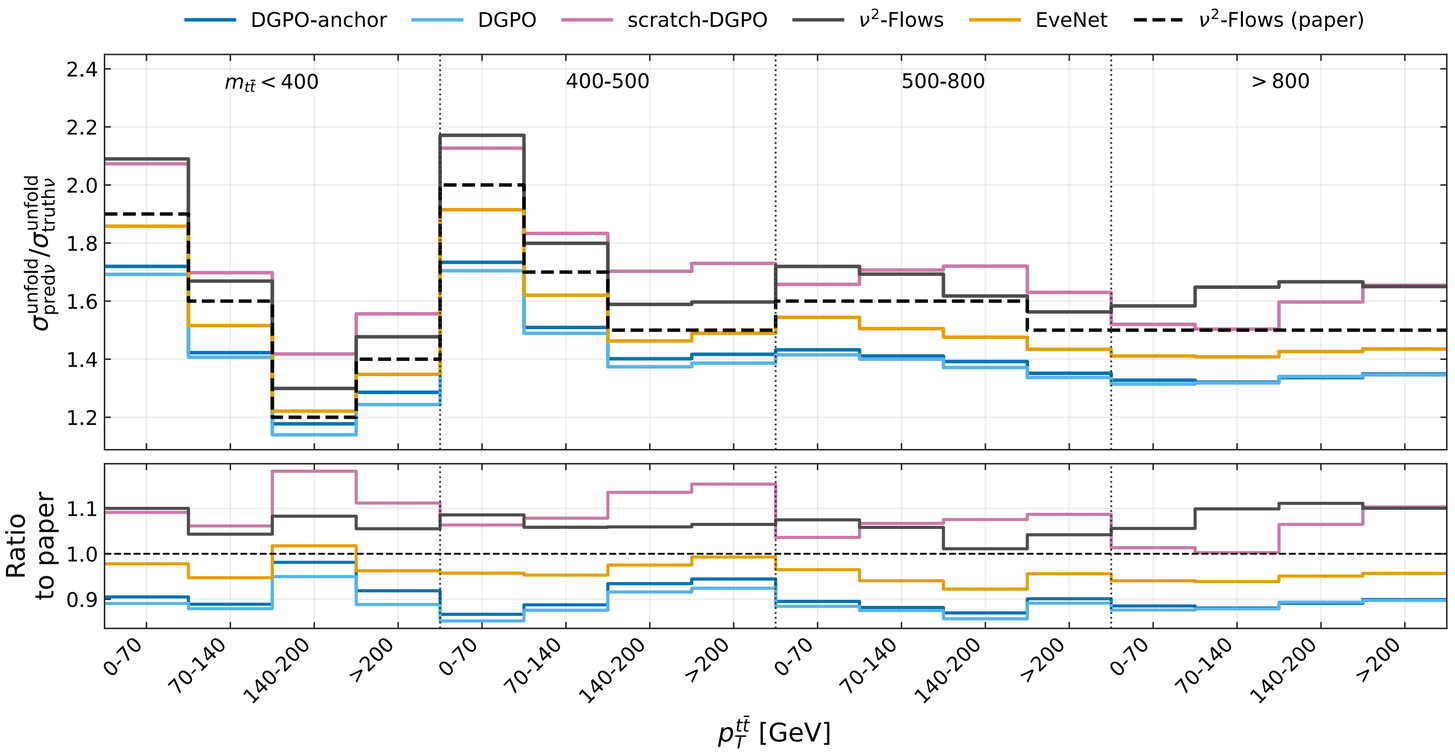


Lesson we learned: Truth-Guided Preferences

- Event-level reconstruction improves, but the generated population drifts.
- A distribution-level anchor suppresses this bias while preserving most of the reconstruction gain.

This reward may not reflect our final scientific objective.



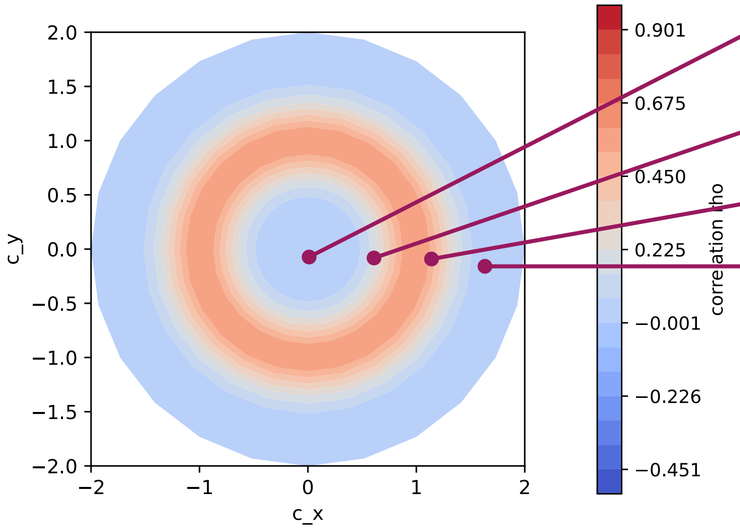


Lesson we learned: Truth Guided Preferences Failure Mode

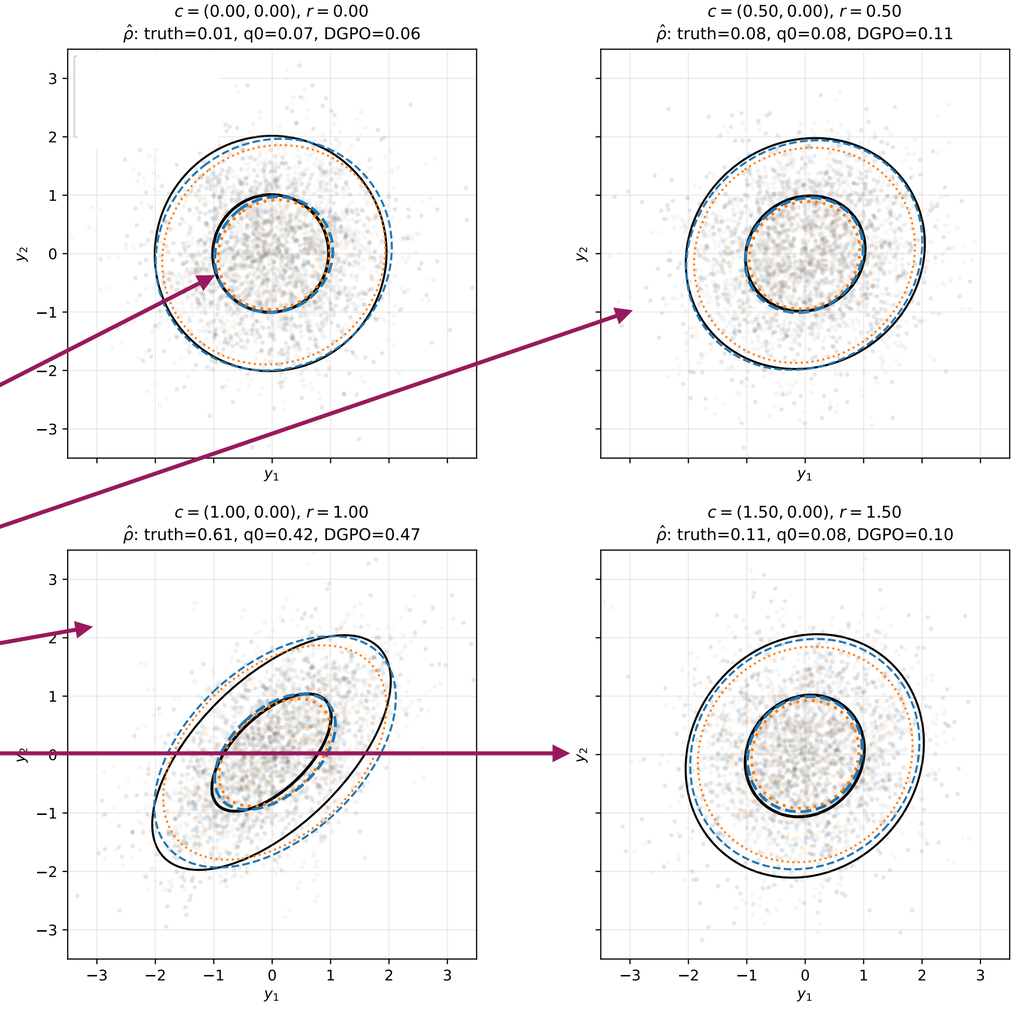
Toy Model: Conditional Distribution

$$c = (c_x, c_y) \rightarrow \rho(c) \rightarrow p(y_1, y_2 \,|\, c)$$

One condition does not correspond to a unique solution; it defines a conditional distribution.



Conditional correlation $\rho(c)$

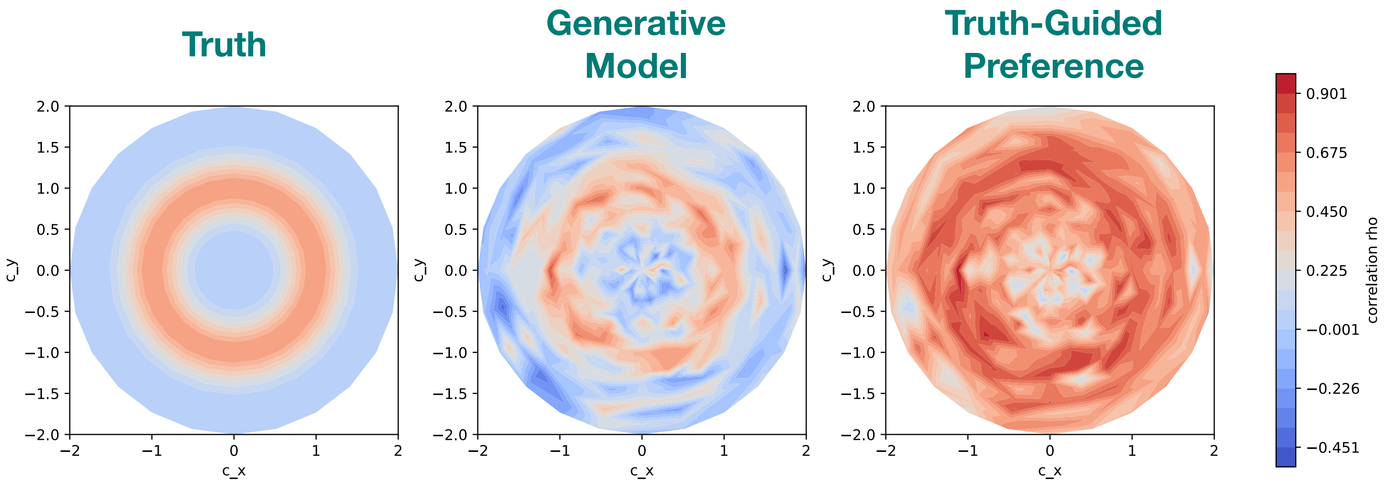


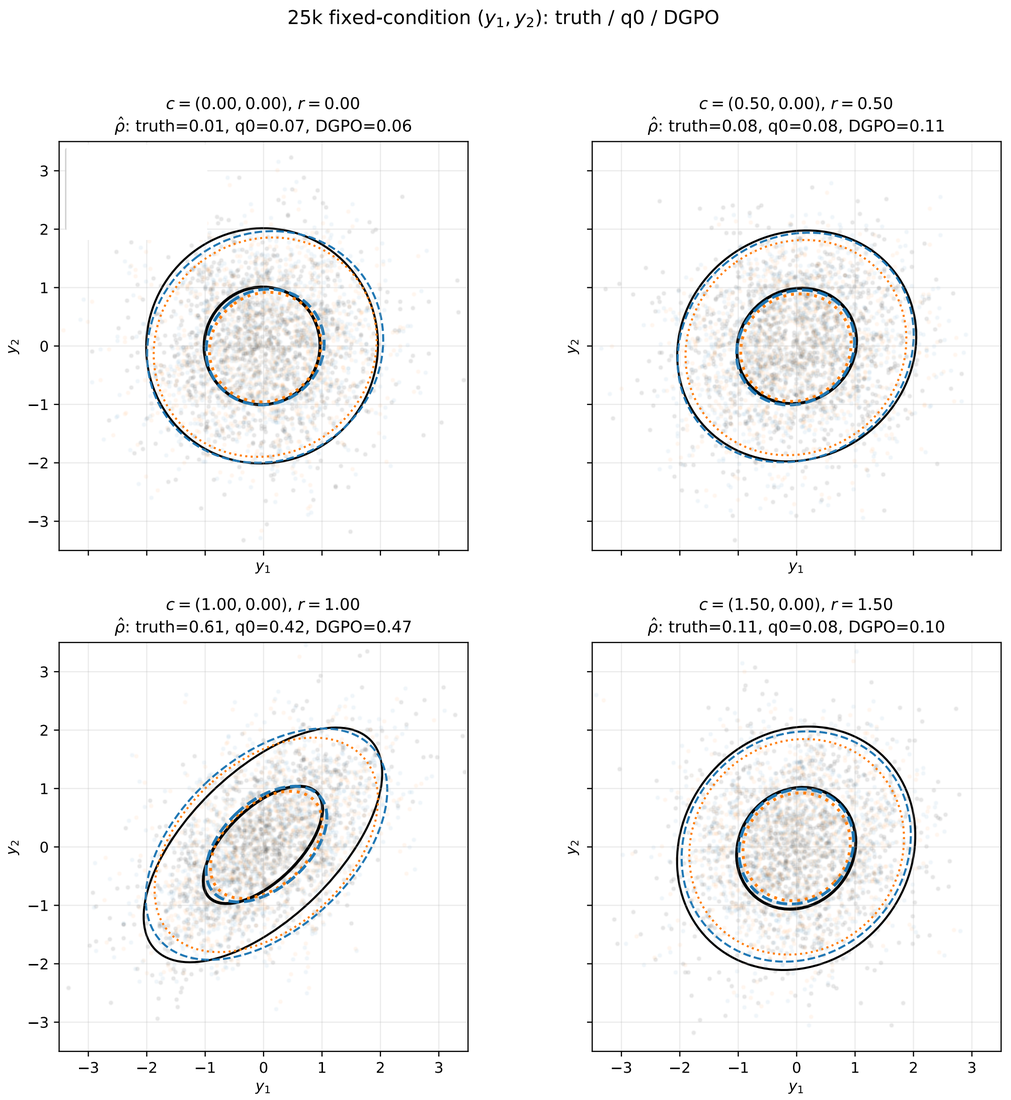
At each fixed c, there is an entire distribution of valid y_1, y_2

Lesson we learned: Truth Guided Preferences Failure Mode

Toy Model: Conditional Distribution

- Truth-guided preference creates a regression-like pull toward a **single realized truth**.
- But the scientific target is the full conditional distribution $p(y_1, y_2 \mid c)$
- Point-wise optimization can therefore distort the learned conditional structure





Population-Level Information as Reward

1. Reframe the target

$$p_{gen}(z\,|\,c) \rightarrow p_{target}(z\,|\,c)$$

The object we want to correct is the conditional distribution

2. Estimate distributional mismatch

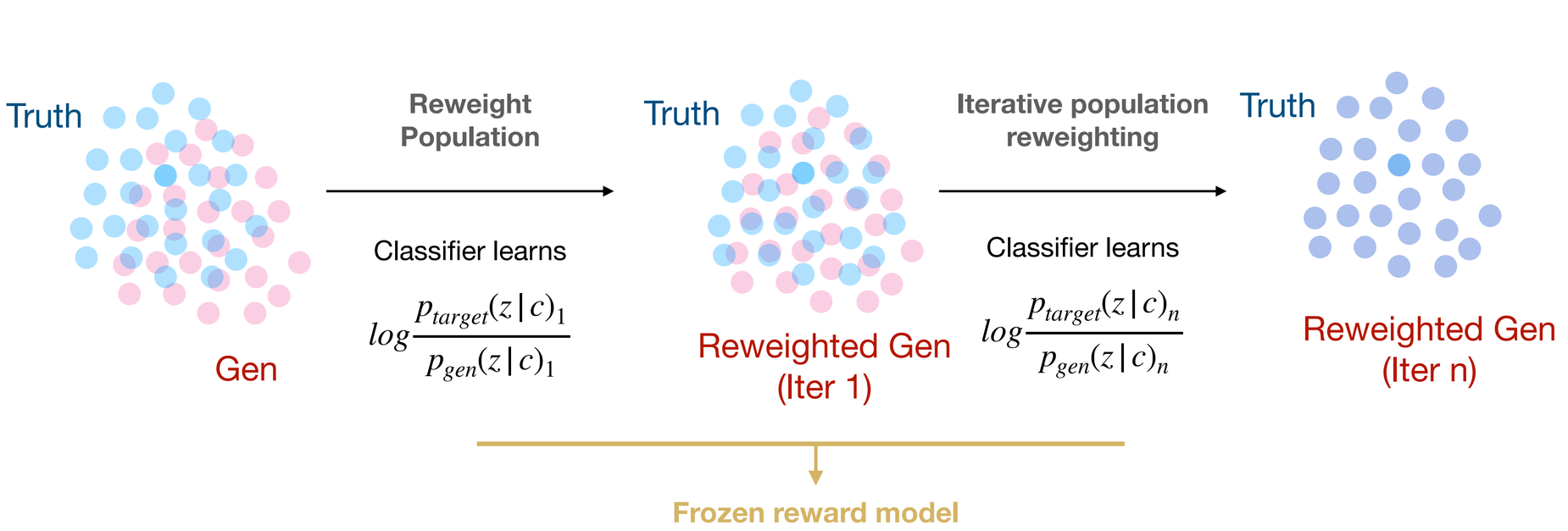
$$\text{classifiers} \rightarrow log\frac{p_{target}(z\,|\,c)}{p_{gen}(z\,|\,c)}$$

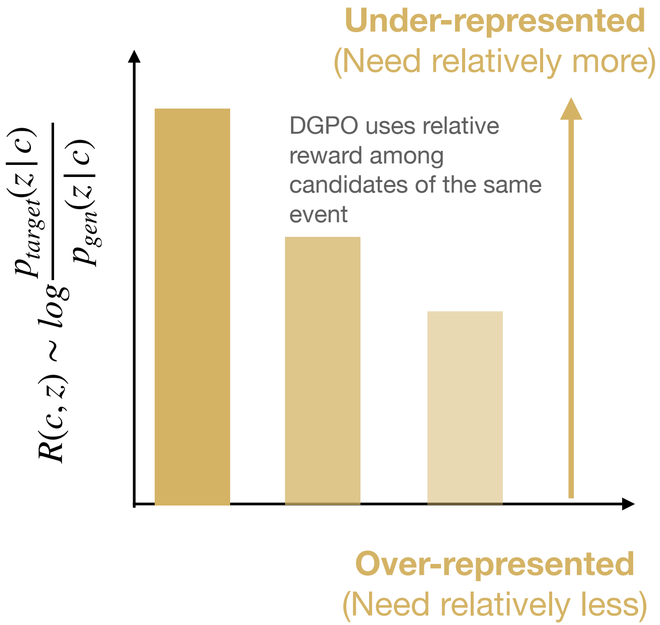
Classifiers learn the conditional density ratio.

3. Turn the likelihood ratio into preference

$$R(c,z) \sim log\frac{p_{target}(z\,|\,c)}{p_{gen}(z\,|\,c)}$$

Learn the preference from populations, apply it event by event.





Population-Level Information as Reward

1. Reframe the target

$$p_{gen}(z\,|\,c) \rightarrow p_{target}(z\,|\,c)$$

The object we want to correct is the conditional distribution

2. Estimate distributional mismatch

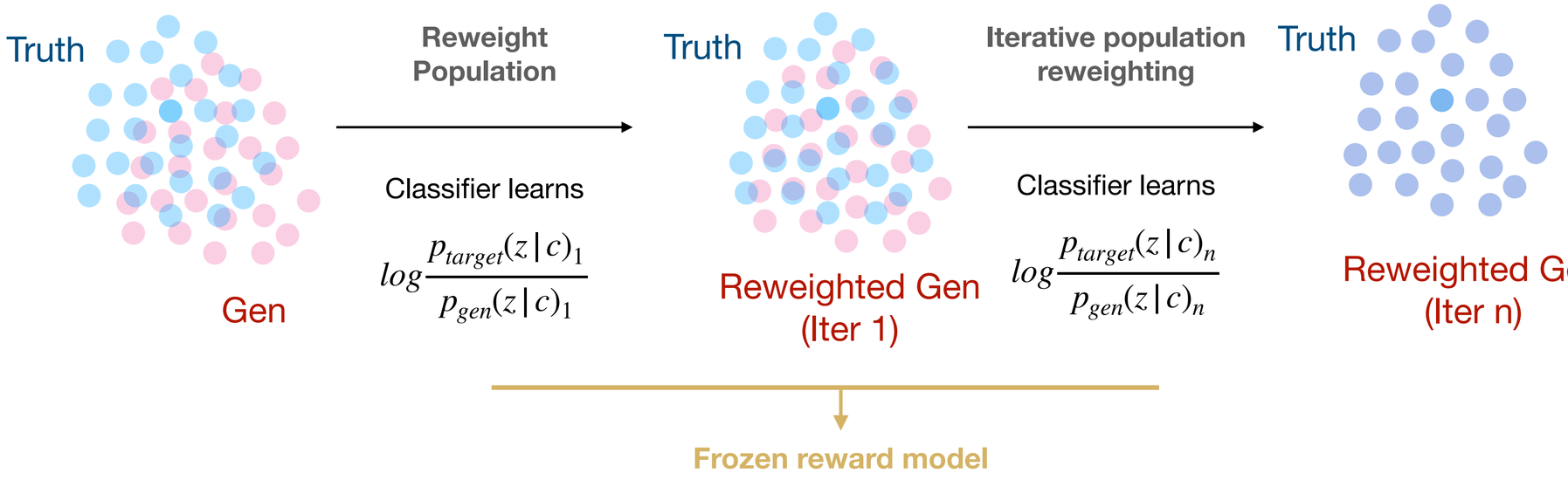
$$\text{classifiers} \rightarrow log\frac{p_{target}(z\,|\,c)}{p_{gen}(z\,|\,c)}$$

Classifiers learn the conditional density ratio.

3. Turn the likelihood ratio into preference

$$R(c,z) \sim log\frac{p_{target}(z\,|\,c)}{p_{gen}(z\,|\,c)}$$

Learn the preference from populations, apply it event by event.



Beyond Hand-Designed Physics-Consistent Rewards

- The likelihood-ratio classifier can exploit **multivariate physics correlations**.
- The reward can therefore capture **physics-relevant structure**

Population-Level Information as Reward

1. Reframe the target

$$p_{gen}(z\,|\,c) \rightarrow p_{target}(z\,|\,c)$$

The object we want to correct is the conditional distribution

2. Estimate distributional mismatch

$$\text{classifiers} \rightarrow log\frac{p_{target}(z\,|\,c)}{p_{gen}(z\,|\,c)}$$

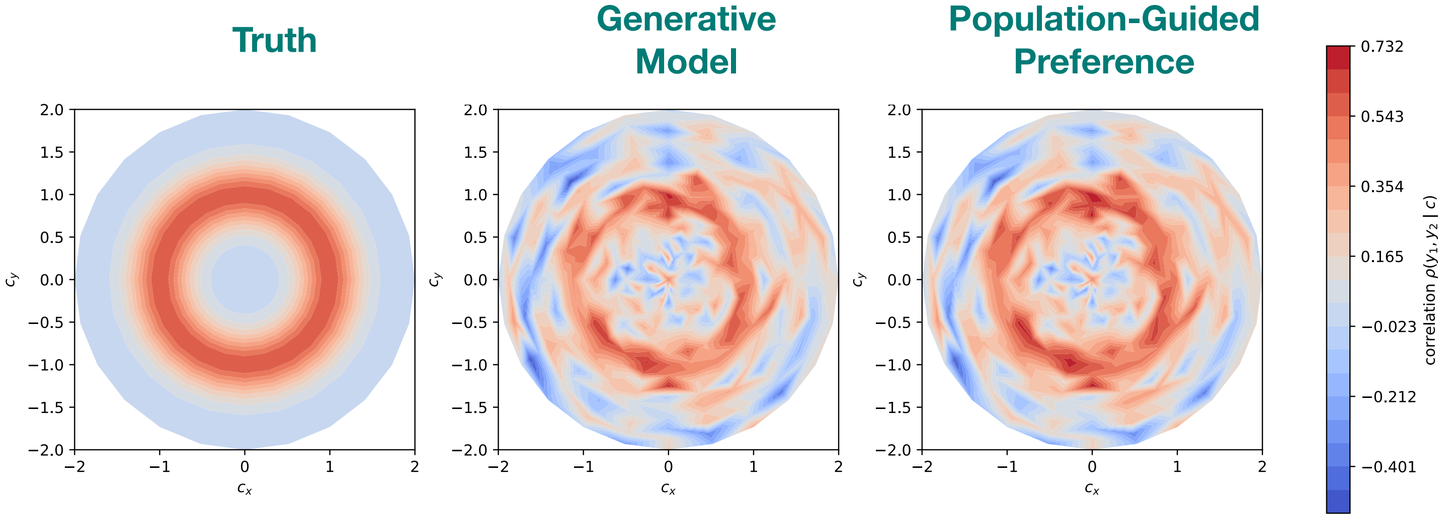
Classifiers learn the conditional density ratio.

3. Turn the likelihood ratio into preference

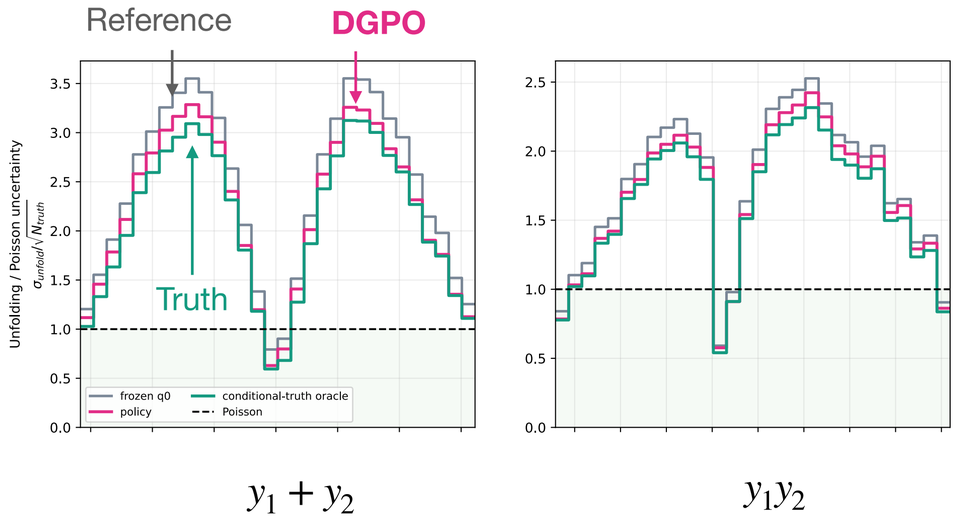
$$R(c,z) \sim log\frac{p_{target}(z\,|\,c)}{p_{gen}(z\,|\,c)}$$

Learn the preference from populations, apply it event by event.

Population-guided preference restores the underlying conditional structure.

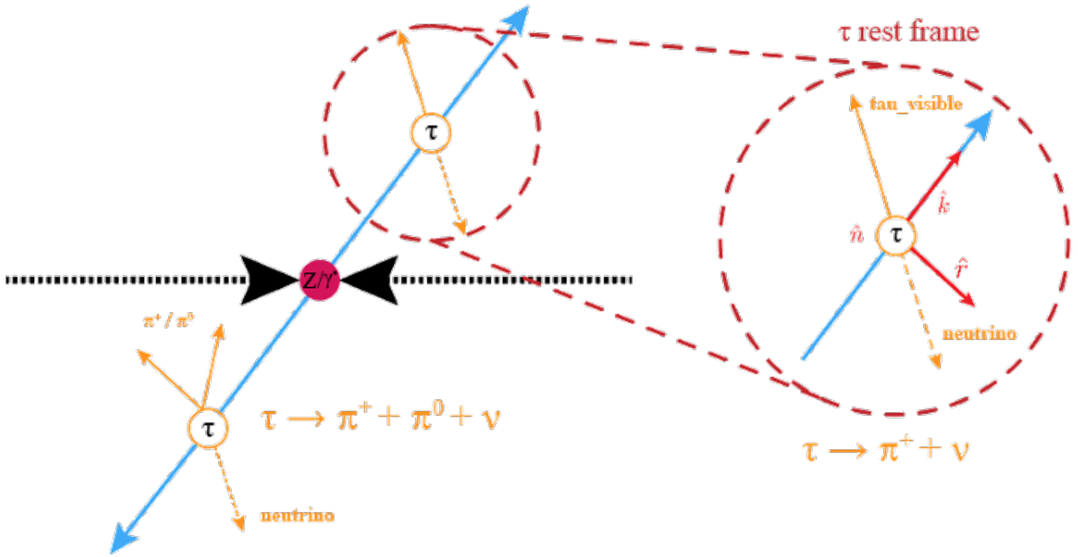


Correlation-sensitive observables unfold more precisely

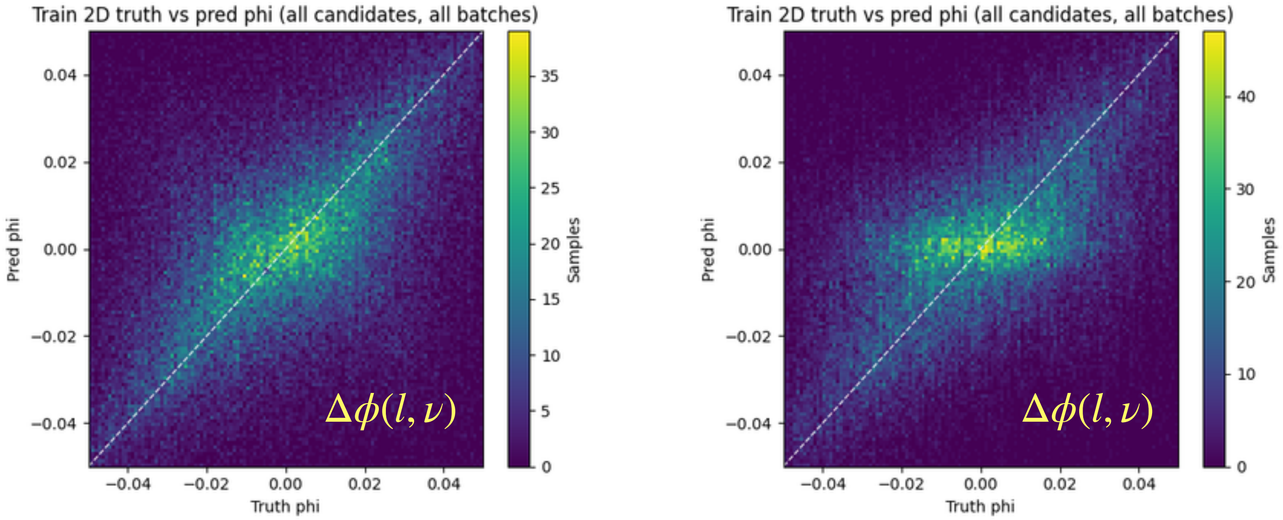


Pressure Test: $Z \to \tau\tau$ Quantum Entanglement Measurement

- Neutrinos carry a substantial fraction of the τ momentum, making the reconstructed $\tau\tau$ system highly sensitive to neutrino reconstruction.
- The EveNet + truth-guided DGPO + anchor strategy does not transfer well to this $Z \to \tau\tau$ topology.
- This provides a stringent pressure test for population-guided post-training.



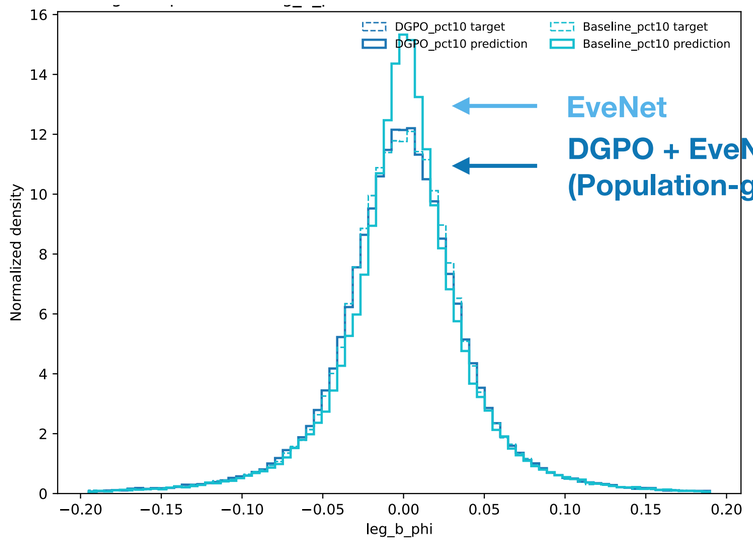
EveNet → **DGPO + EveNet (Truth-Guided)**



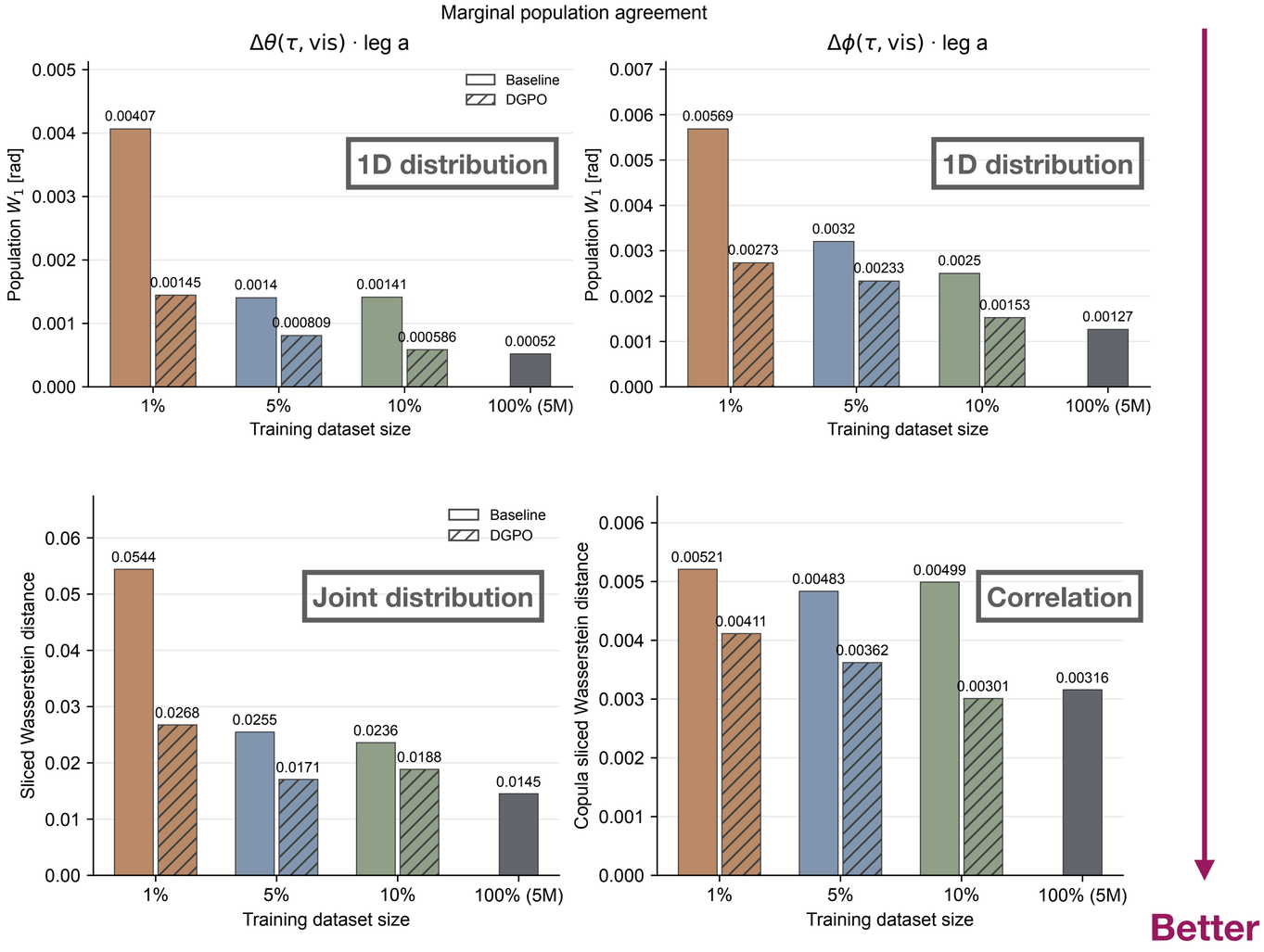
*For the full $Z \to \tau\tau$ quantum-entanglement analysis with DELPHI data, see **Chen-Hua's talk.***

$Z \to \tau\tau$ QE measurement Results

- Population-guided preference optimization reduces the joint-distribution mismatch by **20–50%** across the limited-data regimes.
- The improvement persists after removing the 1D marginals, indicating better multivariate dependence rather than marginal reshaping alone.



Trained with 10% dataset
(Full dataset: 5M statistics)

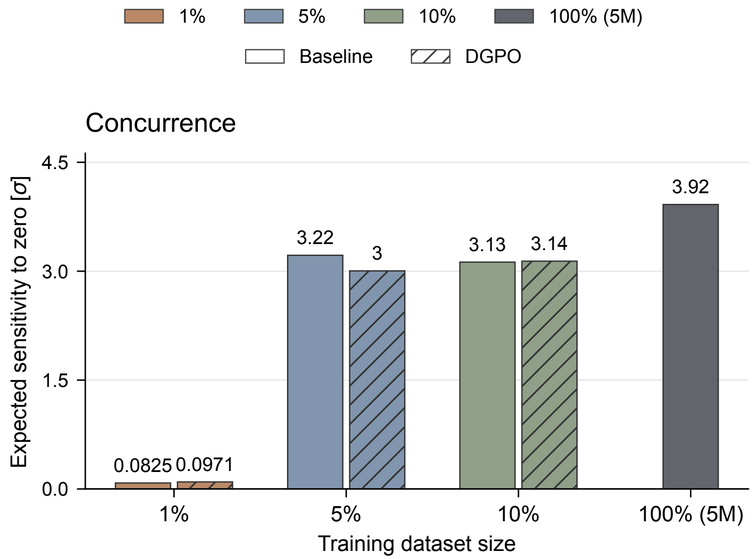


Conclusions

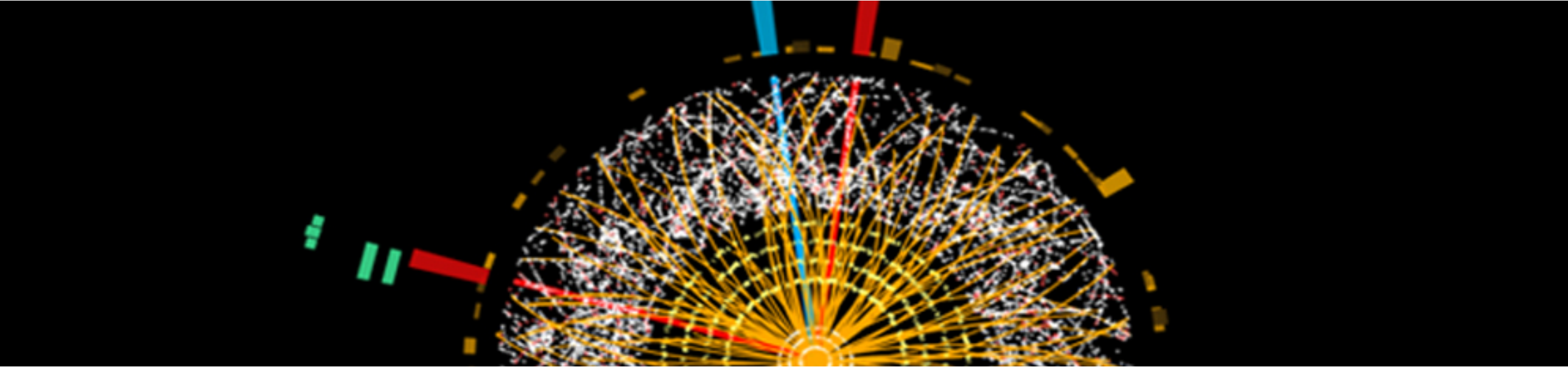
- Large-scale **pretraining** provides a broad generative solution space, but scientific post-training requires a well-defined notion of **“better.”**
- **Truth-guided preferences** improve event-level reconstruction, but can distort the **conditional distribution** in underconstrained inference.
- **Population-level likelihood information** provides a more suitable feedback signal, translating distributional mismatch into **event-wise preferences**.
- Population-level optimization improves distribution fidelity, but does not consistently improve downstream physics sensitivity, since the current reward is not directly aligned with the final measurement.

Could we align post-training directly with inference-level objectives ?

Event level → population level → inference level

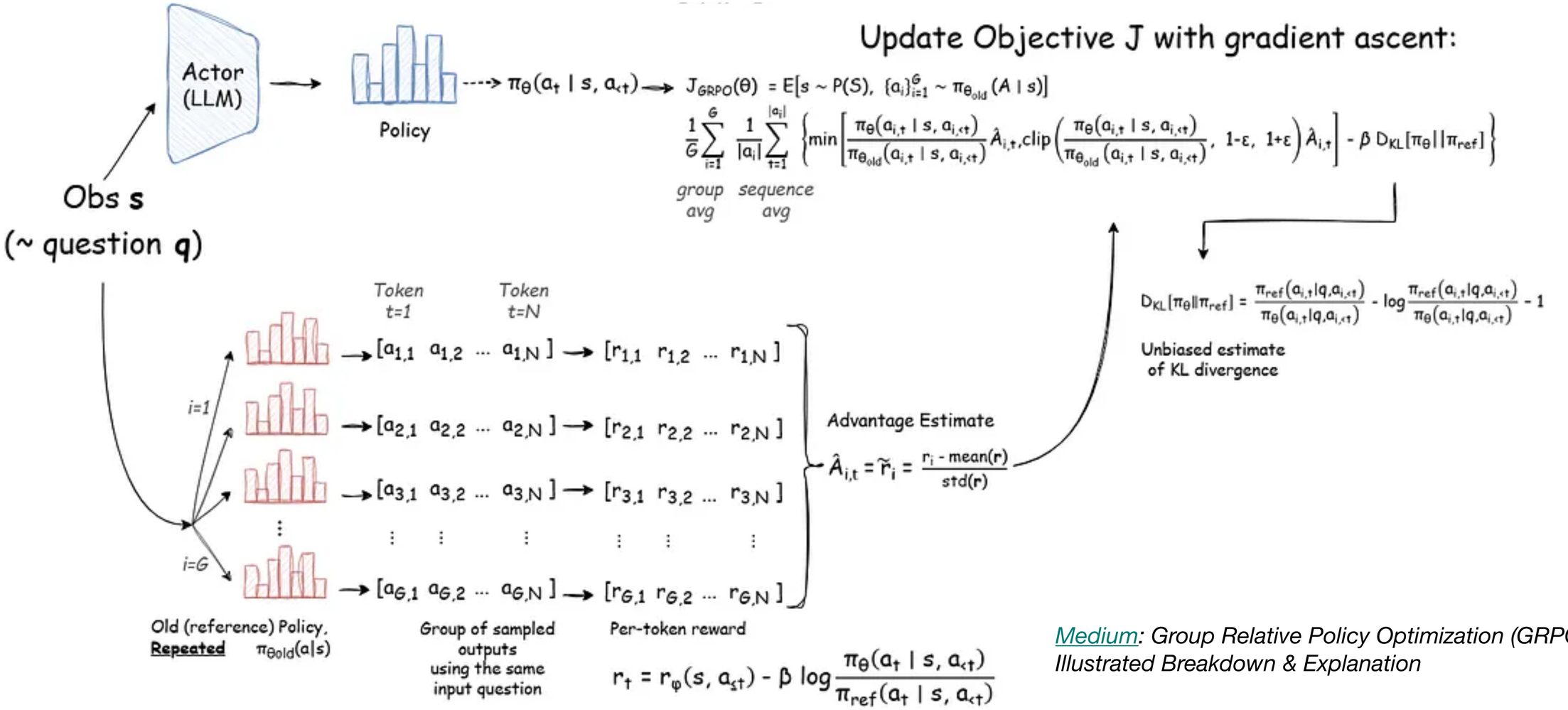


The choice of reward is itself part of the scientific inference problem.



BackUp

Conceptual Origin: Group-Relative Policy Optimization



$$J_{GRPO}(\theta) = \mathbb{E}[s \sim P(S),\ \{a_i\}_{i=1}^{G} \sim \pi_{\theta_{old}}(A \mid s)]$$

$$\frac{1}{G}\sum_{i=1}^{G}\frac{1}{|a_i|}\sum_{t=1}^{|a_i|}\left\{\min\left[\frac{\pi_\theta(a_{i,t} \mid s, a_{i,<t})}{\pi_{\theta_{old}}(a_{i,t} \mid s, a_{i,<t})}\hat{A}_{i,t}, \text{clip}\left(\frac{\pi_\theta(a_{i,t} \mid s, a_{i,<t})}{\pi_{\theta_{old}}(a_{i,t} \mid s, a_{i,<t})}, 1-\varepsilon, 1+\varepsilon\right)\hat{A}_{i,t}\right] - \beta\, D_{KL}[\pi_\theta || \pi_{ref}]\right\}$$

$$D_{KL}[\pi_\theta \| \pi_{ref}] = \frac{\pi_{ref}(a_{i,t}|q,a_{i,<t})}{\pi_\theta(a_{i,t}|q,a_{i,<t})} - \log\frac{\pi_{ref}(a_{i,t}|q,a_{i,<t})}{\pi_\theta(a_{i,t}|q,a_{i,<t})} - 1$$

$$\hat{A}_{i,t} = \tilde{r}_i = \frac{r_i - \text{mean}(r)}{\text{std}(r)}$$

$$r_t = r_\varphi(s, a_{\le t}) - \beta \log\frac{\pi_\theta(a_t \mid s, a_{<t})}{\pi_{ref}(a_t \mid s, a_{<t})}$$

Medium: Group Relative Policy Optimization (GRPO) Illustrated Breakdown & Explanation

DGPO as Policy Optimization for Diffusion Models

Policy Gradient

$$J(\theta) = E_{x\sim\pi_\theta}[R(x)]$$

$$\nabla\theta J(\theta) = E[A(x)\nabla\theta \log \pi_\theta(x)]$$

Positive advantage increases the sample probability, negative advantage suppresses it.

$$A > 0 \Rightarrow \log \pi_\theta(x) \uparrow$$

$$A < 0 \Rightarrow \log \pi_\theta(x) \downarrow$$

Diffusion Bridge

$$-\log p_\theta \propto l_\theta = \|v_\theta(x_t, c, t) - v^*\|^2$$

For a Gaussian reverse diffusion step, the denoising residual acts as a local negative-log-probability surrogate

$$\nabla_\theta \log p_\theta \propto -\nabla_\theta l_\theta$$

DGPO Update

$$\nabla J \propto E[-A\nabla l_\theta]$$

$$L_{policy} = E[A l_\theta]$$

DGPO implements policy-gradient learning by using the diffusion residual as a surrogate for sample log-probability.

$$A_k > 0 \Rightarrow l_k \downarrow \Rightarrow p_\theta(x_k) \uparrow$$

$$A_k < 0 \Rightarrow l_k \uparrow \Rightarrow p_\theta(x_k) \downarrow$$

From Group Preference to the Full DGPO Objective

1. Sample a group

For the same event

$$x_0^{(1,e)}, \ldots, x_0^{(k,e)} \sim \pi_\theta(\cdot \mid c_e)$$

And use the same (t, ϵ)

$$x_t^{(k,e)} = \alpha_t x_0^{(k,e)} + \sigma_t \epsilon$$

2. Relative preference → Advantage

$$R_{k,e} \to A_{k,e}$$

$A_{k,e}$: relative preference within the same event

$A_{k,e} > 0$: prefered

$A_{k,e} < 0$: suppressed

3. Compare current and reference policies

$$l_\theta^{(k,e)} = \|v_\theta - v^*\|^2$$
$$l_{ref}^{(k,e)} = \|v_{ref} - v^*\|^2$$
$$\Delta_{k,e} \equiv l_\theta^{(k,e)} - l_{ref}^{(k,e)} \sim \log \frac{\pi_\theta(x_k \mid c_e)}{\pi_{ref}(x_k \mid c_e)}$$

4. Event-level DGPO Gate

$$M_e = \frac{\beta}{K} \sum_k A_{k,e} \Delta_{k,e}$$
$$w_e = \sigma(M_e)$$

Large w_e: preference ordering is not yet satisfied

Small w_e: current policy already favors the prefered candidates.

5. Final Update

$$L_{DGPO} = \frac{1}{BK} \sum_e w_e \sum_k A_{k,e} l_\theta^{(k,e)}$$
$$L_{total} = L_{DGPO} + \lambda_{KL} L_{KL}$$
$$L_{KL} = \frac{1}{BK} \sum_{e,k} \|v_\theta - v_{ref}\|^2$$