2007.10356

2211.02058

2308.05704

Applications of Likelihood Learning

Benasque 2023

Alfredo Glioti & Jaco ter Hoeve

IPhT & VU Amsterdam

Recap: likelihood ratio from ML

- Starting from two balanced datasets $\mathscr{D}_{\mathrm{SM}}$ and $\mathscr{D}_{\mathrm{EFT}}$ drawn from $f(\boldsymbol{x}\,|\,\mathrm{SM})$ and $f(\boldsymbol{x}\,|\,\mathrm{EFT})$, we minimise e.g. the cross-entropy loss

$$L[g(\boldsymbol{x})] = -\frac{1}{N}\sum_{e\in\mathscr{D}_{\mathrm{EFT}}} w_e \log(1-g(\boldsymbol{x}_e)) - \frac{1}{N}\sum_{\mathscr{D}_{\mathrm{SM}}} w_e \log g(\boldsymbol{x}_e)$$

$\{m_{t\bar{t}}, \eta_l, \Delta\phi, \dots\}$

Event weights

- The learned decision boundary $g(\boldsymbol{x})$ is one-to-one with the likelihood ratio (LR) as $N \to \infty$

$$\frac{\delta L}{\delta g} = 0 \implies \hat{g}(\boldsymbol{x}) = \left(1 + \frac{f(\boldsymbol{x}\,|\,\mathrm{EFT})}{f(\boldsymbol{x}\,|\,\mathrm{SM})}\right)^{-1} \equiv \frac{1}{1+r(\boldsymbol{x})}$$

Parameterise with NNs

The experimental pipeline

We are progressively moving through the simulation chain (latent space)

$$p(x|c) \sim \int dz_{\text{det}} dz_{\text{shower}} dz_{\text{parton}} p(x|z_{\text{det}}) p(z_{\text{det}}|z_{\text{shower}}) p(z_{\text{shower}}|z_{\text{parton}}) p(z_{\text{parton}}|c)$$

DELPHES fast simulation

SMEFT@NLO

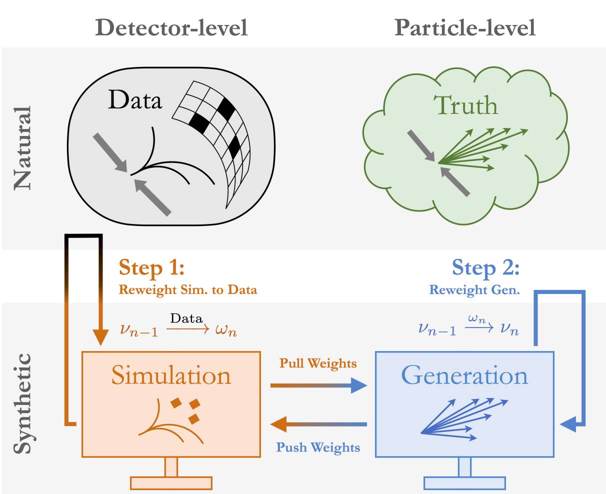
(Likelihood free) Inference Observed $\longrightarrow$ Hidden/latent variables $\longrightarrow$ POI

Generation Observed $\longleftarrow$ Hidden/latent variables $\longleftarrow$ POI

The experimental pipeline

We are progressively moving through the simulation chain (latent space)

$$p(x|c) \sim \int dz_{\text{det}} dz_{\text{shower}} dz_{\text{parton}} p(x|z_{\text{det}}) p(z_{\text{det}}|z_{\text{shower}}) p(z_{\text{shower}}|z_{\text{parton}}) p(z_{\text{parton}}|c)$$



Unbinned unfolding, Omnifold [1911.09107]

SMEFT@NLO

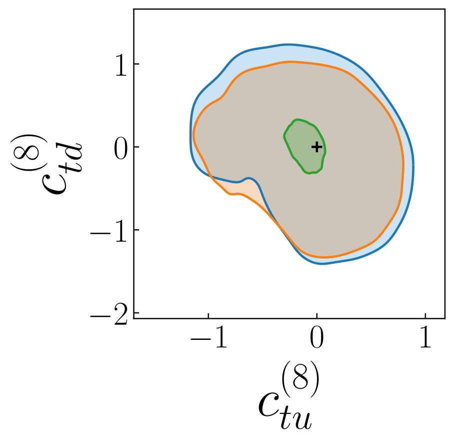
(Likelihood free) Inference	Observed	→	Hidden/latent variables	→ POI
Generation	Observed	←	Hidden/latent variables	← POI

Why interesting for us?

- Global efforts **reinterpret** existing "SM measurements" in an EFT context
- But which measurements are the most **sensitive** to EFT parameters?
  - Inclusive, single to multi-differential (which variables)
  - Binned or unbinned, which binning?

Framework needed to integrate unbinned multivariate observables into **global** SMEFT fits

- **Optimal** bounds on the EFT parameters
- Useful **diagnosis tool** to assess information loss



Applications of likelihood learning

Focusses on global EFT fits

Nikhef-2022-015

Unbinned multivariate observables for global SMEFT analyses from machine learning

Raquel Gomez Ambrosio,[1,2] Jaco ter Hoeve,[3,4] Maeve Madigan,[5] Juan Rojo,[3,4] and Veronica Sanz[6,7]

[1] *Dipartimento di Fisica "G. Occhialini", Universita degli Studi di Milano-Bicocca, and INFN, Sezione di Milano Bicocca, Piazza della Scienza 3, I – 20126 Milano, Italy*
[2] *Dipartimento di Fisica, Universitá di Torino, and INFN, Sezione di Torino, Via P. Giuria 1, 10125 Torino, Italy*
[3] *Department of Physics and Astronomy, VU Amsterdam, 1081HV Amsterdam, The Netherlands*
[4] *Nikhef Theory Group, Science Park 105, 1098 XG Amsterdam, The Netherlands*
[5] *DAMTP, University of Cambridge, Wilberforce Road, Cambridge CB3 0WA, UK*
[6] *Instituto de Física Corpuscular (IFIC), Universidad de Valencia-CSIC, E-46980 Valencia, Spain*
[7] *Department of Physics and Astronomy, University of Sussex, Brighton BN1 9QH, UK*

arXiv:2211.02058v2 [hep-ph] 23 May 2023

Abstract

Theoretical interpretations of particle physics data, such as the determination of the Wilson coefficients of the Standard Model Effective Field Theory (SMEFT), often involve the inference of multiple parameters from a global dataset. Optimizing such interpretations requires the identification of observables that exhibit the highest possible sensitivity to the underlying theory parameters. In this work we develop a flexible open source framework, ML4EFT, enabling the integration of unbinned multivariate observables into global SMEFT fits. As compared to traditional measurements, such observables enhance the sensitivity to the theory parameters by preventing the information loss incurred when binning in a subset of final-state kinematic variables. Our strategy combines machine learning regression and classification techniques to parameterize high-dimensional likelihood ratios, using the Monte Carlo replica method to estimate and propagate methodological uncertainties. As a proof of concept we construct unbinned multivariate observables for top-quark pair and Higgs+Z production at the LHC, demonstrate their impact on the SMEFT parameter space as compared to binned measurements, and study the improved constraints associated to multivariate inputs. Since the number of neural networks to be trained scales quadratically with the number of parameters and can be fully parallelized, the ML4EFT framework is well-suited to construct unbinned multivariate observables which depend on up to tens of EFT coefficients, as required in global fits.

1

Reweighting for more accurate learning

Boosting likelihood learning with event reweighting

Siyu Chen[1], Alfredo Glioti[2], Giuliano Panico[3,4], and Andrea Wulzer[5,6]

[1] *Institut de Théorie des Phénomenes Physiques, EPFL, Lausanne, Switzerland*
[2] *Université Paris-Saclay, CNRS, CEA, Institut de Physique Théorique, 91191, Gif-sur-Yvette, France*
[3] *Dipartimento di Fisica e Astronomia, Università di Firenze, Via G. Sansone 1, 50019 Sesto Fiorentino, Italy*
[4] *INFN, Sezione di Firenze, Via G. Sansone 1, 50019 Sesto Fiorentino, Italy*
[5] Institut de Física d'Altes Energies (IFAE), The Barcelona Institute of Science and Technology (BIST), Campus UAB, 08193 Bellaterra, Barcelona, Spain
[6] ICREA, Institució Catalana de Recerca i Estudis Avançats, Passeig de Lluís Companys 23, 08010 Barcelona, Spain

arXiv:2308.05704v1 [hep-ph] 10 Aug 2023

Abstract

Extracting maximal information from experimental data requires access to the likelihood function, which however is never directly available for complex experiments like those performed at high energy colliders. Theoretical predictions are obtained in this context by Monte Carlo events, which do furnish an accurate but abstract and implicit representation of the likelihood. Strategies based on statistical learning are currently being developed to infer the likelihood function explicitly by training a continuous-output classifier on Monte Carlo events. In this paper, we investigate the usage of Monte Carlo events that incorporate the dependence on the parameters of interest by reweighting. This enables more accurate likelihood learning with less training data and a more robust learning scheme that is more suited for automation and extensive deployment. We illustrate these advantages in the context of LHC precision probes of new Effective Field Theory interactions.

Alfredo

Applications of likelihood learning

Focusses on global EFT fits

Nikhef-2022-015

arXiv:2211.02058v2 [hep-ph] 23 May 2023

Unbinned multivariate observables for global SMEFT analyses from machine learning

Raquel Gomez Ambrosio,[1,2] Jaco ter Hoeve,[3,4] Maeve Madigan,[5] Juan Rojo,[3,4] and Veronica Sanz[6,7]

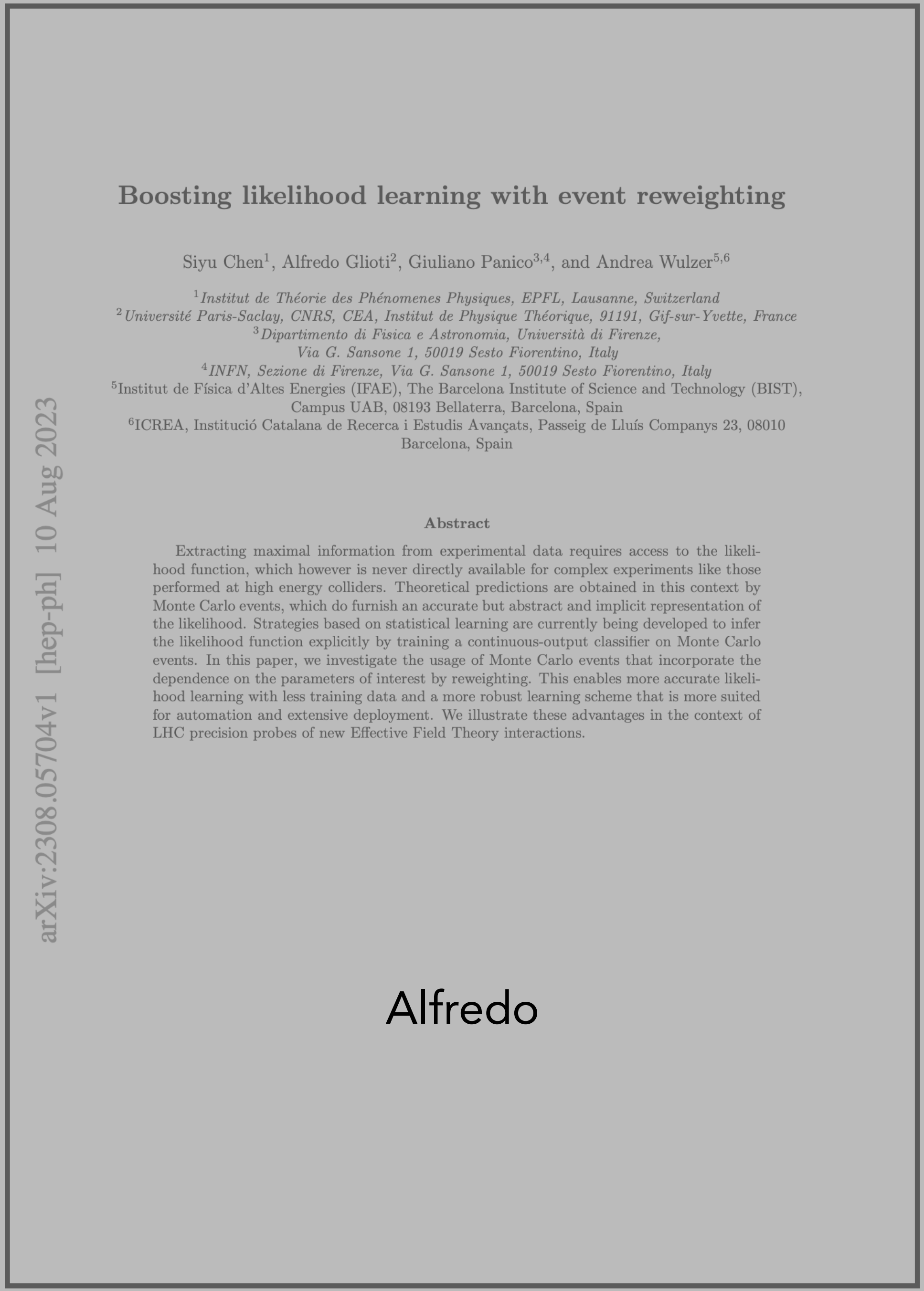
[1] *Dipartimento di Fisica "G. Occhialini", Universita degli Studi di Milano-Bicocca, and INFN, Sezione di Milano Bicocca, Piazza della Scienza 3, I – 20126 Milano, Italy*
[2] *Dipartimento di Fisica, Universitá di Torino, and INFN, Sezione di Torino, Via P. Giuria 1, 10125 Torino, Italy*
[3] *Department of Physics and Astronomy, VU Amsterdam, 1081HV Amsterdam, The Netherlands*
[4] *Nikhef Theory Group, Science Park 105, 1098 XG Amsterdam, The Netherlands*
[5] *DAMTP, University of Cambridge, Wilberforce Road, Cambridge CB3 0WA, UK*
[6] *Instituto de Física Corpuscular (IFIC), Universidad de Valencia-CSIC, E-46980 Valencia, Spain*
[7] *Department of Physics and Astronomy, University of Sussex, Brighton BN1 9QH, UK*

Abstract

Theoretical interpretations of particle physics data, such as the determination of the Wilson coefficients of the Standard Model Effective Field Theory (SMEFT), often involve the inference of multiple parameters from a global dataset. Optimizing such interpretations requires the identification of observables that exhibit the highest possible sensitivity to the underlying theory parameters. In this work we develop a flexible open source framework, ML4EFT, enabling the integration of unbinned multivariate observables into global SMEFT fits. As compared to traditional measurements, such observables enhance the sensitivity to the theory parameters by preventing the information loss incurred when binning in a subset of final-state kinematic variables. Our strategy combines machine learning regression and classification techniques to parameterize high-dimensional likelihood ratios, using the Monte Carlo replica method to estimate and propagate methodological uncertainties. As a proof of concept we construct unbinned multivariate observables for top-quark pair and Higgs+Z production at the LHC, demonstrate their impact on the SMEFT parameter space as compared to binned measurements, and study the improved constraints associated to multivariate inputs. Since the number of neural networks to be trained scales quadratically with the number of parameters and can be fully parallelized, the ML4EFT framework is well-suited to construct unbinned multivariate observables which depend on up to tens of EFT coefficients, as required in global fits.

1

Reweighting for more accurate learning



arXiv:2308.05704v1 [hep-ph] 10 Aug 2023

Boosting likelihood learning with event reweighting

Siyu Chen[1], Alfredo Glioti[2], Giuliano Panico[3,4], and Andrea Wulzer[5,6]

[1] *Institut de Théorie des Phénomenes Physiques, EPFL, Lausanne, Switzerland*
[2] *Université Paris-Saclay, CNRS, CEA, Institut de Physique Théorique, 91191, Gif-sur-Yvette, France*
[3] *Dipartimento di Fisica e Astronomia, Università di Firenze, Via G. Sansone 1, 50019 Sesto Fiorentino, Italy*
[4] *INFN, Sezione di Firenze, Via G. Sansone 1, 50019 Sesto Fiorentino, Italy*
[5] Institut de Física d'Altes Energies (IFAE), The Barcelona Institute of Science and Technology (BIST), Campus UAB, 08193 Bellaterra, Barcelona, Spain
[6] ICREA, Institució Catalana de Recerca i Estudis Avançats, Passeig de Lluís Companys 23, 08010 Barcelona, Spain

Abstract

Extracting maximal information from experimental data requires access to the likelihood function, which however is never directly available for complex experiments like those performed at high energy colliders. Theoretical predictions are obtained in this context by Monte Carlo events, which do furnish an accurate but abstract and implicit representation of the likelihood. Strategies based on statistical learning are currently being developed to infer the likelihood function explicitly by training a continuous-output classifier on Monte Carlo events. In this paper, we investigate the usage of Monte Carlo events that incorporate the dependence on the parameters of interest by reweighting. This enables more accurate likelihood learning with less training data and a more robust learning scheme that is more suited for automation and extensive deployment. We illustrate these advantages in the context of LHC precision probes of new Effective Field Theory interactions.

Alfredo

The ML4EFT framework

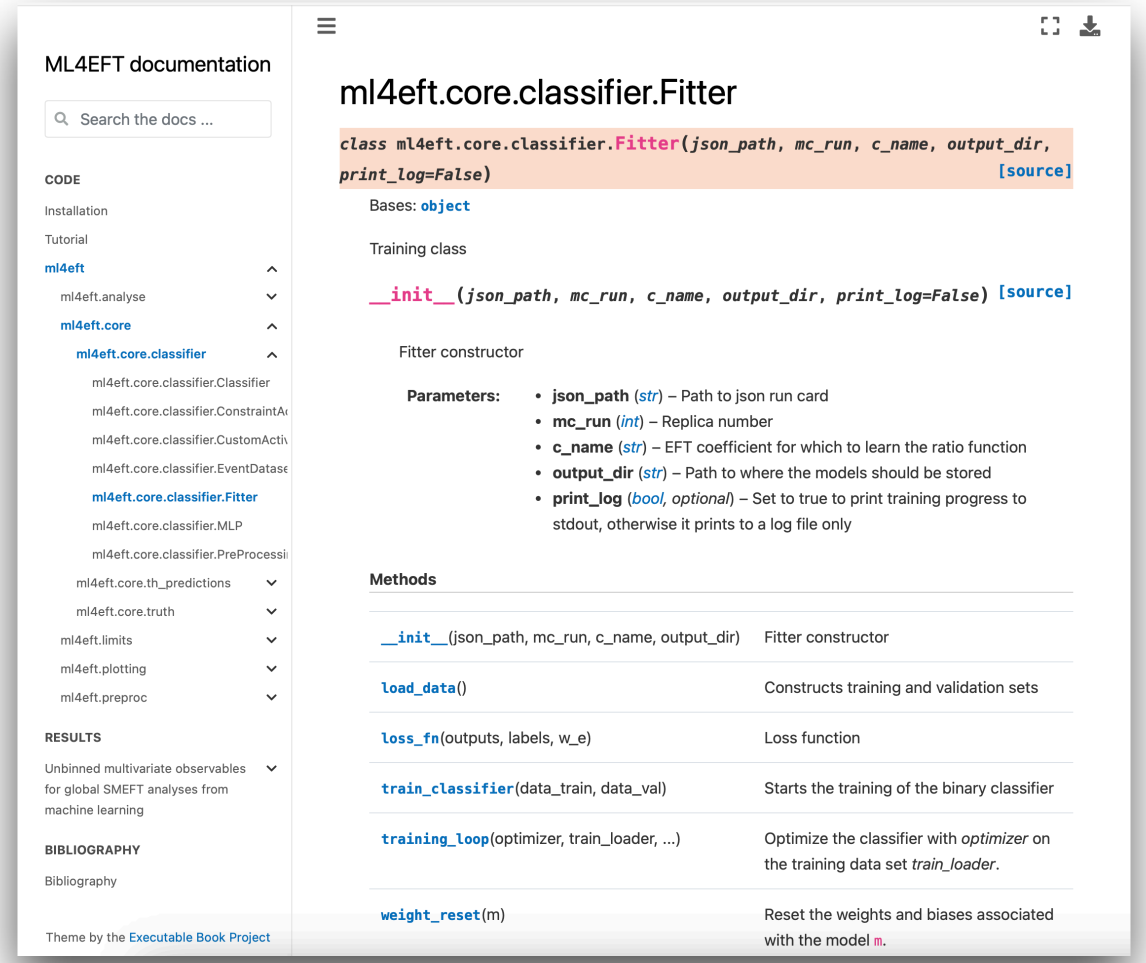
`pip install ml4eft`

https://lhcfitnikhef.github.io/ML4EFT

2211.02058 R. Gomez Ambrosio, JtH, M. Madigan, J. Rojo, V.Sanz

Open-source NN-based python framework for the integration of unbinned multivariate observables into global SMEFT fits

- **Goal**: provide optimal constraints on the SMEFT
- **Diagnostic tool:** what is the information loss incurred by a particular choice of bins?
- **Projections:** how will SMEFT constraints improve if unbinned data are made available?



Modular structure, easy to maintain, well documented

Anticipating global fits

- Global EFT fits typically feature ~50 WCs and thus efficient scaling with the number of WCs becomes essential

- **ML4EFT 1.0:** learn the coefficient functions separately and combine afterwards

Assumes no sign flips in interferences
Fix is part of **ML4EFT2.0**

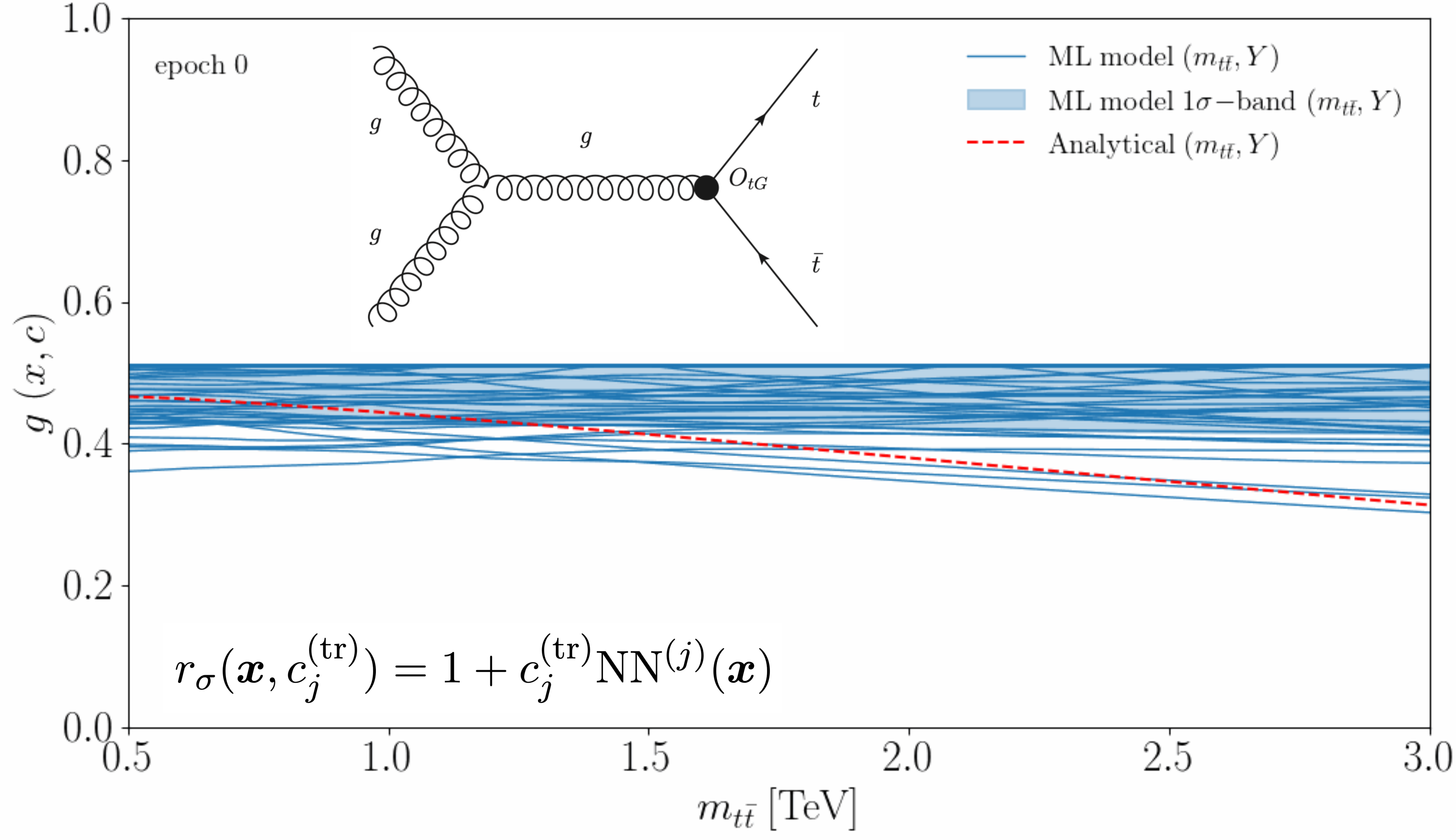
$$r(\boldsymbol{x}, \boldsymbol{c}) = 1 + \sum_{j=1}^{n_{\rm eft}} r^{(j)}(\boldsymbol{x}) c_j + \sum_{j=1}^{n_{\rm eft}} \sum_{k \geq j}^{n_{\rm eft}} r^{(j,k)}(\boldsymbol{x}) c_j c_k$$

Example: to learn a single $r^{(j)}$, generate $\mathscr{D}_{\rm sm}$ and $\mathscr{D}_{\rm eft}$ at c_j up to $\mathcal{O}(\Lambda^{-2})$. Then $r(\boldsymbol{x}, \boldsymbol{c}) = 1 + r^{(j)}(\boldsymbol{x}) c_j^{(\rm tr)}$ and training means

$$g(\boldsymbol{x}, c_j^{(\rm tr)}) = \left(1 + \left[1 + c_j^{(\rm tr)} \cdot \mathrm{NN}^{(j)}(\boldsymbol{x})\right]\right)^{-1} \qquad \mathrm{NN}^{(j)}(\boldsymbol{x}) \to r^{(j)}(\boldsymbol{x})$$

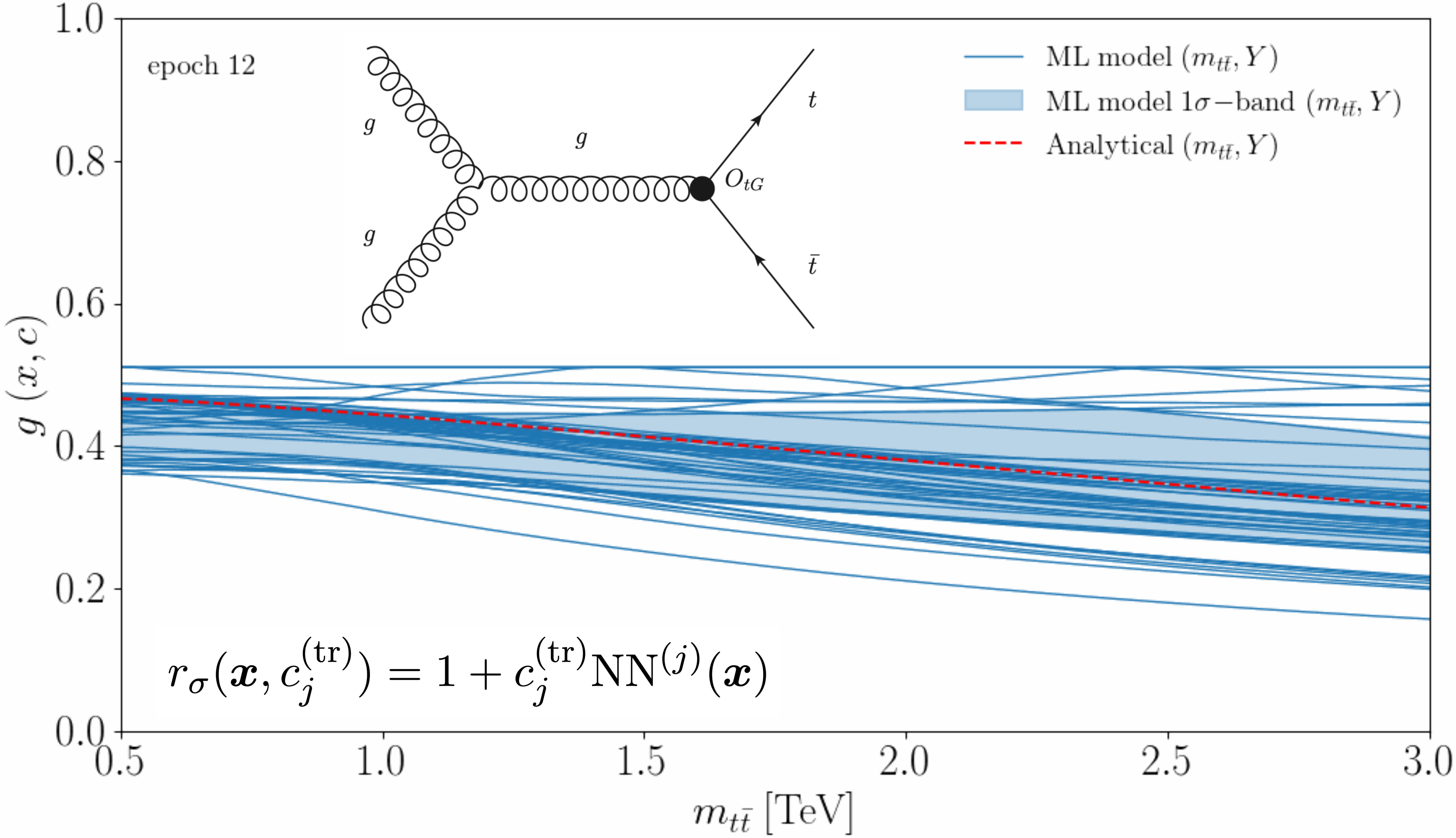
Toy example

$$L[g(\boldsymbol{x},\boldsymbol{c})] = -\frac{1}{N}\sum_{e\in\mathcal{D}_{\rm eft}} w_e \log(1-g(\boldsymbol{x}_e,\boldsymbol{c})) - \frac{1}{N}\sum_{e\in\mathcal{D}_{\rm sm}} w_e \log g(\boldsymbol{x}_e,\boldsymbol{c})$$



Toy example

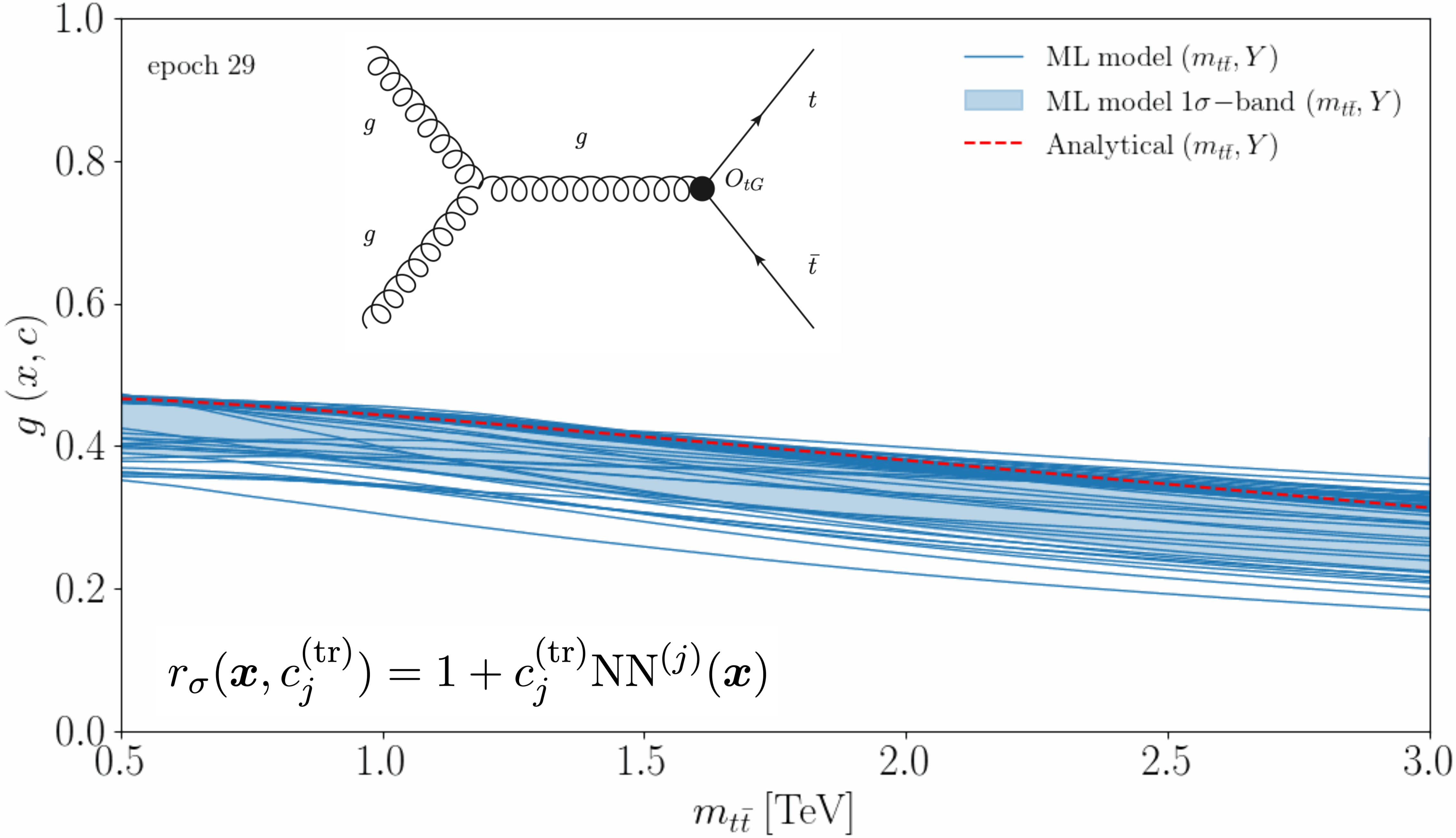
$$L[g(\boldsymbol{x},\boldsymbol{c})] = -\frac{1}{N}\sum_{e\in\mathcal{D}_{\rm eft}} w_e \log(1-g(\boldsymbol{x}_e,\boldsymbol{c})) - \frac{1}{N}\sum_{e\in\mathcal{D}_{\rm sm}} w_e \log g(\boldsymbol{x}_e,\boldsymbol{c})$$



$$r_\sigma(\boldsymbol{x}, c_j^{(\rm tr)}) = 1 + c_j^{(\rm tr)}\mathrm{NN}^{(j)}(\boldsymbol{x})$$

Toy example

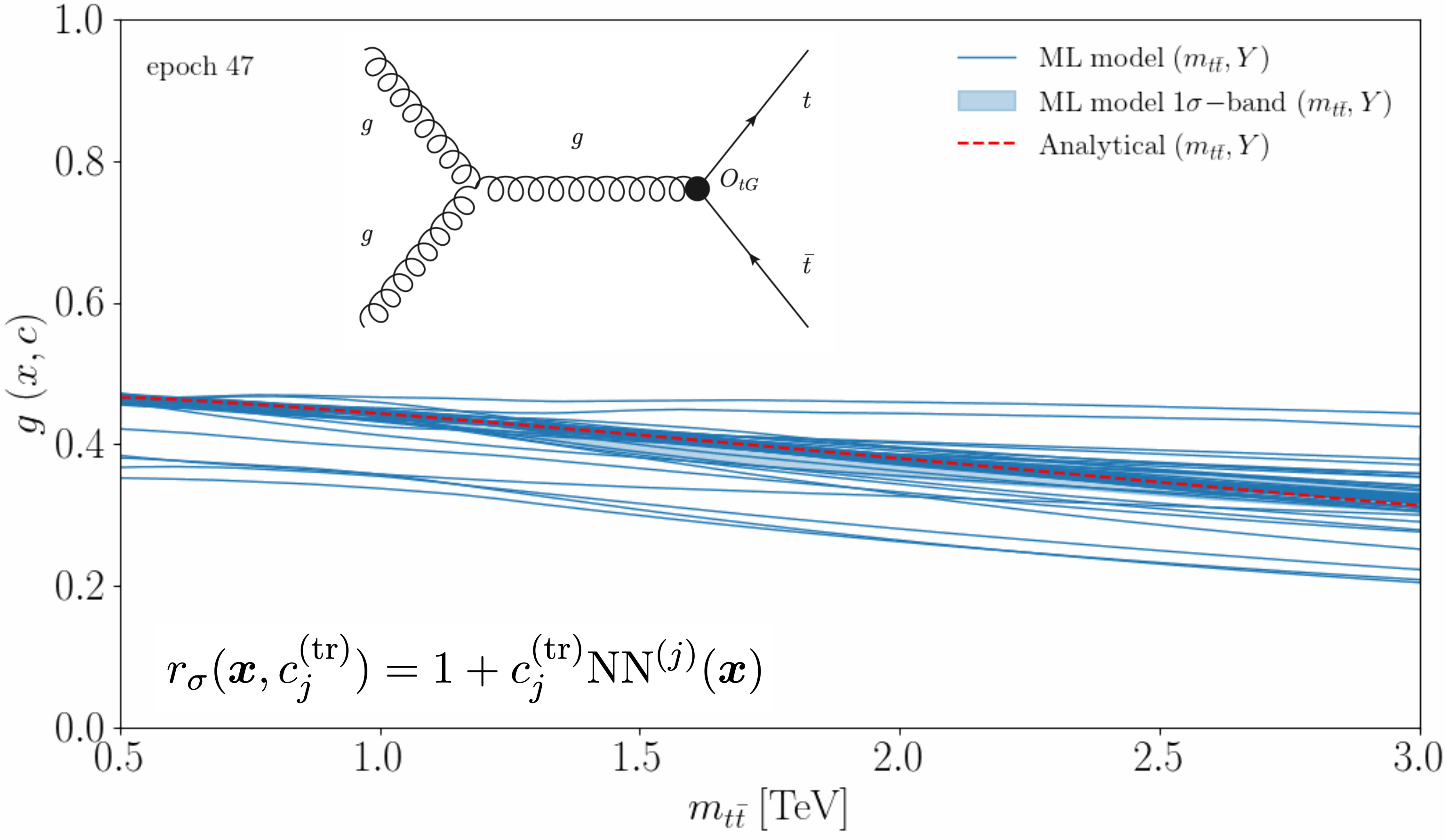
$$L[g(\boldsymbol{x},\boldsymbol{c})] = -\frac{1}{N}\sum_{e\in\mathcal{D}_{\rm eft}} w_e \log(1-g(\boldsymbol{x}_e,\boldsymbol{c})) - \frac{1}{N}\sum_{e\in\mathcal{D}_{\rm sm}} w_e \log g(\boldsymbol{x}_e,\boldsymbol{c})$$



$$r_\sigma(\boldsymbol{x}, c_j^{(\rm tr)}) = 1 + c_j^{(\rm tr)}\mathrm{NN}^{(j)}(\boldsymbol{x})$$

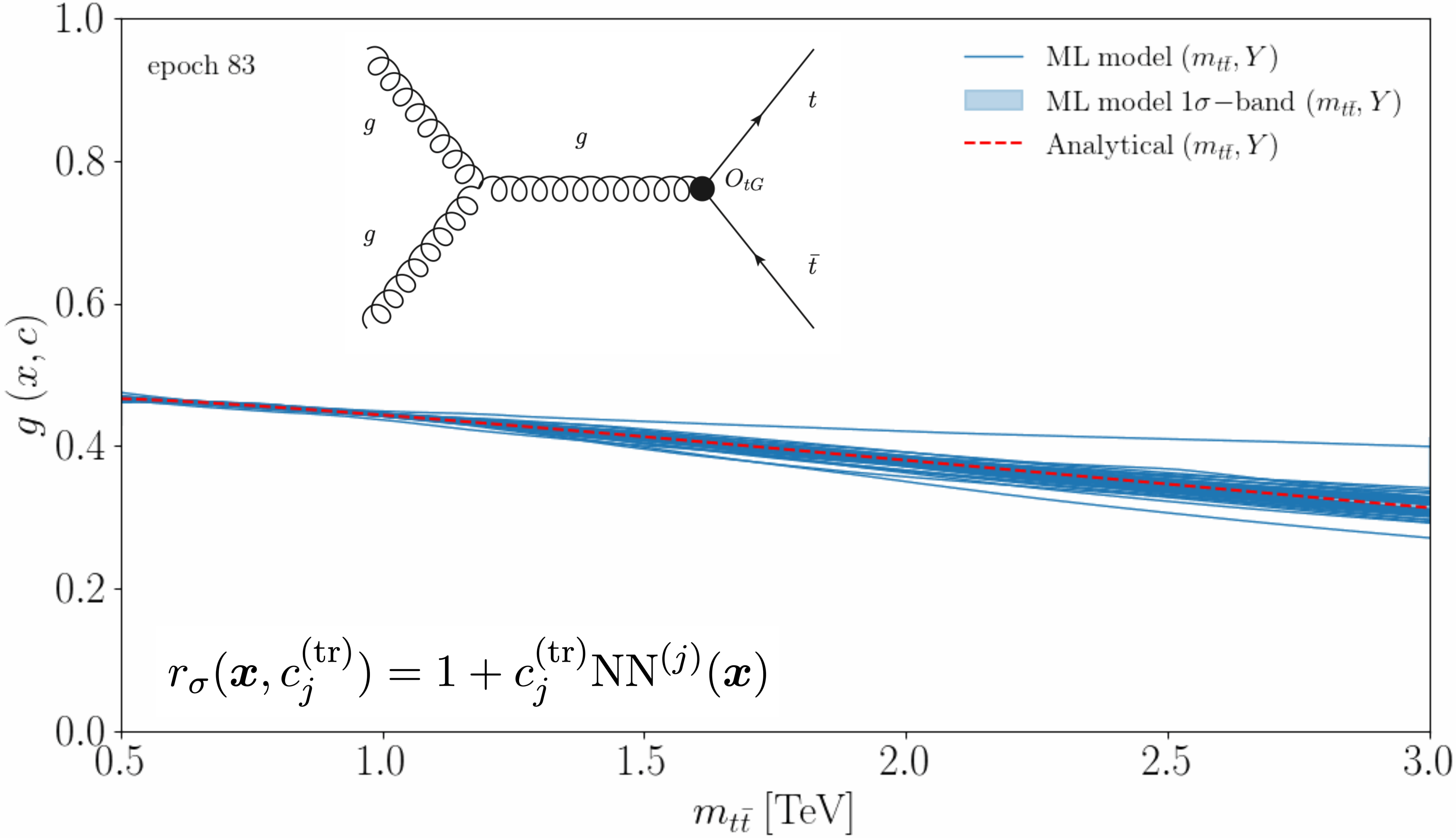
Toy example

$$L[g(\boldsymbol{x},\boldsymbol{c})] = -\frac{1}{N}\sum_{e\in\mathscr{D}_{\rm eft}} w_e \log(1-g(\boldsymbol{x}_e,\boldsymbol{c})) - \frac{1}{N}\sum_{e\in\mathscr{D}_{\rm sm}} w_e \log g(\boldsymbol{x}_e,\boldsymbol{c})$$



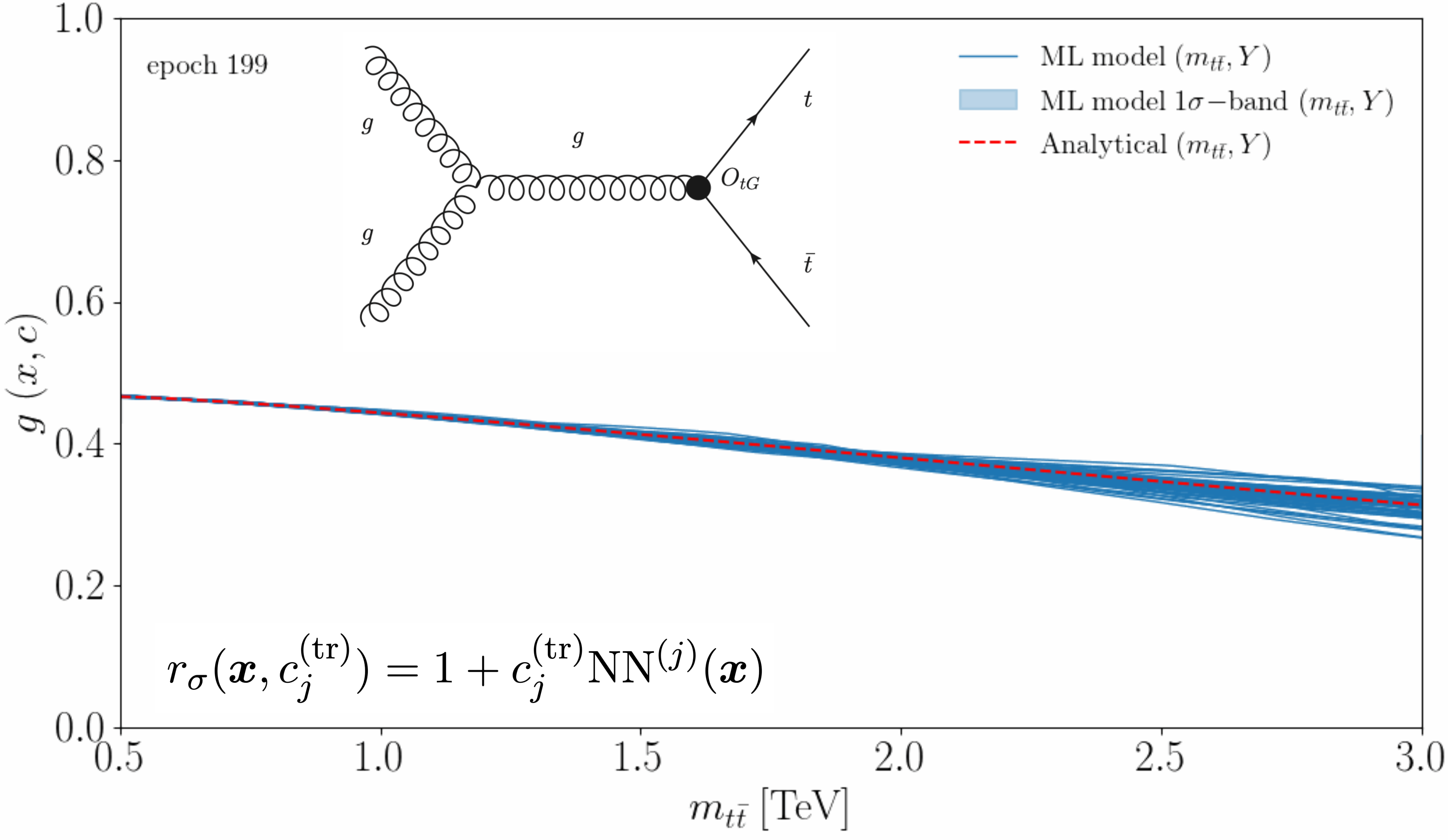
Toy example

$$L[g(\boldsymbol{x},\boldsymbol{c})] = -\frac{1}{N}\sum_{e\in\mathscr{D}_{\rm eft}} w_e \log(1-g(\boldsymbol{x}_e,\boldsymbol{c})) - \frac{1}{N}\sum_{e\in\mathscr{D}_{\rm sm}} w_e \log g(\boldsymbol{x}_e,\boldsymbol{c})$$



Toy example

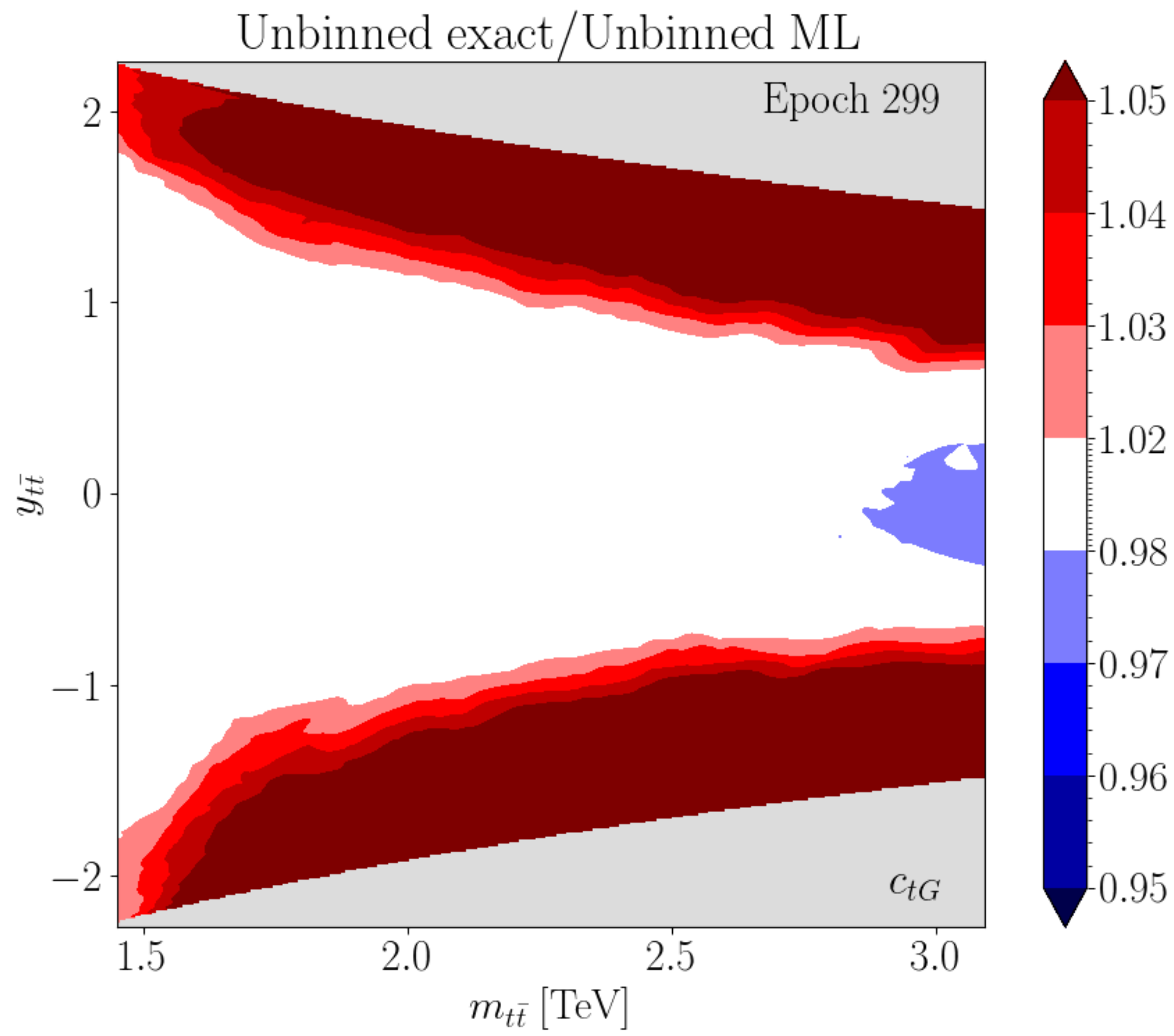
$$L[g(\boldsymbol{x}, \boldsymbol{c})] = -\frac{1}{N} \sum_{e \in \mathcal{D}_{\text{eft}}} w_e \log(1 - g(\boldsymbol{x}_e, \boldsymbol{c})) - \frac{1}{N} \sum_{e \in \mathcal{D}_{\text{sm}}} w_e \log g(\boldsymbol{x}_e, \boldsymbol{c})$$

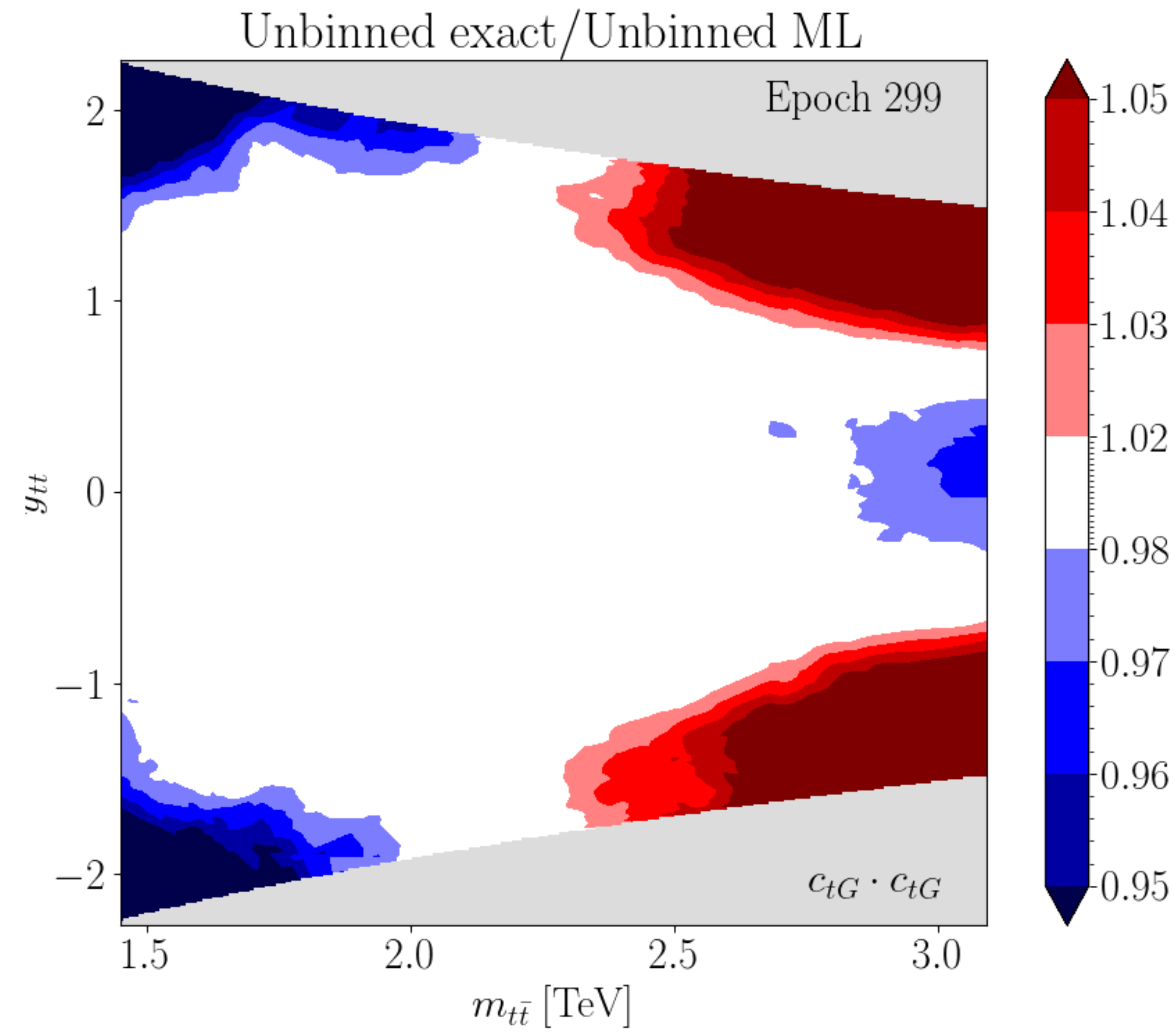


$$r_\sigma(\boldsymbol{x}, c_j^{(\text{tr})}) = 1 + c_j^{(\text{tr})} \text{NN}^{(j)}(\boldsymbol{x})$$

Toy example

2 kinematic features

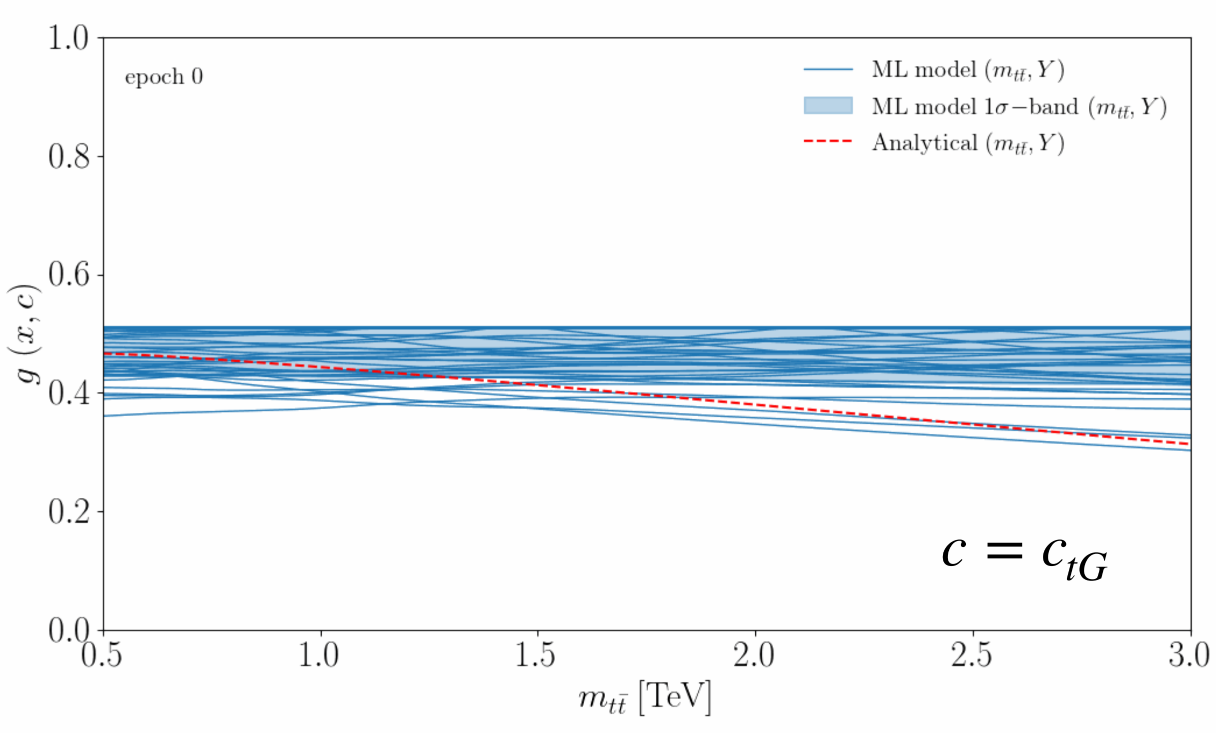




Uncertainty treatment

- We only have finite training data and NNs are subject to methodological uncertainties
- Propagate uncertainties as well as finite training set effects to the space of models by training multiple replicas

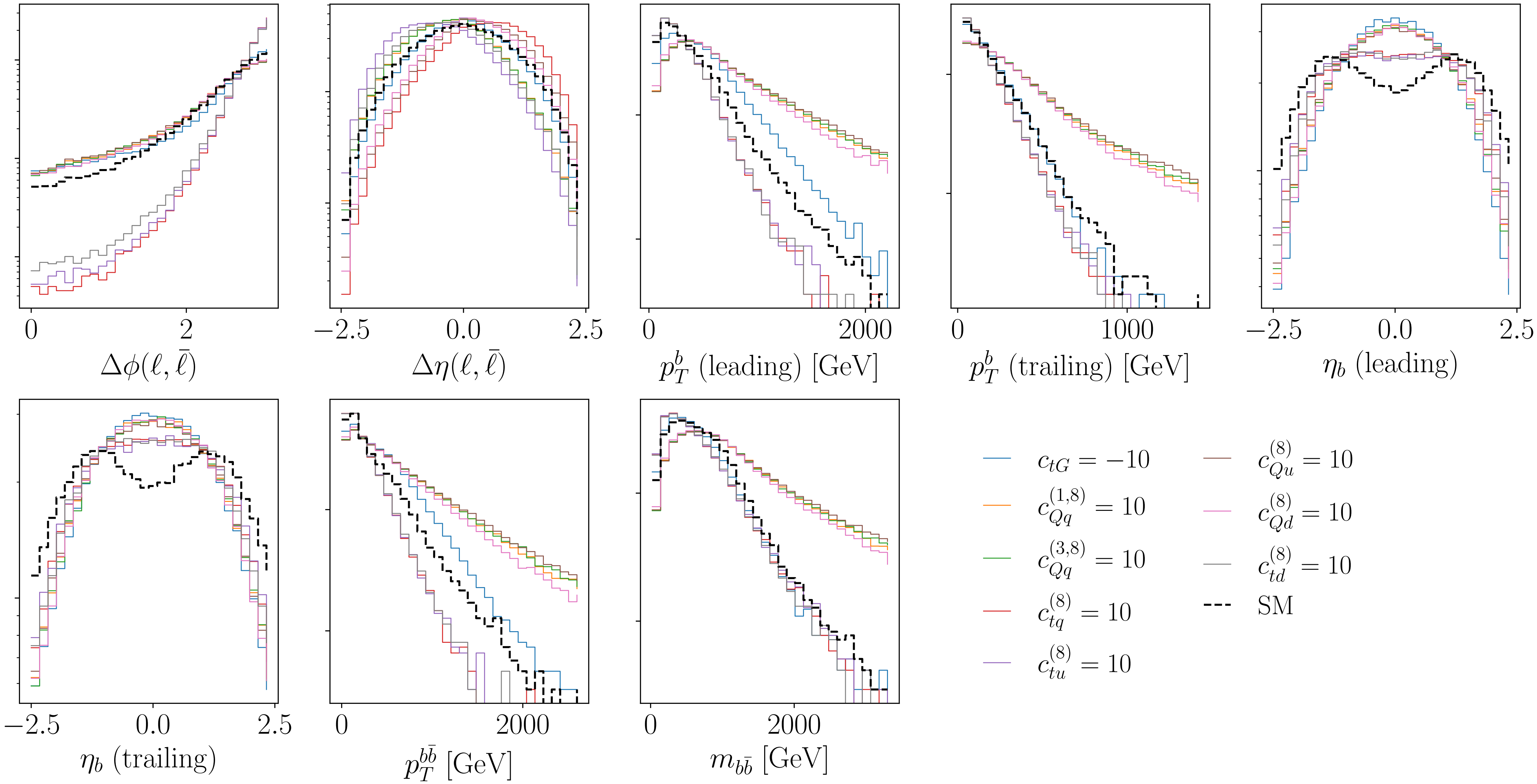
$$\hat{r}^{(i)}(\boldsymbol{x}, \boldsymbol{c}) \equiv 1 + \sum_{j=1}^{n_{\mathrm{eft}}} \mathrm{NN}_i^{(j)}(\boldsymbol{x}) c_j + \sum_{j=1}^{n_{\mathrm{eft}}} \sum_{k \geq j}^{n_{\mathrm{eft}}} \mathrm{NN}_i^{(j,k)}(\boldsymbol{x}) c_j c_k \,, \qquad i = 1, \ldots, N_{\mathrm{rep}}$$



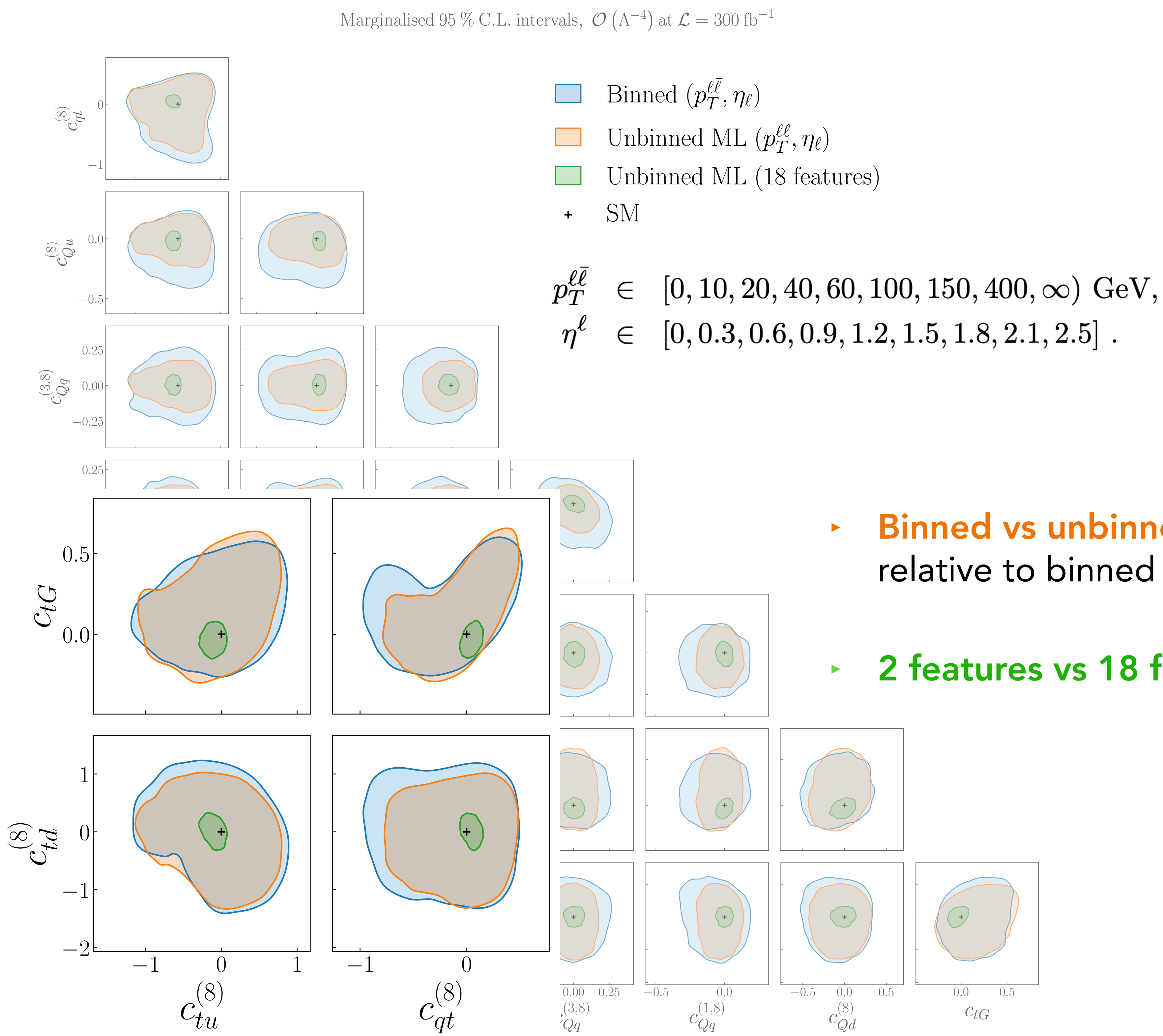
Process	N_{rep}	$\widetilde{N}_{\mathrm{ev}}$ (per replica)	N_{nn}	#trainings
$pp \to t\bar{t}$	50	10^5	4	200
$pp \to t\bar{t} \to b\bar{b}\ell^+\ell^-\nu_\ell\bar{\nu}_\ell$	25	10^5	40	1000

Let's go multivariate

- $pp \to t\bar{t} \to b\bar{b}\ell^+\ell^-\nu_\ell\bar{\nu}_\ell$: 18 features, 8 EFT coefficients
- $pp \to hZ \to b\bar{b}\ell^+\ell^-$: 7 features, 7 EFT coefficients

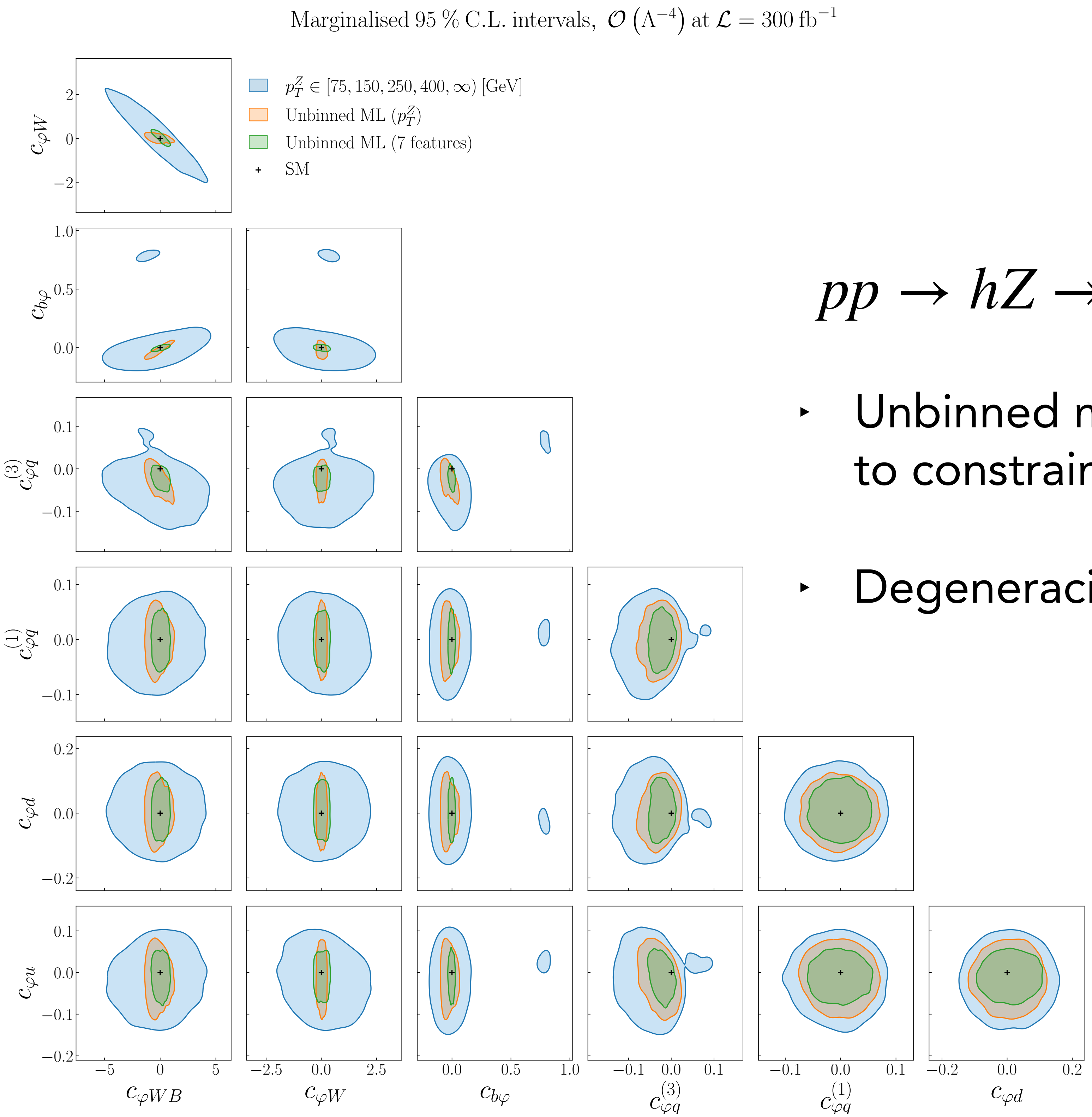


Unbinned observables in the top sector



- **Binned vs unbinned** in $(p_T^{\ell\bar{\ell}}, \eta_\ell)$ small improvement relative to binned setup
- **2 features vs 18 features:** big increase in sensitivity

Unbinned observables in Higgs + Z associated production

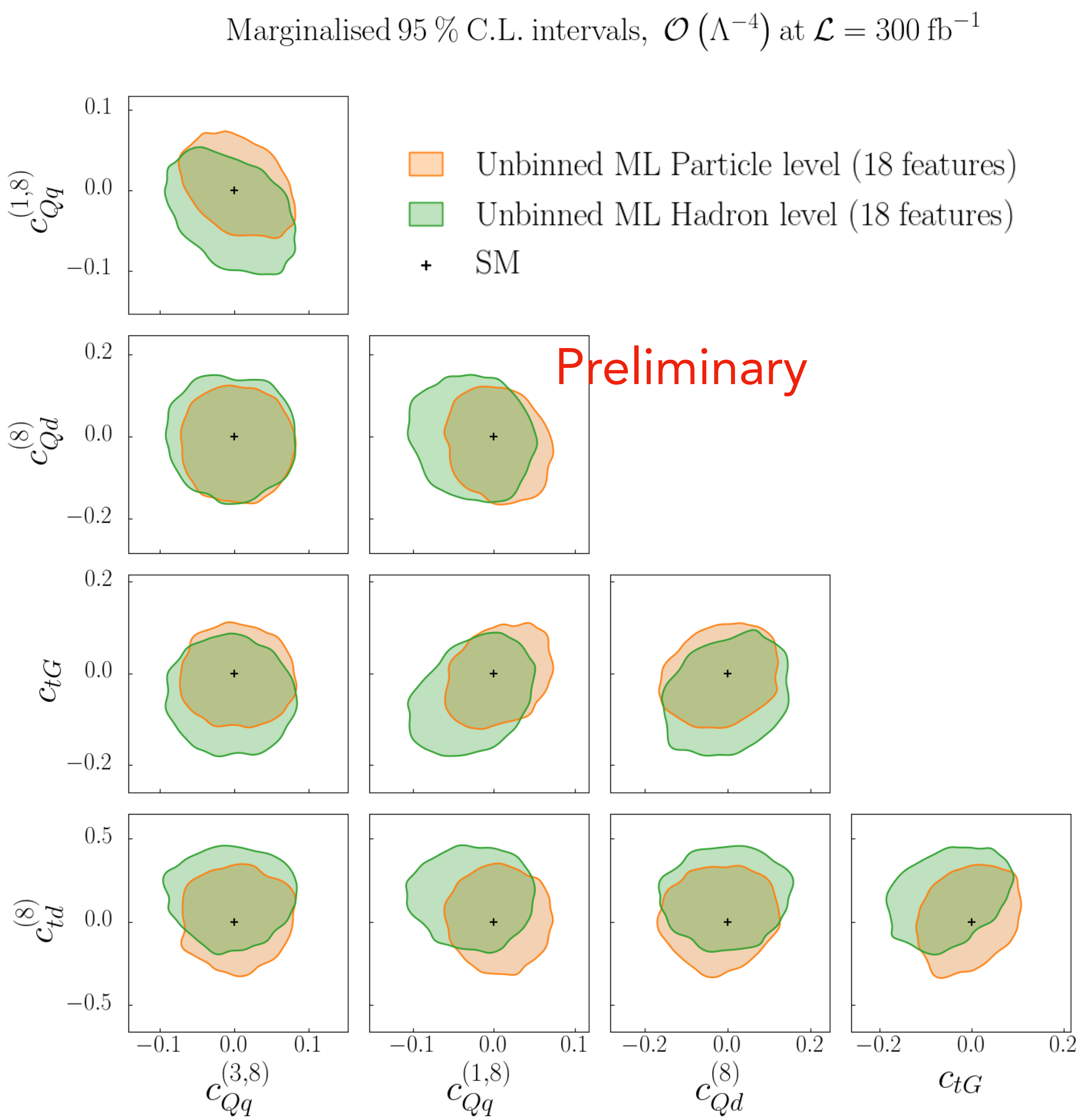


$$pp \to hZ \to b\bar{b}\ell^+\ell^-$$

- Unbinned multivariate data is advantageous to constrain the EFT parameter space
- Degeneracies get lifted

Ongoing efforts

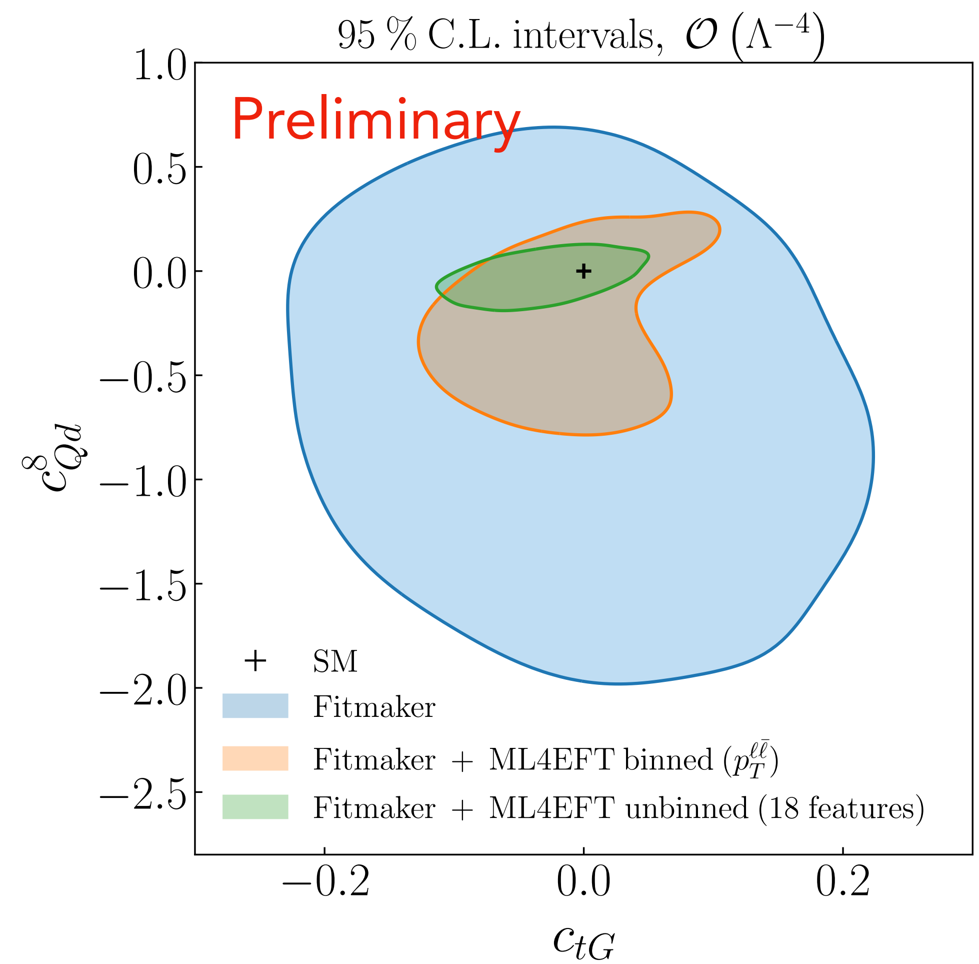
1. Hadronised level



MSc project by Pim Herbschleb

2. Integration into global fits

$$\log\mathcal{L}(c) = \sum_{k=1}^{N_D^{(\text{unbinned})}} \log\mathcal{L}_k^{\text{unbinned}}(c) + \sum_{k=1}^{N_D^{(\text{binned})}} \log\mathcal{L}_k^{\text{binned}}(c)$$



Applications of likelihood learning

Focusses on global EFT fits

Nikhef-2022-015

Unbinned multivariate observables for global SMEFT analyses from machine learning

Raquel Gomez Ambrosio,[1,2] Jaco ter Hoeve,[3,4] Maeve Madigan,[5] Juan Rojo,[3,4] and Veronica Sanz[6,7]

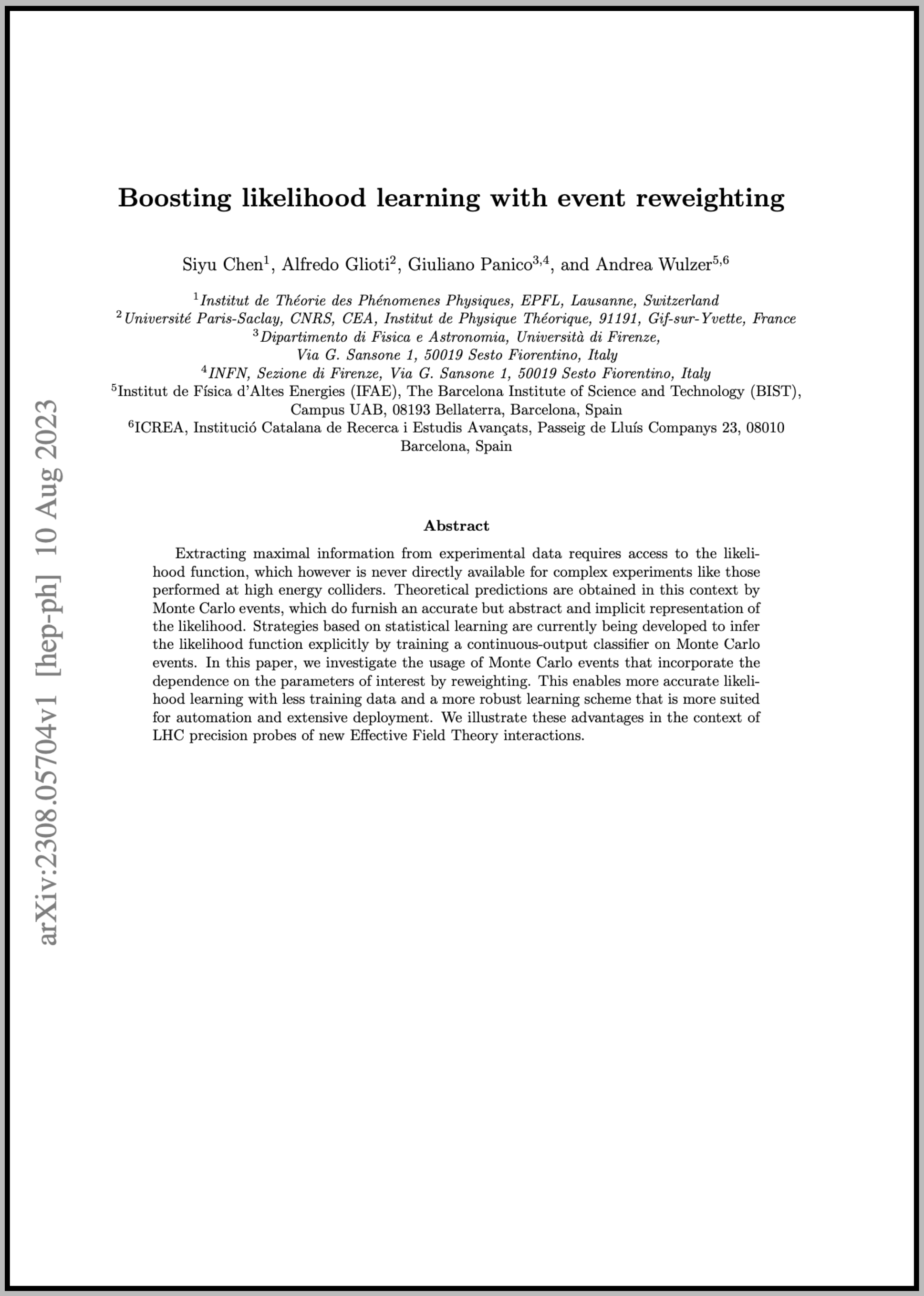
[1] *Dipartimento di Fisica "G. Occhialini", Universita degli Studi di Milano-Bicocca, and INFN, Sezione di Milano Bicocca, Piazza della Scienza 3, I – 20126 Milano, Italy*
[2] *Dipartimento di Fisica, Universitá di Torino, and INFN, Sezione di Torino, Via P. Giuria 1, 10125 Torino, Italy*
[3] *Department of Physics and Astronomy, VU Amsterdam, 1081HV Amsterdam, The Netherlands*
[4] *Nikhef Theory Group, Science Park 105, 1098 XG Amsterdam, The Netherlands*
[5] *DAMTP, University of Cambridge, Wilberforce Road, Cambridge CB3 0WA, UK*
[6] *Instituto de Física Corpuscular (IFIC), Universidad de Valencia-CSIC, E-46980 Valencia, Spain*
[7] *Department of Physics and Astronomy, University of Sussex, Brighton BN1 9QH, UK*

Abstract

Theoretical interpretations of particle physics data, such as the determination of the Wilson coefficients of the Standard Model Effective Field Theory (SMEFT), often involve the inference of multiple parameters from a global dataset. Optimizing such interpretations requires the identification of observables that exhibit the highest possible sensitivity to the underlying theory parameters. In this work we develop a flexible open source framework, ML4EFT, enabling the integration of unbinned multivariate observables into global SMEFT fits. As compared to traditional measurements, such observables enhance the sensitivity to the theory parameters by preventing the information loss incurred when binning in a subset of final-state kinematic variables. Our strategy combines machine learning regression and classification techniques to parameterize high-dimensional likelihood ratios, using the Monte Carlo replica method to estimate and propagate methodological uncertainties. As a proof of concept we construct unbinned multivariate observables for top-quark pair and Higgs+Z production at the LHC, demonstrate their impact on the SMEFT parameter space as compared to binned measurements, and study the improved constraints associated to multivariate inputs. Since the number of neural networks to be trained scales quadratically with the number of parameters and can be fully parallelized, the ML4EFT framework is well-suited to construct unbinned multivariate observables which depend on up to tens of EFT coefficients, as required in global fits.

1

Reweighting for more accurate learning



Boosting likelihood learning with event reweighting

Siyu Chen[1], Alfredo Glioti[2], Giuliano Panico[3,4], and Andrea Wulzer[5,6]

[1] *Institut de Théorie des Phénomenes Physiques, EPFL, Lausanne, Switzerland*
[2] *Université Paris-Saclay, CNRS, CEA, Institut de Physique Théorique, 91191, Gif-sur-Yvette, France*
[3] *Dipartimento di Fisica e Astronomia, Università di Firenze, Via G. Sansone 1, 50019 Sesto Fiorentino, Italy*
[4] *INFN, Sezione di Firenze, Via G. Sansone 1, 50019 Sesto Fiorentino, Italy*
[5] Institut de Física d'Altes Energies (IFAE), The Barcelona Institute of Science and Technology (BIST), Campus UAB, 08193 Bellaterra, Barcelona, Spain
[6] ICREA, Institució Catalana de Recerca i Estudis Avançats, Passeig de Lluís Companys 23, 08010 Barcelona, Spain

Abstract

Extracting maximal information from experimental data requires access to the likelihood function, which however is never directly available for complex experiments like those performed at high energy colliders. Theoretical predictions are obtained in this context by Monte Carlo events, which do furnish an accurate but abstract and implicit representation of the likelihood. Strategies based on statistical learning are currently being developed to infer the likelihood function explicitly by training a continuous-output classifier on Monte Carlo events. In this paper, we investigate the usage of Monte Carlo events that incorporate the dependence on the parameters of interest by reweighting. This enables more accurate likelihood learning with less training data and a more robust learning scheme that is more suited for automation and extensive deployment. We illustrate these advantages in the context of LHC precision probes of new Effective Field Theory interactions.