Modern Machine Learning Tools for Unfolding

Javier Mariño Villadamigo

In collaboration with Nathan Huetsch, Anja Butter, Theo Heimel, Tilman Plehn

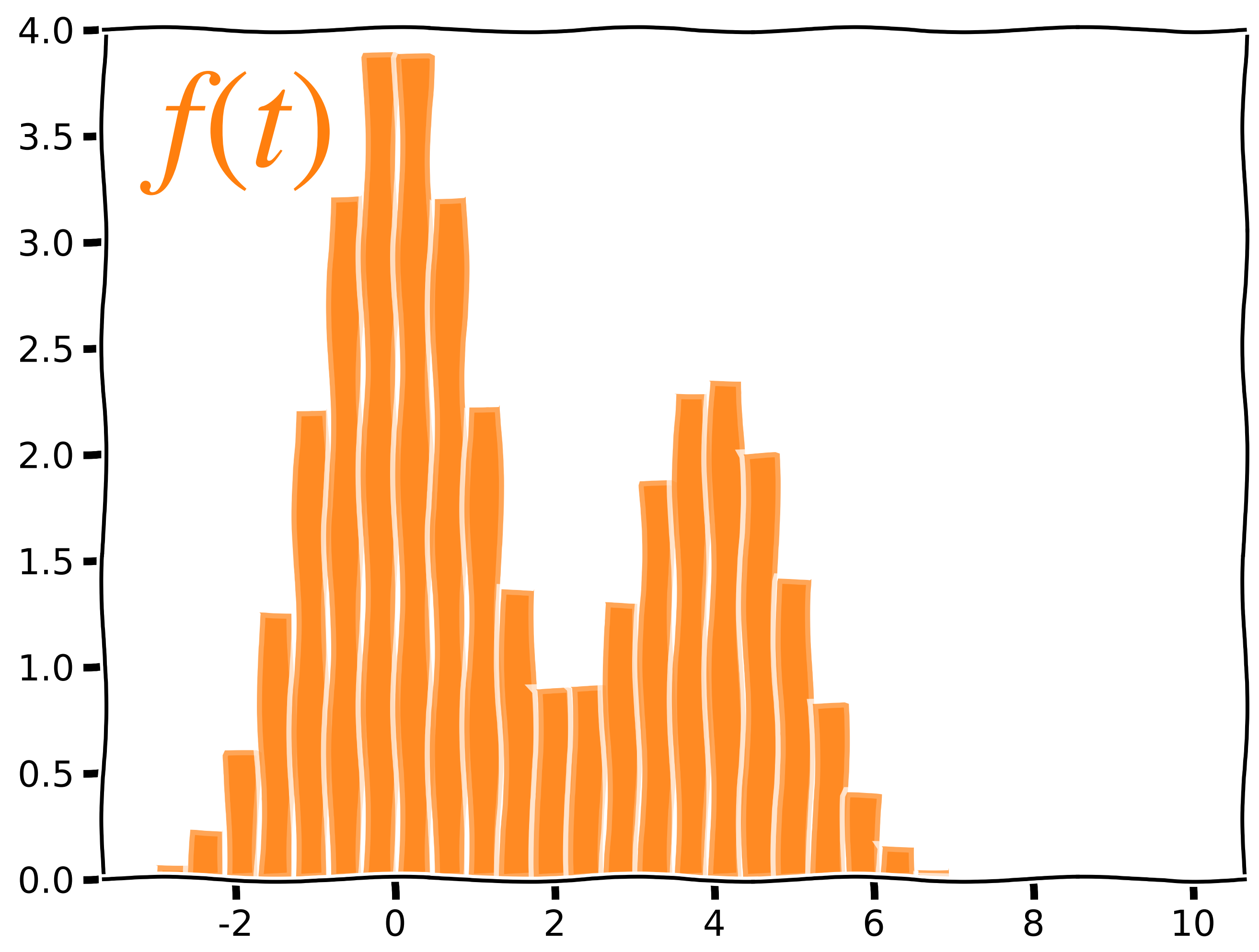
ACAT March 12th 2024

Institut für Theoretische Physik - University of Heidelberg

UNIVERSITÄT HEIDELBERG ZUKUNFT SEIT 1386

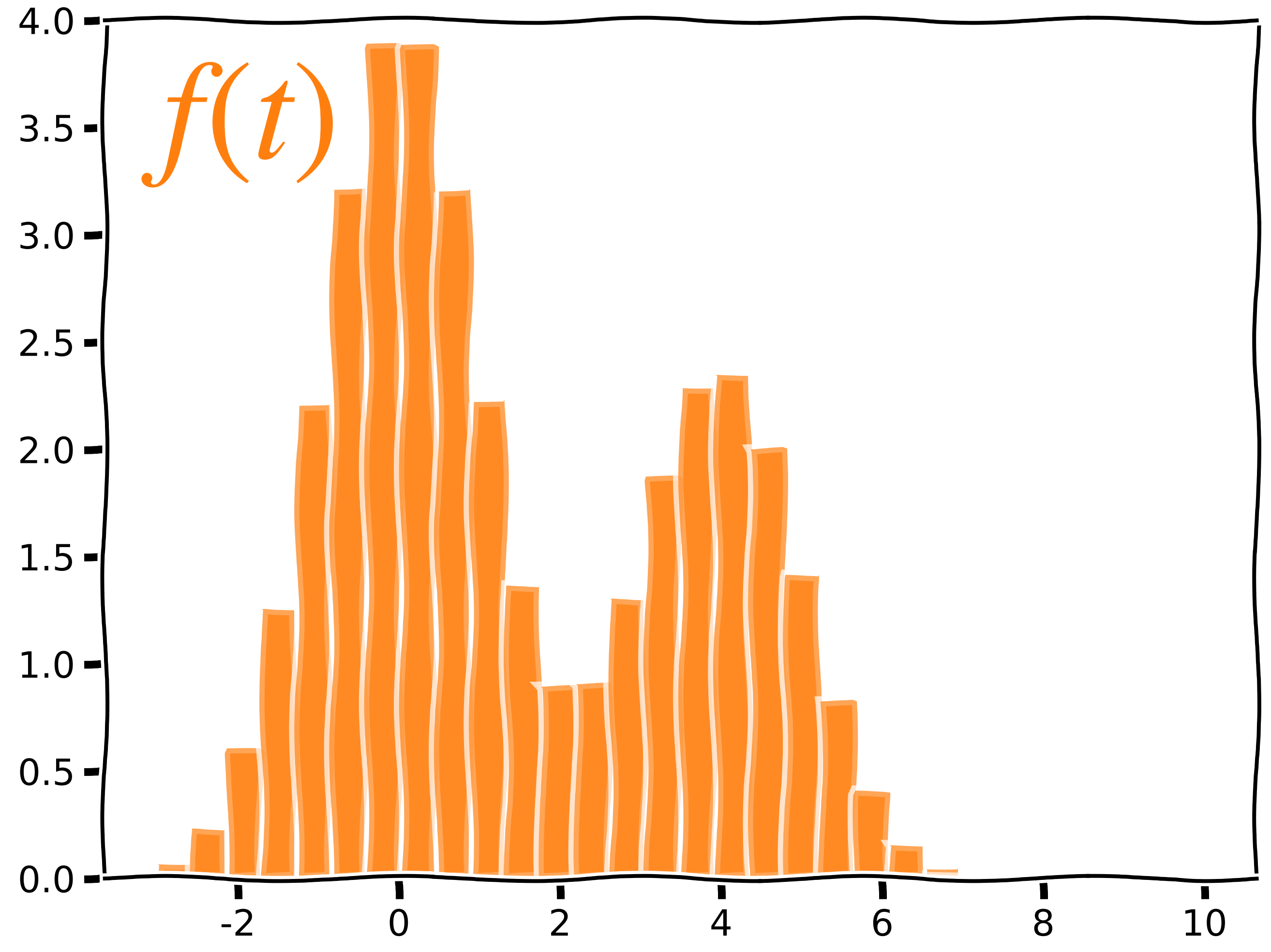
Fundamentals of unfolding

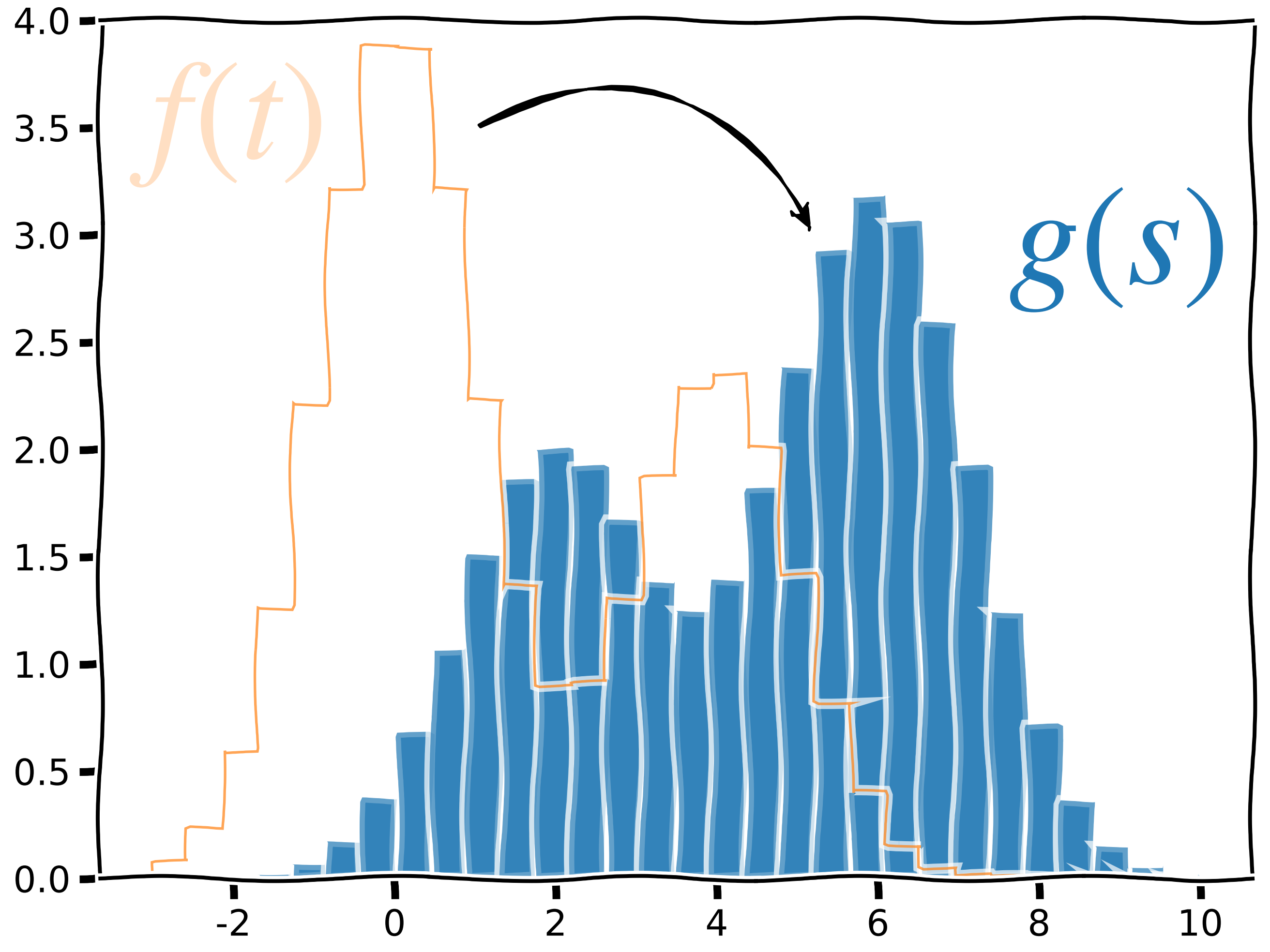
- Distributions $f(t)$ of a physics variable t to be measured in particle physics experiments are often not directly accessible.



Fundamentals of unfolding

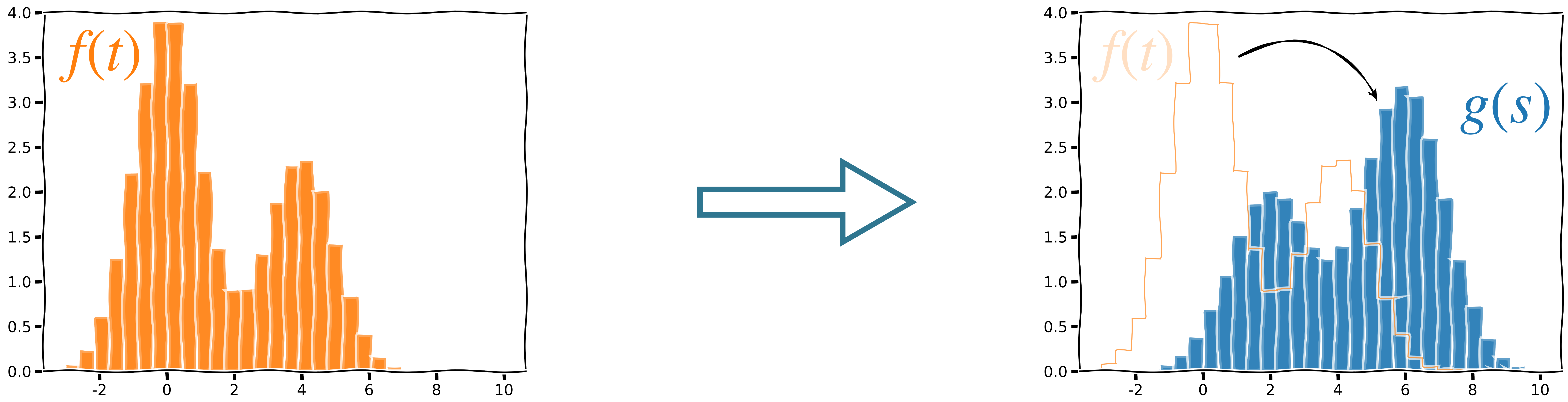
- Distributions $f(t)$ of a physics variable t to be measured in particle physics experiments are often not directly accessible.





Fundamentals of unfolding

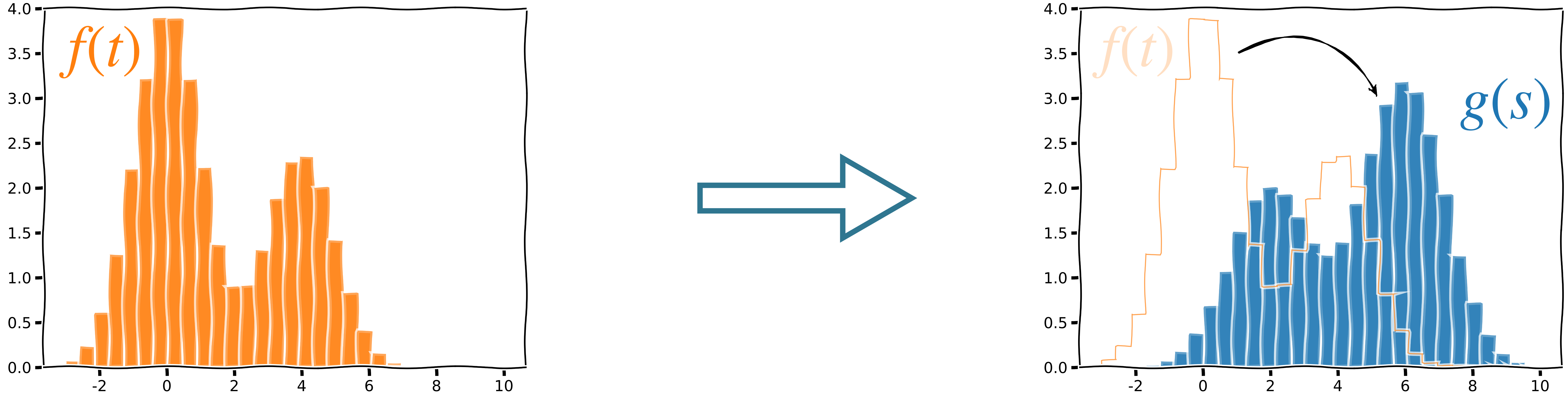
- Distributions $f(t)$ of a physics variable t to be measured in particle physics experiments are often not directly accessible.



- Using Monte Carlo (MC) methods, the direct process from an assumption $f(t)^{\text{model}}$ to the expected measured distribution $g(s)$ can be simulated.

Fundamentals of unfolding

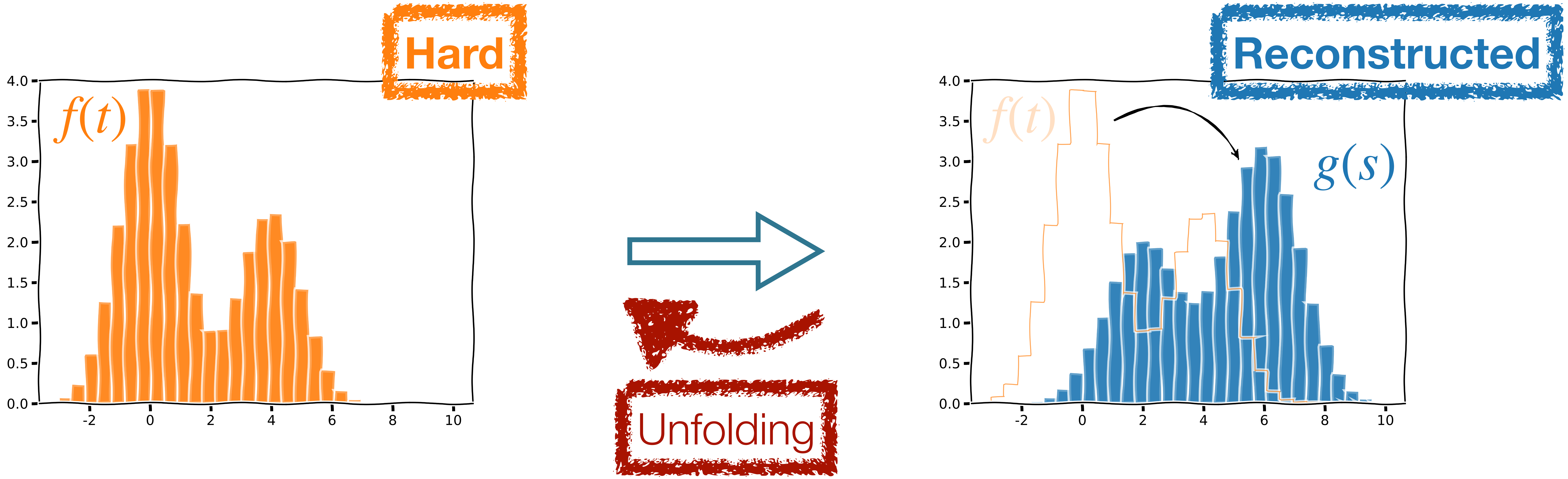
- Distributions $f(t)$ of a physics variable t to be measured in particle physics experiments are often not directly accessible.



- Using Monte Carlo (MC) methods, the direct process from an assumption $f(t)^{\text{model}}$ to the expected measured distribution $g(s)$ can be simulated.
- The inverse process from the actual measured distribution $g(s)$ to the true distribution $f(t)$ is difficult and ill-posed: small changes in $g(s)$ can cause large modifications in the reconstructed $\tilde{f}(t)$.

Fundamentals of unfolding

- Distributions $f(t)$ of a physics variable t to be measured in particle physics experiments are often not directly accessible.



- Using Monte Carlo (MC) methods, the direct process from an assumption $f(t)^{\text{model}}$ to the expected measured distribution $g(s)$ can be simulated.
- The inverse process from the actual measured distribution $g(s)$ to the true distribution $f(t)$ is difficult and ill-posed: small changes in $g(s)$ can cause large modifications in the reconstructed $\tilde{f}(t)$.

Why neural networks?

Traditionally:

- **Matrix-based** unfolding

$$g(s) = \int R(s\,|\,t)\; f(t)\; \mathrm{d}t$$

Why neural networks?

Traditionally:

- **Matrix-based** unfolding

$$g(s) = \int \underbrace{\boxed{R(s\,|\,t)}}_{\text{detector response matrix}} f(t)\ \mathrm{d}t$$

Why neural networks?

Traditionally:

- **Matrix-based** unfolding

$$g(s) = \int \underbrace{\boxed{R(s\,|\,t)}}_{\text{detector response matrix}} f(t)\ \mathrm{d}t \quad \xrightarrow{\text{binning}} \quad r_i = \sum_j R_{ij} \cdot t_j$$

Why neural networks?

Traditionally:

- **Matrix-based** unfolding

$$g(s) = \int \boxed{R(s\,|\,t)}\, f(t)\ \mathrm{d}t \xrightarrow{\text{binning}} r_i = \sum_j R_{ij} \cdot t_j$$

detector response matrix

- Various ways to invert the detector response matrix: SVD, IBU, IDS, etc.

Why neural networks?

Traditionally:

- **Matrix-based** unfolding

$$g(s) = \int \boxed{R(s\,|\,t)}\, f(t)\ \mathrm{d}t \xrightarrow{\text{binning}} r_i = \sum_j R_{ij} \cdot t_j$$

detector response matrix

  - Various ways to invert the detector response matrix: SVD, IBU, IDS, etc.
  - General need for regularization: trade-off between bias and statistical uncertainty

Why neural networks?

Traditionally:

- **Matrix-based** unfolding

$$g(s) = \int \underbrace{\boxed{R(s\,|\,t)}}_{\text{detector response matrix}} f(t)\ \mathrm{d}t \quad \xrightarrow[\text{binning}]{} \quad r_i = \sum_j R_{ij} \cdot t_j$$

  - Various ways to invert the detector response matrix: SVD, IBU, IDS, etc.
  - General need for regularization: trade-off between bias and statistical uncertainty
  - Requires binning and can only unfold a few dimensions

Why neural networks?

Traditionally:

- **Matrix-based** unfolding

$$g(s) = \int \underbrace{\boxed{R(s\,|\,t)}}_{\text{detector response matrix}} f(t)\ \mathrm{d}t \xrightarrow{\text{binning}} r_i = \sum_j R_{ij} \cdot t_j$$

  - Various ways to invert the detector response matrix: SVD, IBU, IDS, etc.
  - General need for regularization: trade-off between bias and statistical uncertainty
  - Requires binning and can only unfold a few dimensions

With neural networks:

- **ML-based** unfolding

Why neural networks?

Traditionally:

- **Matrix-based** unfolding

$$g(s) = \int \underbrace{\boxed{R(s\,|\,t)}}_{\text{detector response matrix}} f(t)\ \mathrm{d}t \xrightarrow[\text{binning}]{} r_i = \sum_j R_{ij} \cdot t_j$$

  - Various ways to invert the detector response matrix: SVD, IBU, IDS, etc.
  - General need for regularization: trade-off between bias and statistical uncertainty
  - Requires binning and can only unfold a few dimensions

With neural networks:

- **ML-based** unfolding
  - Unbinned: advantageous if one wants to derive quantities from the unfolding observables

Why neural networks?

Traditionally:

- **Matrix-based** unfolding

$$g(s) = \int \underbrace{\boxed{R(s\,|\,t)}}_{\text{detector response matrix}} f(t)\ \mathrm{d}t \quad \xrightarrow[\text{binning}]{} \quad r_i = \sum_j R_{ij} \cdot t_j$$

  - Various ways to invert the detector response matrix: SVD, IBU, IDS, etc.
  - General need for regularization: trade-off between bias and statistical uncertainty
  - Requires binning and can only unfold a few dimensions

With neural networks:

- **ML-based** unfolding
  - Unbinned: advantageous if one wants to derive quantities from the unfolding observables
  - Allows to unfold (and account for correlations in) many dimensions

Why neural networks?

Traditionally:

- **Matrix-based** unfolding

$$g(s) = \int \underbrace{\boxed{R(s\,|\,t)}}_{\text{detector response matrix}} f(t)\ \mathrm{d}t \quad \xrightarrow{\text{binning}} \quad r_i = \sum_j R_{ij} \cdot t_j$$

  - Various ways to invert the detector response matrix: SVD, IBU, IDS, etc.
  - General need for regularization: trade-off between bias and statistical uncertainty
  - Requires binning and can only unfold a few dimensions

With neural networks:

- **ML-based** unfolding
  - Unbinned: advantageous if one wants to derive quantities from the unfolding observables
  - Allows to unfold (and account for correlations in) many dimensions
  - Some methods *allow* for independent single-event unfolding

Several approaches

Event reweighting

- Omnifold [1911.09107]
- (*)

Distribution mapping

- Direct Diffusion [2311.17175]
- Schrödinger Bridge [2308.12351]
- (*)

Conditional phase space sampling

- GANs [1912.00477]
- Latent Diffusion [2305.10399]
- Conditional Flow Matching [2305.10475]
- cINN [2212.08674, 2006.06685]
- (*)

(*) These are not comprehensive lists. For a more extensive catalogue see for example the HEP ML Living Review

Several approaches

Event reweighting

- Omnifold [1911.09107]
- (*)

Distribution mapping

- Direct Diffusion [2311.17175]
- Schrödinger Bridge [2308.12351]
- (*)

Conditional phase space sampling

- GANs [1912.00477]
- Latent Diffusion [2305.10399]
- Conditional Flow Matching [2305.10475]
- cINN [2212.08674, 2006.06685]
- (*)

(*) These are not comprehensive lists. For a more extensive catalogue see for example the HEP ML Living Review

Several approaches

Event reweighting

- Omnifold [1911.09107]
- (*)

Distribution mapping

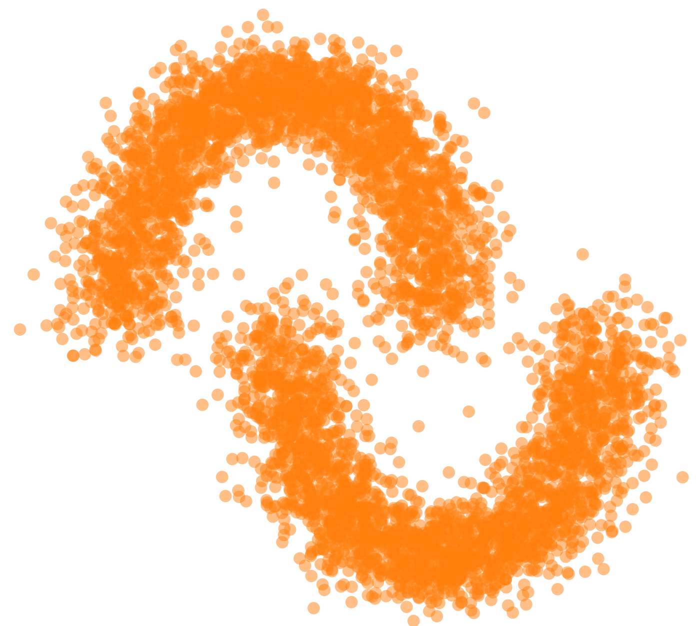
- Direct Diffusion [2311.17175]
- Schrödinger Bridge [2308.12351]
- (*)

Conditional phase space sampling

- GANs [1912.00477]
- Latent Diffusion [2305.10399]
- Conditional Flow Matching [2305.10475]
- cINN [2212.08674, 2006.06685]
- (*)

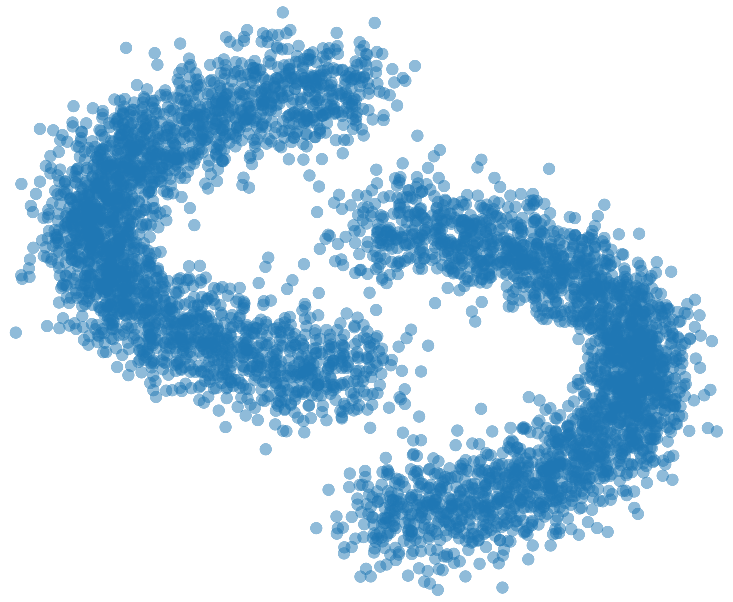
(*) These are not comprehensive lists. For a more extensive catalogue see for example the HEP ML Living Review

Direct Diffusion (DiDi)



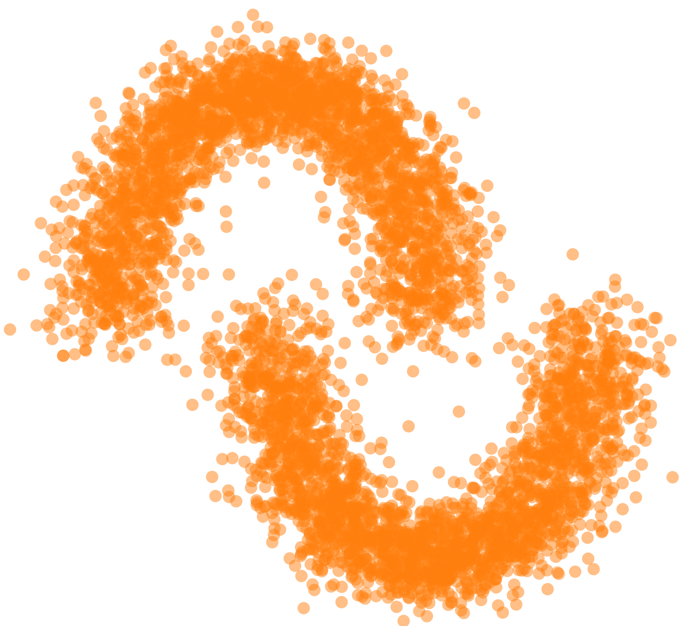
$x_0 \sim p_{\text{model}}(x_{\text{hard}})$

$$\frac{\mathrm{d}x(t)}{\mathrm{d}t} = v_\theta(x(t), t)$$

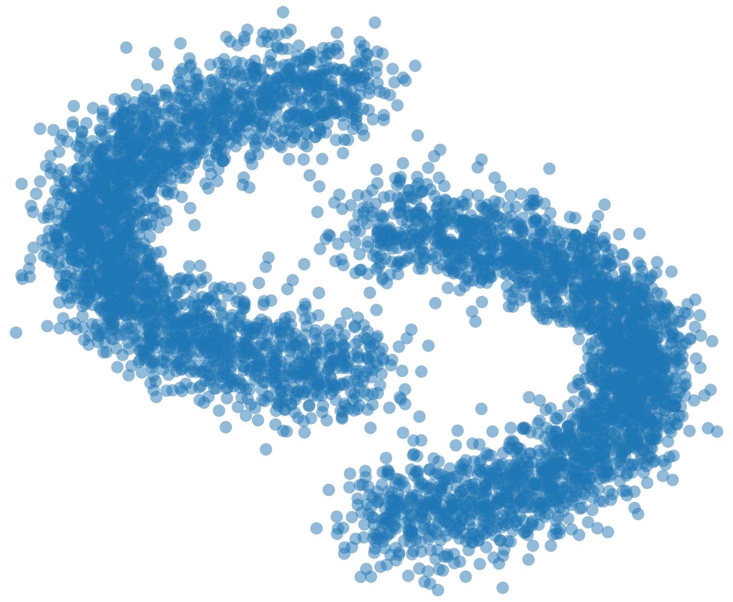


$x_1 \sim p_{\text{reco}}(x_{\text{reco}})$

Direct Diffusion (DiDi)



$$\frac{\mathrm{d}x(t)}{\mathrm{d}t} = v_\theta(x(t), t)$$

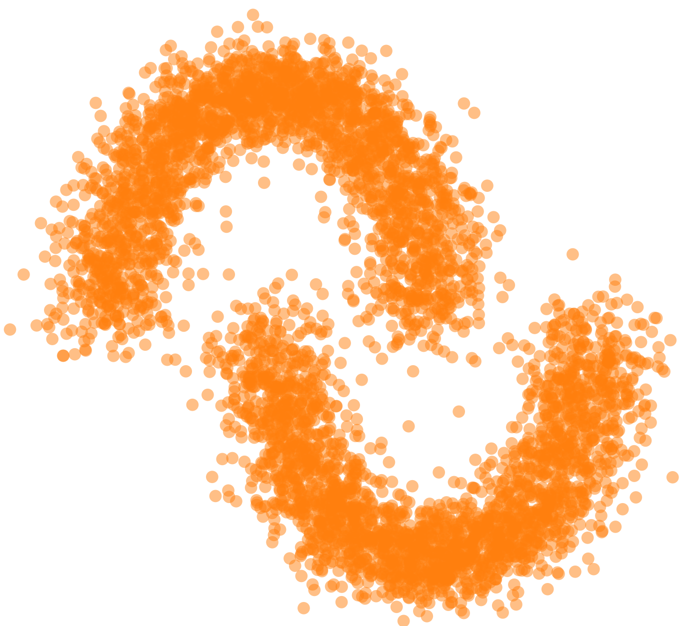


$x_0 \sim p_{\mathrm{model}}(x_{\mathrm{hard}})$

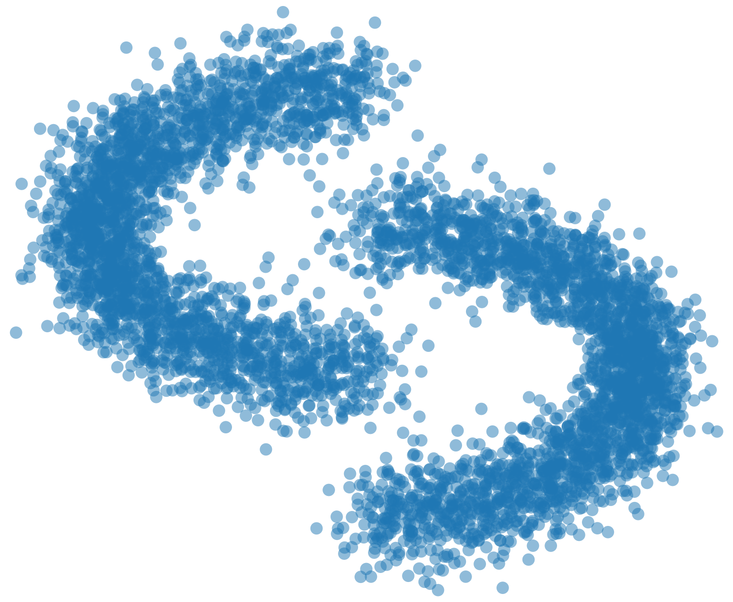
$x_1 \sim p_{\mathrm{reco}}(x_{\mathrm{reco}})$

- Connect x_0 and x_1 with a linear trajectory: $x(t) = (1 - t)x_0 + tx_1$

Direct Diffusion (DiDi)



$$\frac{\mathrm{d}x(t)}{\mathrm{d}t} = v_\theta(x(t), t)$$

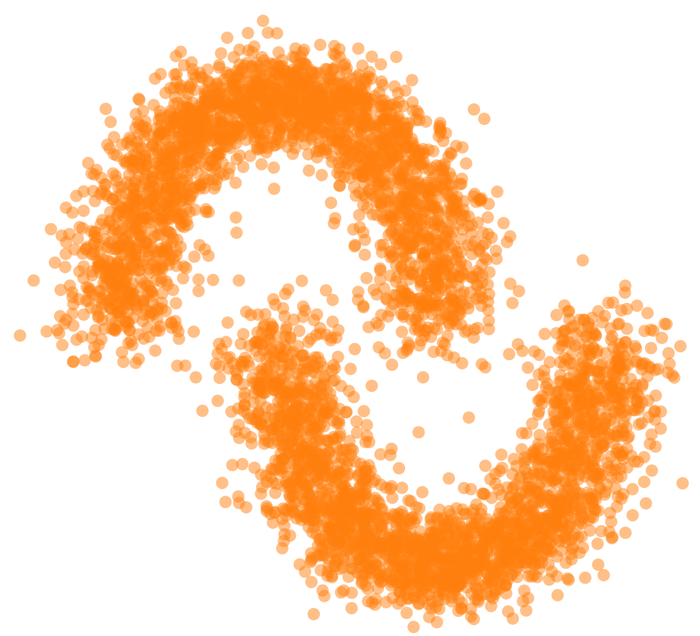


$x_0 \sim p_{\mathrm{model}}(x_{\mathrm{hard}})$

$x_1 \sim p_{\mathrm{reco}}(x_{\mathrm{reco}})$

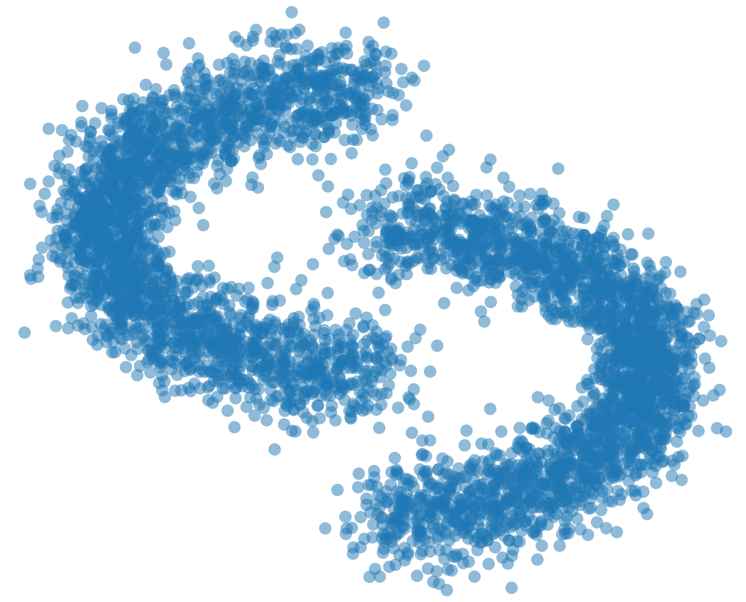
- Connect x_0 and x_1 with a linear trajectory: $x(t) = (1 - t)x_0 + tx_1$
- The NN is regressed to predict the velocity field: $v_\theta(x(t), t) \approx \frac{\mathrm{dx(t)}}{\mathrm{dt}} = x_1 - x_0$

Direct Diffusion (DiDi)



$x_0 \sim p_{\text{model}}(x_{\text{hard}})$

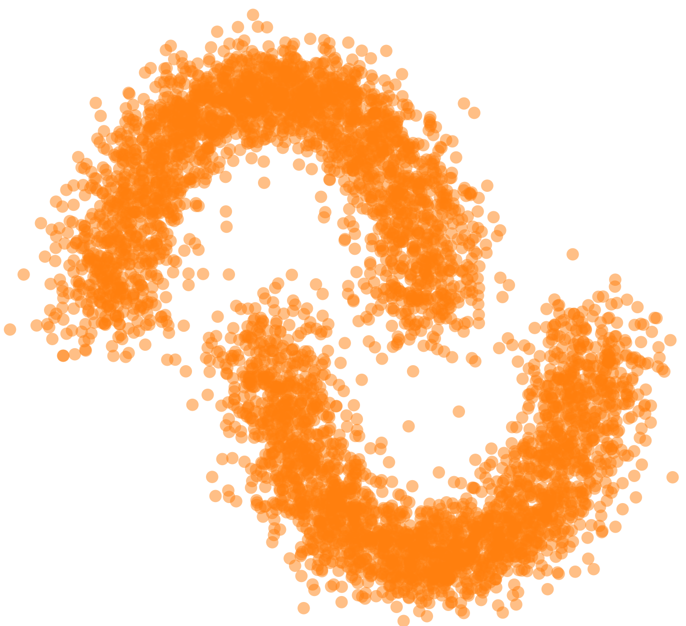
$$\frac{\mathrm{d}x(t)}{\mathrm{d}t} = v_\theta(x(t), t)$$



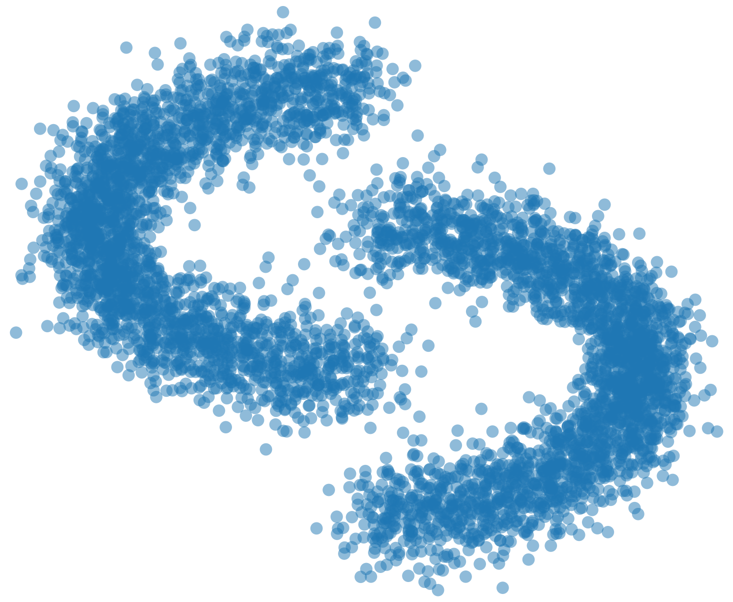
$x_1 \sim p_{\text{reco}}(x_{\text{reco}})$

- Connect x_0 and x_1 with a linear trajectory: $x(t) = (1 - t)x_0 + tx_1$
- The NN is regressed to predict the velocity field: $v_\theta(x(t), t) \approx \frac{\mathrm{dx(t)}}{\mathrm{dt}} = x_1 - x_0$
- For sampling, solve ODE starting from x_1: $x_0 = x_1 + \int_1^0 v_\theta(x(t), t)\mathrm{dt}$

Direct Diffusion (DiDi)



$$\frac{\mathrm{d}x(t)}{\mathrm{d}t} = v_\theta(x(t), t)$$



$x_0 \sim p_{\mathrm{model}}(x_{\mathrm{hard}})$

$x_1 \sim p_{\mathrm{reco}}(x_{\mathrm{reco}})$

- Connect x_0 and x_1 with a linear trajectory: $x(t) = (1-t)x_0 + tx_1$
- The NN is regressed to predict the velocity field: $v_\theta(x(t), t) \approx \frac{\mathrm{dx(t)}}{\mathrm{dt}} = x_1 - x_0$
- For sampling, solve ODE starting from x_1: $x_0 = x_1 + \int_1^0 v_\theta(x(t), t)\mathrm{dt}$
- Loss: $\mathscr{L}_{\mathrm{DiDi}} = \left\langle [v_\theta((1-t)x_0 + tx_1,\ t) - (x_1 - x_0)]^2 \right\rangle_{t\sim\mathscr{U}([0,1]),\ (x_0,x_1)\sim p(x_{\mathrm{hard}},x_{\mathrm{reco}})}$

Several methods

Event reweighting

- Omnifold [1911.09107]
- (*)

Distribution mapping

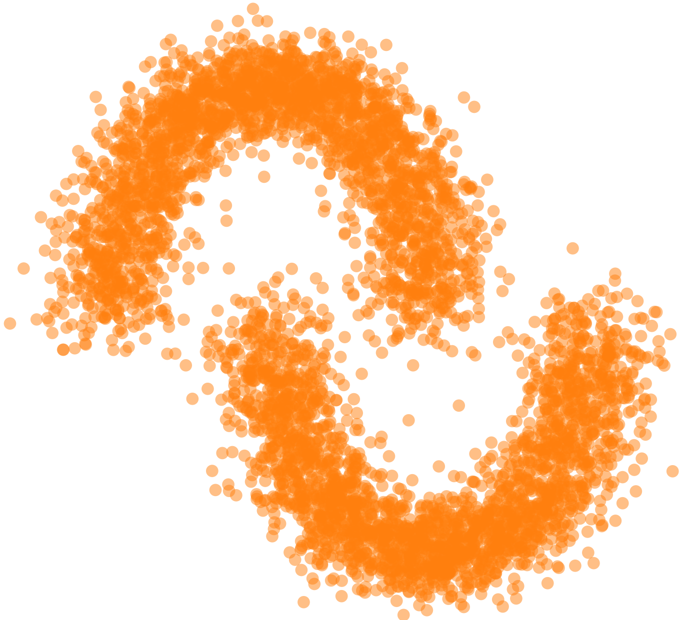
- Direct Diffusion [2311.17175]
- Schrödinger Bridge [2308.12351]
- (*)

Conditional phase space sampling

- GANs [1912.00477]
- Latent Diffusion [2305.10399]
- Conditional Flow Matching [2305.10475]
- cINN [2212.08674, 2006.06685]
- (*)

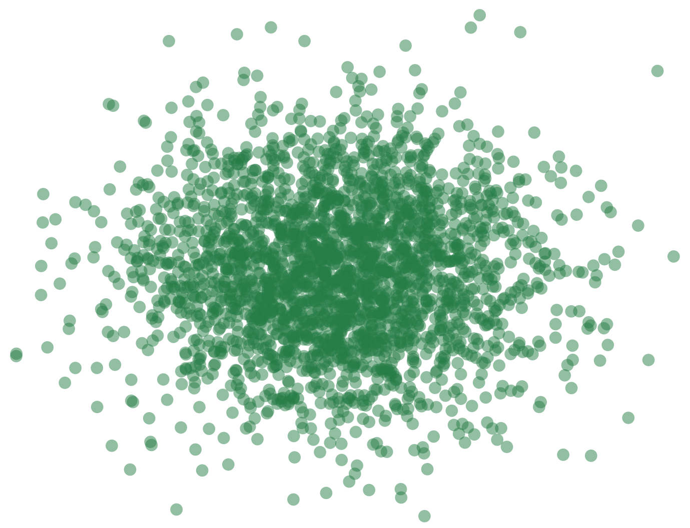
(*) These are not comprehensive lists. For a more extensive catalogue see for example the HEP ML Living Review

Conditional Flow Matching (CFM)



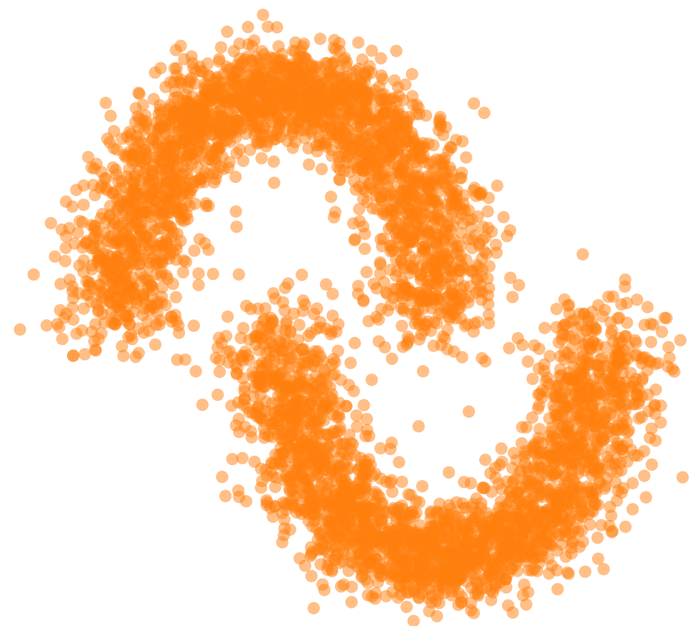
$x_0 \sim p_{\text{model}}(x_{\text{hard}} \mid x_{\text{reco}})$

$$\frac{\mathrm{d}x(t)}{\mathrm{d}t} = v_\theta(x(t), t \mid x_{\text{reco}})$$



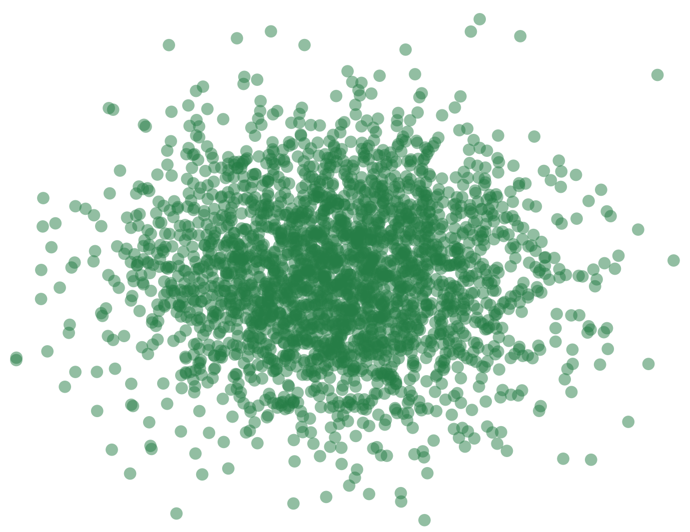
$\epsilon = z \sim p_{\text{latent}}(z)$

Conditional Flow Matching (CFM)



$$x_0 \sim p_{\text{model}}(x_{\text{hard}} \,|\, x_{\text{reco}})$$

$$\frac{\mathrm{d}x(t)}{\mathrm{d}t} = v_\theta(x(t), t \,|\, x_{\text{reco}})$$



$$\epsilon = z \sim p_{\text{latent}}(z)$$

- Connect x_0 and ϵ with a linear trajectory: $x(t) = (1-t)x_0 + t\epsilon$
- The NN is regressed to predict the velocity field: $v_\theta(x(t), t \,|\, x_{\text{reco}}) \approx \frac{\mathrm{dx(t)}}{\mathrm{dt}} = \epsilon - x_0$
- For sampling, solve ODE starting from ϵ: $x_0 = \epsilon + \int_1^0 v_\theta(x(t), t \,|\, x_{\text{reco}})\mathrm{dt}$
- Loss: $\mathscr{L}_{\text{CFM}} = \left\langle [v_\theta((1-t)x_0 + t\epsilon,\ t,\ x_{\text{reco}}) - (\epsilon - x_0)]^2 \right\rangle_{t\sim\mathscr{U}([0,1]),\ (x_0, x_{\text{reco}})\sim p(x_{\text{hard}}, x_{\text{reco}}),\ \epsilon\sim\mathscr{N}(0,1)}$

Several methods

Event reweighting

- Omnifold [1911.09107]
- (*)

Distribution mapping

- Direct Diffusion [2311.17175]
- Schrödinger Bridge [2308.12351]
- (*)

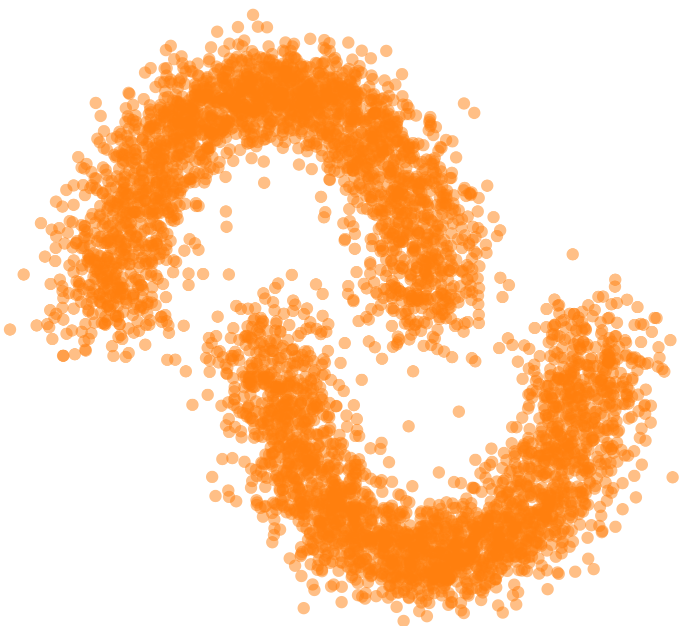
Conditional phase space sampling

- GANs [1912.00477]
- Latent Diffusion [2305.10399]
- Conditional Flow Matching [2305.10475]
- cINN [2212.08674, 2006.06685]
- (*)

(*) These are not comprehensive lists. For a more extensive catalogue see for example the HEP ML Living Review

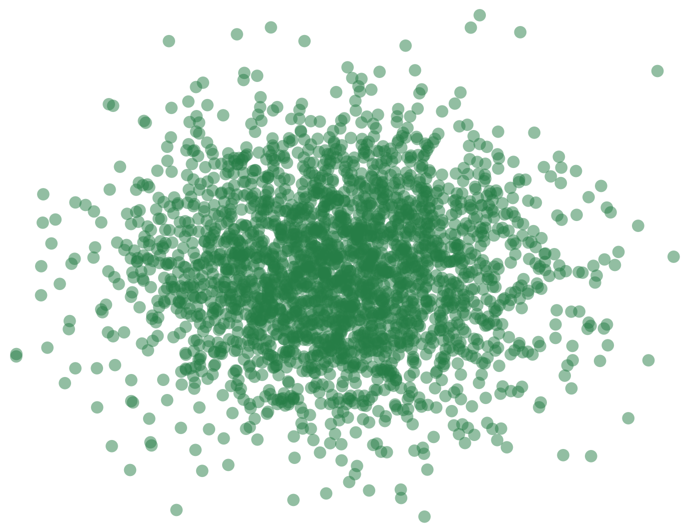
Conditional INN (cINN)

$$g_\theta(x_{\text{hard}} \,|\, x_{\text{reco}})$$



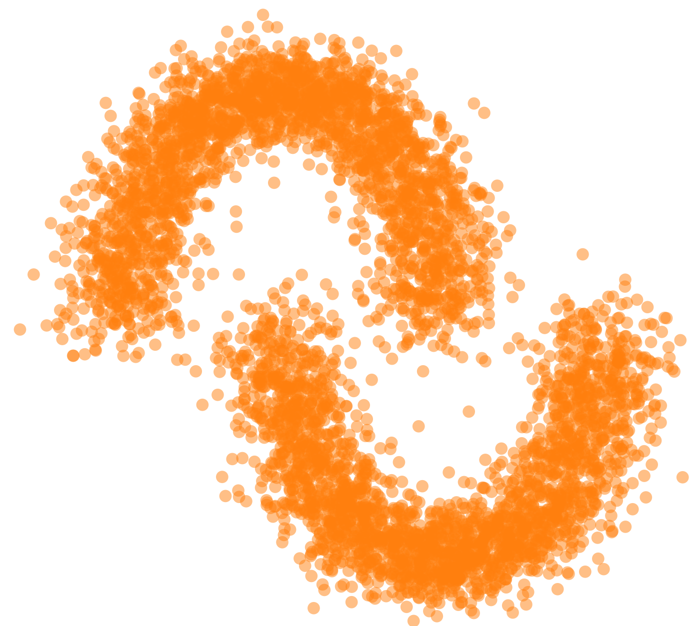
$$x_0 \sim p_{\text{model}}(x_{\text{hard}} \,|\, x_{\text{reco}})$$

$$g_\theta^{-1}(z \,|\, x_{\text{reco}})$$



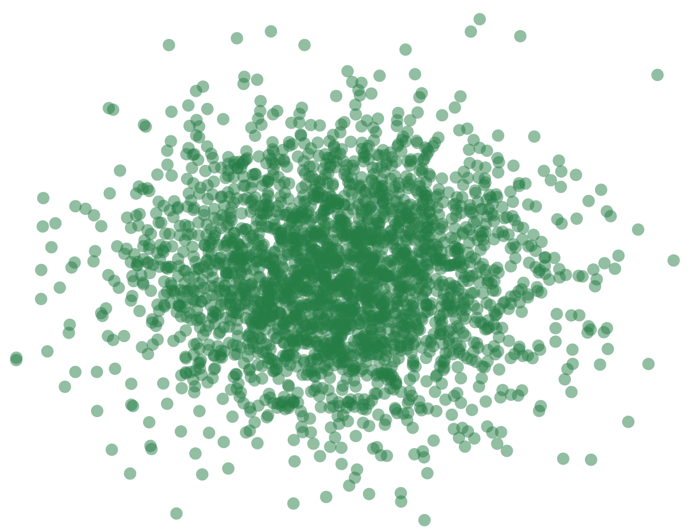
$$z \sim p_{\text{latent}}(z)$$

Conditional INN (cINN)



$$g_\theta(x_{\text{hard}}\,|\,x_{\text{reco}})$$

$$g_\theta^{-1}(z\,|\,x_{\text{reco}})$$

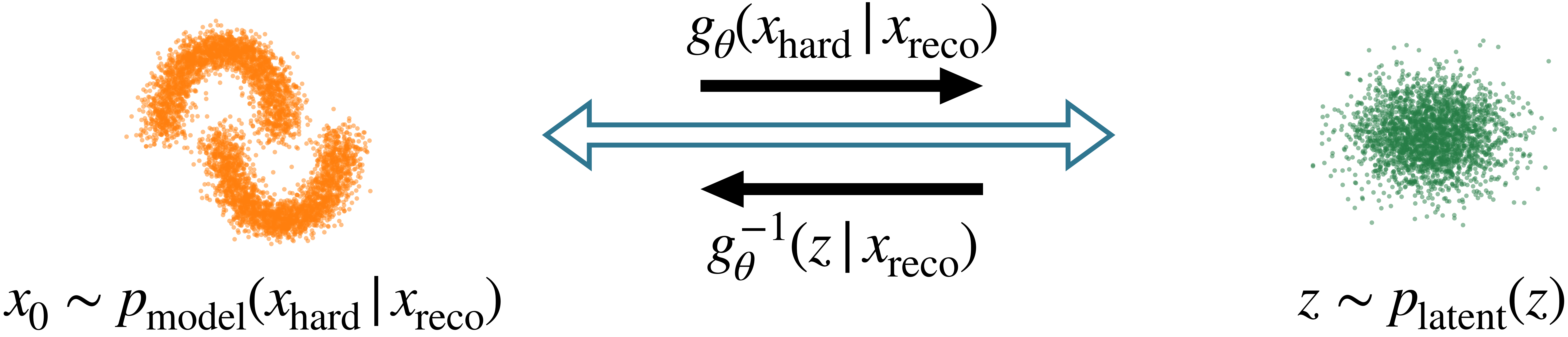


$x_0 \sim p_{\text{model}}(x_{\text{hard}}\,|\,x_{\text{reco}})$ $\qquad$ $z \sim p_{\text{latent}}(z)$

- Bijective function between $p_{\text{latent}}(z)$ and $p_{\text{model}}(x_{\text{hard}}\,|\,x_{\text{reco}})$:

$$p_{\text{model}}(x_{\text{hard}}\,|\,x_{\text{reco}}) = p_{\text{latent}}(z)\left|\det\frac{\partial g_\theta(x_{\text{h}},x_{\text{r}})}{\partial x_{\text{hard}}}\right| = p_{\text{lat.}}(z)\;\left|\det J_{g_\theta}\right|$$

Conditional INN (cINN)

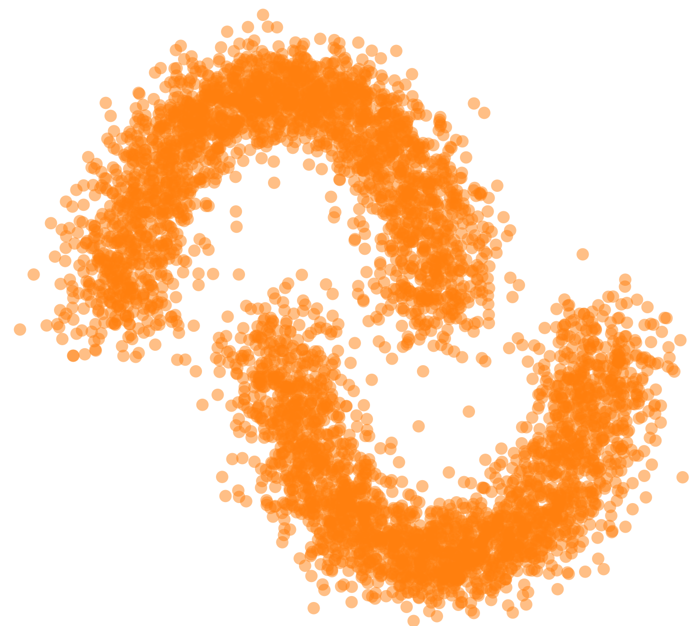


- Bijective function between $p_{\text{latent}}(z)$ and $p_{\text{model}}(x_{\text{hard}} \,|\, x_{\text{reco}})$:

$$p_{\text{model}}(x_{\text{hard}} \,|\, x_{\text{reco}}) = p_{\text{latent}}(z) \left| \det \frac{\partial g_\theta(x_{\text{h}}, x_{\text{r}})}{\partial x_{\text{hard}}} \right| = p_{\text{lat.}}(z) \quad \left| \det J_{g_\theta} \right|$$

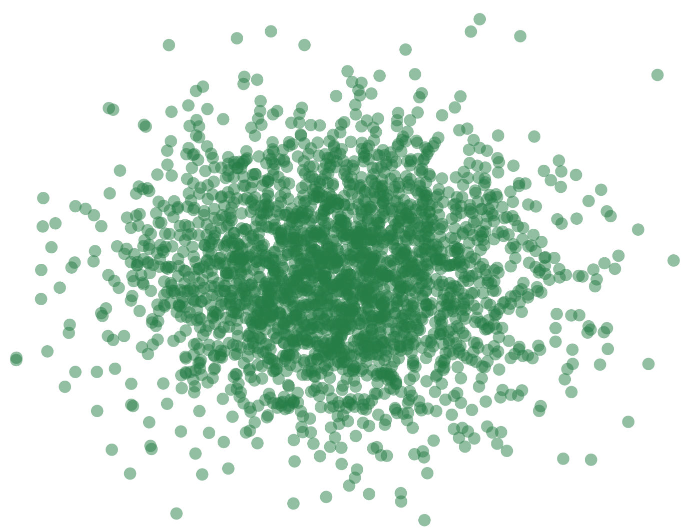
- Pairs $(x_{\text{hard}},\, x_{\text{reco}})$ are passed through the NN to the latent space: $\quad z = g_\theta(x_{\text{hard}} \,|\, x_{\text{reco}})$

Conditional INN (cINN)



$$g_\theta(x_{\text{hard}} \,|\, x_{\text{reco}})$$

$$g_\theta^{-1}(z \,|\, x_{\text{reco}})$$



$x_0 \sim p_{\text{model}}(x_{\text{hard}} \,|\, x_{\text{reco}})$ $\qquad$ $z \sim p_{\text{latent}}(z)$

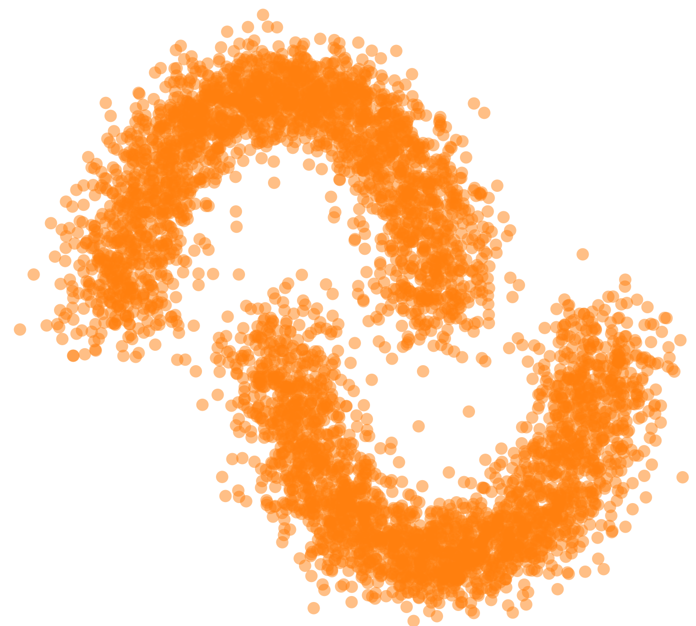
- Bijective function between $p_{\text{latent}}(z)$ and $p_{\text{model}}(x_{\text{hard}} \,|\, x_{\text{reco}})$:

$$p_{\text{model}}(x_{\text{hard}} \,|\, x_{\text{reco}}) = p_{\text{latent}}(z) \left| \det \frac{\partial g_\theta(x_{\text{h}}, x_{\text{r}})}{\partial x_{\text{hard}}} \right| = p_{\text{lat.}}(z) \;\; \left| \det J_{g_\theta} \right|$$

- Pairs $(x_{\text{hard}},\, x_{\text{reco}})$ are passed through the NN to the latent space: $\quad z = g_\theta(x_{\text{hard}} \,|\, x_{\text{reco}})$

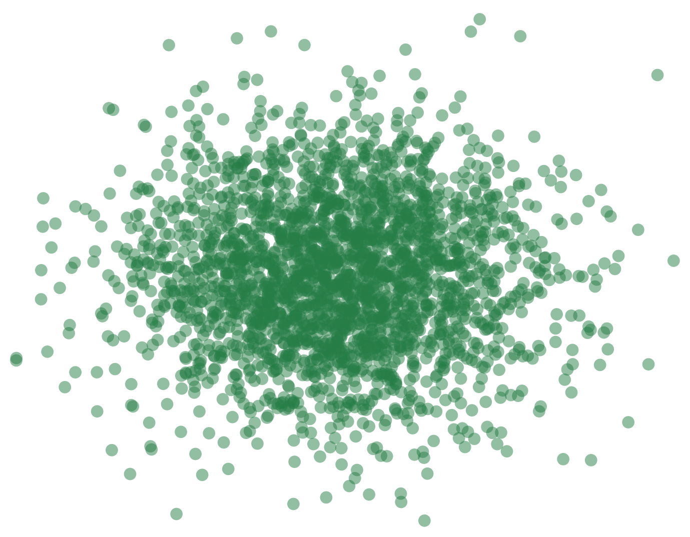
- Once trained, one can sample -conditioned on reco- from the latent: $\quad p_{\text{hard}}(x) \approx p_{\text{model}}(x_{\text{hard}} \,|\, x_{\text{reco}})$

Conditional INN (cINN)



$$g_\theta(x_{\text{hard}}\,|\,x_{\text{reco}})$$

$$g_\theta^{-1}(z\,|\,x_{\text{reco}})$$



$x_0 \sim p_{\text{model}}(x_{\text{hard}}\,|\,x_{\text{reco}})$ $\qquad$ $z \sim p_{\text{latent}}(z)$

- Bijective function between $p_{\text{latent}}(z)$ and $p_{\text{model}}(x_{\text{hard}}\,|\,x_{\text{reco}})$:

$$p_{\text{model}}(x_{\text{hard}}\,|\,x_{\text{reco}}) = p_{\text{latent}}(z)\left|\det \frac{\partial g_\theta(x_{\text{h}}, x_{\text{r}})}{\partial x_{\text{hard}}}\right| = p_{\text{lat.}}(z)\ \left|\det J_{g_\theta}\right|$$

- Pairs $(x_{\text{hard}},\, x_{\text{reco}})$ are passed through the NN to the latent space: $\quad z = g_\theta(x_{\text{hard}}\,|\,x_{\text{reco}})$
- Once trained, one can sample -conditioned on reco- from the latent: $\quad p_{\text{hard}}(x) \approx p_{\text{model}}(x_{\text{hard}}\,|\,x_{\text{reco}})$
- Loss: $\quad \mathscr{L}_{\text{cINN}} = -\langle \log p_{\text{model}}(x_{\text{hard}}\,|\,x_{\text{reco}})\rangle_{(x_0,x_1)\sim p(x_{\text{hard}},x_{\text{reco}})}$

Z + jets events

$Z(p_T > 200 \text{ GeV})$ + jets events generated at $\sqrt{s} = 14 \text{ TeV}$ with Pythia 8.244 and Delphes simulation 3.5.0 available on Zenodo. Slight modification from [1911.09107] dataset

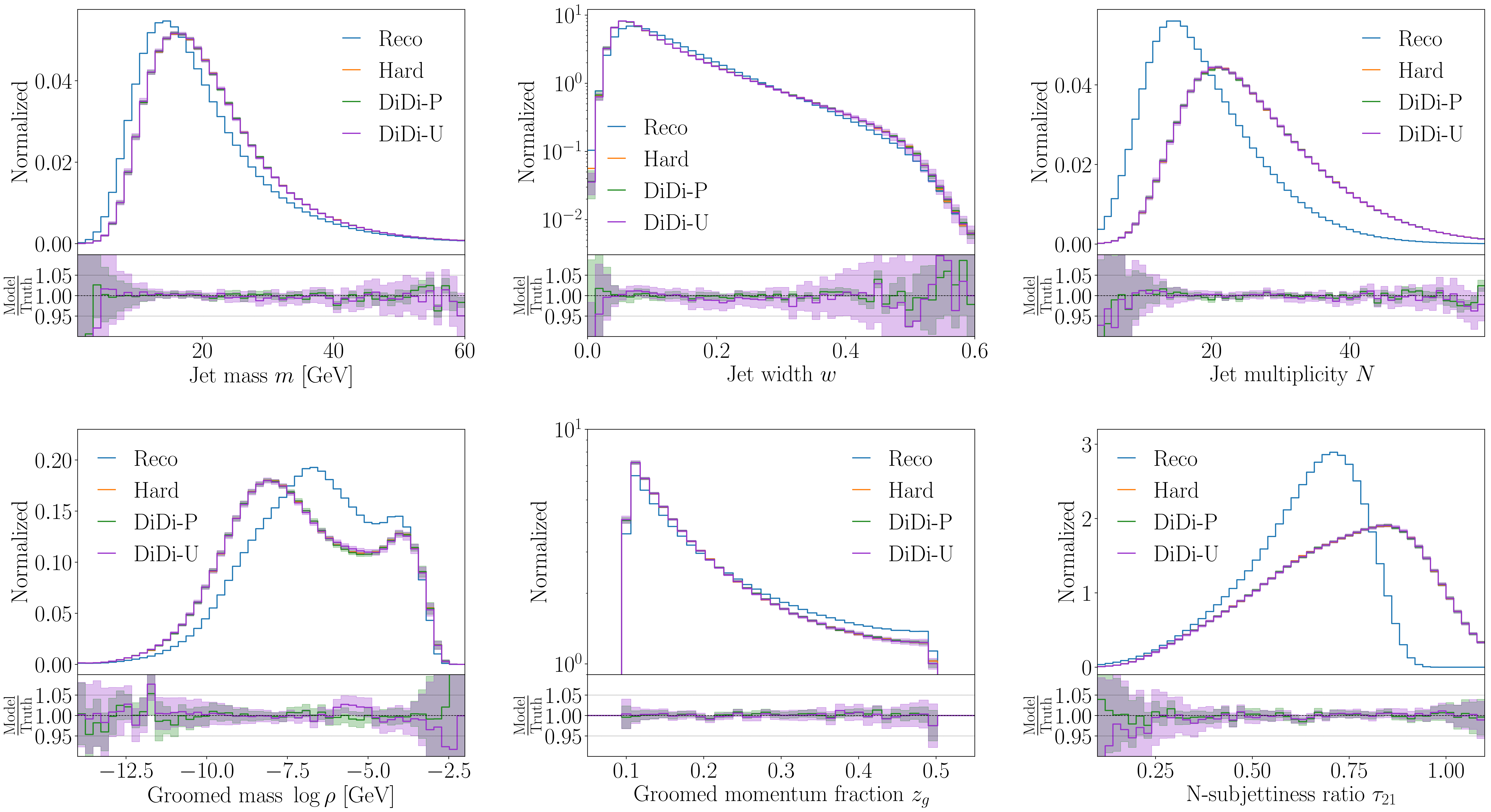
Z + jets events

$Z(p_T > 200\ \mathrm{GeV})$ + jets events generated at $\sqrt{s} = 14\ \mathrm{TeV}$ with Pythia 8.244 and Delphes simulation 3.5.0 available on Zenodo. Slight modification from [1911.09107] dataset

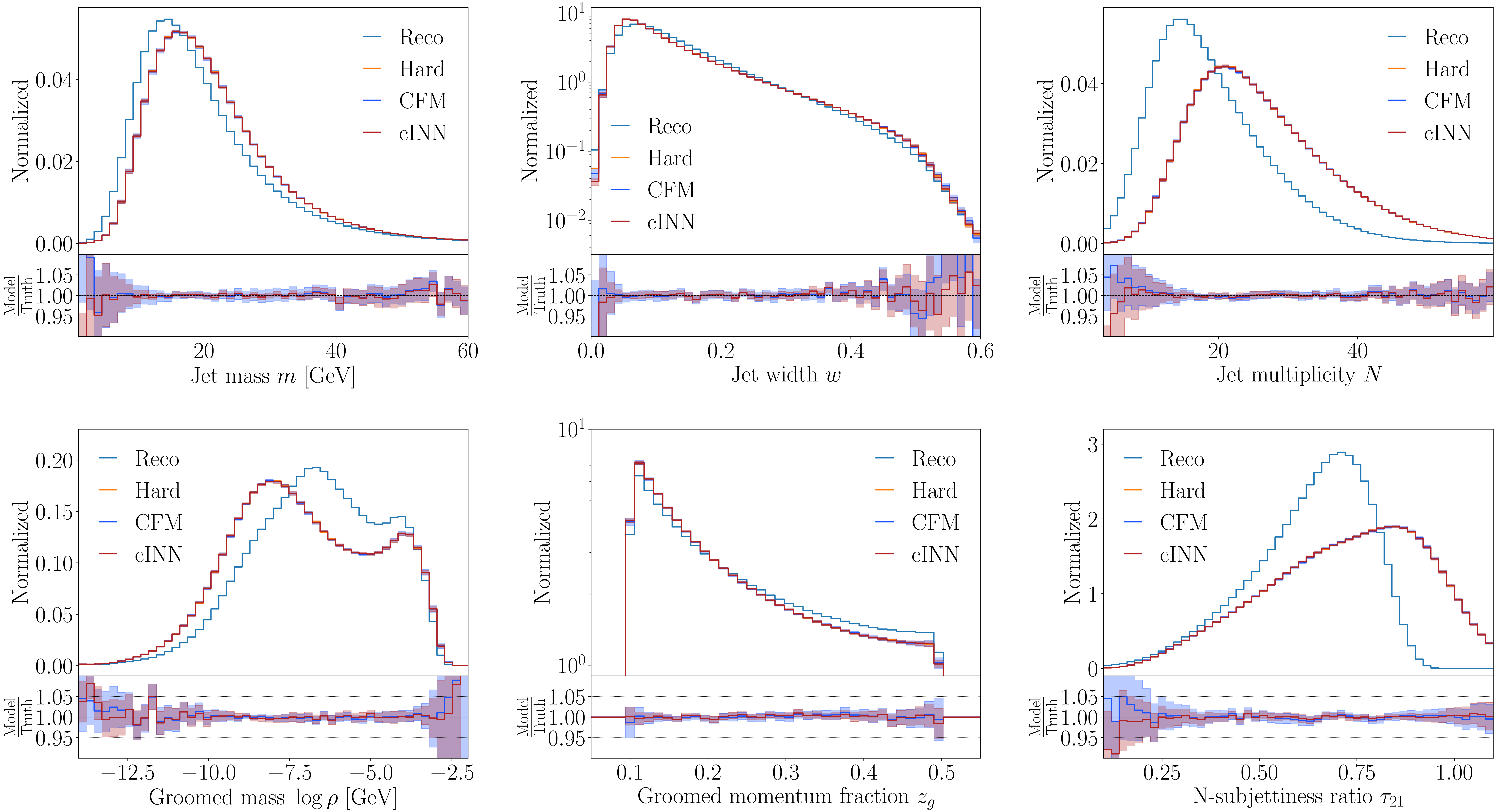
Six widely-used jet substructure observables:

- Jet mass m
- Jet width w
- Jet constituents multiplicity N
- Groomed mass $\log \rho = 2 \log\ (m_{\mathrm{SD}}\ /\ p_T)$
- Groomed momentum fraction $z_g = \tau_1^{\beta=1}$
- N-subjettiness ratio $\tau_{21} = \tau_2^{\beta=1}/\tau_1^{\beta=1}$

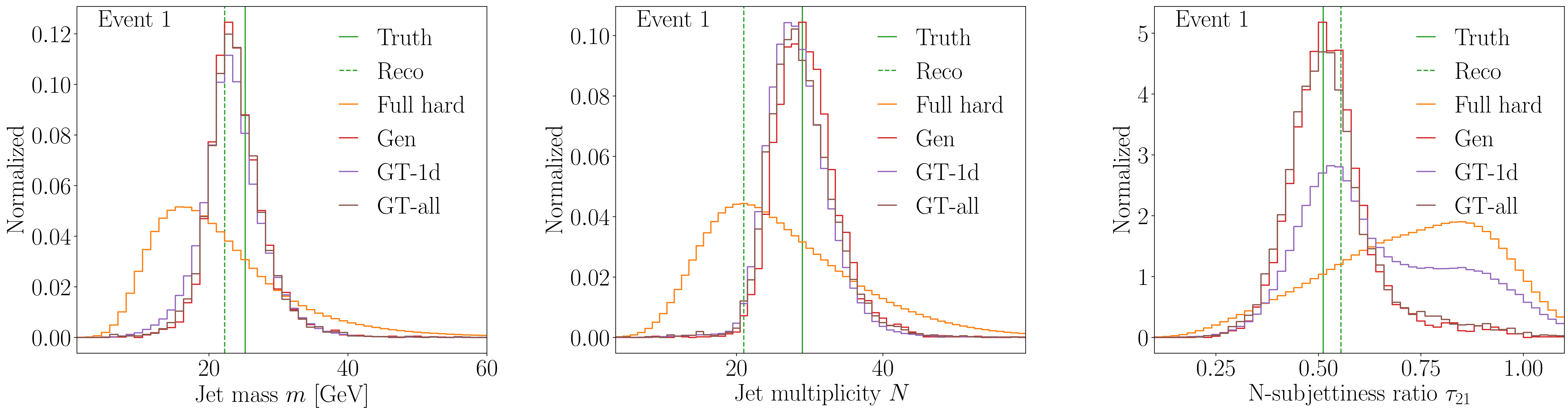
Results (DiDi)



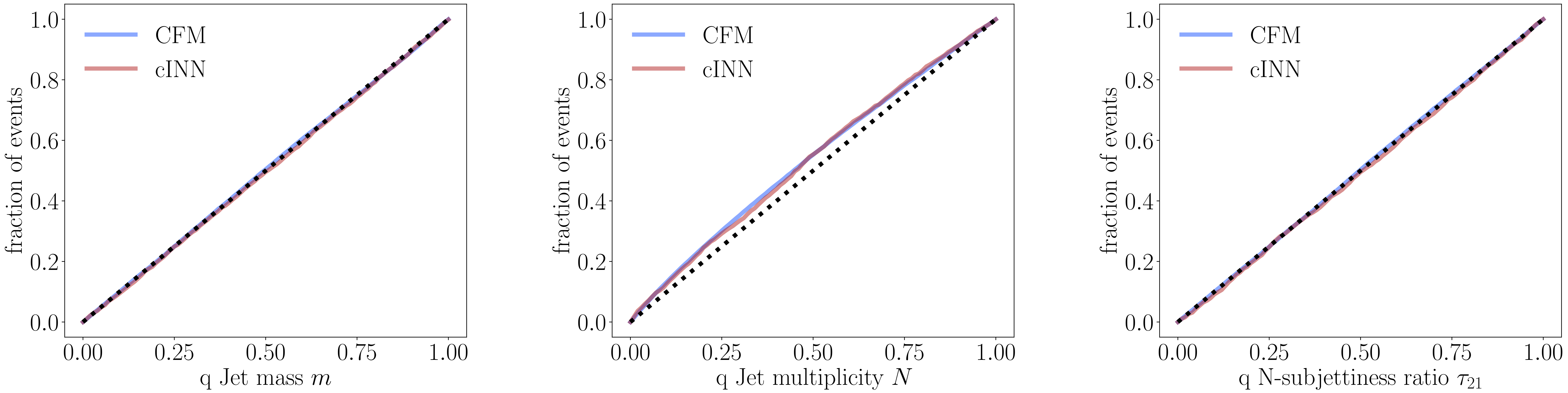
Results (CFM & cINN)



Single event unfolding



Calibration



Top-pair events: unfolding to parton-level

Matrix elements are evaluated at $\sqrt{s} = 13\ \mathrm{TeV}$ using MadGraph_aMC@NLO. Showering and hadronization are simulated with Pythia8, and detector response is simulated with Delphes with the standard CMS card. For a detailed description see [2305.10399].

Top-pair events: unfolding to parton-level

Matrix elements are evaluated at $\sqrt{s} = 13\ \mathrm{TeV}$ using MadGraph_aMC@NLO. Showering and hadronization are simulated with Pythia8, and detector response is simulated with Delphes with the standard CMS card. For a detailed description see [2305.10399].

Unfolding from 6 final-state particles $(bl\nu)(bqq)$:

- 4 DoFs for the lepton
- 3 DoFs for the missing p_T^ν
- 5 DoFs per jet (4-momentum + b-tag)

Top-pair events: unfolding to parton-level

Matrix elements are evaluated at $\sqrt{s} = 13\ \mathrm{TeV}$ using MadGraph_aMC@NLO. Showering and hadronization are simulated with Pythia8, and detector response is simulated with Delphes with the standard CMS card. For a detailed description see [2305.10399].

Unfolding from 6 final-state particles $(bl\nu)(bqq)$:

- 4 DoFs for the lepton
- 3 DoFs for the missing p_T^ν
- 5 DoFs per jet (4-momentum + b-tag)

Total: 27 DoFs at reco-level and 19 DoFs at parton-level

Top-pair events: unfolding to parton-level

Much harder problem:

- Unfolding to parton-level is not only inverting detector effects, but rather inverting the entire forward simulation chain

Top-pair events: unfolding to parton-level

Much harder problem:

- Unfolding to parton-level is not only inverting detector effects, but rather inverting the entire forward simulation chain
- Faithful modeling of complex correlations at parton-level, i.e., W boson and top mass resonances

Top-pair events: unfolding to parton-level

Much harder problem:

- Unfolding to parton-level is not only inverting detector effects, but rather inverting the entire forward simulation chain
- Faithful modeling of complex correlations at parton-level, i.e., W boson and top mass resonances
- Non-trivial combinatorics between physics objects at both levels

Top-pair events: unfolding to parton-level

Much harder problem:

- Unfolding to parton-level is not only inverting detector effects, but rather inverting the entire forward simulation chain
- Faithful modeling of complex correlations at parton-level, i.e., W boson and top mass resonances
- Non-trivial combinatorics between physics objects at both levels

Adding transformers:

Top-pair events: unfolding to parton-level

Much harder problem:

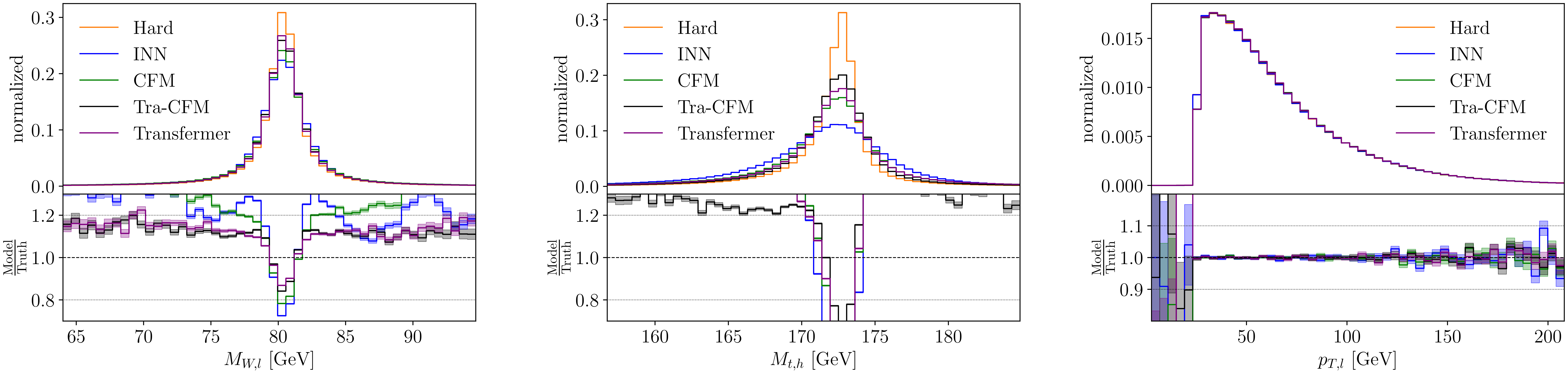
- Unfolding to parton-level is not only inverting detector effects, but rather inverting the entire forward simulation chain
- Faithful modeling of complex correlations at parton-level, i.e., W boson and top mass resonances
- Non-trivial combinatorics between physics objects at both levels

Adding transformers:

- Transfermer and Tra-CFM as an extension to the cINN and CFM [2310.07752]. A transformer is employed to encode correlations at reco and parton-level.

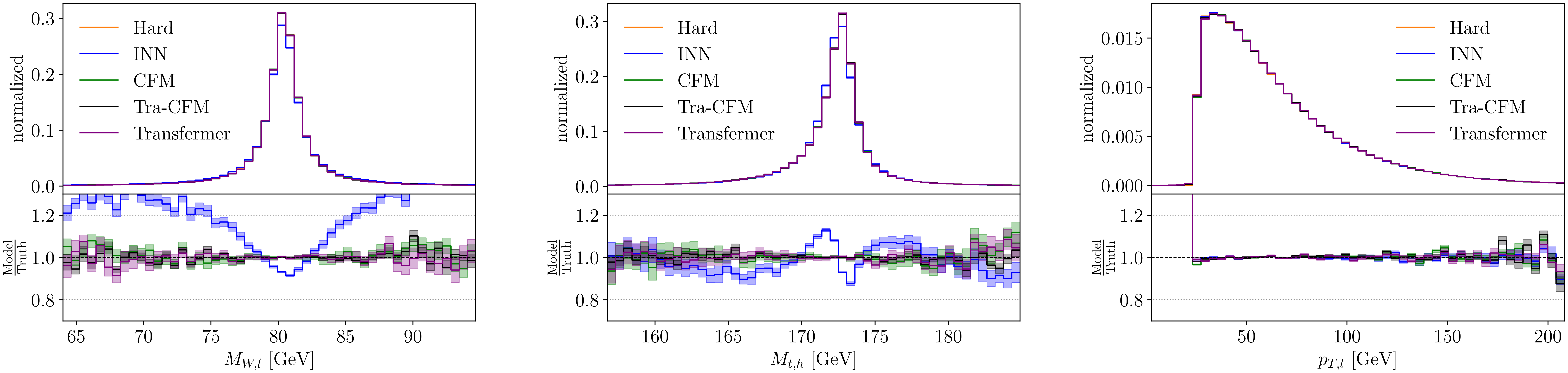
Results: naive parametrization

- Unfold: $(p_{T,b_l}, \eta_{b_l}, \phi_{b_l}, p_{T,l}, \eta_l, \phi_l, p_{T,\nu}, \eta_\nu, \phi_\nu, p_{T,b_h}, \eta_{b_h}, \phi_{b_h}, m_{q_1}, p_{T,q_1}, \eta_{q_1}, \phi_{q_1}, p_{T,q_2}, \eta_{q_2}, \phi_{q_2})$



Results: mass parametrization

- Unfold: $(m_t, p_{T,t}^L, \eta_t^L, \phi_t^L, m_W, \eta_W^T, \phi_W^T, (m_{d_1}^W), \eta_{d_1}^W, \phi_{d_1}^W)$



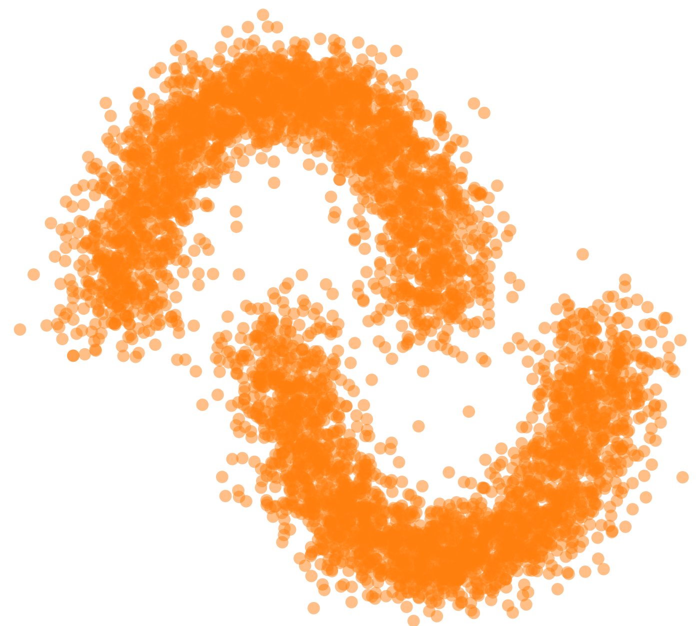
Summary and outlook

- ML-based unfolding is an unbinned transformative analysis tool capable of dealing with correlations accross many dimensions.
- Distribution mapping can be trained on matched and unmatched data and is relatively fast to train
- CFM and cINN both learn the phase space probabilities for each event, so it is best suited to describe complex detector effects, but their also more complex architectures to train.
- Parton-level unfolding is a reasonably complicated task, but transformers help greatly in accounting for correlations and resonances reconstruction
- The mass parametrization allows for more efficient unfolding without the need of very large networks
- Single-event unfolding, calibrated posteriors, compare to other models…

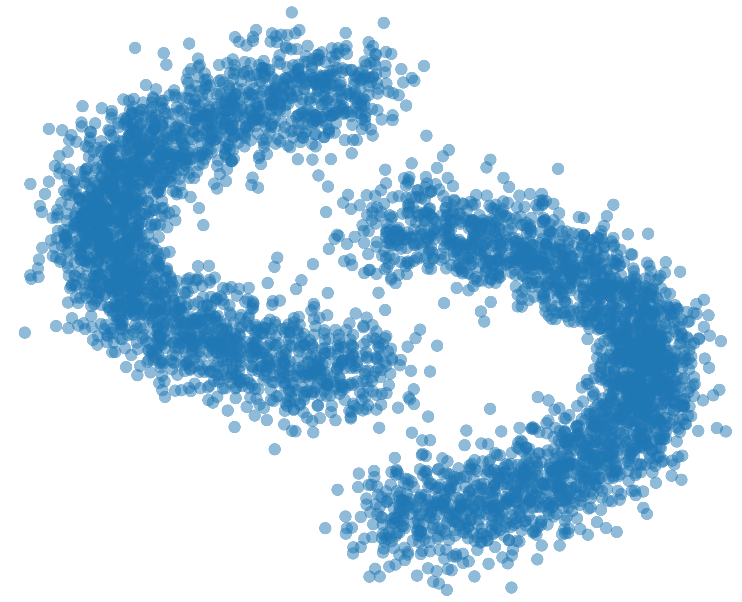
Thanks for your attention!

Backup

Direct Diffusion (DiDi)



$$\frac{\mathrm{d}x(t)}{\mathrm{d}t} = v_\theta(x(t), t)$$



$x_0 \sim p_{\text{model}}(x_{\text{hard}})$ $\qquad$ $x_1 \sim p_{\text{reco}}(x_{\text{reco}})$

- Connect x_0 and x_1 with a linear trajectory: $\quad x(t) = (1-t)x_0 + tx_1$
- The NN is regressed to predict the velocity field: $\quad v_\theta(x(t), t) \approx \frac{\mathrm{dx(t)}}{\mathrm{dt}} = x_1 - x_0$
- For sampling, solve ODE starting from x_1: $\quad x_0 = x_1 + \int_1^0 v_\theta(x(t), t)\mathrm{d}t$
- Loss:

$$\mathscr{L}_{\text{DiDi-P}} = \left\langle [v_\theta((1-t)x_0 + tx_1,\ t) - (x_1 - x_0)]^2 \right\rangle_{t\sim\mathscr{U}([0,1]),\ (x_0,x_1)\sim p(x_{\text{hard}},x_{\text{reco}})}$$

$$\mathscr{L}_{\text{DiDi-U}} = \left\langle [v_\theta((1-t)x_0 + tx_1,\ t) - (x_1 - x_0)]^2 \right\rangle_{t\sim\mathscr{U}([0,1]),\ x_0\sim p(x_{\text{hard}}),x_1\sim p(x_{\text{reco}})}$$

Z + jets events

$Z(p_T > 200\ \mathrm{GeV})$ + jets events generated at $\sqrt{s} = 14\ \mathrm{TeV}$ with Pythia 8.244 and Delphes simulation 3.5.0 available on Zenodo

Six widely-used jet substructure observables:

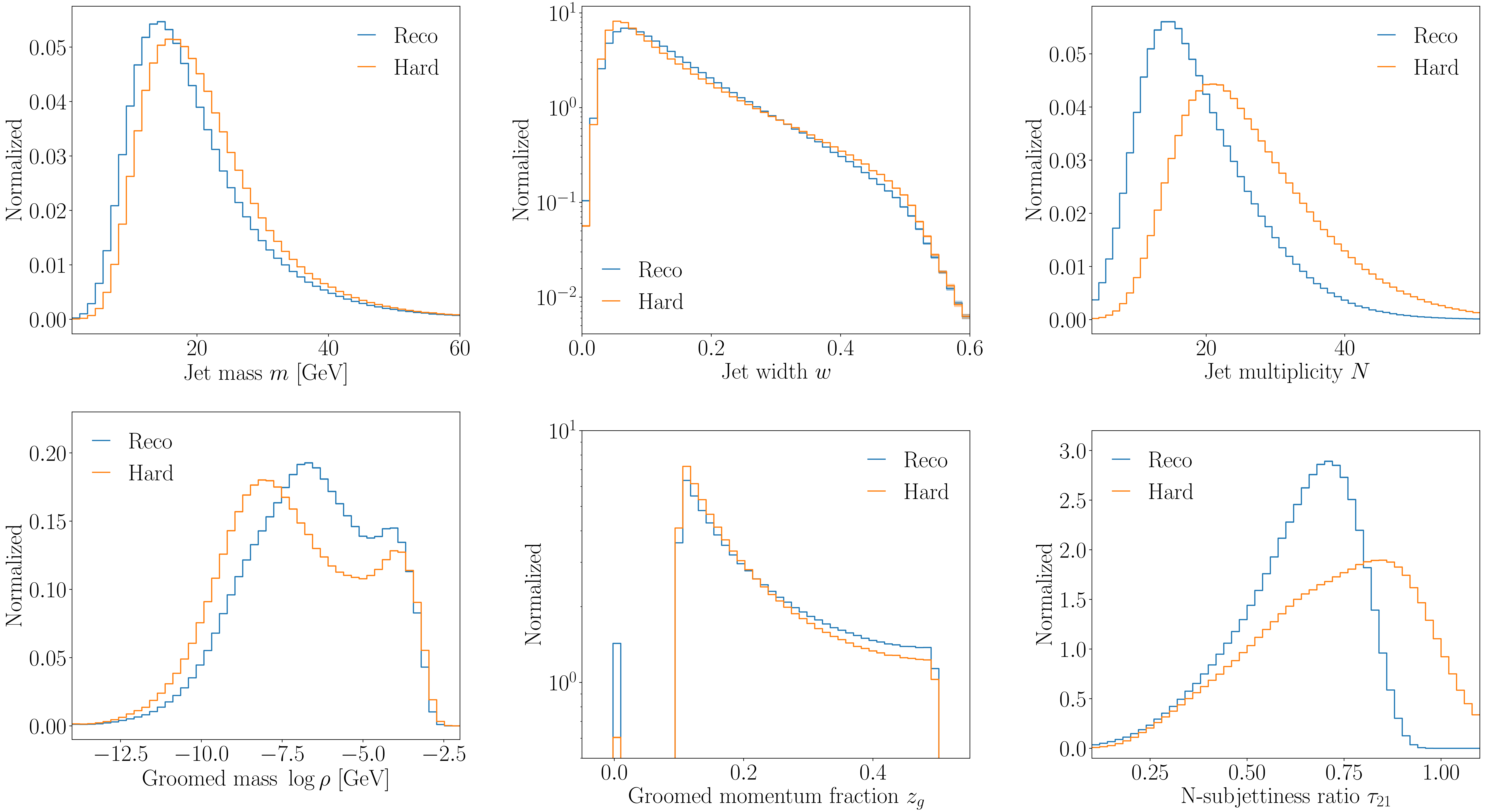
- Jet mass m
- Jet width w
- Jet constituents multiplicity N
- Groomed mass $\log \rho = 2 \log\ (m_{\mathrm{SD}}\ /\ p_T)$
- Groomed momentum fraction $z_g = \tau_1^{\beta=1}$
- N-subjettiness ratio $\tau_{21} = \tau_2^{\beta=1}/\tau_1^{\beta=1}$

Networks of ~3M parameters

19M training events and 1M validation events

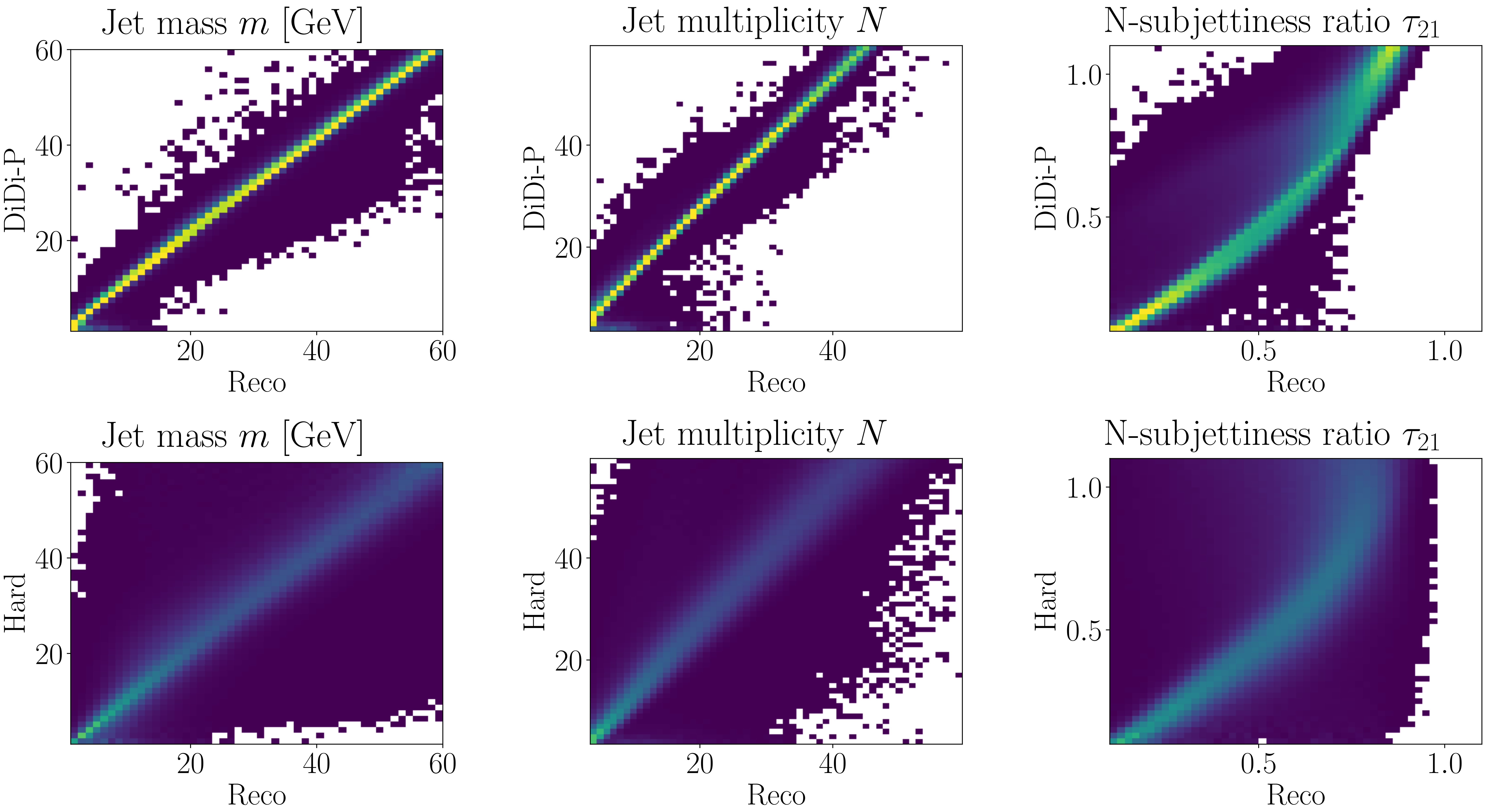
~4M events for testing

The dataset

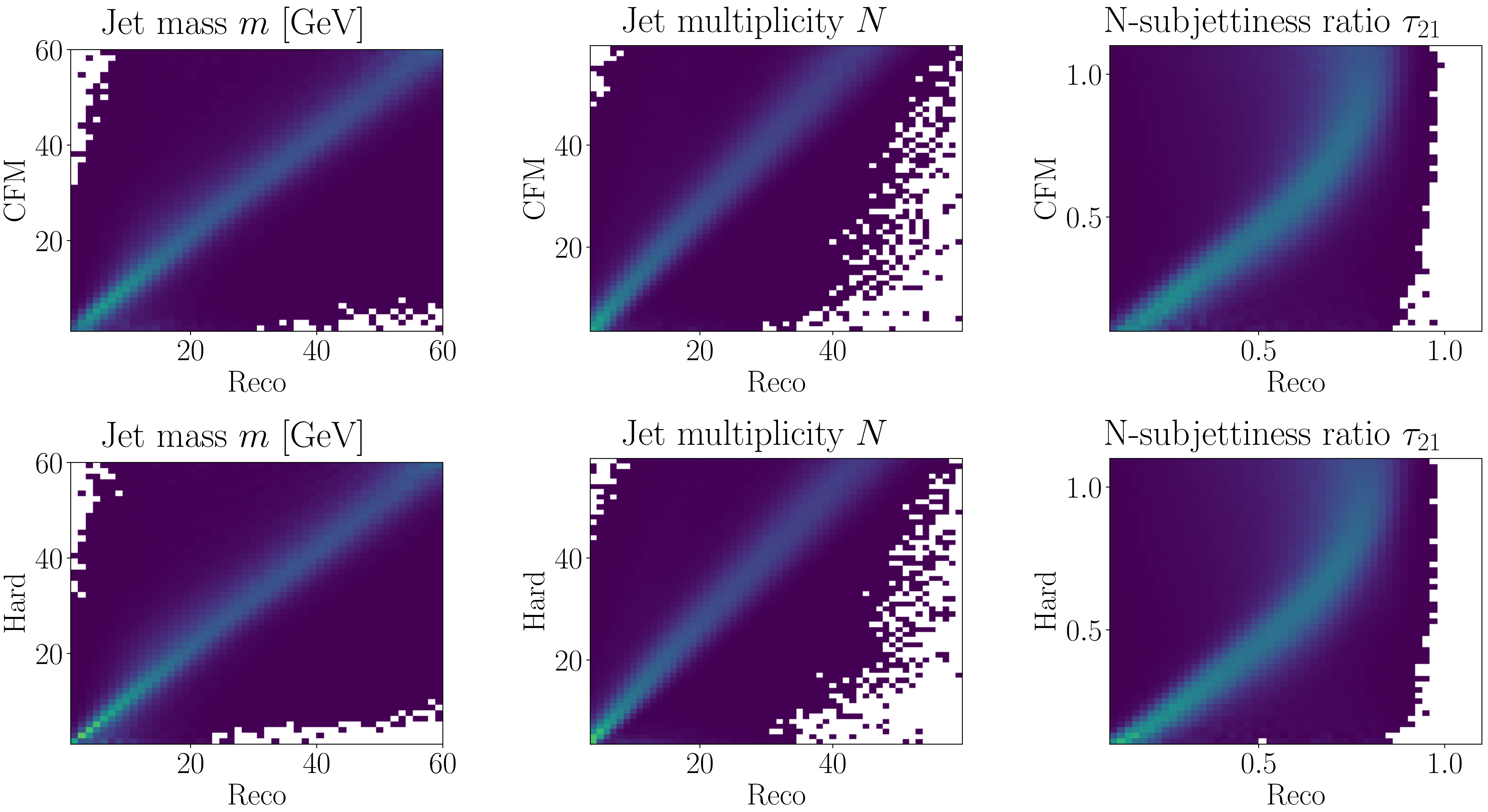


Reco
Hard
Normalized
Jet mass m [GeV]
Jet width w
Jet multiplicity N
Groomed mass $\log\rho$ [GeV]
Groomed momentum fraction z_g
N-subjettiness ratio τ_{21}

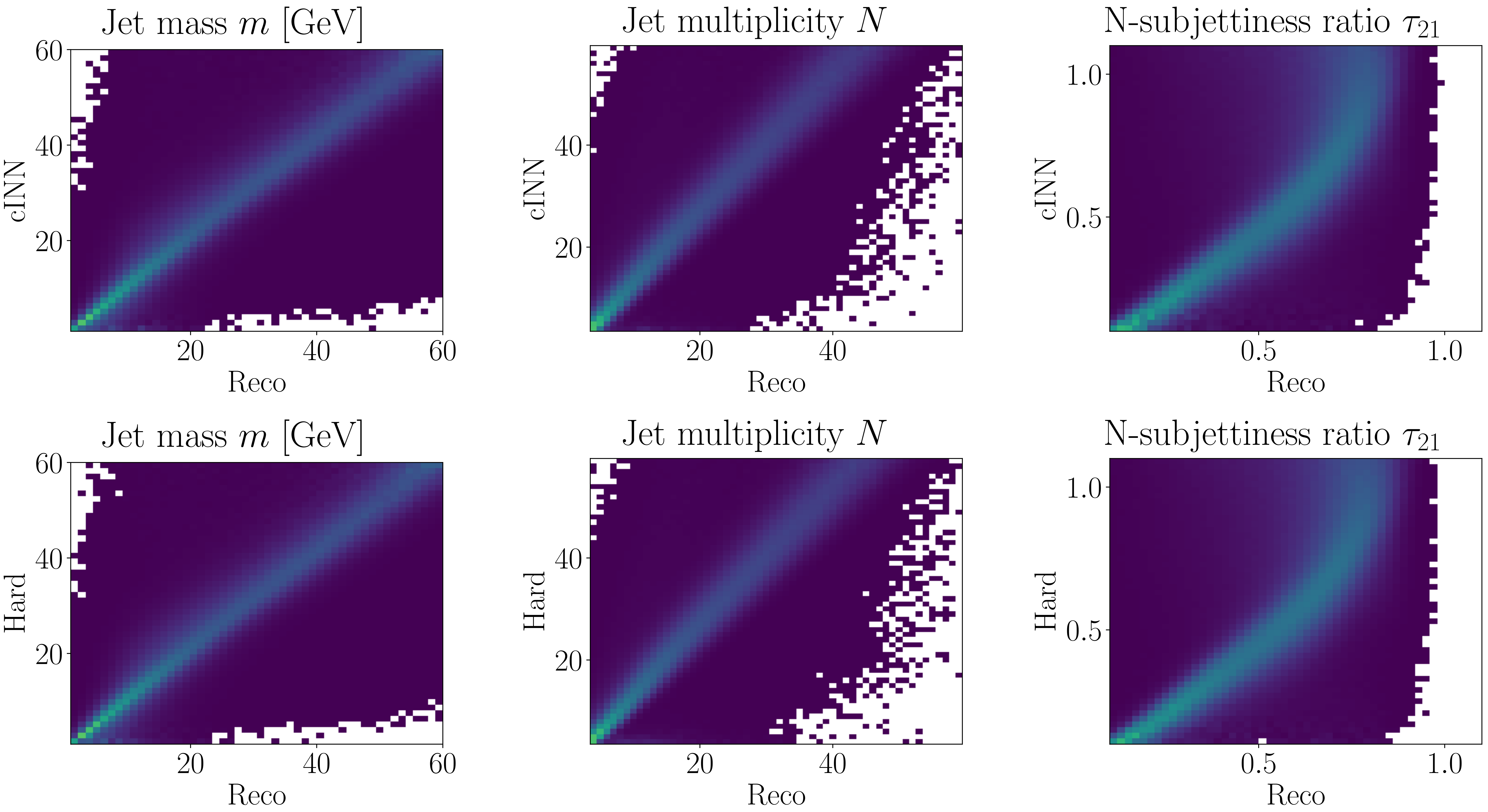
Optimal transport (DiDi)



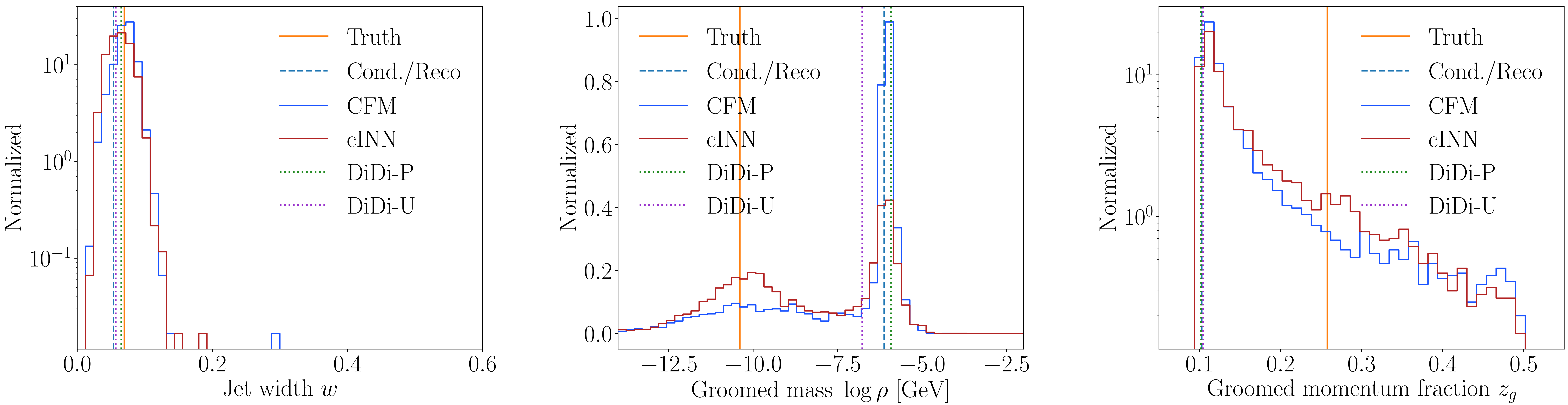
Optimal transport (CFM)



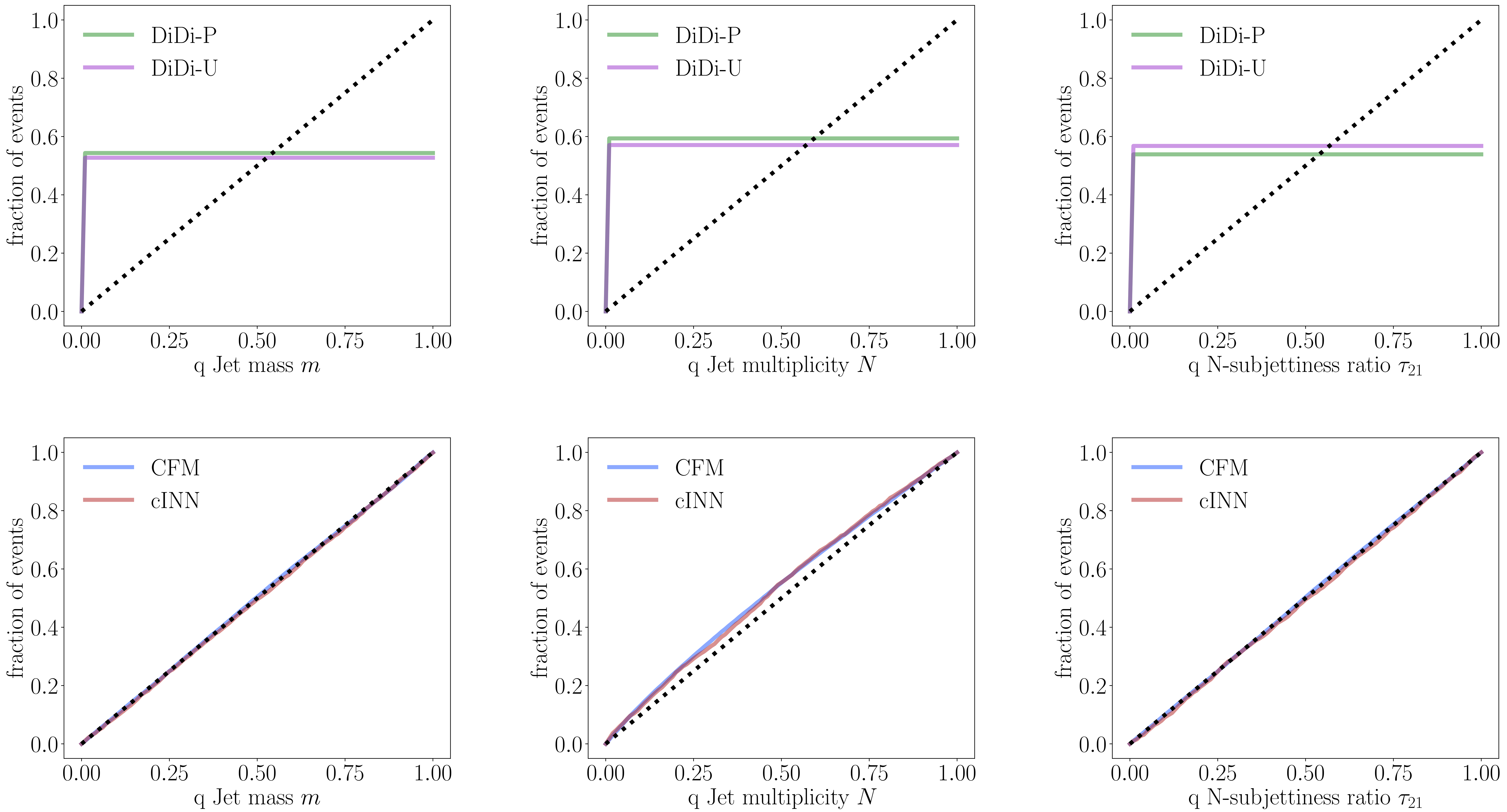
Optimal transport (cINN)



Single events



Calibration



Top-pair events: unfolding to parton-level

Matrix elements are evaluated at $\sqrt{s} = 13\ \mathrm{TeV}$ using MadGraph_aMC@NLO. Showering and hadronization are simulated with Pythia8, and detector response is simulated with Delphes with the standard CMS card. For a detailed description see [2305.10399].

Unfolding from 6 final-state particles $(bl\nu)(bqq)$:

- 4 DoFs for the lepton
- 3 DoFs for the missing p_T^ν
- 5 DoFs per jet (4-momentum + b-tag)

Total: 27 DoFs at reco-level and 19 DoFs at parton-level

Non-bayesian networks

cINN ~ 8M parameters, CFM ~ 6M, Tra-CFM, Transfermer ~ 3M

10M training events and 1M testing events

Top-pair events: unfolding to parton-level

Adding transformers:

- For transfermer, likelihoods are factorized autoregressively on all previous parton-level dimensions and reco-level event:

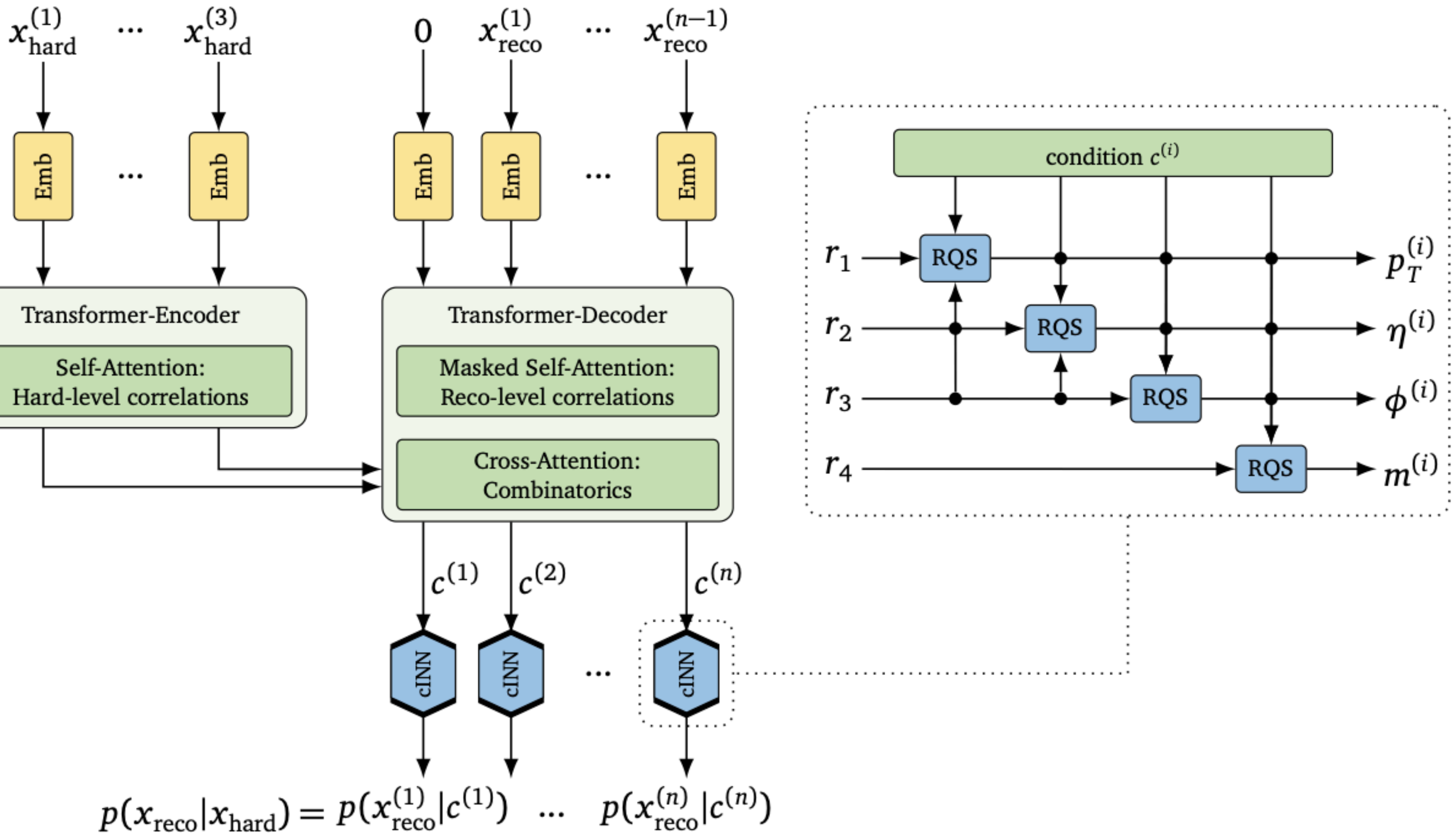
$$p_{\text{model}}(x_{\text{part}} \,|\, x_{\text{reco}}) = \prod_{i=1}^{n} p_{\text{model}}(x_{\text{part}}^{(i)} \mid c(x_{\text{part}}^{(0)}, \ldots, x_{\text{part}}^{(i-1)}, x_{\text{reco}}))$$

- For Tra-CFM, the transformer is made time-dependent and a small CFM predicts velocities at each different dimension:

$$v(x_{\text{part}}(t), t \,|\, x_{\text{reco}}) = \left(v^{(1)}(c^{(1)}, t), \; \ldots, v^{(n)}(c^{(n)}, t)\right)$$

Top-pair events: unfolding to parton-level

Transfermer



Top-pair events: unfolding to parton-level

Tra-CFM