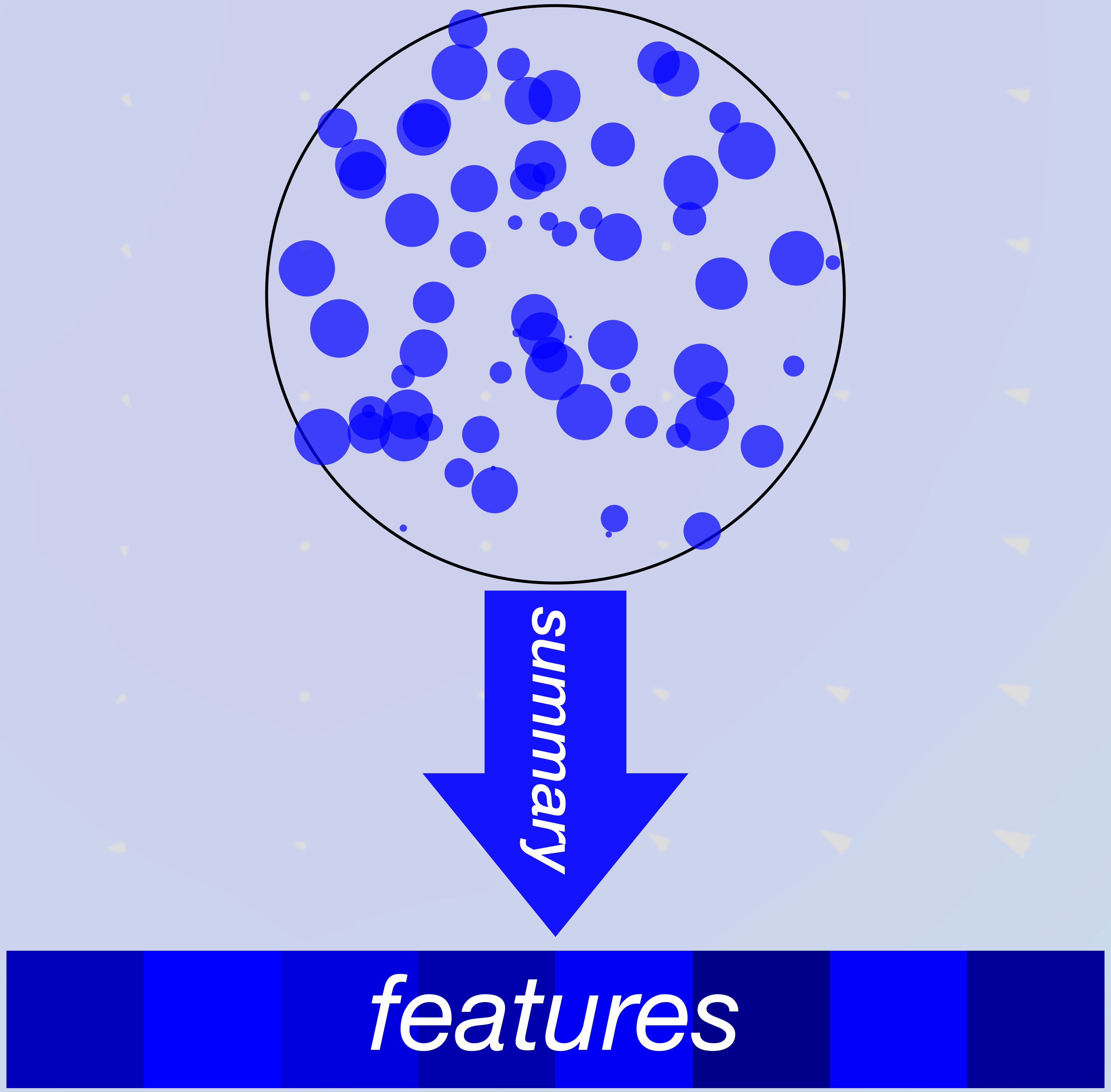
a continuous calibration of the ATLAS flavor-tagging classifiers via optimal transportation maps

Chris Pollard, Warwick
for the ATLAS collaboration

incredible progress in recent years…

incredible progress in recent years…

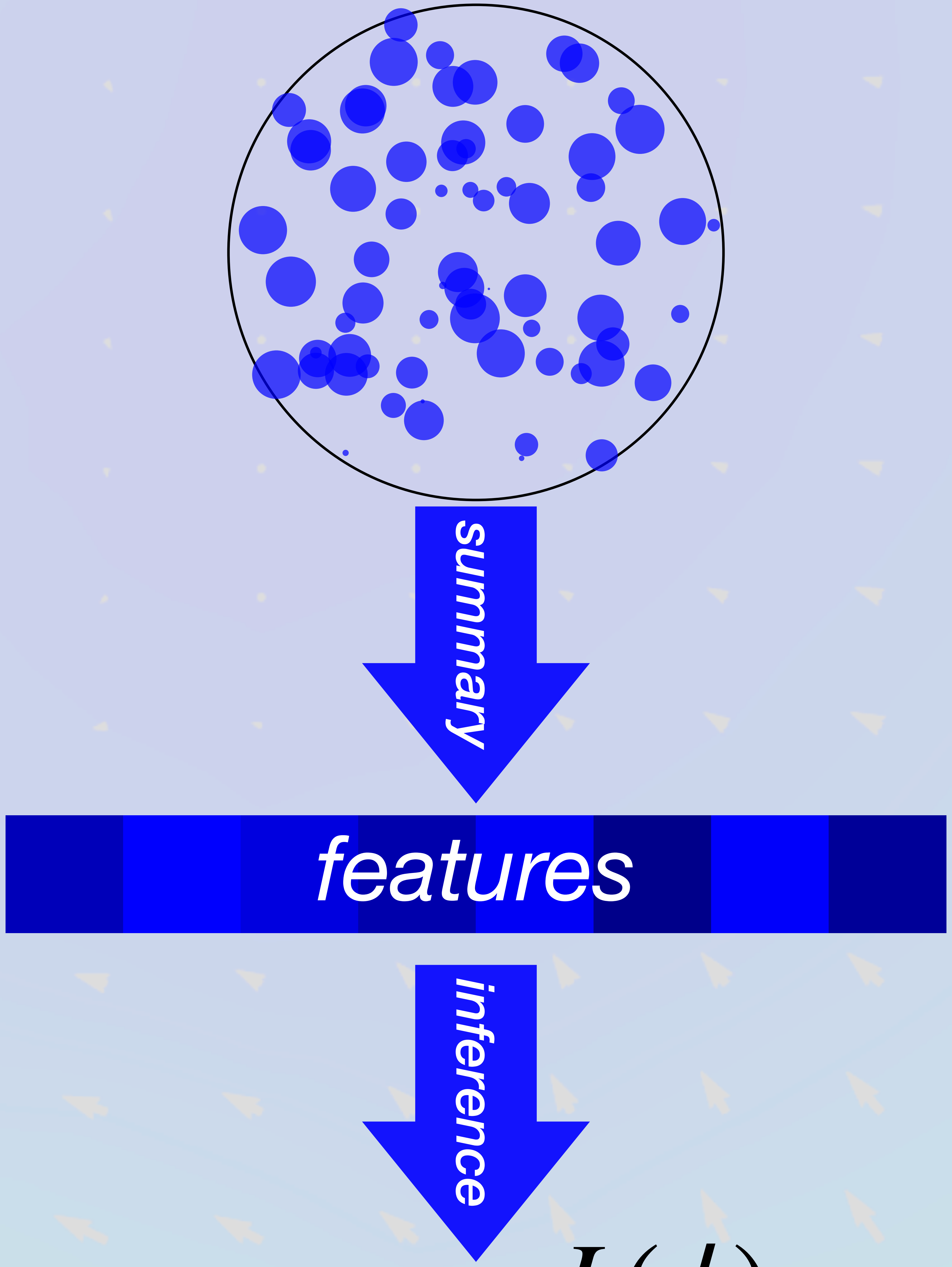
- *more informative* observables for both *specific* and *general* inference tasks



incredible progress in recent years…

- *more informative* observables for both *specific* and *general* inference tasks

- huge growth in constraining power of observed data



$$p(\phi) \text{ or } \frac{L(\phi)}{L(\hat{\phi})}$$

why do we need summaries / embeddings / “physics objects”?

why do we need
summaries / embeddings / “physics objects”?

tractability: ~millions of detector channels to read out per LHC bunch crossing.

why do we need
summaries / embeddings / “physics objects”?

tractability: ~millions of detector channels to read out per LHC bunch crossing.

correctness: we don’t have a simulator that *adequately describes all details* of the data!

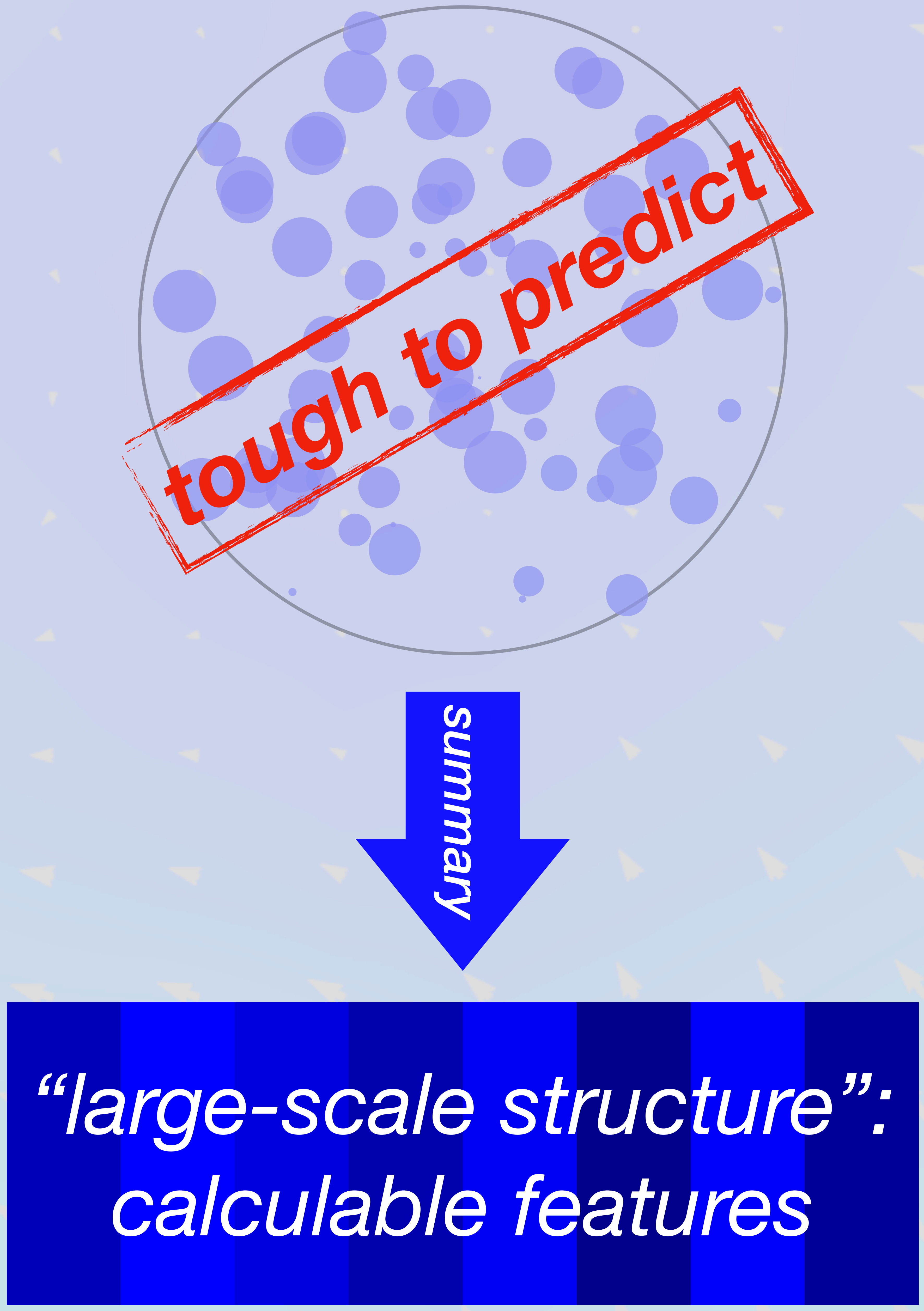
indeed, this is perhaps the main “point” of constructing jets:

indeed, this is perhaps the main "point" of constructing jets:

- we cannot correctly predict the details of QCD with arbitrary accuracy;

indeed, this is perhaps the main "point" of constructing jets:

- we cannot correctly predict the details of QCD with arbitrary accuracy;
- we *can* predict the "large-scale structure" of the fragmentation of partons → jets.



even so, ***useful details*** enable better inference.

- we can often *measure* the density over some informative features better than we can predict it.

even so, ***useful details*** enable better inference.

- we can often *measure* the density over some informative features better than we can predict it.
- a *calibration region* may be used to measure said density and the simulation *corrected* based on this measurement.

even so, *useful details* enable better inference.

- we can often *measure* the density over some informative features better than we can predict it.
- a *calibration region* may be used to measure said density and the simulation *corrected* based on this measurement.
- this is inherently a *mis-specification* problem, and one we're trying to learn how to solve generally!

we have a recipe for reweighting a density $p(\vec{x})$ (e.g. sim) to $q(\vec{x})$ (e.g. data):

- $w(x) \approx \exp D_p^q(\vec{x})$ for a data vs. simulation discriminator, $D_p^q(\vec{x})$.

we have a recipe for reweighting a density $p(\vec{x})$ (e.g. sim) to $q(\vec{x})$ (e.g. data):

- $w(x) \approx \exp D_p^q(\vec{x})$ for a data vs. simulation discriminator, $D_p^q(\vec{x})$.

it is often preferable to *move* events:

$$T : \vec{x} \to \vec{x} \implies T_\# p \approx q$$

we have a recipe for reweighting a density $p(\vec{x})$ (e.g. sim) to $q(\vec{x})$ (e.g. data):

- $w(x) \approx \exp D_p^q(\vec{x})$ for a data vs. simulation discriminator, $D_p^q(\vec{x})$.

it is often preferable to *move* events:

$$T : \vec{x} \to \vec{x} \implies T_{\#}p \approx q$$

- for "reasonably behaved" $\vec{x}$, p, and q, at least one T exists.
- usually want to change the simulation *as little as possible*:
  - i.e. find the $\hat{T}$ that minimally (or "optimally") morphs p into q;
  - this is an *optimal transport* (OT) problem.

we have been implicitly using OT for decades in HEP:

- correct a gaussian density $\leftrightarrow$ alter μ, σ of simulated distribution.

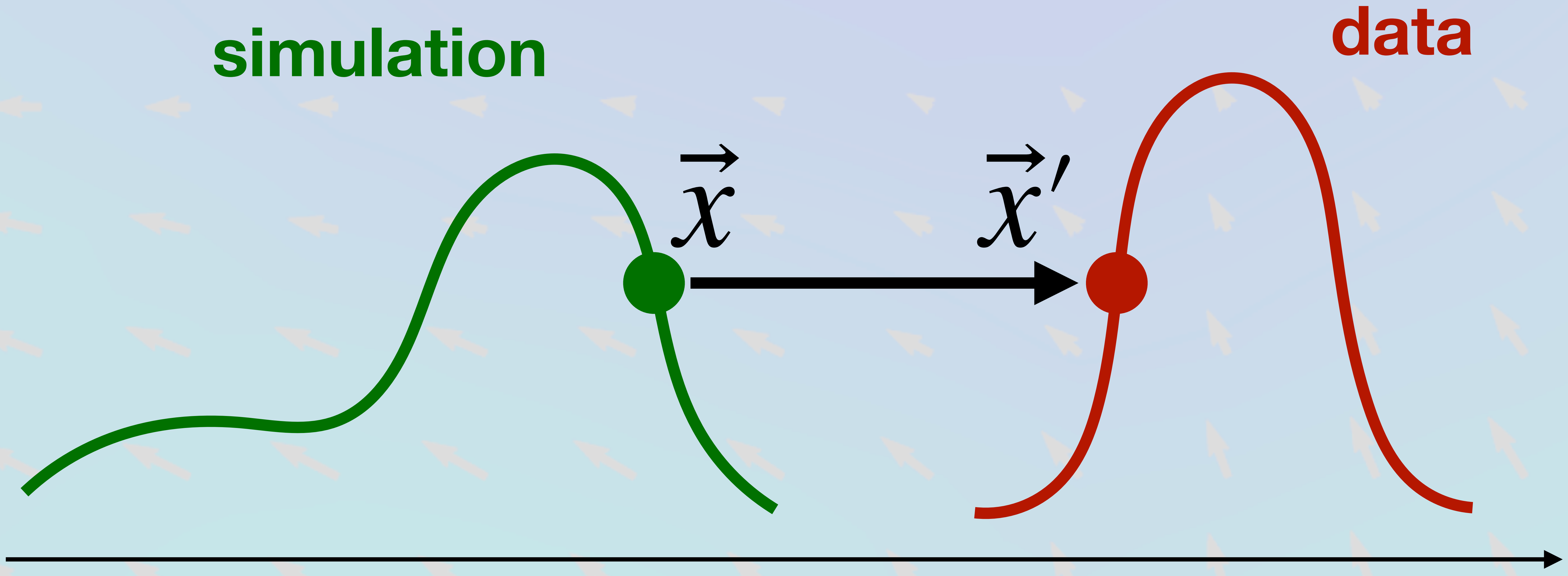
we have been implicitly using OT for decades in HEP:

- correct a gaussian density $\leftrightarrow$ alter μ, σ of simulated distribution.

neural OT generalizes this process.

- for euclidean spaces, the OT map is the gradient of some convex potential:

$$\hat{T}\vec{x} \equiv \vec{\nabla}\phi(\vec{x}).$$



we have been implicitly using OT for decades in HEP:

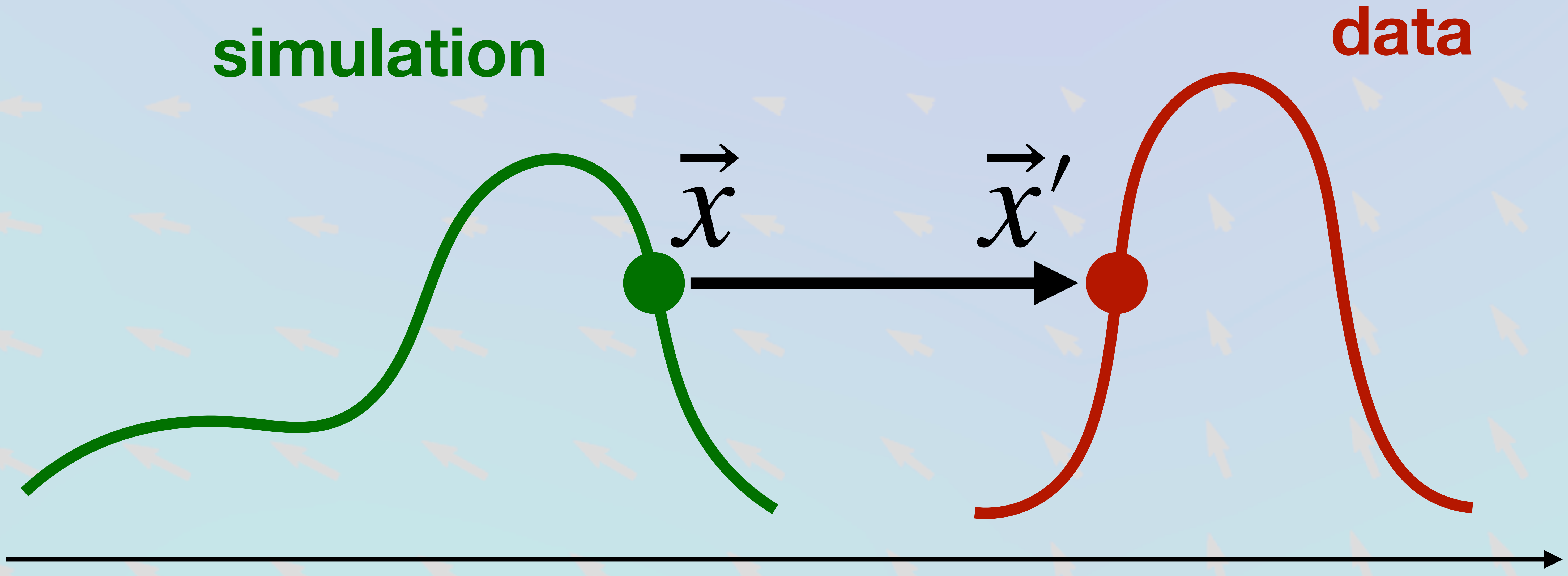
- correct a gaussian density ↔ alter μ, σ of simulated distribution.

neural OT generalizes this process.

- for euclidean spaces, the OT map is the gradient of some convex potential:

$$\hat{T}\vec{x} \equiv \vec{\nabla}\phi(\vec{x}).$$

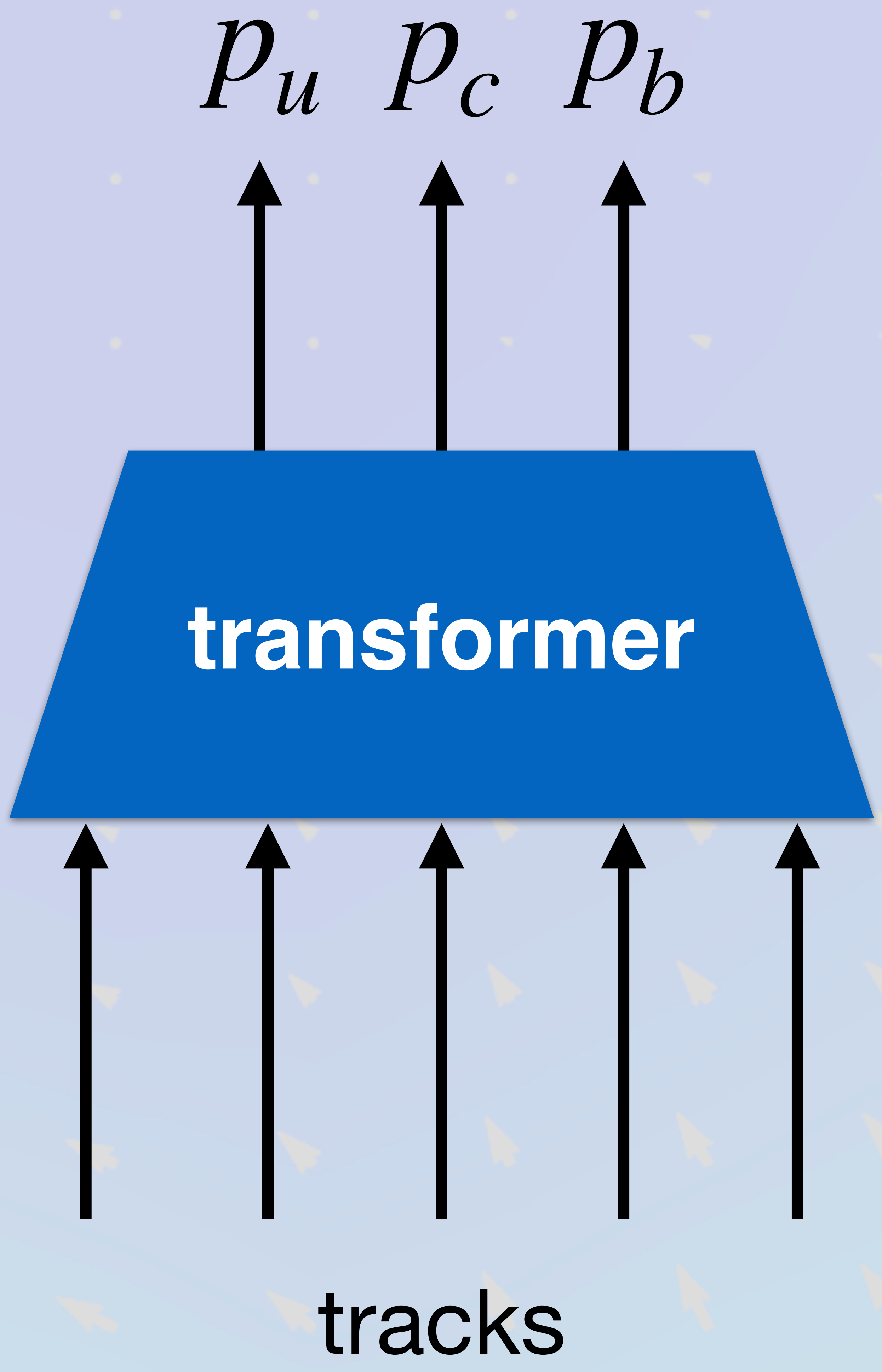
- partially-convex neural networks learn ϕ_z and therefore $\hat{T}_z$:
  - the OT map is *conditional* on z.



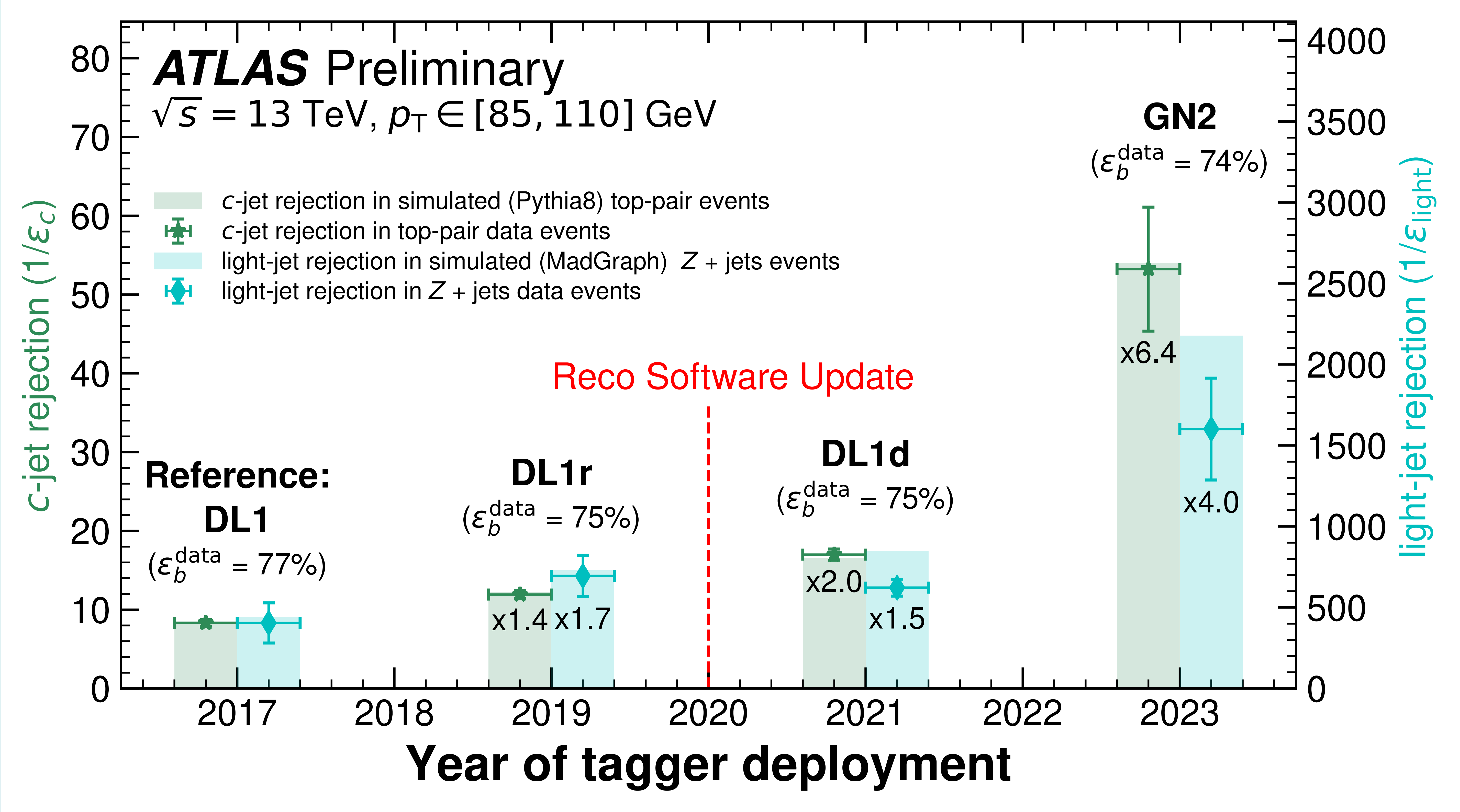
jet flavor-tagging is a classification problem:

- ATLAS's classifiers emit the probability of a jet to contain a b-hadron, c-hadron, or neither (p_b, p_c, p_u).
- modern algorithms are transformer-based:
  - ✦ charged-particle tracks as "point cloud" inputs.
  - ✦ see Greta's overview for more details
- historically, discriminant scores were calibrated in bins

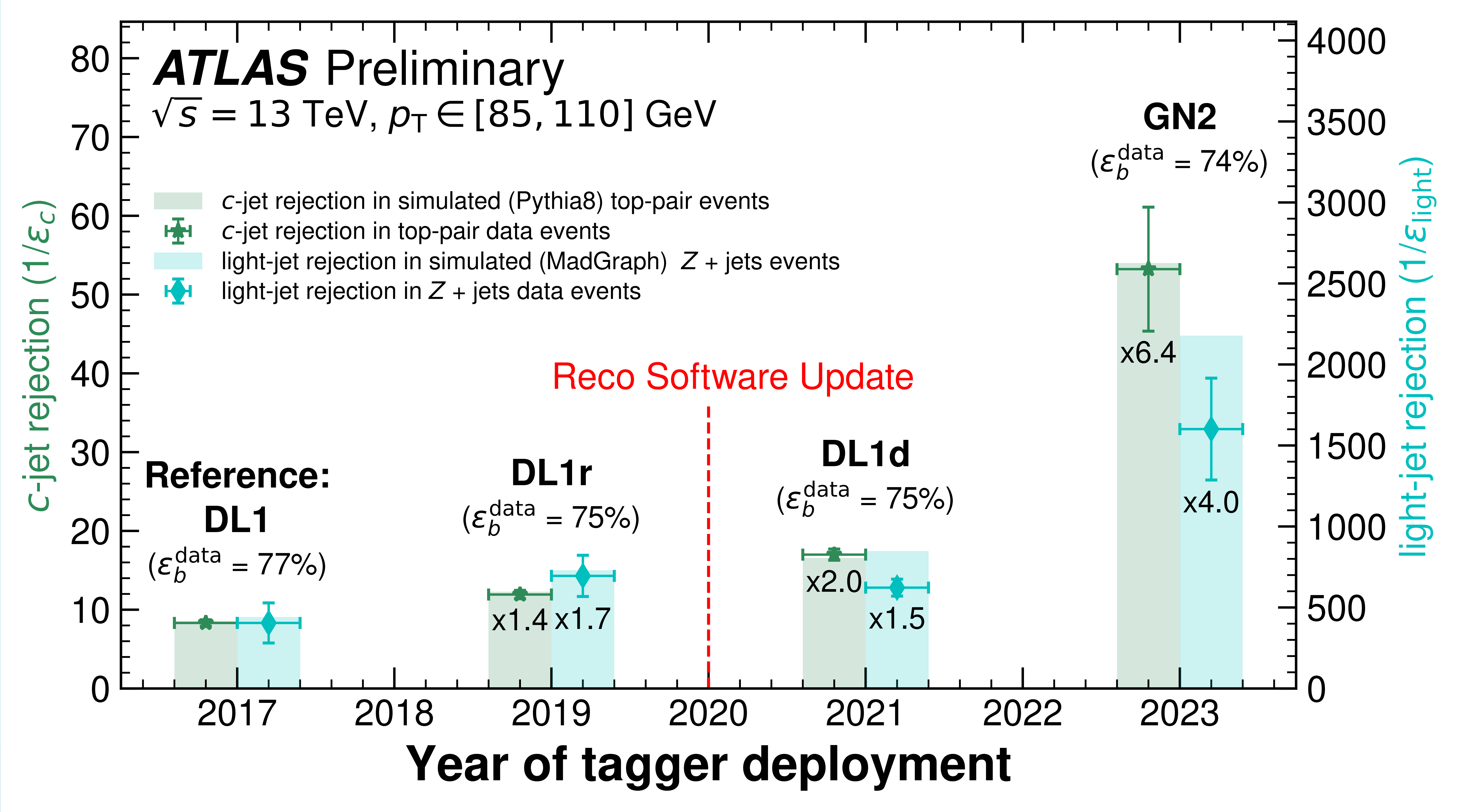
$$D_b \equiv \log \frac{p_b}{f_c p_c + (1 - f_c) p_u}$$



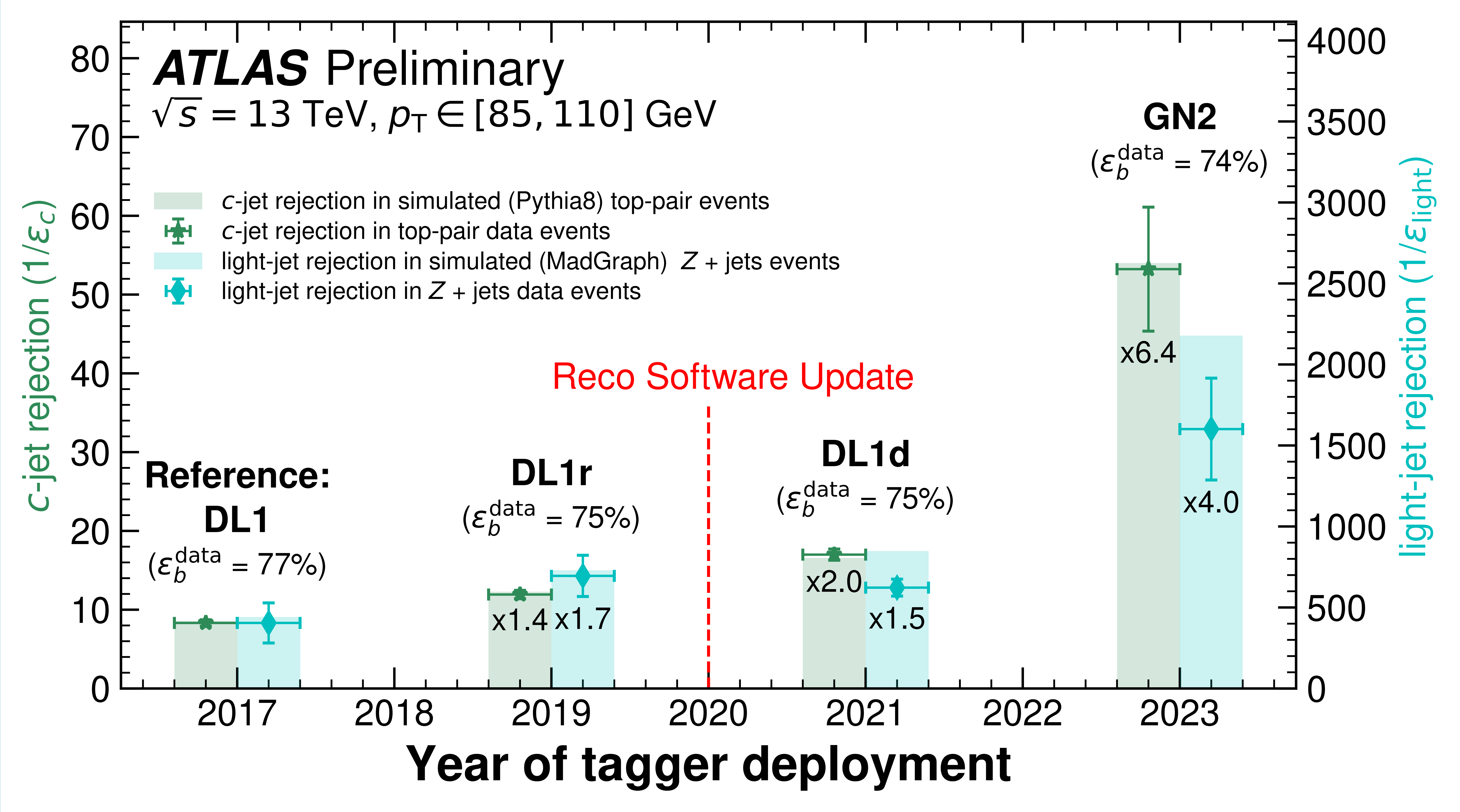
◉ transformer-based discriminators deliver incredible separating power



- ◉ transformer-based discriminators deliver incredible separating power
  - ✦ but some clear mismodeling in the simulation.



- transformer-based discriminators deliver incredible separating power
  - but some clear mismodeling in the simulation.
- until now, we had no direct calibration for the jet flavor probabilities.
  - to do so, we defined $q_i \equiv \mathrm{logit}\, p_i$, treated $\vec{q}$ as euclidean, and calibrated via OT.



notation:

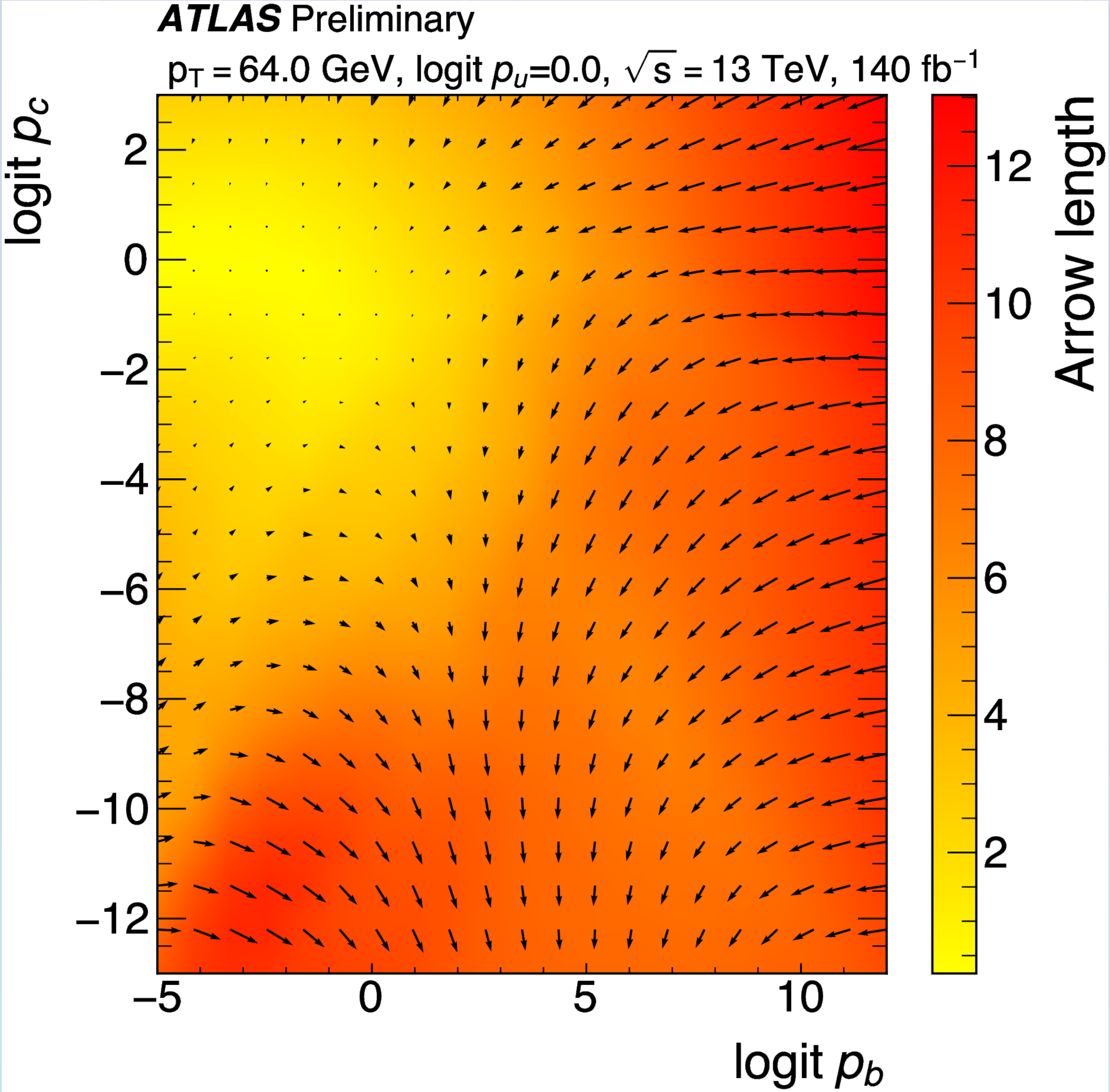
- $q_i \equiv \operatorname{logit} p_i$: flav. class. scores
- $p'_{\rm sim}(\vec{q}, p_T) \equiv p_{\rm sim}(\vec{q}\,|\,p_T)\, p_{\rm data}(p_T)$
- $\hat{T}_\# \equiv p_T$-dependent OT map

results: light-flavor jets

◉ we obtain the full 3D OT maps in $\vec{q}$ space s.t. $\hat{T}_\# p_{\rm sim} \approx p_{\rm data}$,

✦ derived as a function of jet p_T.

◉ here we show a 2D slice for $q_b \times q_c$ at fixed $q_u \times p_T$.



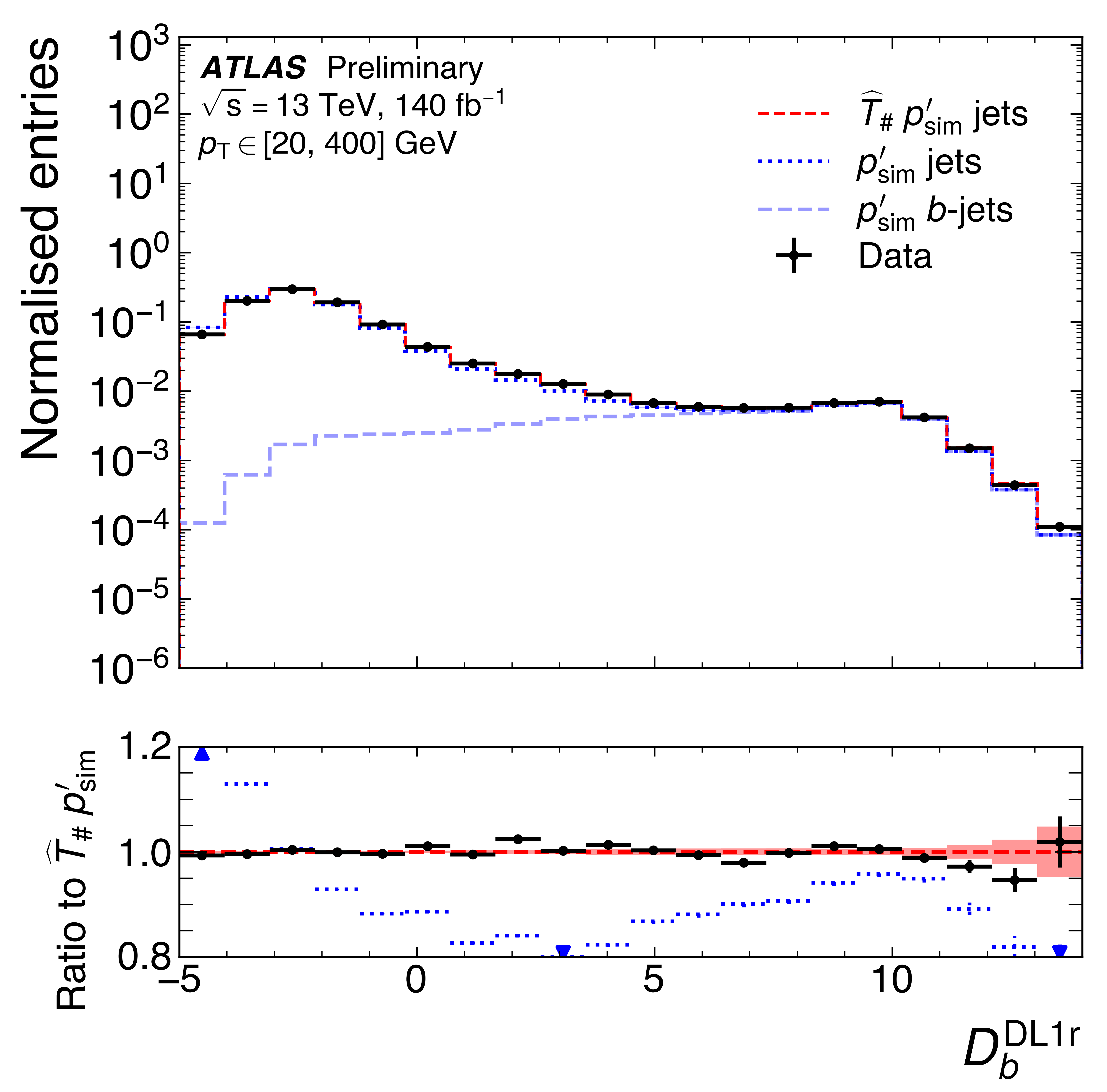
notation:

- $q_i \equiv \operatorname{logit} p_i$: flav. class. scores
- $p'_{\rm sim}(\vec{q}, p_T) \equiv p_{\rm sim}(\vec{q}\,|\,p_T)\, p_{\rm data}(p_T)$
- $\hat{T}_\# \equiv p_T$-dependent OT map

the technology works!

◉ very good agreement after calibration, including for D_b.

◉ D_b was not a calibration target → full space of p_i properly corrected.



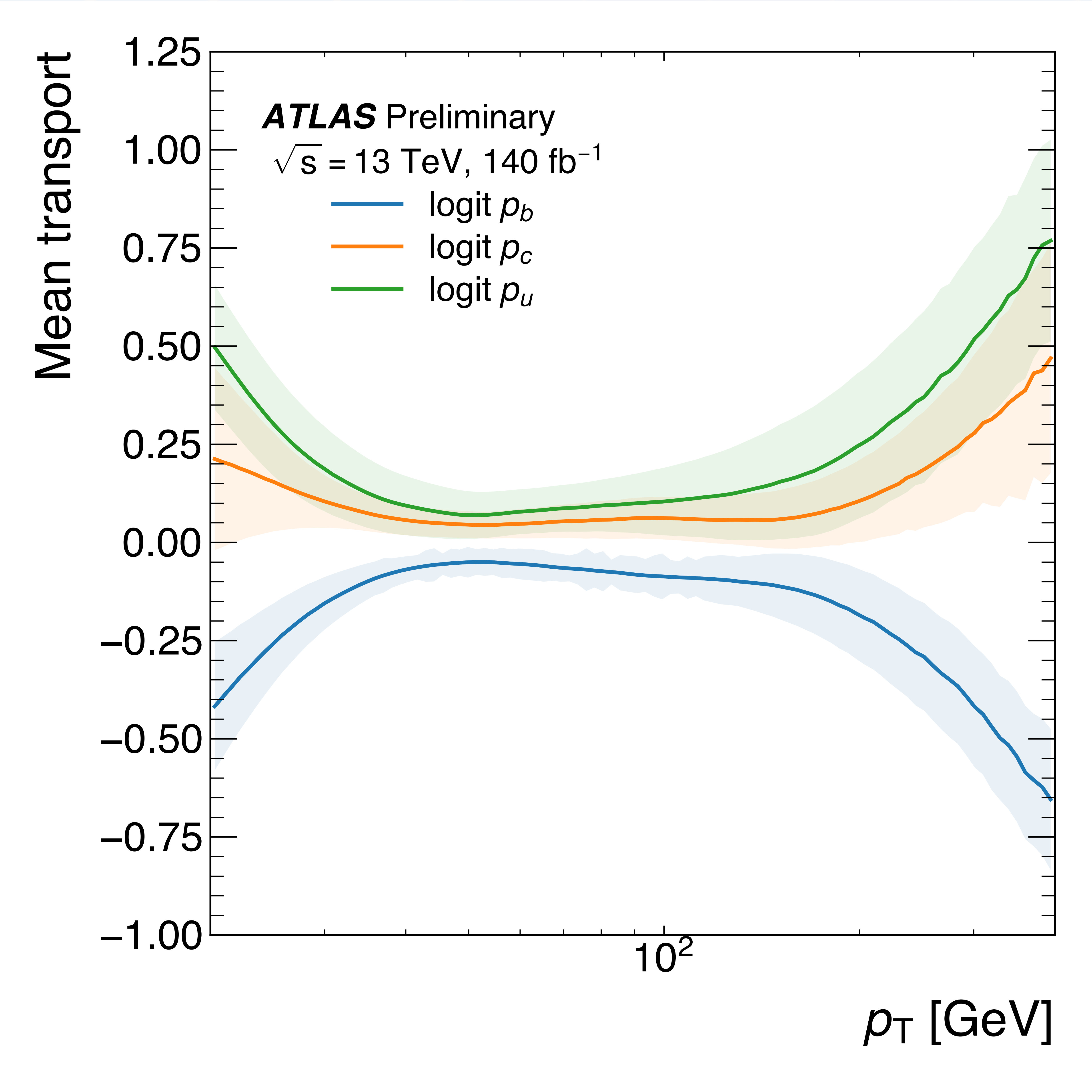
notation:

- $q_i \equiv \operatorname{logit} p_i$: flav. class. scores
- $p'_{\mathrm{sim}}(\vec{q}, p_T) \equiv p_{\mathrm{sim}}(\vec{q}\,|\,p_T)\, p_{\mathrm{data}}(p_T)$
- $\hat{T}_\# \equiv p_T$-dependent OT map

◉ for b-jets: $\hat{T}\,p_b < p_b$, while the reverse is true for p_c and p_u.

✦ the simulation overstates its classification power.

◉ yields a very fine-grained understanding of mismodeling!

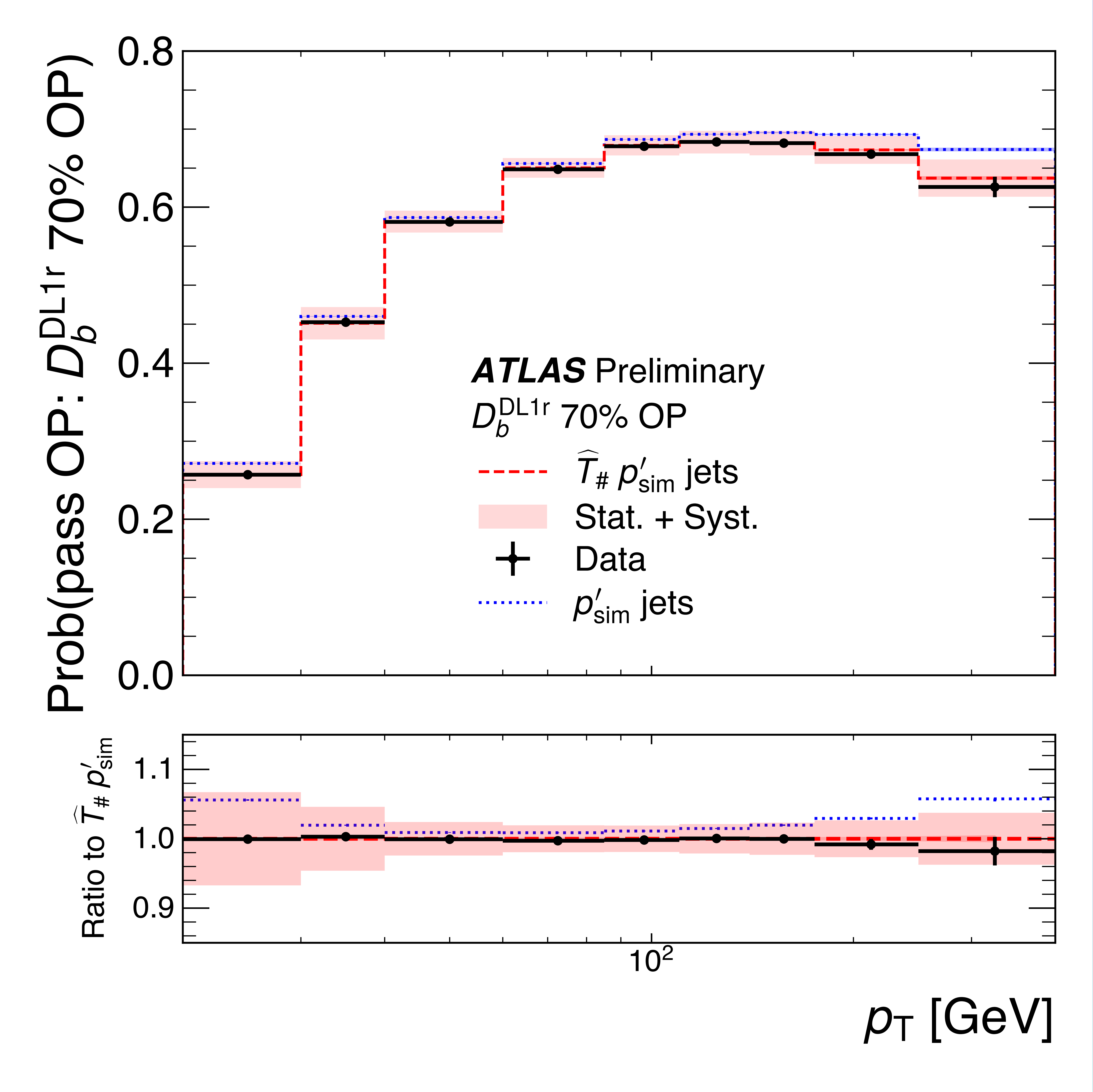


notation:

- $q_i \equiv \text{logit}\, p_i$: flav. class. scores
- $p'_{\text{sim}}(\vec{q}, p_T) \equiv p_{\text{sim}}(\vec{q}\,|\,p_T)\, p_{\text{data}}(p_T)$
- $\hat{T}_\# \equiv p_T$-dependent OT map

conventional operating points are "automatically" corrected.

◉ e.g. this one used for measuring the $H \leftrightarrow b, t$ couplings.



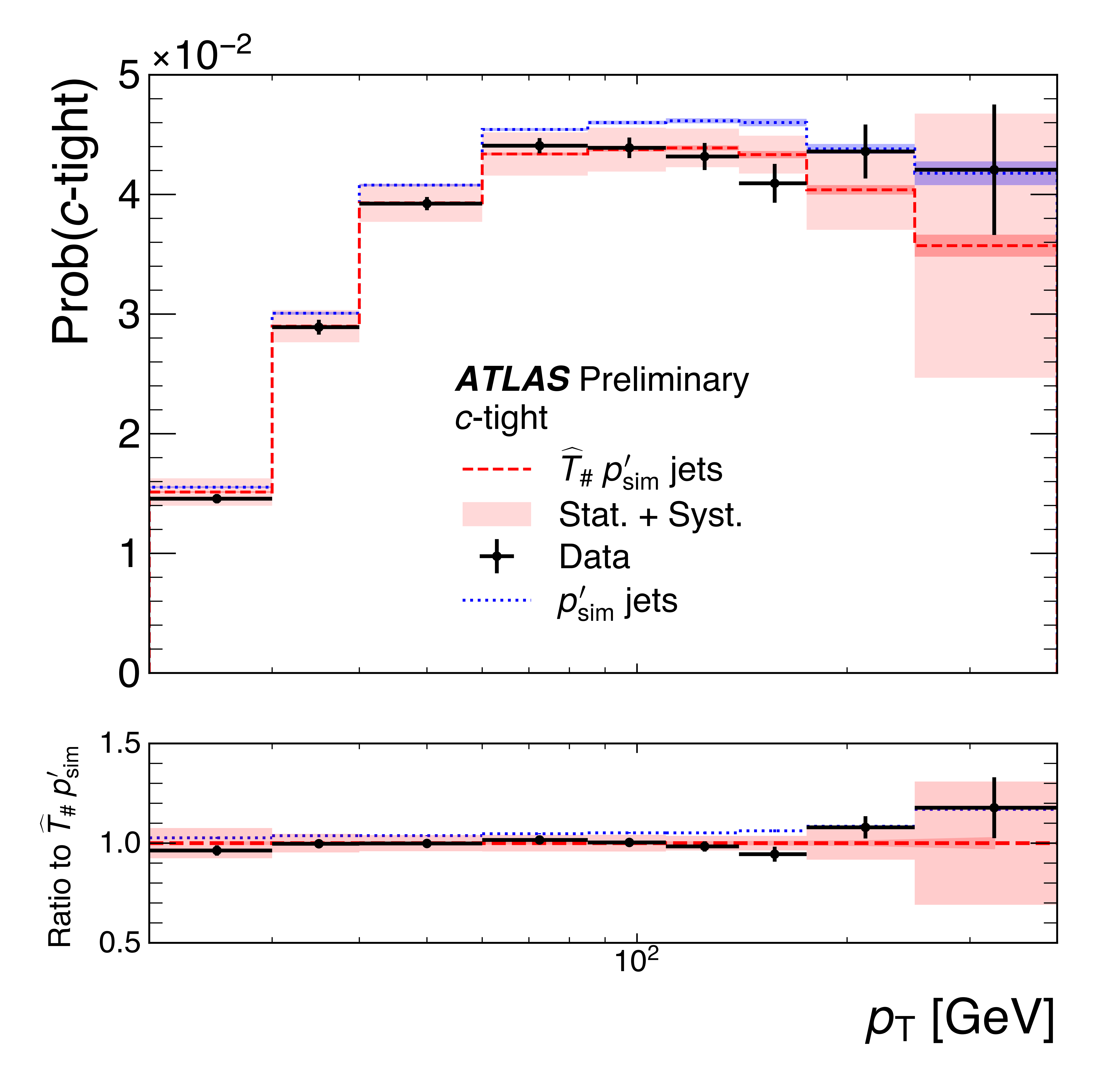
notation:

- $q_i \equiv \text{logit}\, p_i$: flav. class. scores
- $p'_{\text{sim}}(\vec{q}, p_T) \equiv p_{\text{sim}}(\vec{q}\,|\,p_T)\, p_{\text{data}}(p_T)$
- $\hat{T}_\# \equiv p_T$-dependent OT map

more general uses of flavor-tagging become possible.

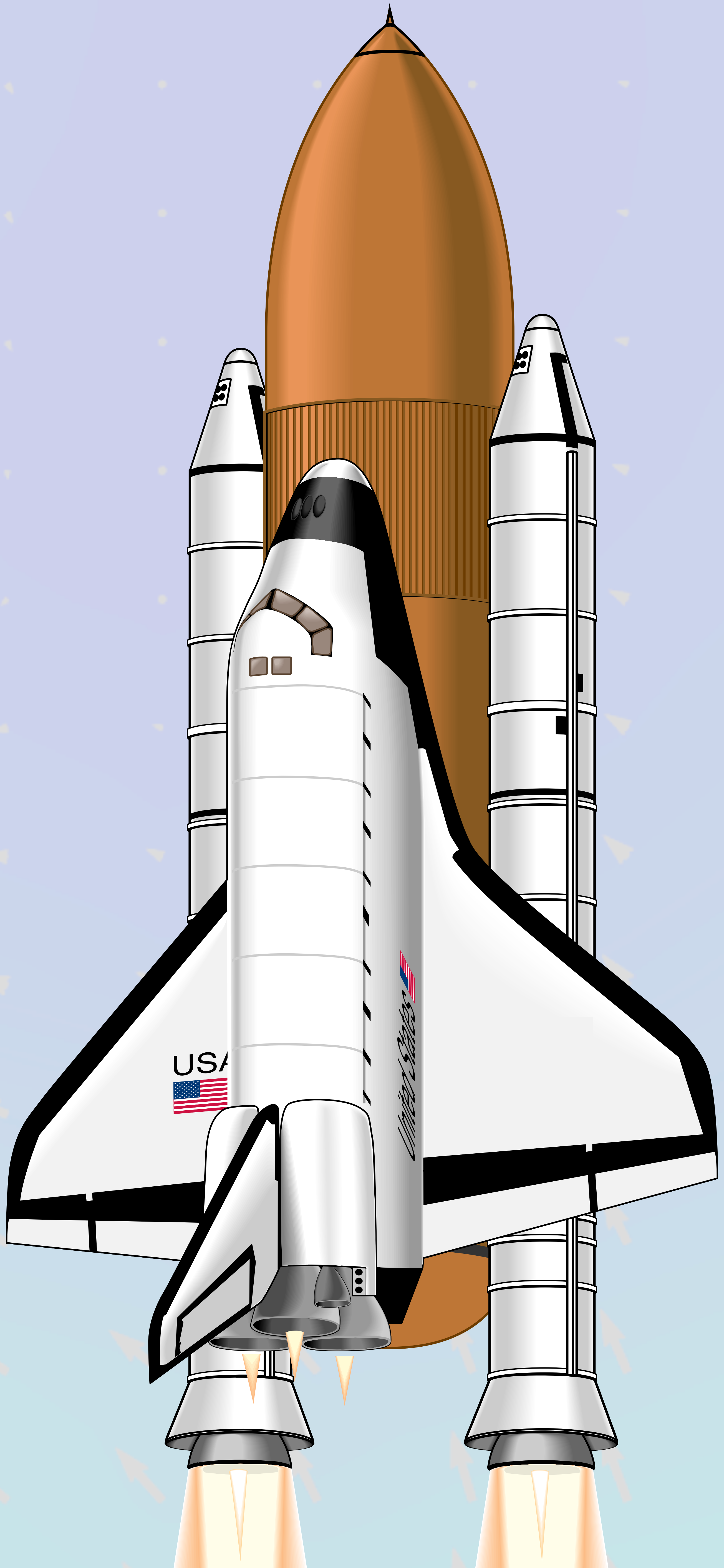
◉ discriminator used to constrain $H \leftrightarrow c$ couplings calibrated "for free".

◉ only because the full 3D density $\hat{T}_\# \, p_{\text{sim}}(\vec{q}\,|\,p_T)$ agrees with data.



the future

- here we've performed a first 3+1D continuous calibration via OT.
  - ✦ details and further background reading [here].
- the technique is *general*: it enables high-dimensional, transport-based calibrations.
- additional (informative) conditionals should result in more *universal* and *precise* calibrations,
  - ✦ allowing *richer* summary statistics for *better inference*.
- *we look forward to seeing what others can do with this technology!*




thank you,
and happy calibrating!