TUHH Hamburg University of Technology

Universität Hamburg
DER FORSCHUNG | DER LEHRE | DER BILDUNG

DASHH Data Science in Hamburg HELMHOLTZ Graduate School for the Structure of Matter

Fast Perfekt: Regression-based refinement of fast simulation

Patrick Connor[1], Samuel Bein[2,3], Peter Schleper[2], **Lars Stietz[2,4]**, Moritz Wolf[2]

[1] Center for Data and Computing in Natural Sciences

[2] University of Hamburg, Institute of Experimental Physics

[3] Université catholique de Louvain

[4] Hamburg University of Technology, Institute of Mathematics, Chair of Computational Mathematics

HELMHOLTZ Grant No. HIDS-0002

Fast Perfekt

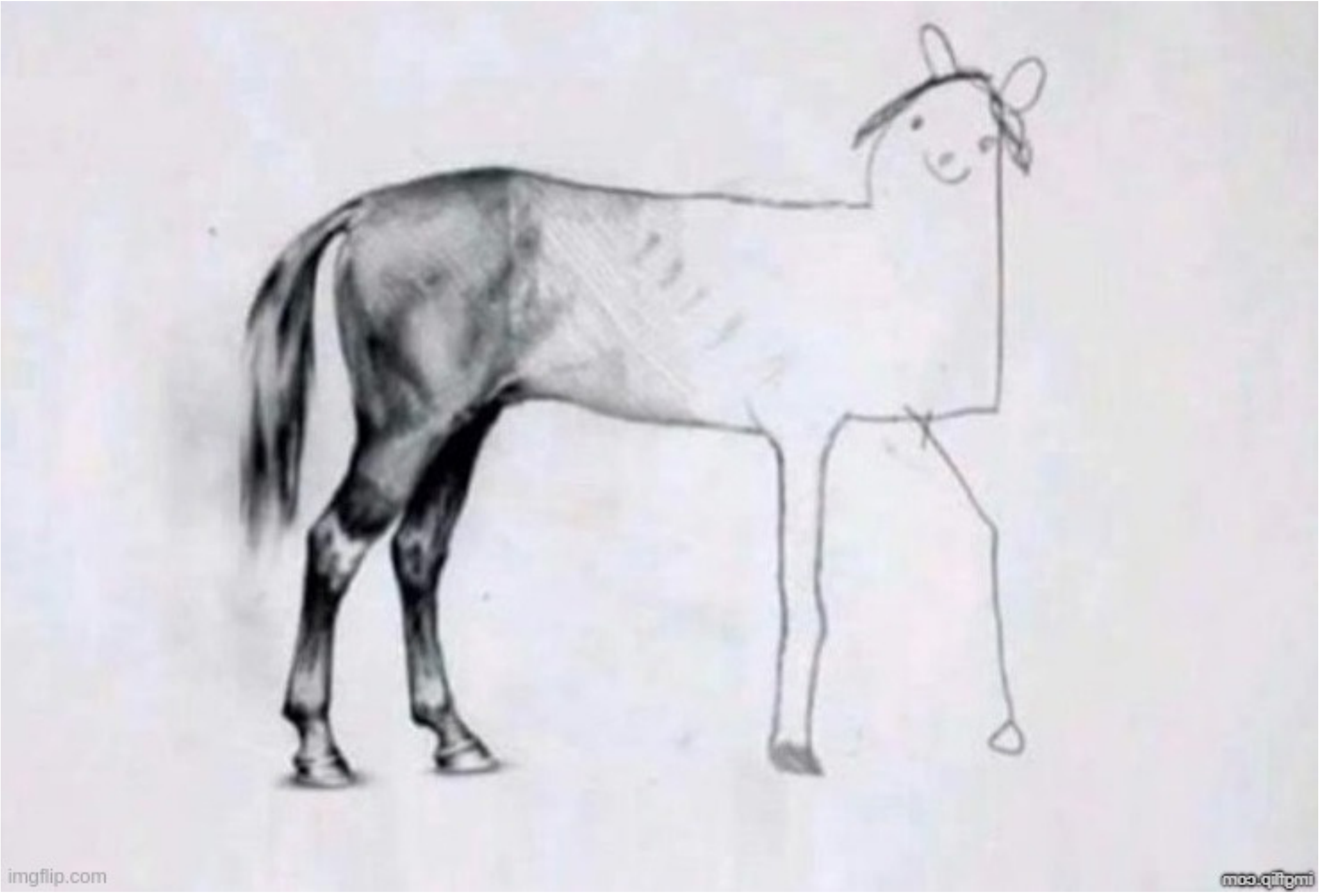
Refining simulation

Hi-Fi Simulation:

- High amount of computational power
- Accurate simulations

Lo-Fi Simulation:

- Low amount of computational power
- Approximate simulations



https://imgflip.com/i/98nedf

Refinement: Use Lo-Fi simulation routine and make the output similar to the Hi-Fi simulation output

Fast Perfekt

Goal

- Refine final analysis observables
- Accurate refinement of the distribution (bulk and tails)
- Modeling correlations among several available and hidden features
- Weightless, preserving statistical precision
- Fast and deterministic to ensure efficiency and traceability

Network

Network

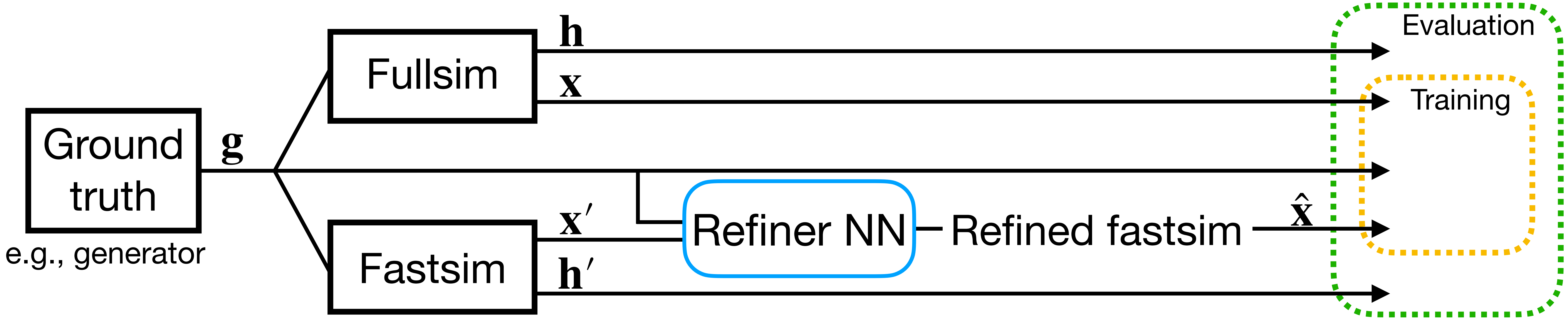
Simulation chain

1. Generator outputs ground truth $\mathbf{g}$

2. Detector simulation output $\mathbf{x}, \mathbf{x}'$ (*available features*) and $\mathbf{h}, \mathbf{h}'$ (*hidden features*)

- **Hidden** features are either **unavailable** or **not chosen** among the features to refine

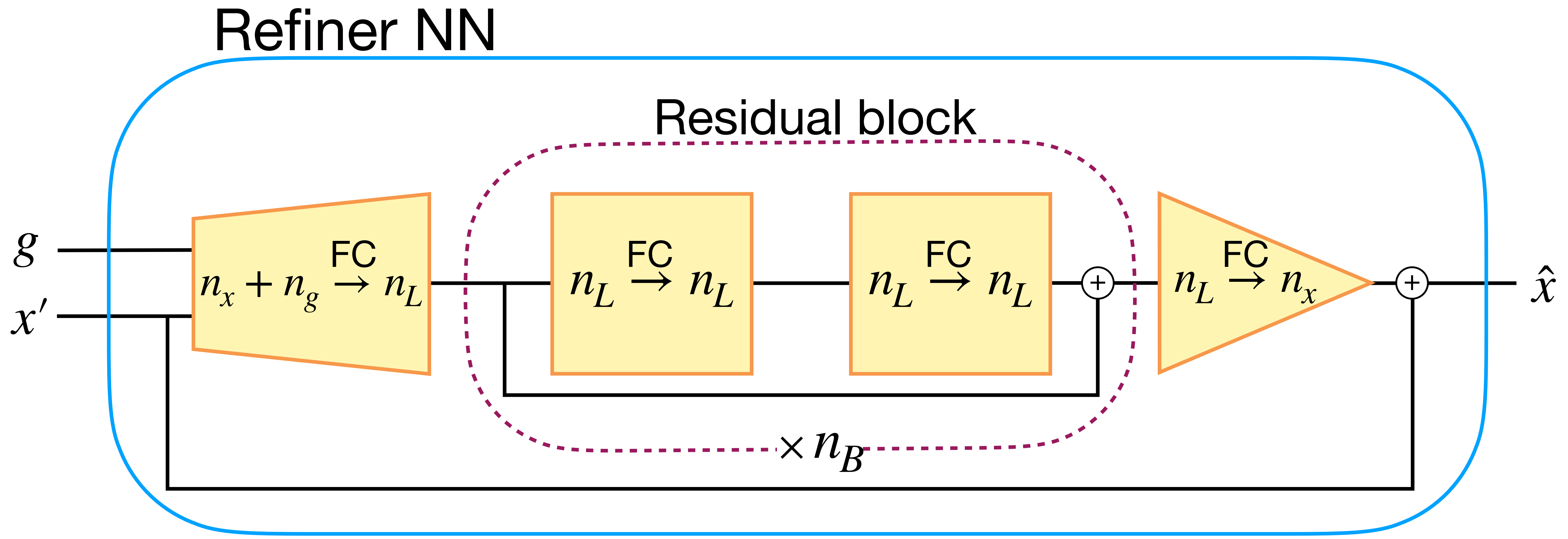
3. Apply Refiner NN to to Fastsim output $\mathbf{x}'$ conditioned on $\mathbf{g}$ to get refined fastsim $\hat{\mathbf{x}}$



Network

Refiner NN

- Architecture inspired by ResNet[1]
- Network constructed to determine residual corrections
- Initialized such that the network behaves as identity function before training
- Fully connected (FC) layers to ensure fast evaluations



[1] Keiming He et al., Deep Residual Learning for Image Recognition, 2015

Analytic Example

Analytic Example

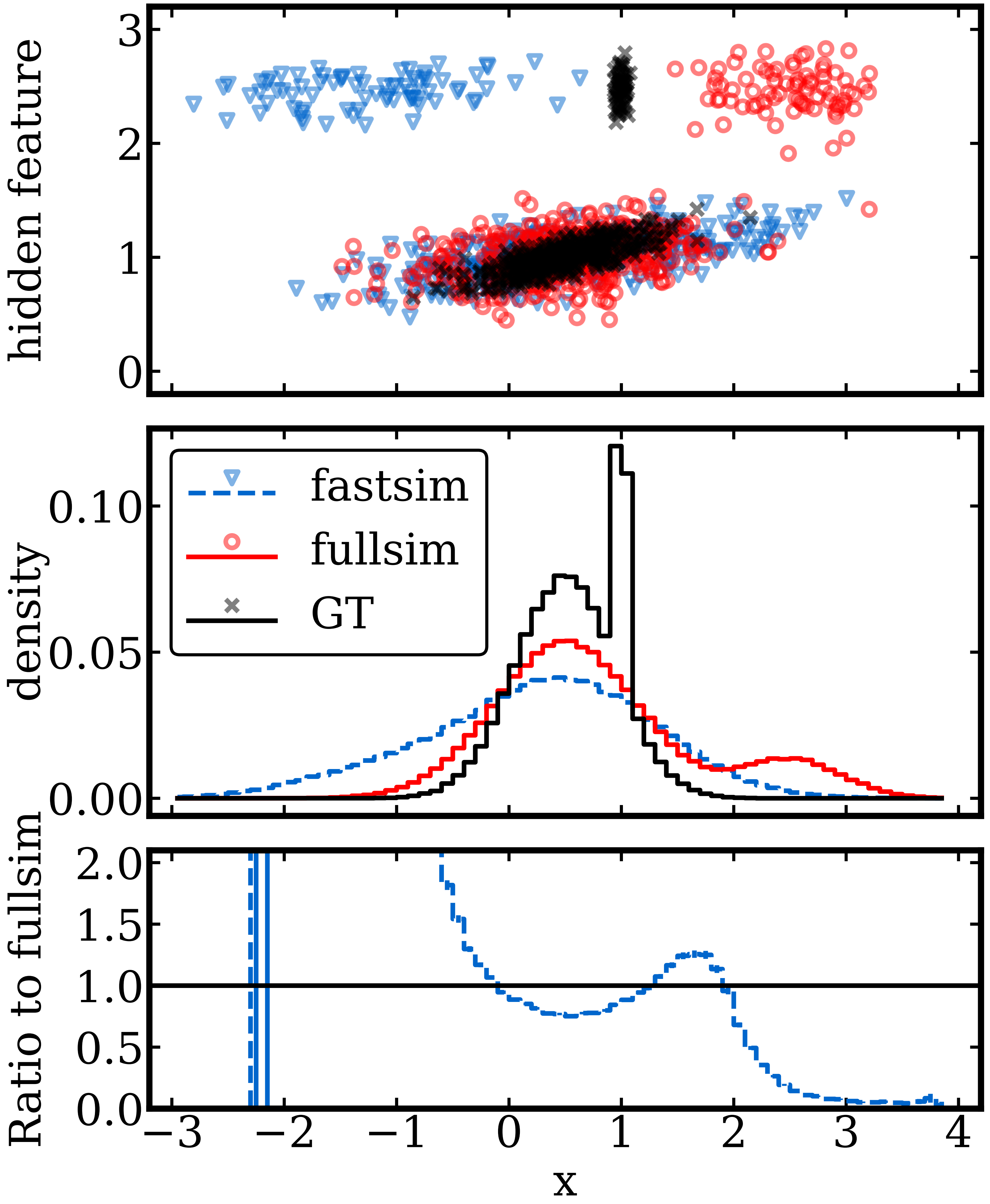
Data Set

- Two populations separated through a hidden feature
- Independent smearing process for fullsim and fastsim

Goal:

- Refine x variable
  - Match the distribution of x
  - Model correlations to the hidden feature correctly, i.e., match the populations

Wich loss do we use?



Analytic Example

Loss Functions

- Mean Squared Error (MSE):
  - Gives pairwise comparison of targets $y \in \mathbb{R}^n$ and output data $f_\theta(x) \in \mathbb{R}^n$

$$\mathrm{MSE}(\theta) = \frac{1}{m} \sum_{i=1}^{m} \|y_i - f_\theta(x_i)\|^2$$

- Maximum Mean Discrepancy (MMD):
  - Gives comparison of ensembles of target $y \sim P$ and output data $f_\theta(x) \sim \hat{P}$, thus compares underlying distributions

$$\mathrm{MMD}(\theta) = \frac{1}{m^2} \sum_{i,j=1}^{m} k(y_i, y_j) + \frac{1}{m^2} \sum_{i,j=1}^{m} k(f_\theta(x_i), f_\theta(x_j)) - \frac{2}{m^2} \sum_{i,j=1}^{m} k(y_i, f_\theta(x_j))$$

$$k(x, y) = \exp\left(- \sum_{l=1}^{n} \frac{(x_l - y_l)^2}{\sigma_l^2} \right)$$

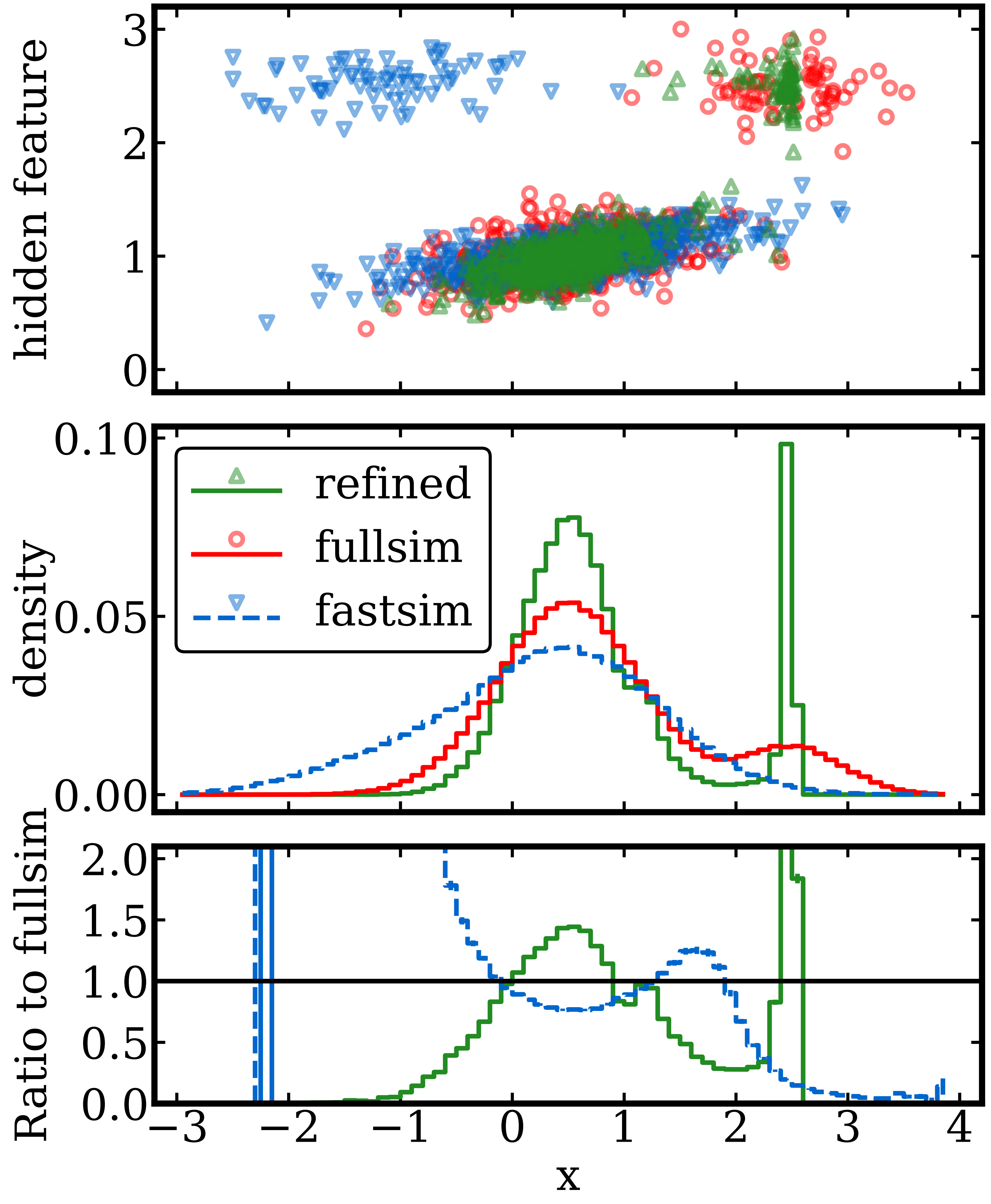
Analytic Example

MSE Result

- **Benefit**:

  ✓Populations are matched

- **Drawback:**
  - Distributions disagree
  - MSE leads to a regression to the mean phenomenon



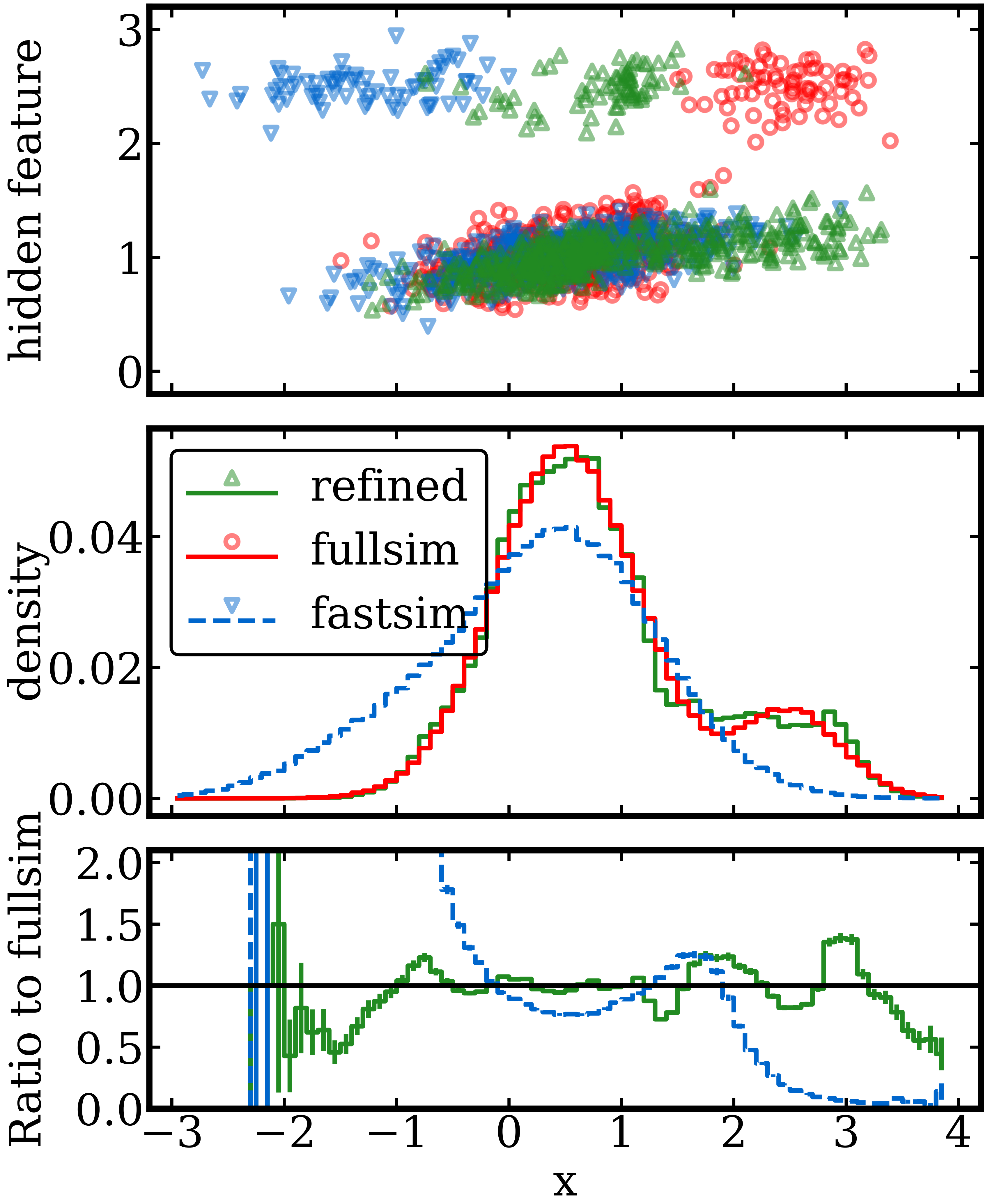
Analytic Example

MMD Result

- **Benefit**:

  ✓Distributions agree

- **Drawback**:
  - The populations are not correctly matched
  - Points from the large population are used for the second mode in the marginal



Analytic Example

2-Stage Training

1. Stage:
   - We combine MSE and MMD to form
   $$\mathscr{L}(\theta,\lambda) = \mathrm{MSE}(\theta) - \lambda\big(\varepsilon - \mathrm{MMD}(\theta)\big) - \frac{\delta}{2}\big(\varepsilon - \mathrm{MMD}(\theta)\big)^2$$
   - Minimize w.r.t. θ, maximize w.r.t. λ (Modified Differential Method of Multipliers[1])
   - Train until convergence

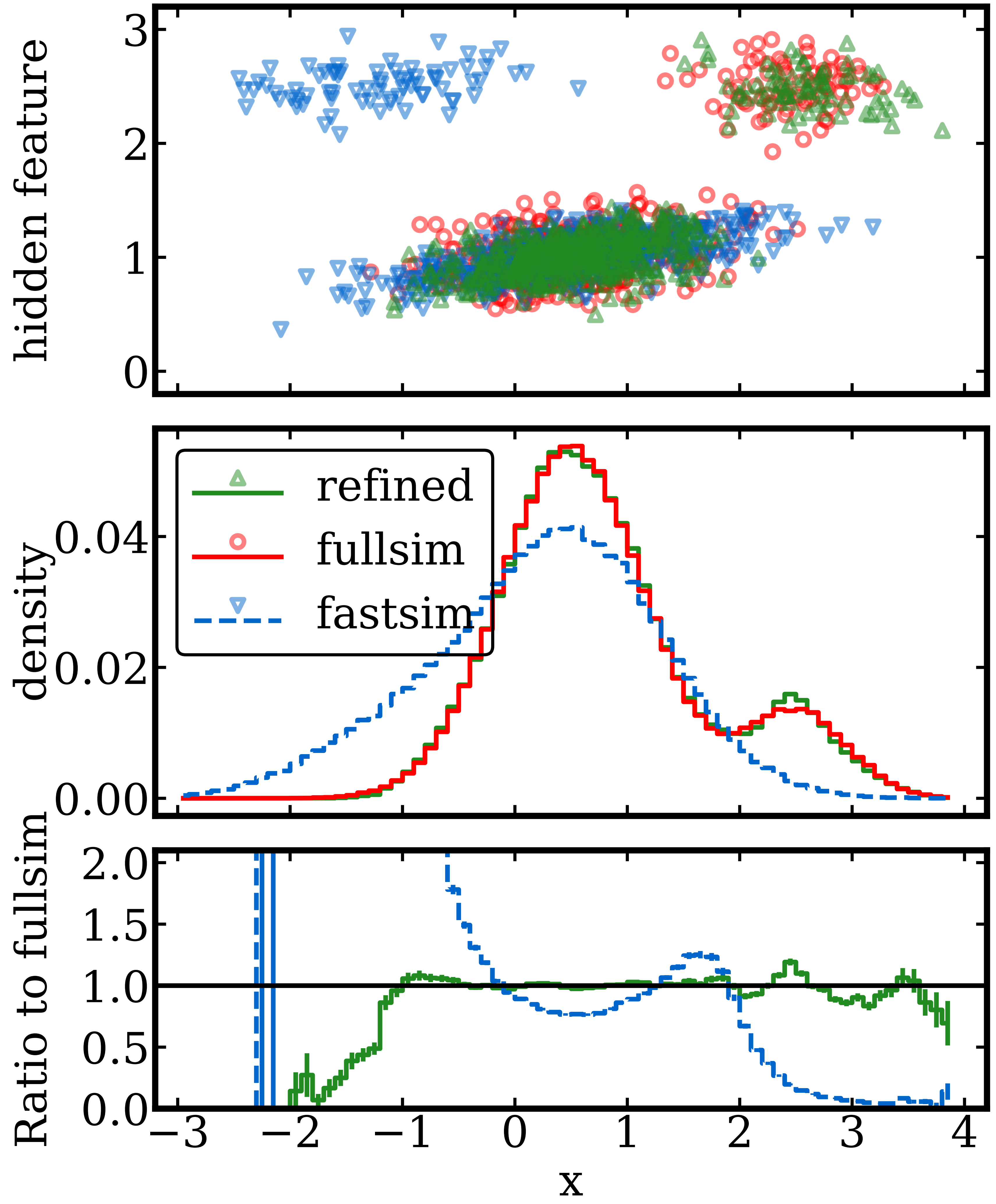
2. Stage:
   - We use MMD only
   - Let the network fine-tune the distribution agreement

[1] John C. Platt & Alan H. Barr, Constrained Differential Optimization, 1988, NIPS

Analytic Example

2-Stage Scheme

- Populations agree
- Marginal distribution agrees
- Drawbacks from 1-stage training eliminated



Analytic Example

Overall Results

Fullsim vs	Fastsim	Refined (MMD-only)	Refined (2-stage)
Omniscient MMD $\times 10^3$	38.39 ± 2.90	45.11 ± 3.03	0.33 ± 0.17
MMD $\times 10^3$	39.07 ± 3.70	0.41 ± 0.30	0.23 ± 0.15
MSE $\times 10^3$	1384 ± 63	602 ± 24	200 ± 6
χ^2/ndof	3493	28	20

- Omniscient MMD includes the hidden feature to the data vector
  → statement on correlations to hidden features
- MMD-only and 2-stage refinement was able to improve all metrics
- Best values in all metrics are achieved by the 2-stage training

Application to collider physics

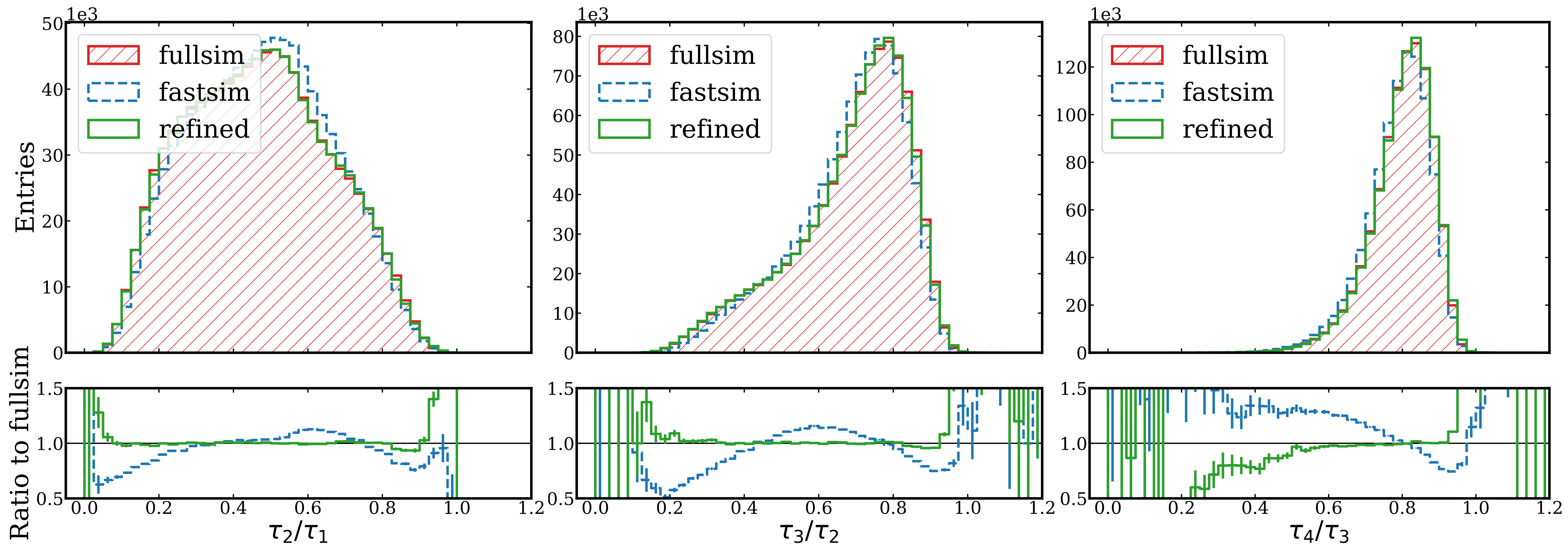
Application to collider physics

Refining analysis-level observables

See next talk by Sam:
Refining CMS Fast Simulation
with ML-based regression

Fullsim vs	Fastsim	Refined (MMD-only)	Refined (2-stage)
Omniscient MMD $\times 10^3$	0.5386 ± 0.0077	0.5147 ± 0.0068	0.5149 ± 0.0066
MMD $\times 10^3$	1.114 ± 0.087	0.305 ± 0.022	0.303 ± 0.024
MSE $\times 10^3$	10.99 ± 0.23	8.25 ± 0.02	8.24 ± 0.02
χ^2/ndof	33.34	1.97	2.52

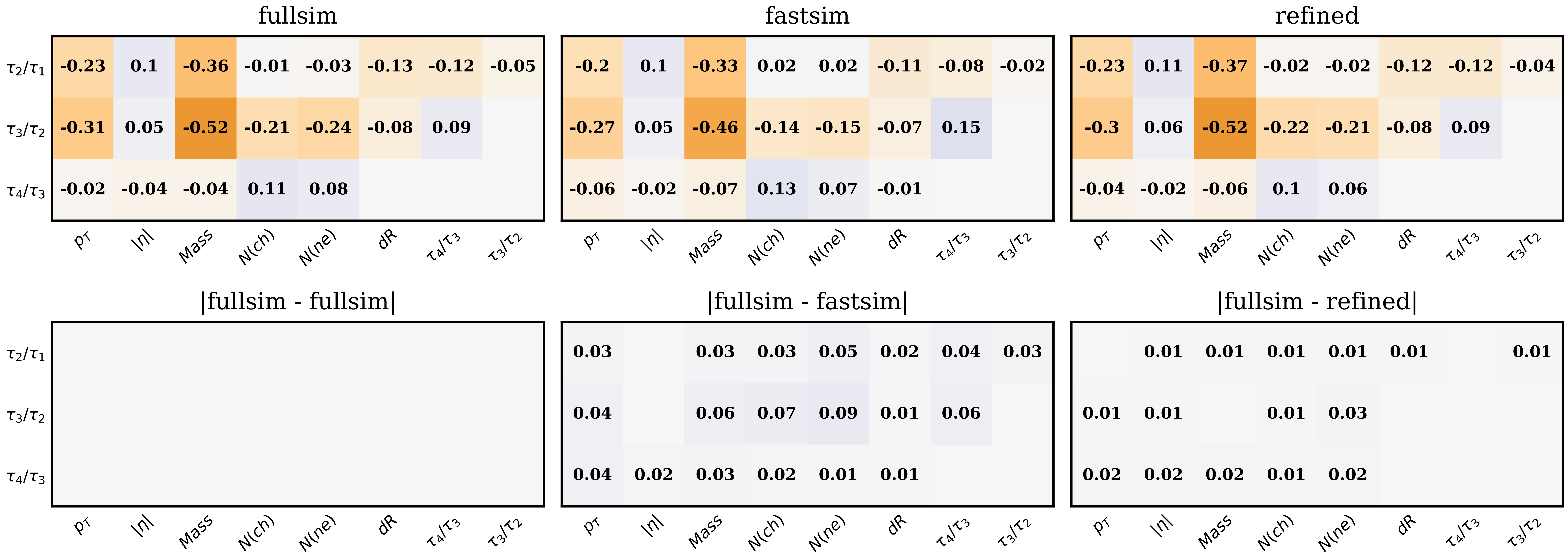
- Similar results for 2-stage and 1-stage approach
- 1-stage approach sufficient if modes already coincide well
- Evaluation time: ~1ms (CPU)



Application to collider physics

Refining analysis-level observables — Correlations

- Pearson-correlation coefficient between different observables, available or hidden
- The refined correlations are improved



Conclusion

Conclusions

- Introduced regression-based refinement/morphing using MMD loss
- Refinement works on the output of fastsim → using existing domain knowledge

- **1-stage** training uses MMD, evaluating distribution similarity
  - Good refinement of the available feature on distribution level
  - Falls short in presence of strong bias in hidden dimension
  - Suitable for unpaired training data

- **2-stage** training includes an additional first stage using a combination of MSE and MMD
  - Improves the accuracy of the refinement
  - Describes the multidimensional domain well

Preprint available: "Fast Perfekt: Regression-based refinement of fast simulation",
Moritz Wolf, Lars O. Stietz, Patrick L.S. Connor, Peter Schleper, Samuel Bein, arXiv:2410.15992

TUHH Hamburg University of Technology

Universität Hamburg
DER FORSCHUNG | DER LEHRE | DER BILDUNG

DASHH Data Science in Hamburg HELMHOLTZ Graduate School for the Structure of Matter

Thank You very much

Lars Stietz — lars.stietz@tuhh.de

TUHH, Institute of Mathematics, Chair for Computational Mathematics
UHH, Institute of Experimental Physics

HELMHOLTZ Grant No. HIDS-0002

TUHH Hamburg University of Technology

Universität Hamburg
DER FORSCHUNG | DER LEHRE | DER BILDUNG

DASHH
Data Science in Hamburg
HELMHOLTZ Graduate School
for the Structure of Matter

Backup

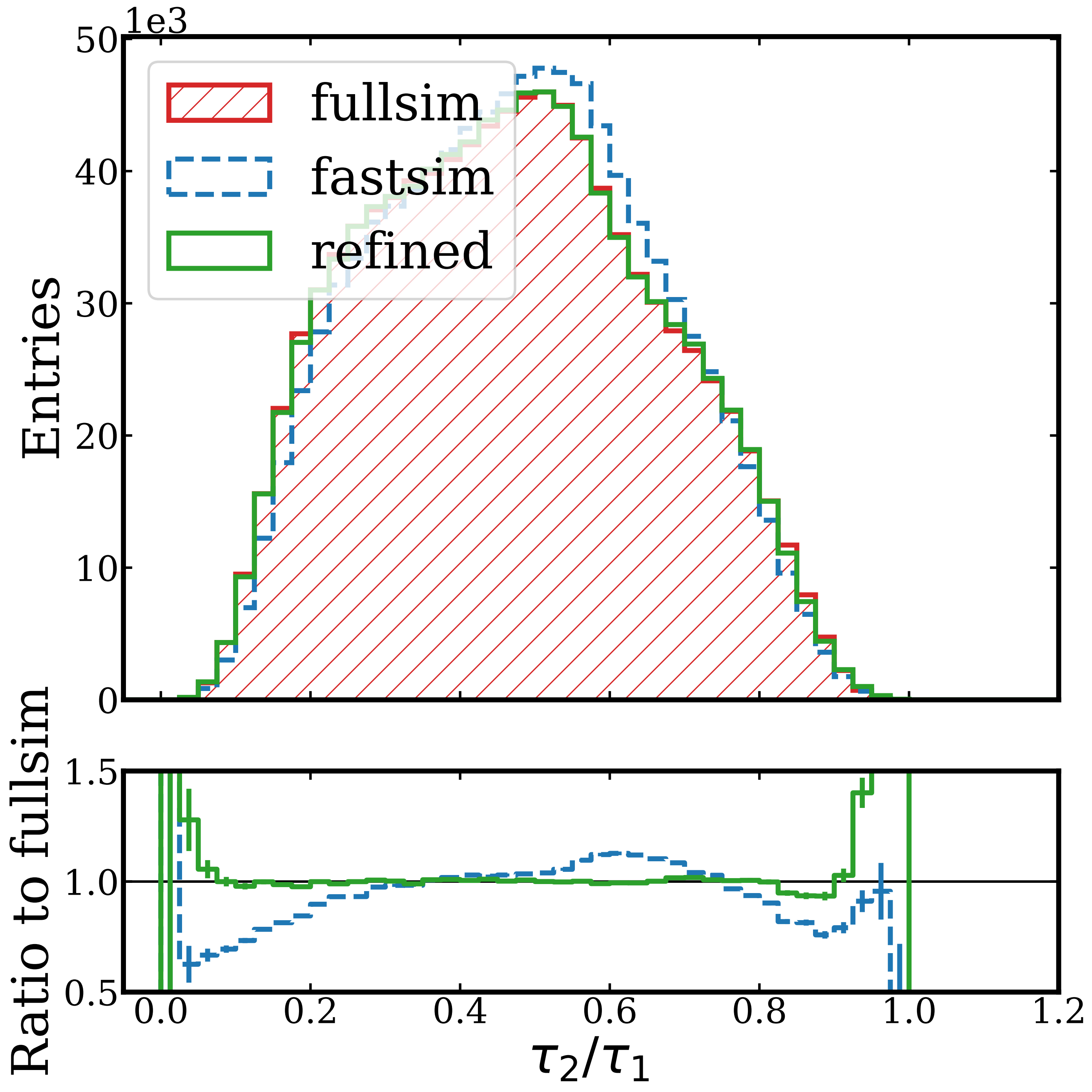
HELMHOLTZ

Application to collider physics

Data set

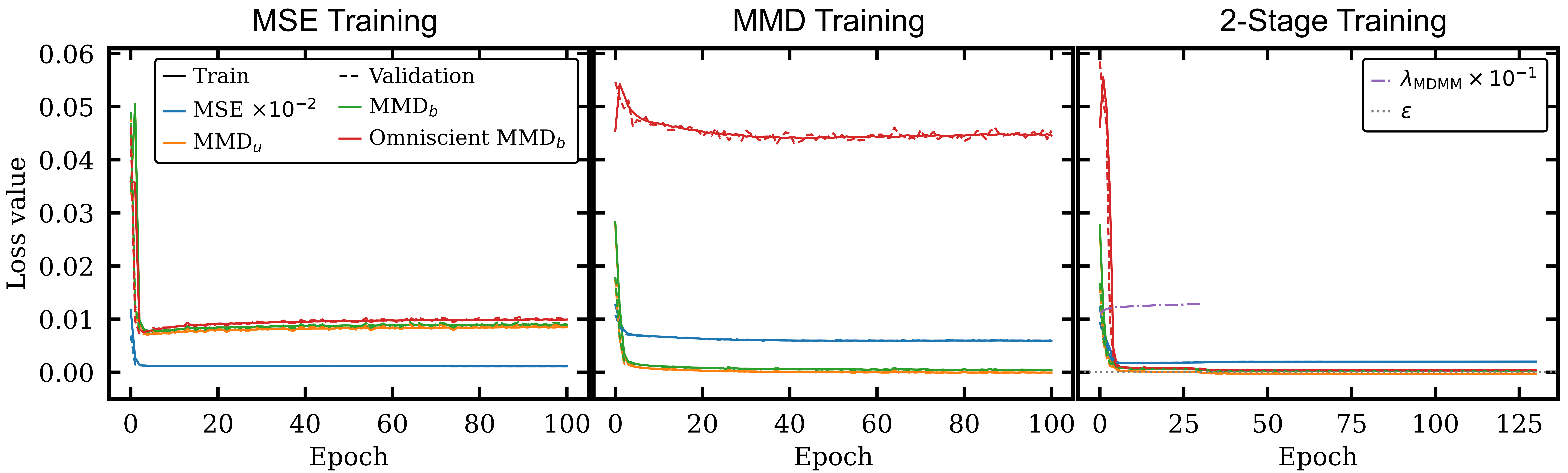
- **Generator:** Pythia 8.1
- **Simulation:** Delphes (Fullsim) vs "flawed" Delphes (Fastsim)
- **Available features**: Jet substructure observables

  N-subjettiness ratios $\mathbf{x}' = \left(\frac{\tau_2}{\tau_1}, \frac{\tau_3}{\tau_2}, \frac{\tau_4}{\tau_3}\right)$
- **Hidden features:** jet mass, p_T, η, $\mathrm{dR} = \sqrt{\mathrm{d}\eta^2 + \mathrm{d}\phi^2}$, number of charged (neutral) jet constituents N(ch) (N(ne))



Analytic Example

Loss Curves



- MSE-only training gets correlations to hidden features right
- MMD-only training reduces correlation informations to hidden feature (Omniscient MMD)
- 2-stage training gets the best of both worlds

Analytic Example

Loss Curves