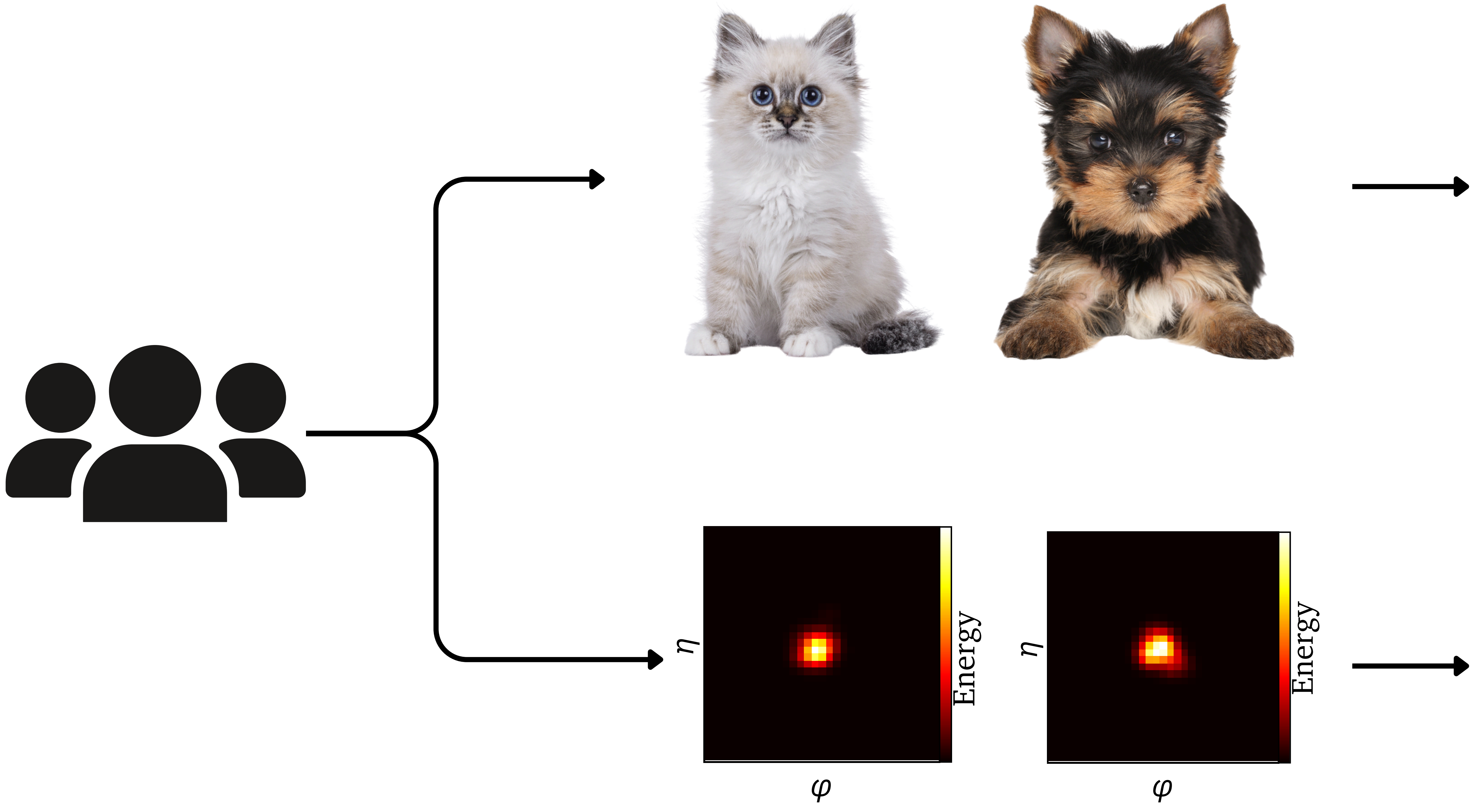
Deep learning the QCD of quark gluon tagging

CRC TRR 257 young scientist meeting 2025 | Heidelberg

Sophia Vent | Institute for theoretical physics Heidelberg
Ramon Winterhalder & Tilman Plehn

UNIVERSITÄT HEIDELBERG
ZUKUNFT SEIT 1386

Motivation: Classification



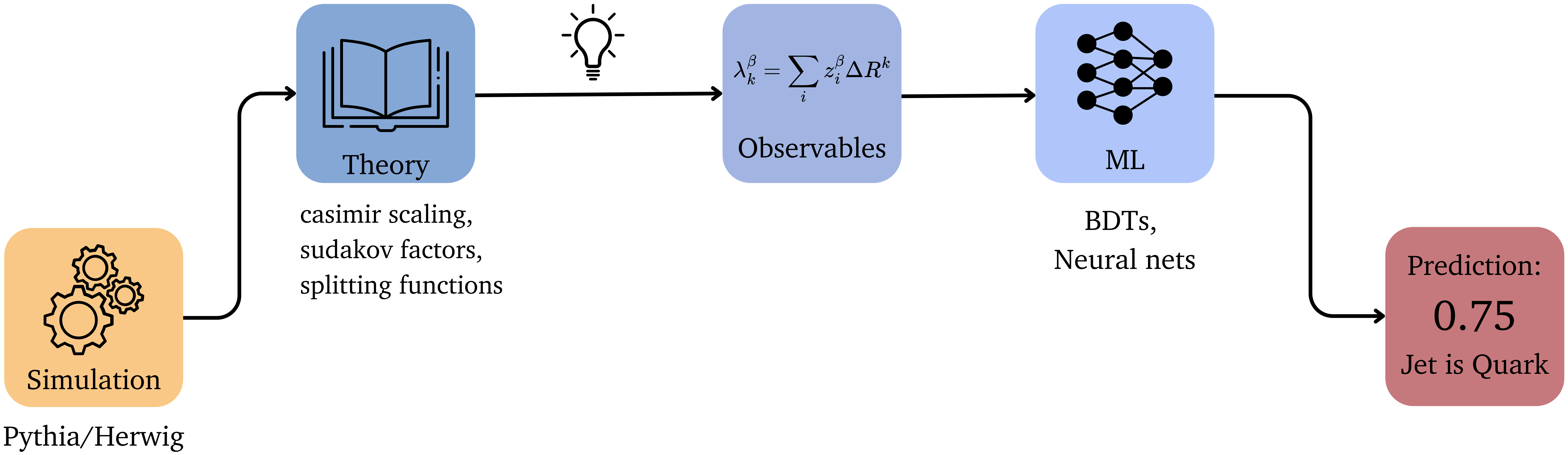
well defined

no cut-ground truth

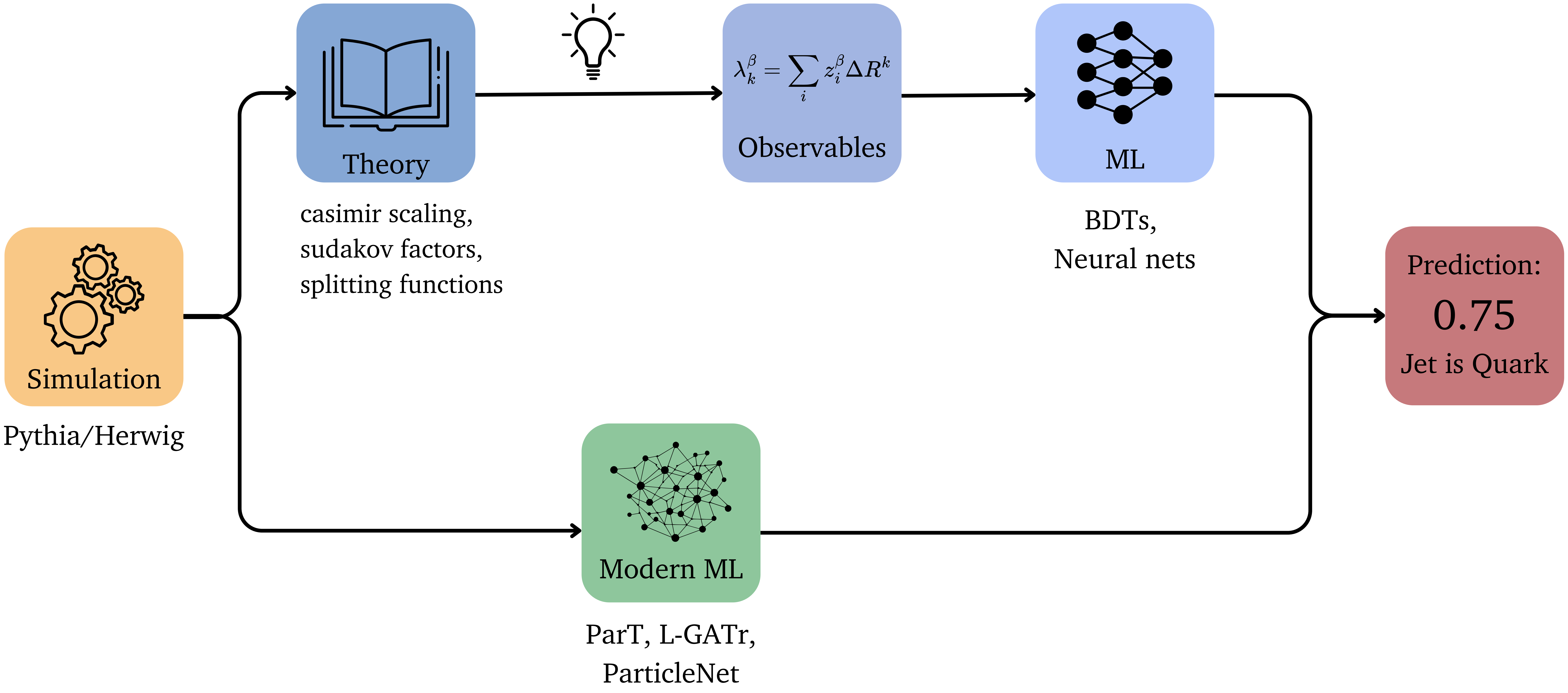
quark and gluon jets have a complex signature

⟶ use Machine learning (ML) techniques

Motivation

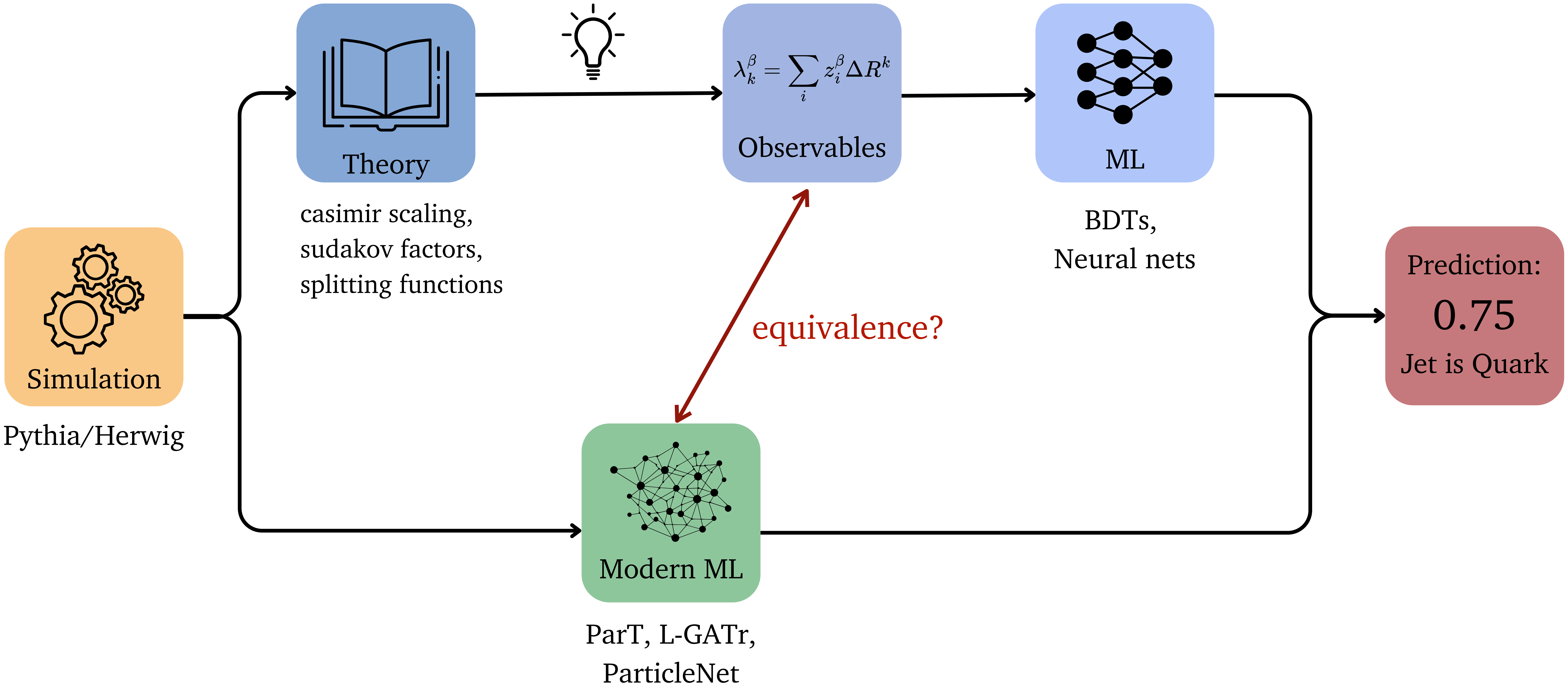

Simulation
Pythia/Herwig
Theory
casimir scaling,
sudakov factors,
splitting functions
$\lambda_k^\beta = \sum_i z_i^\beta \Delta R^k$
Observables
ML
BDTs,
Neural nets
Prediction:
0.75
Jet is Quark

Motivation



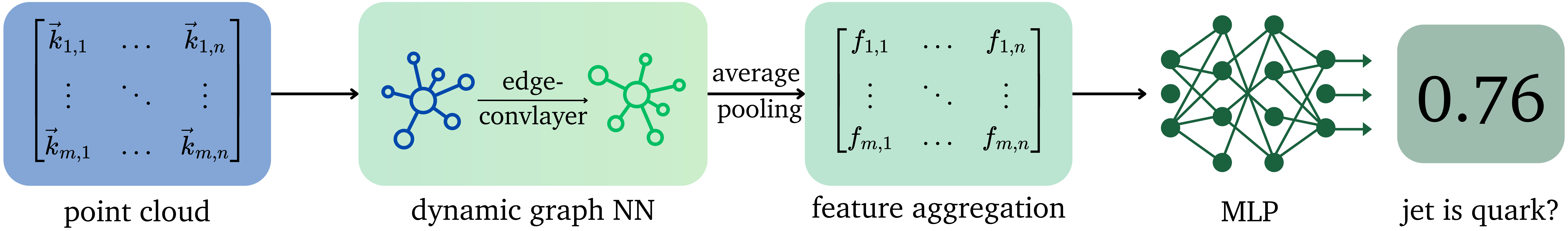
Simulation
Pythia/Herwig
Theory
casimir scaling,
sudakov factors,
splitting functions
$\lambda_k^\beta = \sum_i z_i^\beta \Delta R^k$
Observables
ML
BDTs,
Neural nets
Modern ML
ParT, L-GATr,
ParticleNet
Prediction:
0.75
Jet is Quark

Motivation

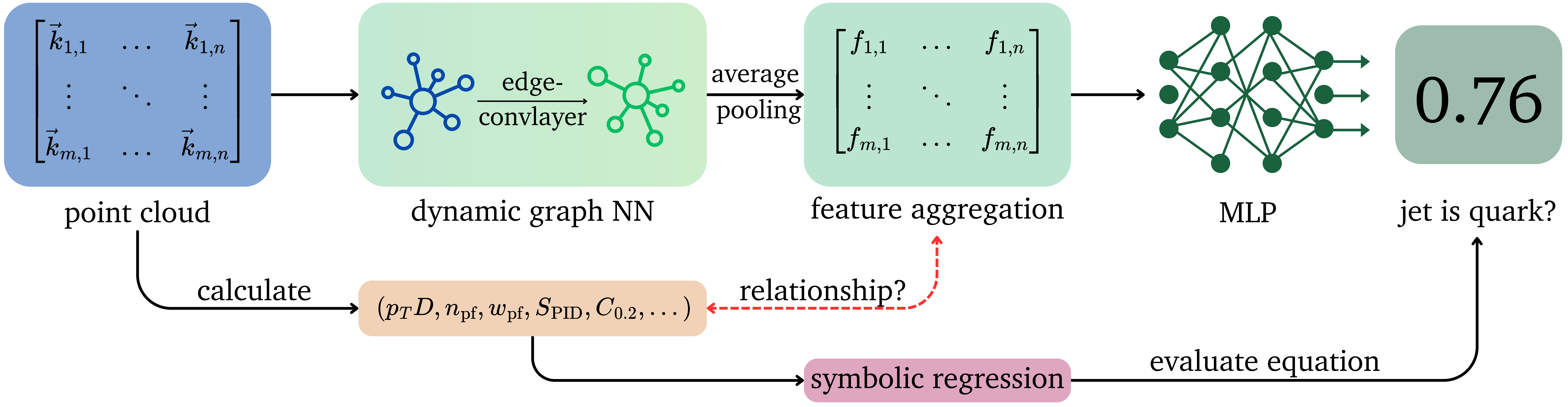


Theory
casimir scaling,
sudakov factors,
splitting functions
$\lambda_k^\beta = \sum_i z_i^\beta \Delta R^k$
Observables
ML
BDTs,
Neural nets
Simulation
Pythia/Herwig
equivalence?
Prediction:
0.75
Jet is Quark
Modern ML
ParT, L-GATr,
ParticleNet

ParticleNet Architecture



ParticleNet architecture



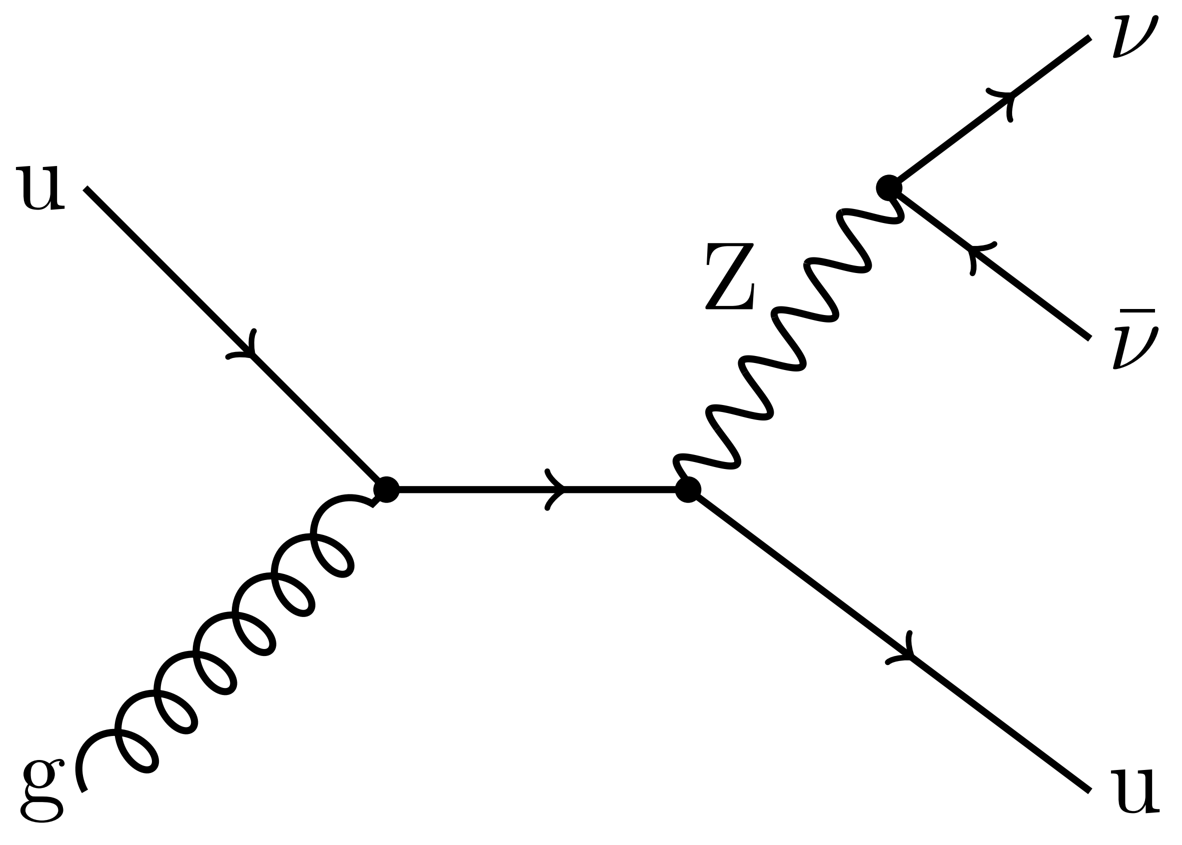
1) analyze latent structure

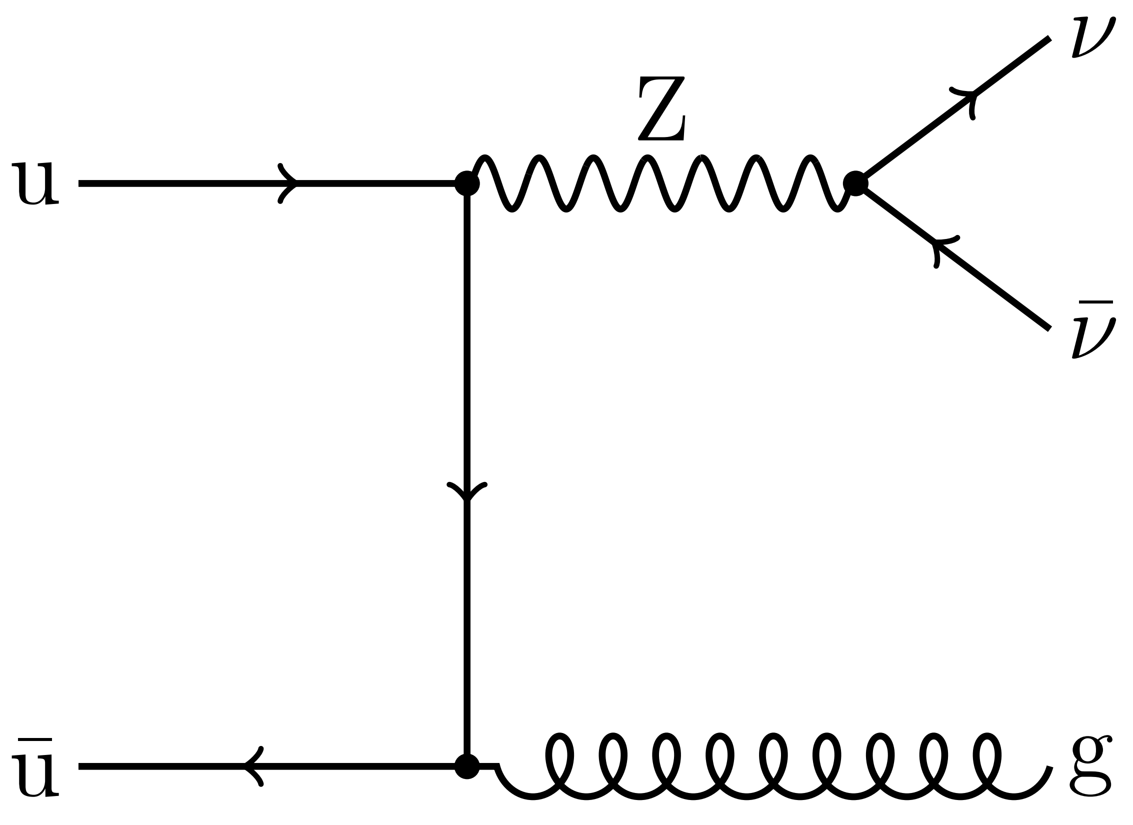
2) turning the neural network into a physics equations

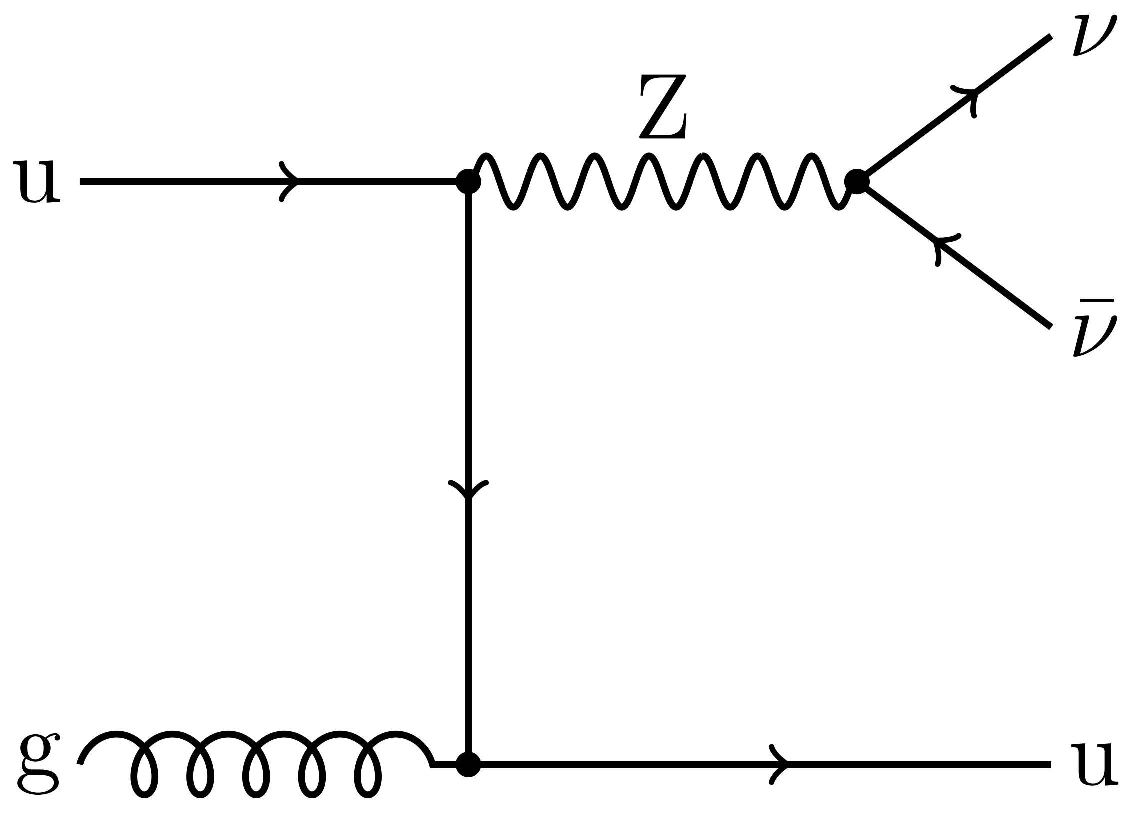
Dataset

2 million jets, quarks and gluons generated with Pythia or Herwig

$q\bar{q} \to Z\,(\to \nu\bar{\nu}) + g$ and $qg \to Z\,(\to \nu\bar{\nu}) + (uds)$

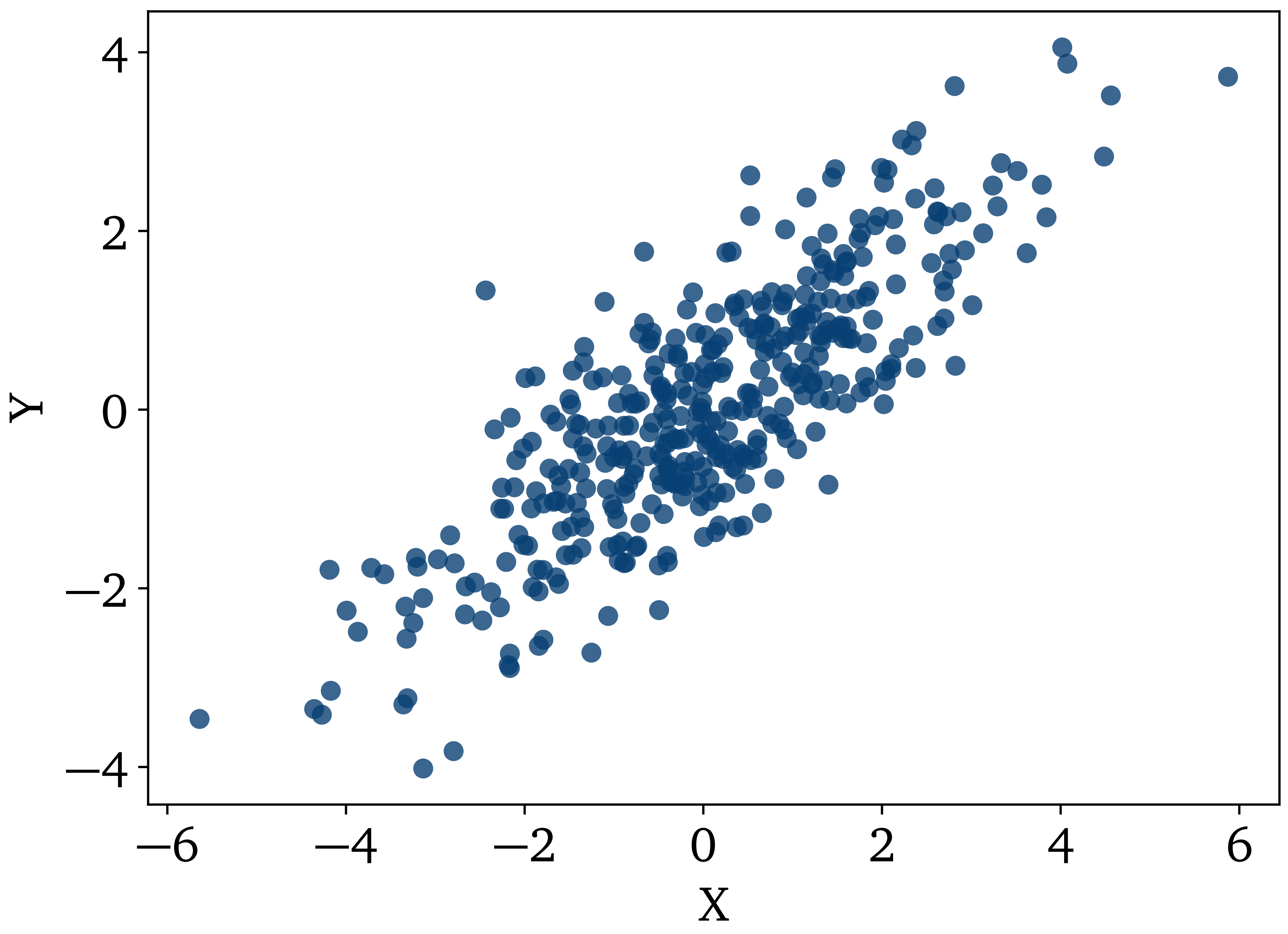




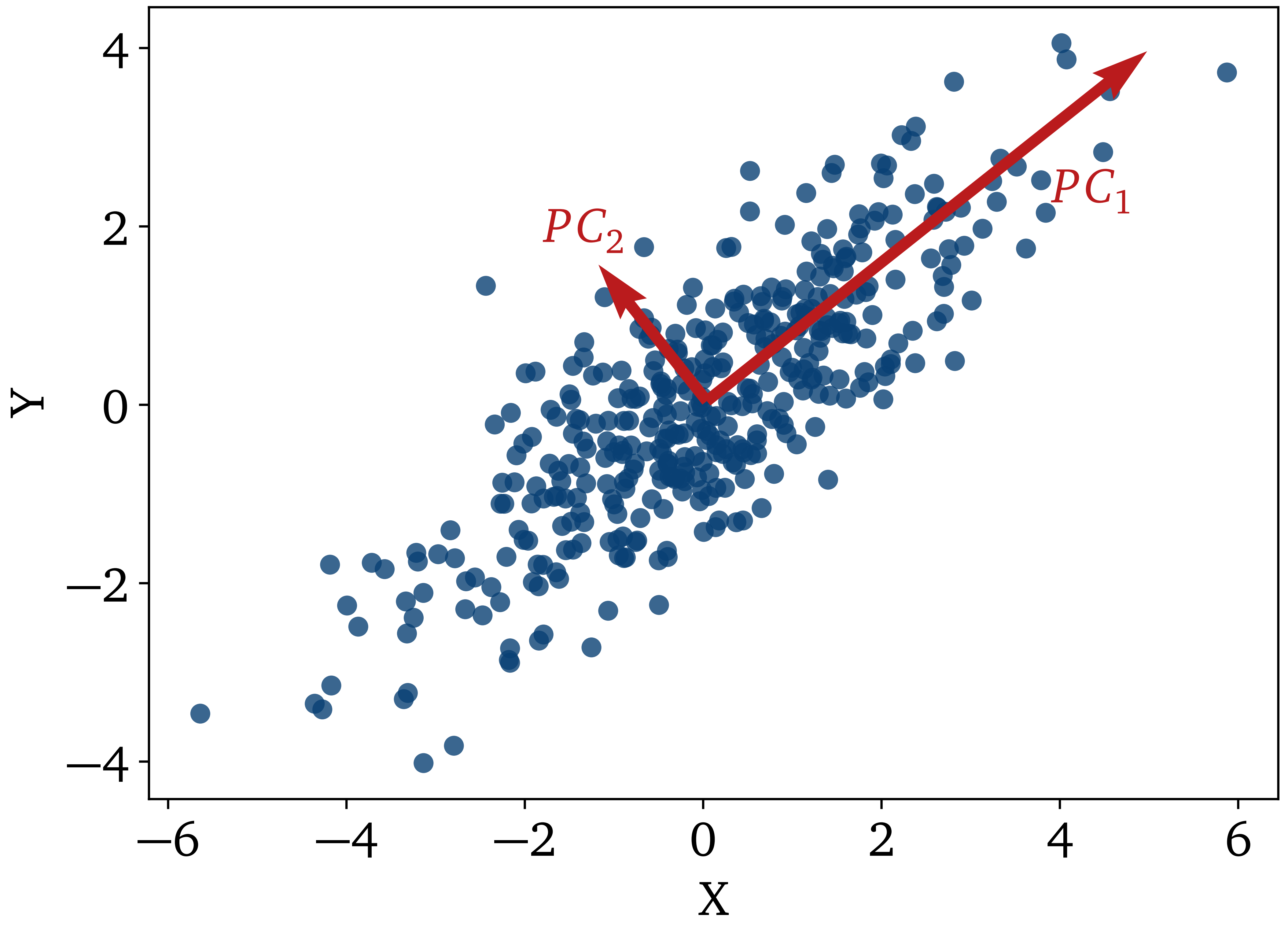


Dimensionality reduction

Principal component analysis (PCA)



Principal component analysis (PCA)



X

Basis transformation:

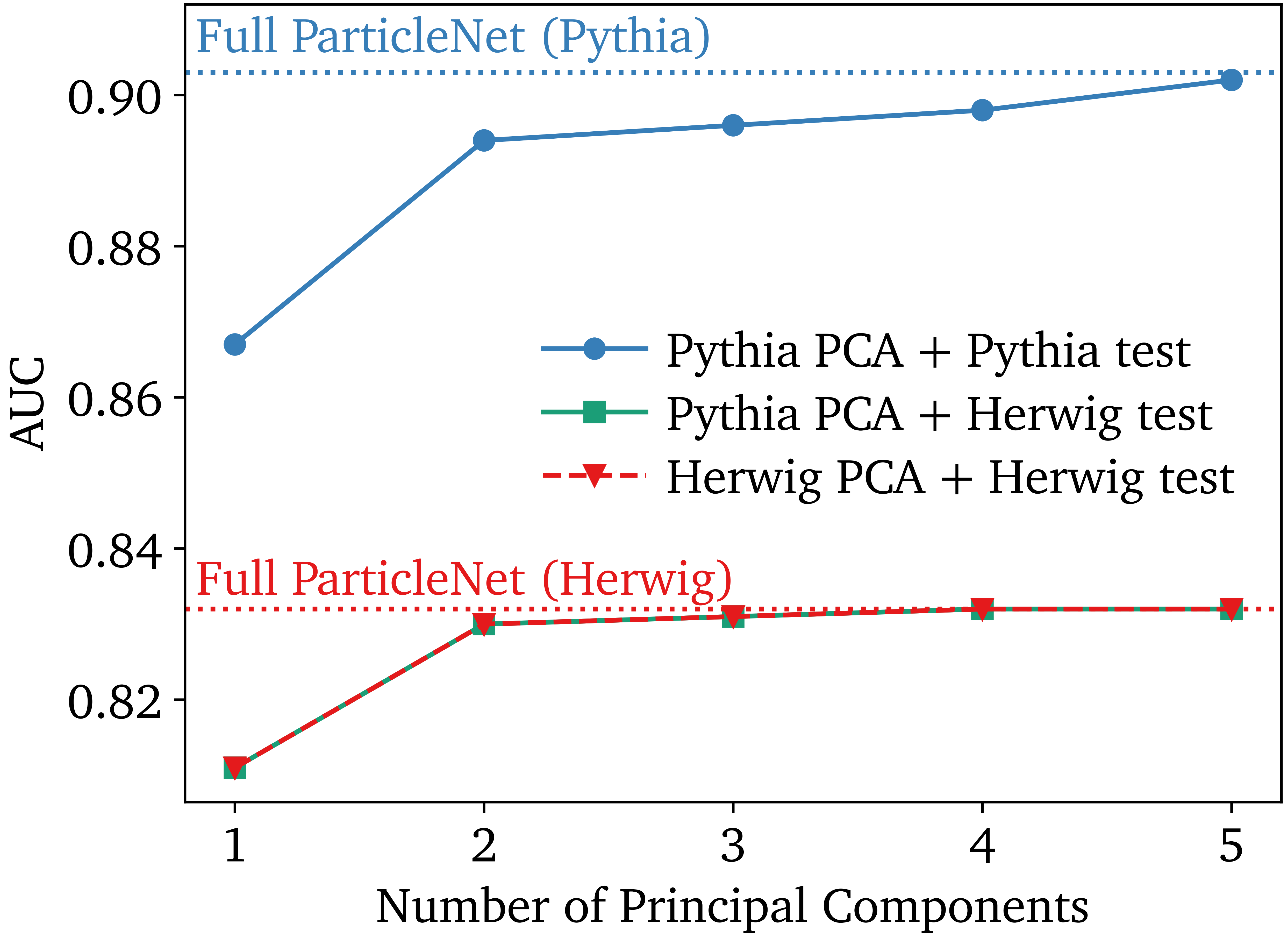
- find new axis that capture the variance
- reduce dimensionality

Dimensionality reduction using PCA

Dimensionality reduction:

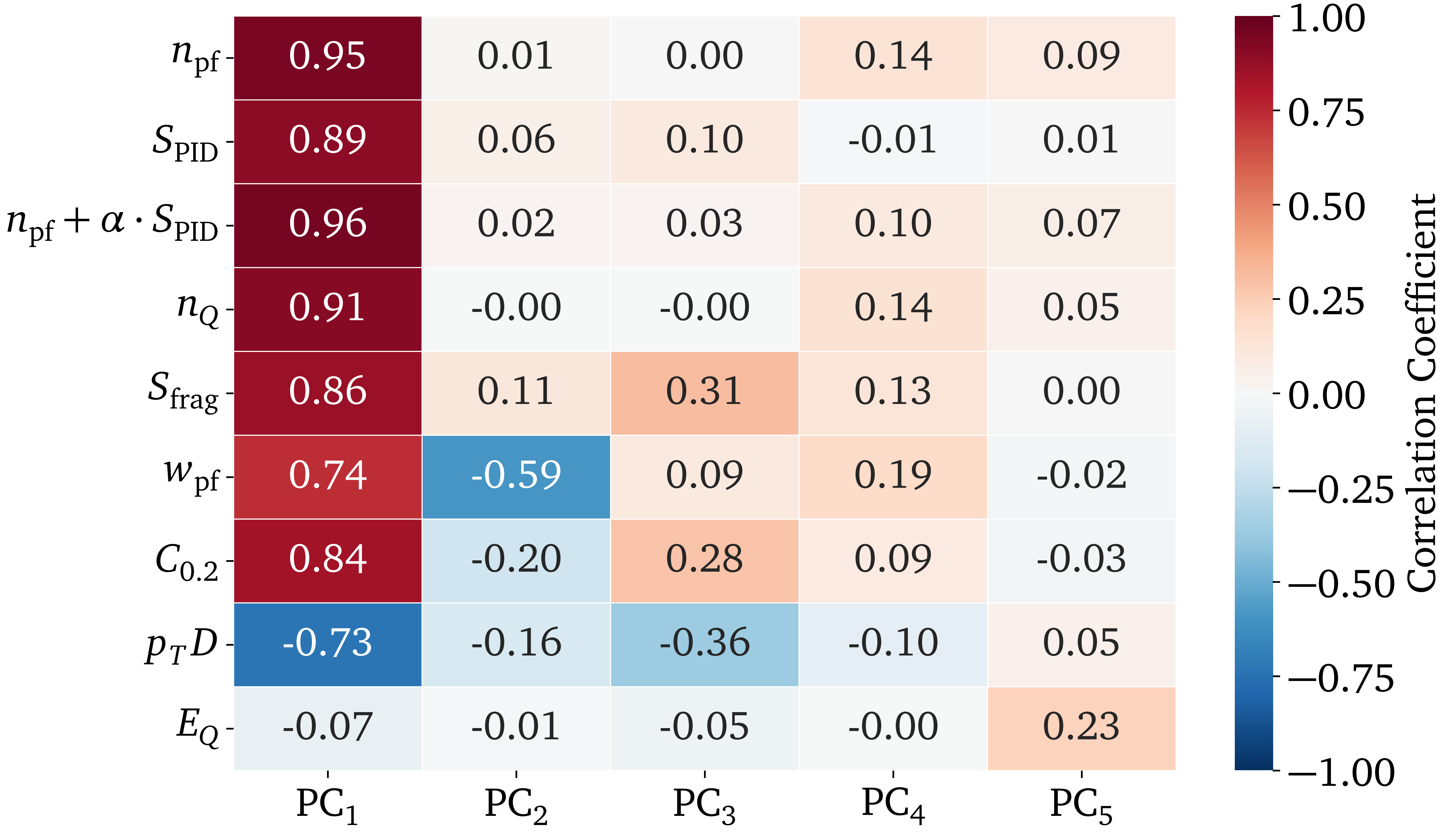
$64 \rightarrow 5$

even across generators



From latent features to observables

From latent features to Observables

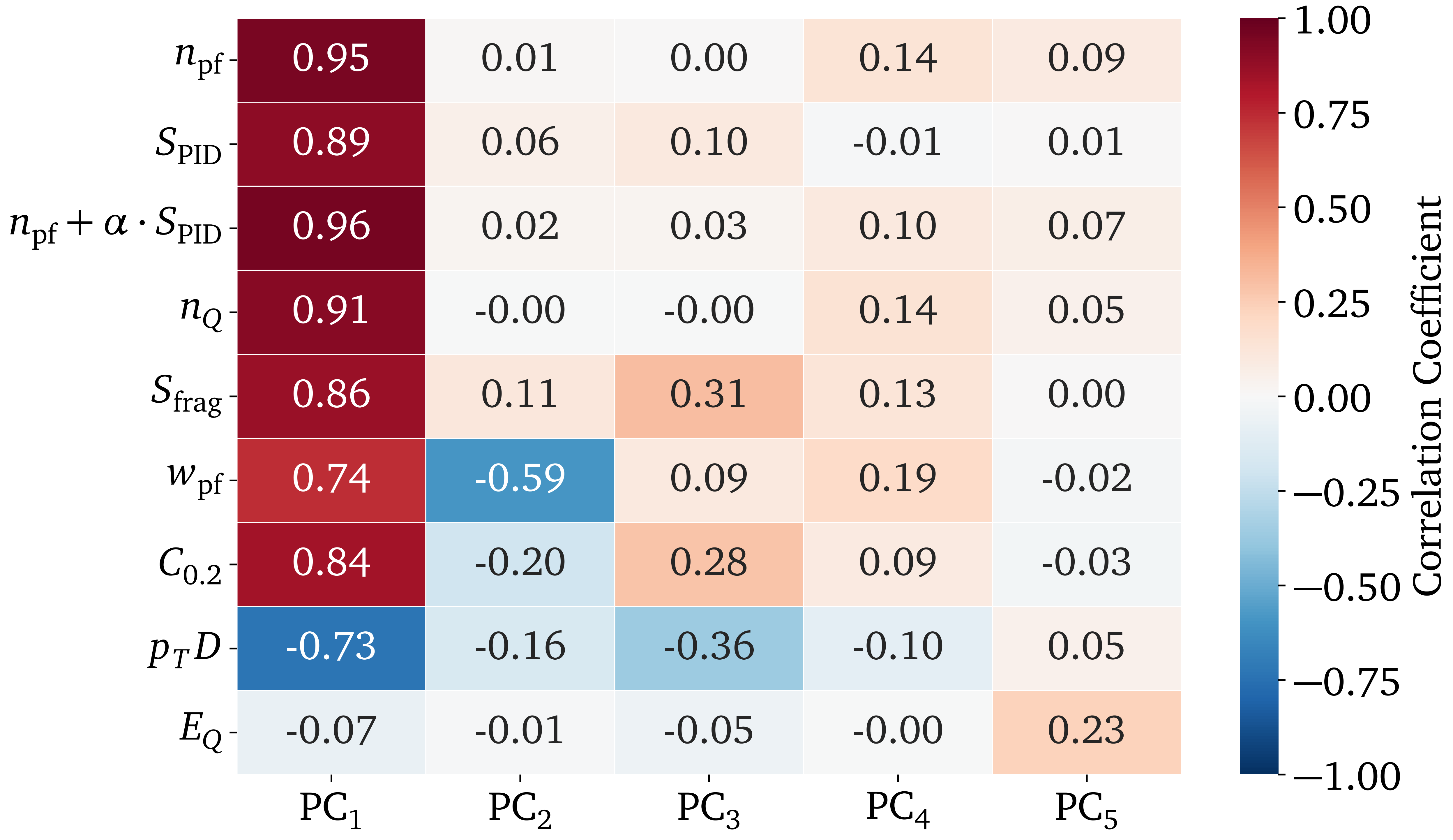


PC1 : particle multiplicity and particle diversity

shannon particle ID entropy :

$$S_{\mathrm{PID}} = \sum p_i \log(p_i)$$

From latent features to Observables



Remaining PC Observables:

- have to be decorrellated from previous PC directions

PC2:

- Shape related feature

PC3:

- Fragmentation related feature

Feature engineering

Generalized angularities:

$$\lambda_k^\beta = \sum_i z_i^\beta \Delta R^k \qquad z_i = \frac{p_{T,i}}{p_{T,\mathrm{jet}}}$$

New observables:

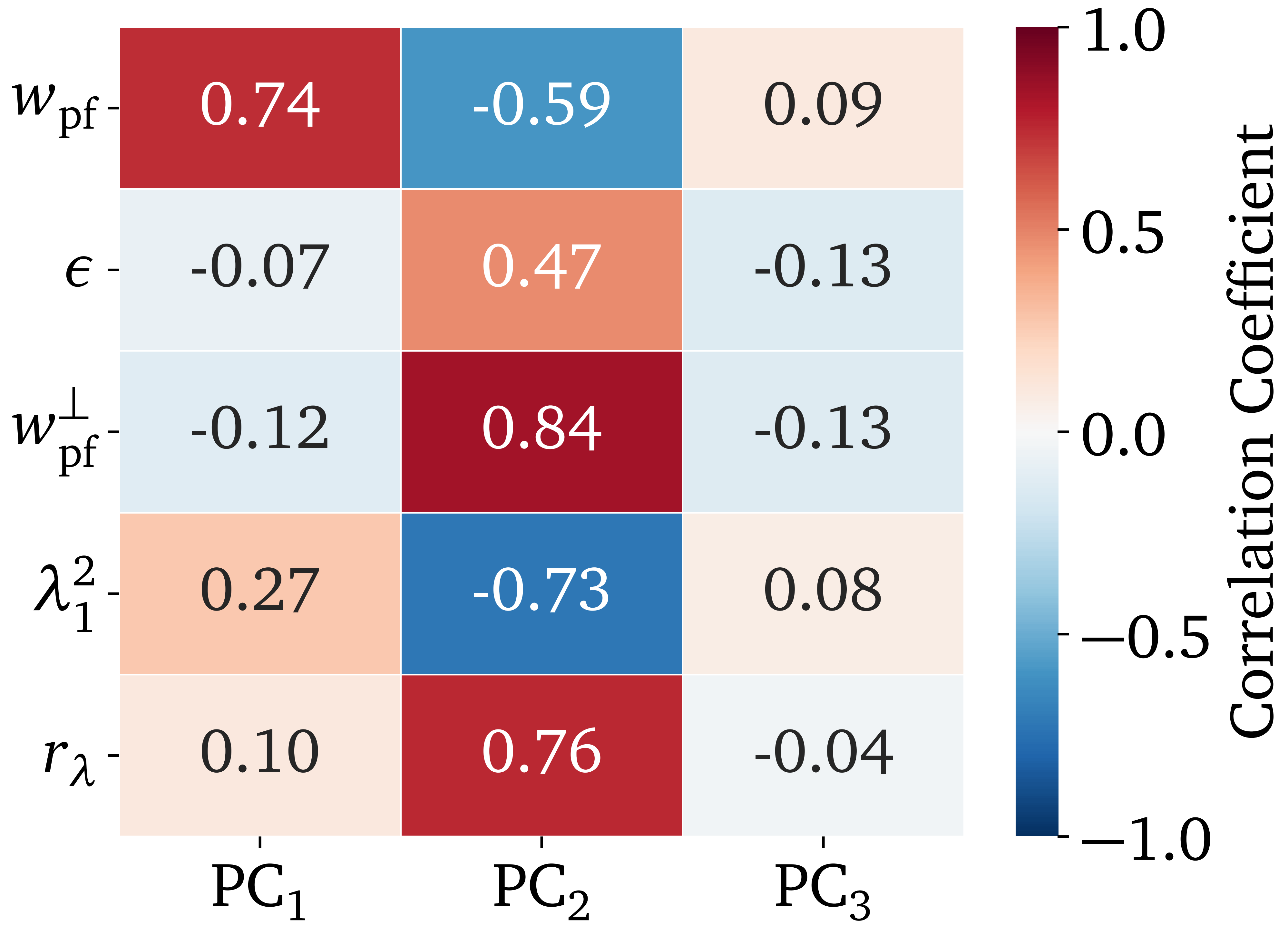
$$r_\lambda = \frac{\lambda_{0.5}^1}{\lambda_1^2}$$

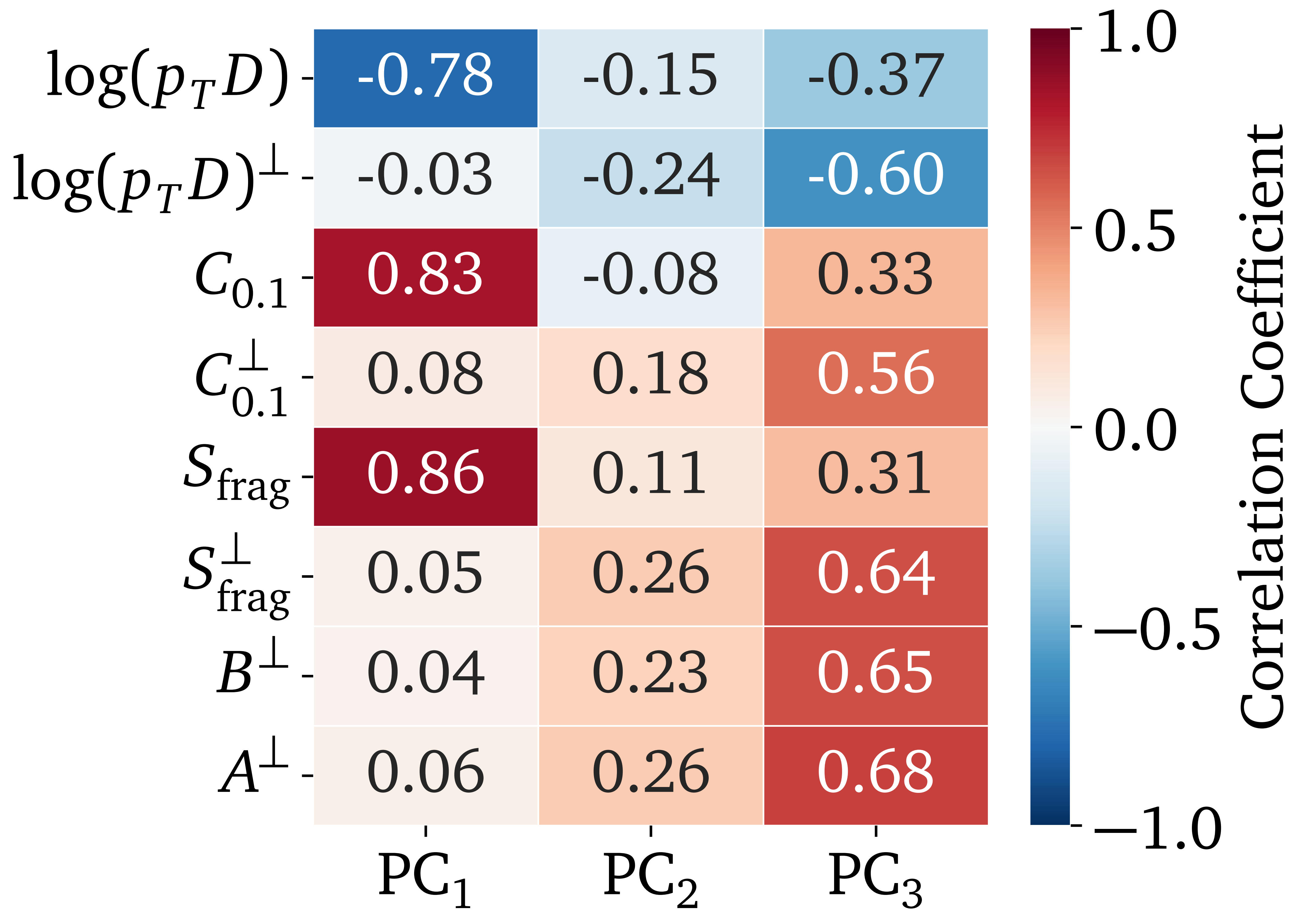
Linear decorellation by hand:

$$O_{\mathrm{PC}k}^\perp = O - \sum_{i<k} c_k \cdot \mathrm{PC}i$$

$$w_{\mathrm{pf}}^\perp = \alpha \cdot n_{\mathrm{pf}} - w_{\mathrm{pf}}$$

Feature engineering





Feature summary

- PC1: multiplicity and diversity
- PC2: shape and radial momentum spread
- PC3: fragmentation pattern
- PC5: charge

Introduction of new observables : S_{PID} & r_λ

Turning Neural Networks into Physics Equations

Symbolic Regression using PySR

A neural network is essentially a **complex function approximation**

The tagger prediction $y(\vec{x})$ maps an input vector $\vec{x}$ to an output

Symbolic regression (via PySR) finds an interpretable analytic equation

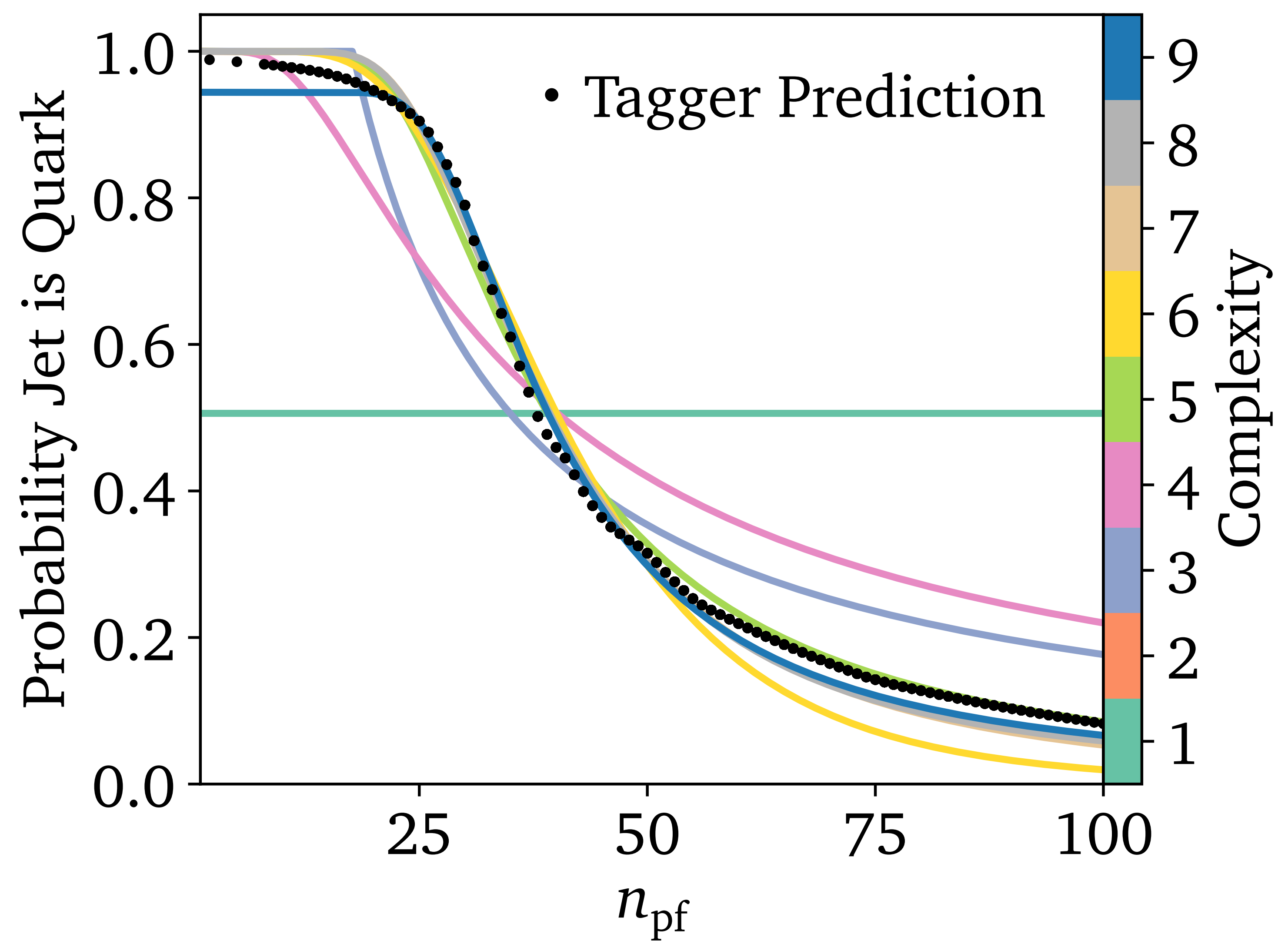
$$y(\vec{x}) \approx f(\vec{x})$$

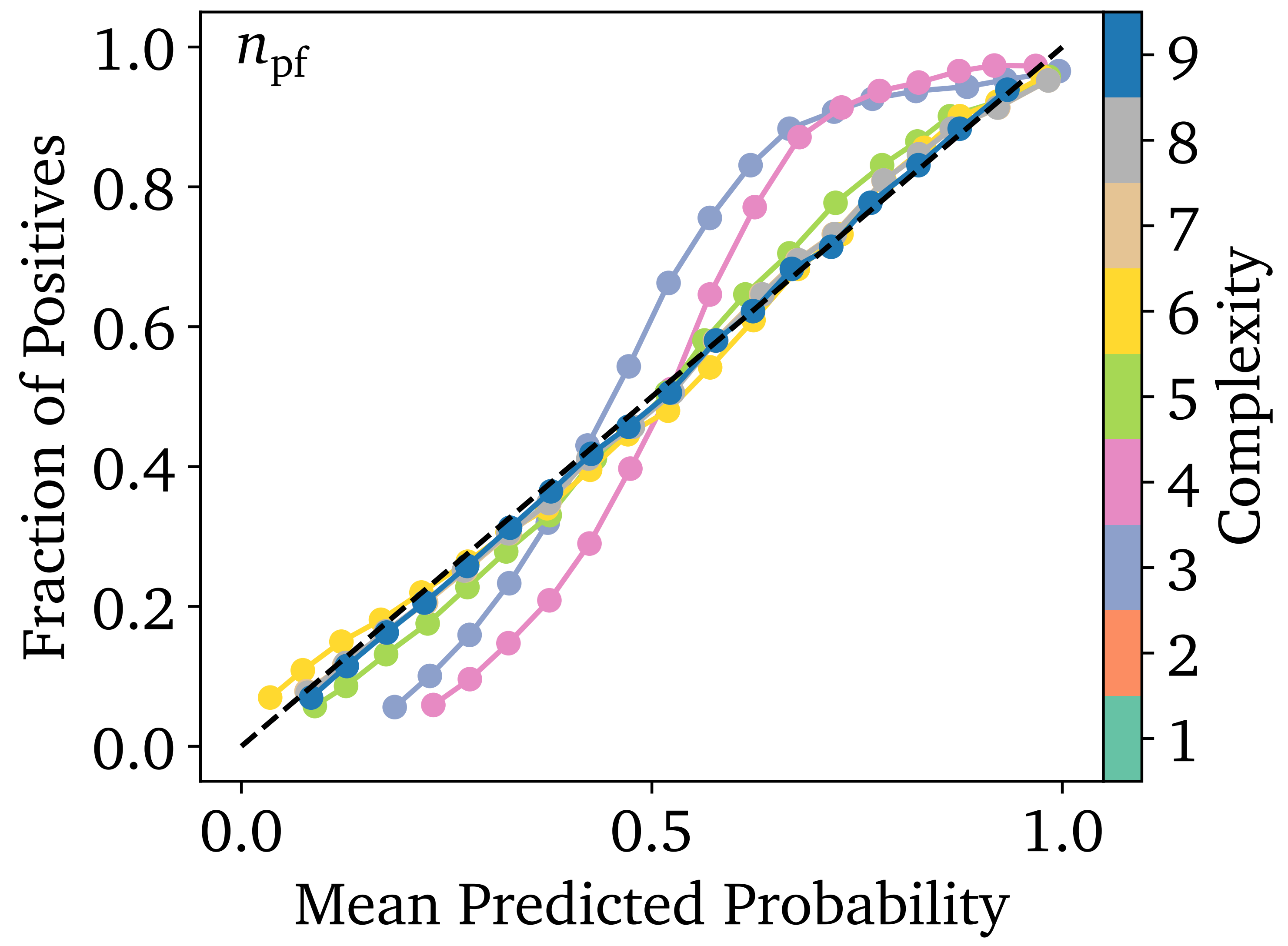
PySR uses a genetic algorithm to evolve symbolic expressions that connect inputs to outputs

One dimensional equation

Workflow:

- Train a 1D neural network
- Apply PySR
- evaluate equation (performance, calibration…)

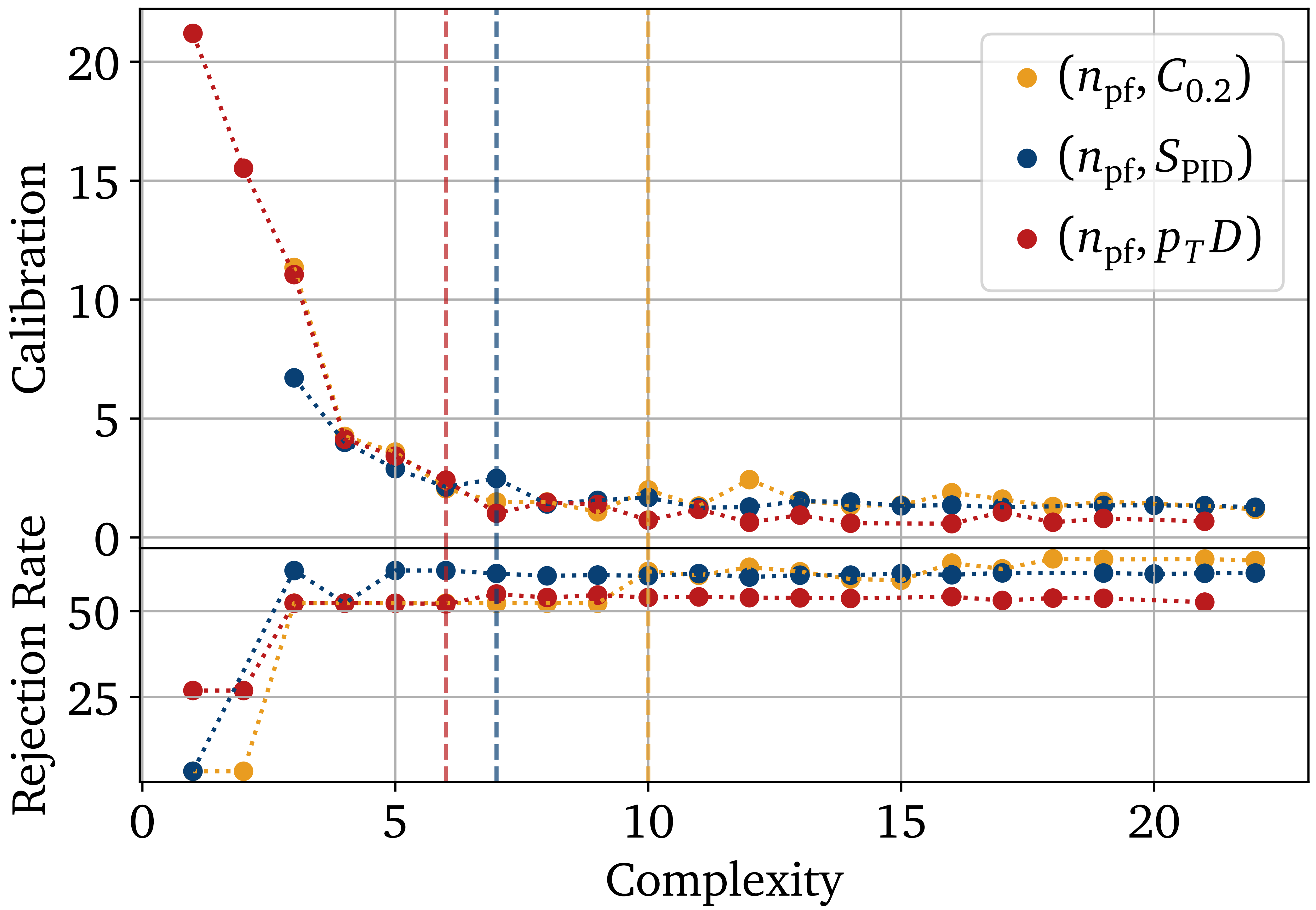


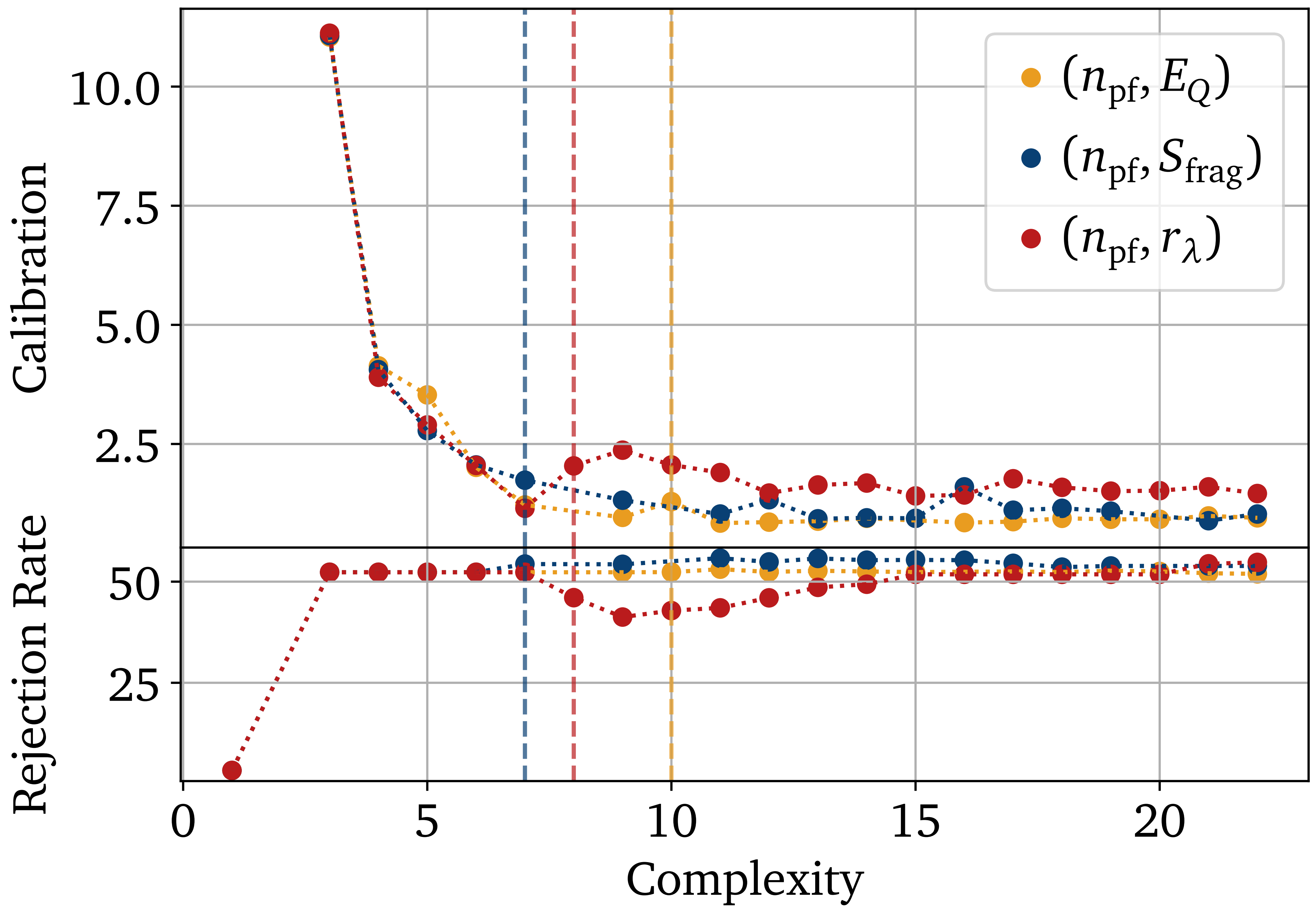


One dimensional equation

Observable	Equation
n_{pf}	$0.94 \cdot \tanh\left(21036 \cdot \left(0.005 + \frac{1}{n_{\mathrm{pf}}}\right)^3\right)$
S_{PID}	$\tanh^2\left(\frac{1.14}{S_{\mathrm{PID}}^3}\right)$
S_{frag}	$\tanh^2\left(\frac{18.08}{S_{\mathrm{frag}}^3}\right)$

Two dimensional equations





Full quark gluon equation

$$\tanh^3\left(0.55 \cdot C_{0.2} + 2 \cdot \left(-0.02 \cdot r_\lambda \cdot \left(C_{0.2} \cdot p_T D \cdot S_{\text{PID}} \cdot S_{\text{frag}} - 0.25\right) + 1\right)^3\right)$$

Performance:

- neural network AUC = 0.872
- SR equation AUC = 0.871

Full quark gluon equation

$$\tanh^3\left(0.55 \cdot C_{0.2} + 2 \cdot \left(-0.02 \cdot r_\lambda \cdot \left(C_{0.2} \cdot p_T D \cdot S_{\mathrm{PID}} \cdot S_{\mathrm{frag}} - 0.25\right) + 1\right)^3\right)$$

Problems:

- High computational cost for high complexities
- Scaling would boost performance but cost generalizability

Summary

- Introduced explainable AI tools
- Identification of latent features
- Introduction of new observables
- Surrogate equations for neural networks

Summary

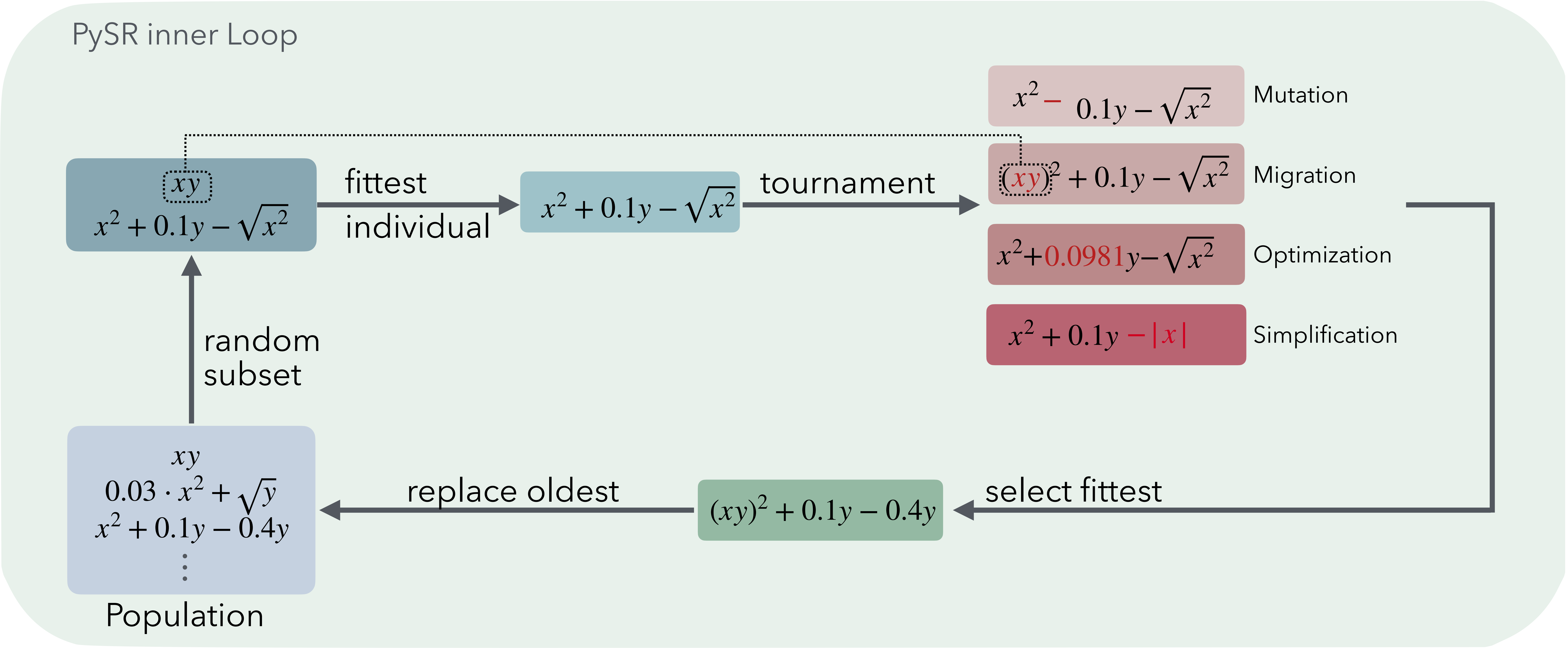
- Introduced explainable AI tools
- Identification of latent features
- Introduction of new observables
- Surrogate equations for neural networks

Stay tuned!

Thank you!

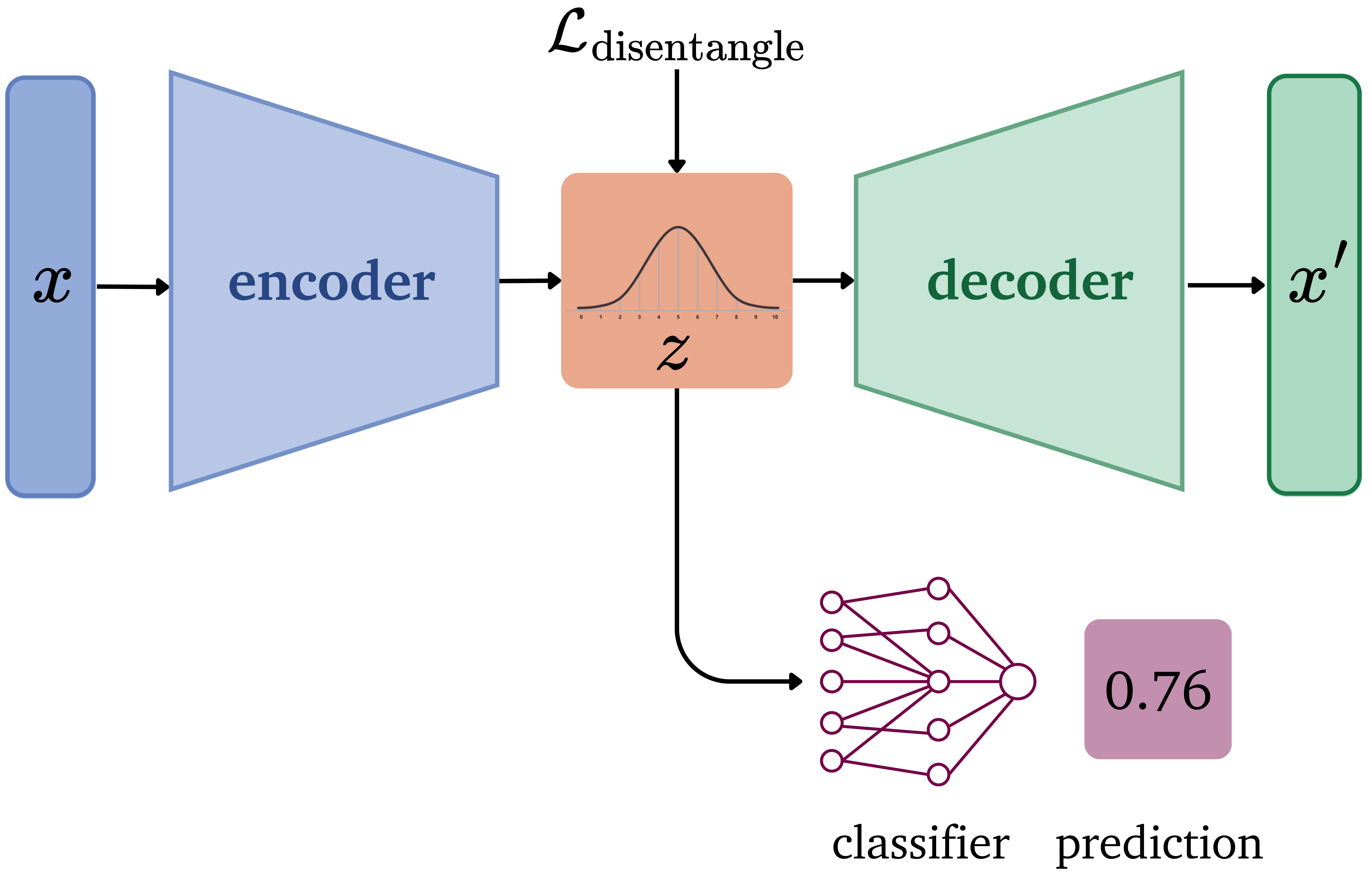
Additional slides

Symbolic regression using PySR



Non linear dimensionality reduction

Decorrelated Latent Classifier



Feature Attribution with SHAP

SHAP (SHapley Additive exPlanations)

Quantify each feature's contribution to the model's prediction

Shapely values from gametheory

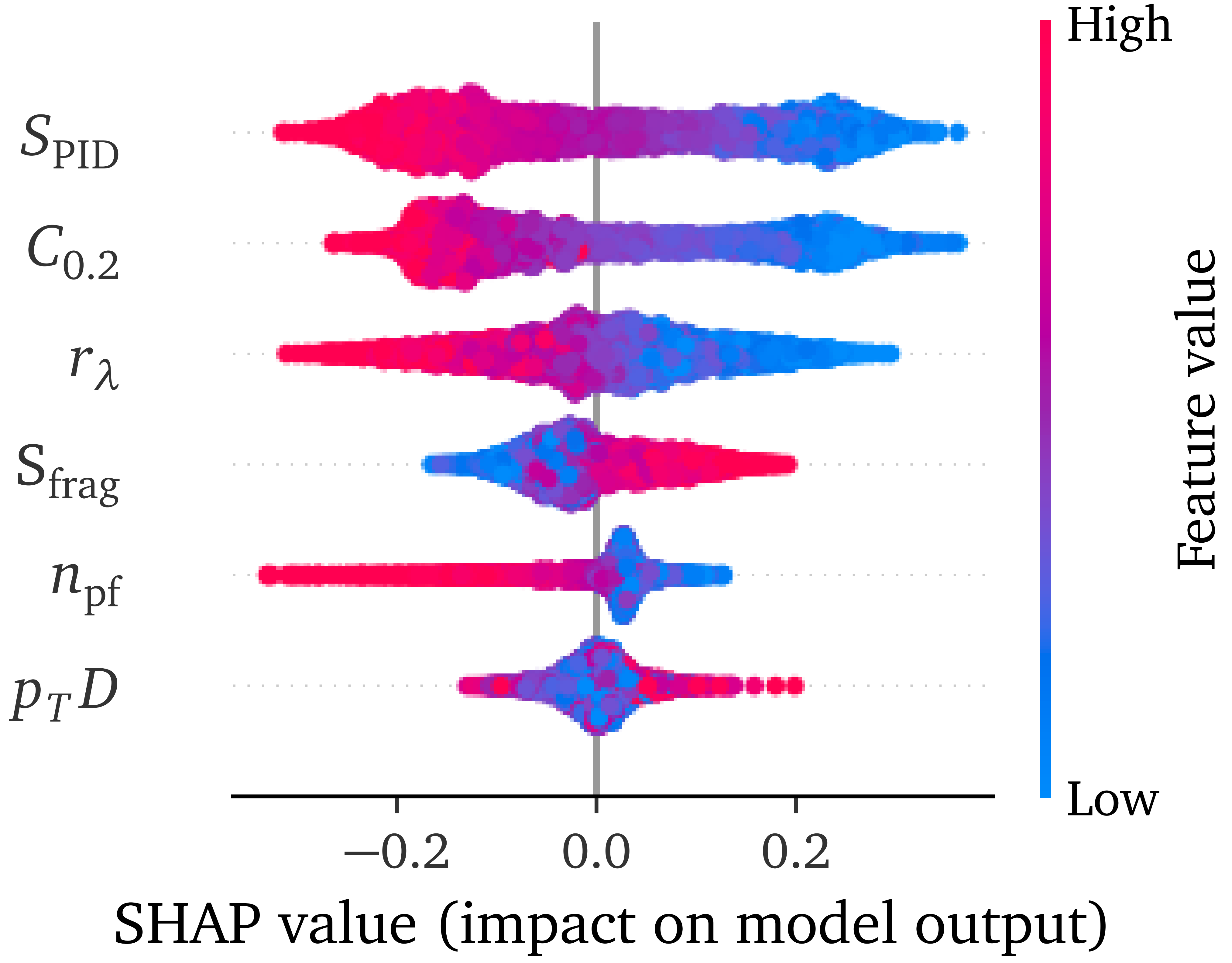
$$\mathscr{V}_i = \sum_{S \subseteq F \setminus \{i\}} \frac{|S|!(|F| - |S| - 1)!}{|F|!} \left[f(S \cup \{i\}) - f(S) \right]$$

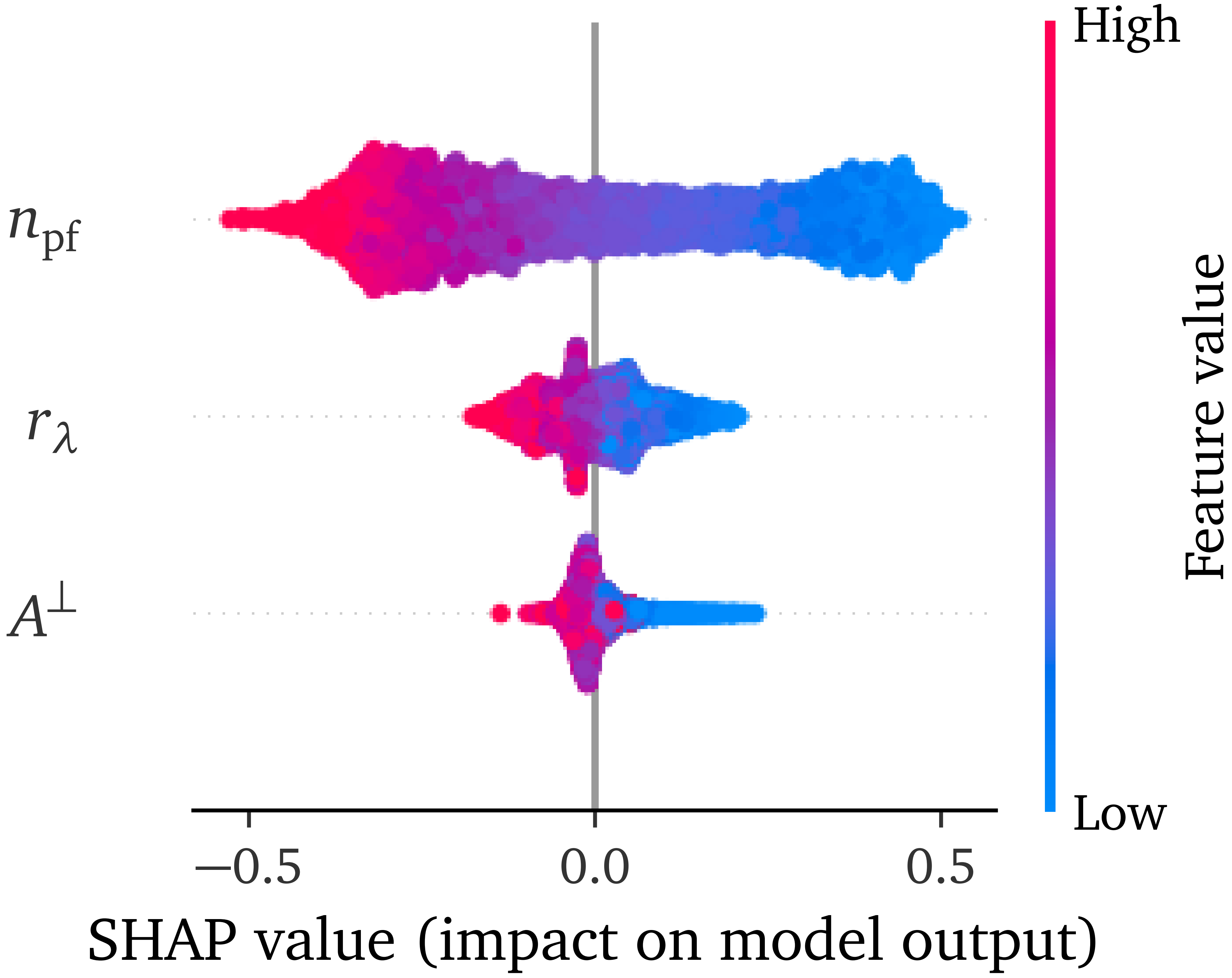
Uncover:

1) Feature importance ranking

2) Identifying which regions of a feature drive predictions

SHAP (SHapley Additive exPlanations)





Doesn’t hold for correlated features ⟶ decorrellated feature set