NEW PHYSICS FROM GEOMETRY

Raffaele Tito D’Agnolo - CEA IPhT Saclay and ENS Paris

There can still be **HUGE** signals hiding in collider data

MY STATISTICAL PROBLEM

Goodness-of-fit test in high-dimensional spaces

MY STATISTICAL PROBLEM

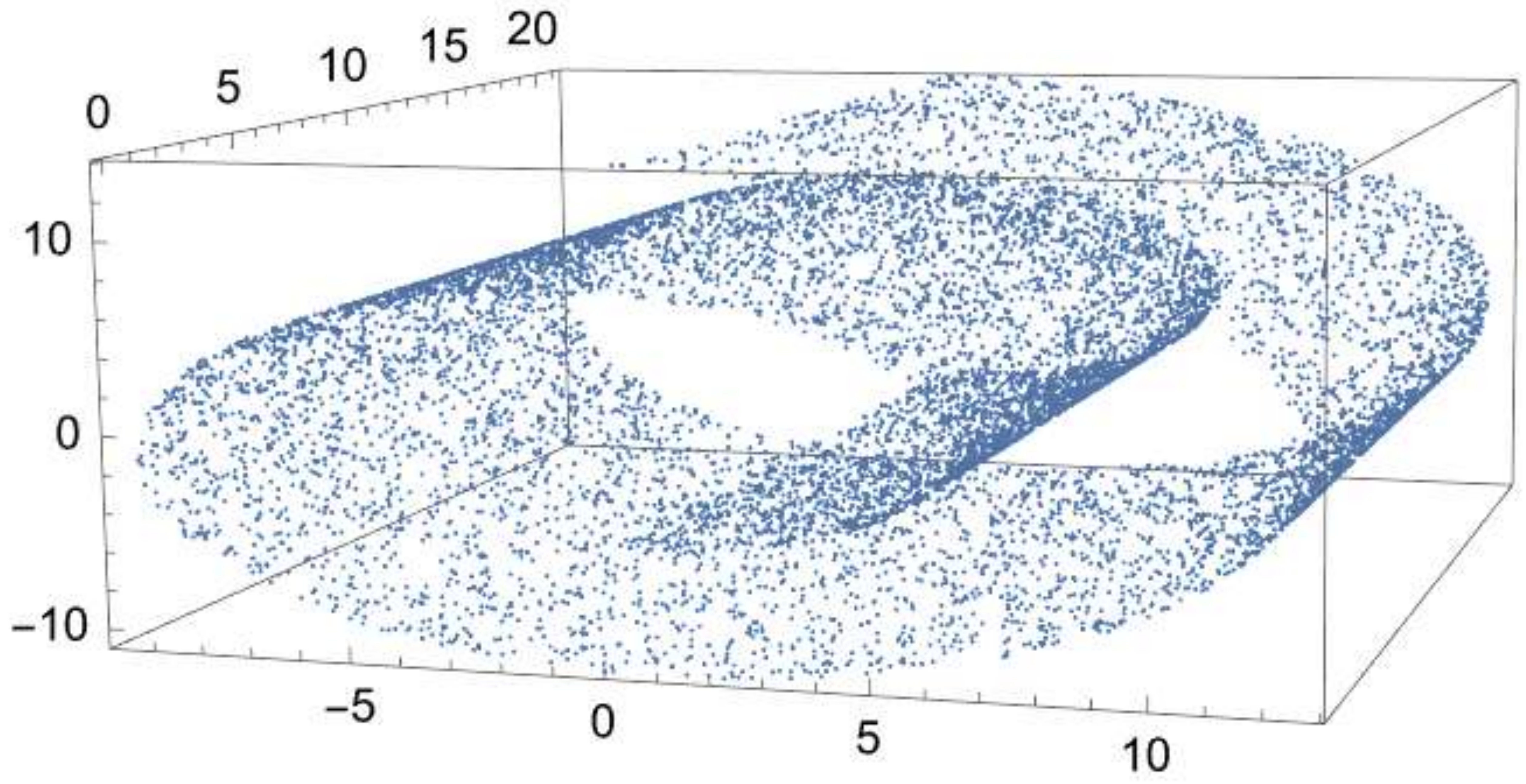
Computing power

Rare, weakly expressed, embedded in the data manifold
(but well-known reference hypothesis)

1 event

=

1 point in D dimensions



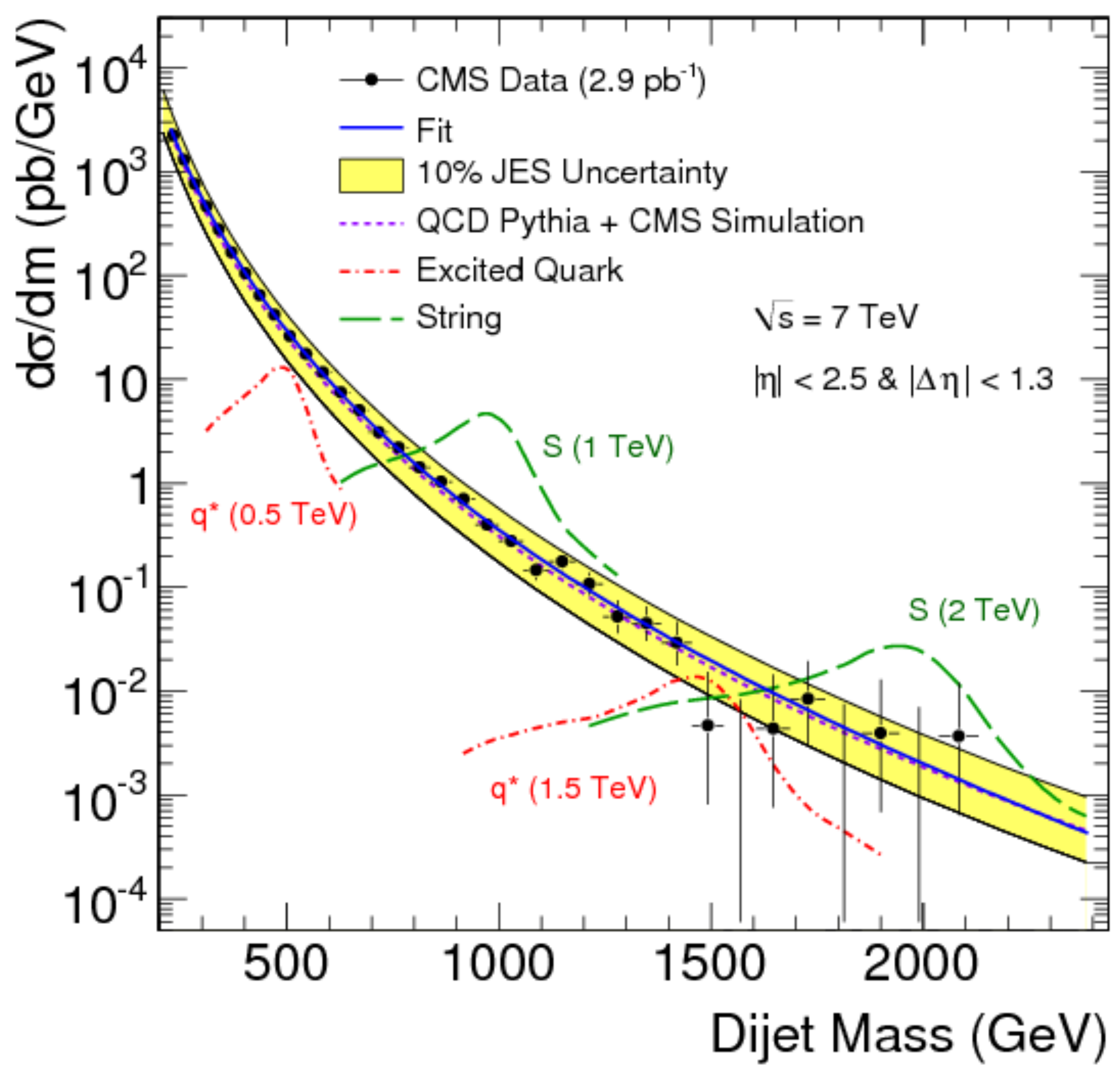
Dataset = data manifold

OUTLINE

1. There can still be huge signals hiding in collider data
2. Why is it challenging to extract them?
3. Intrinsic dimension for new physics
4. Locality for new physics

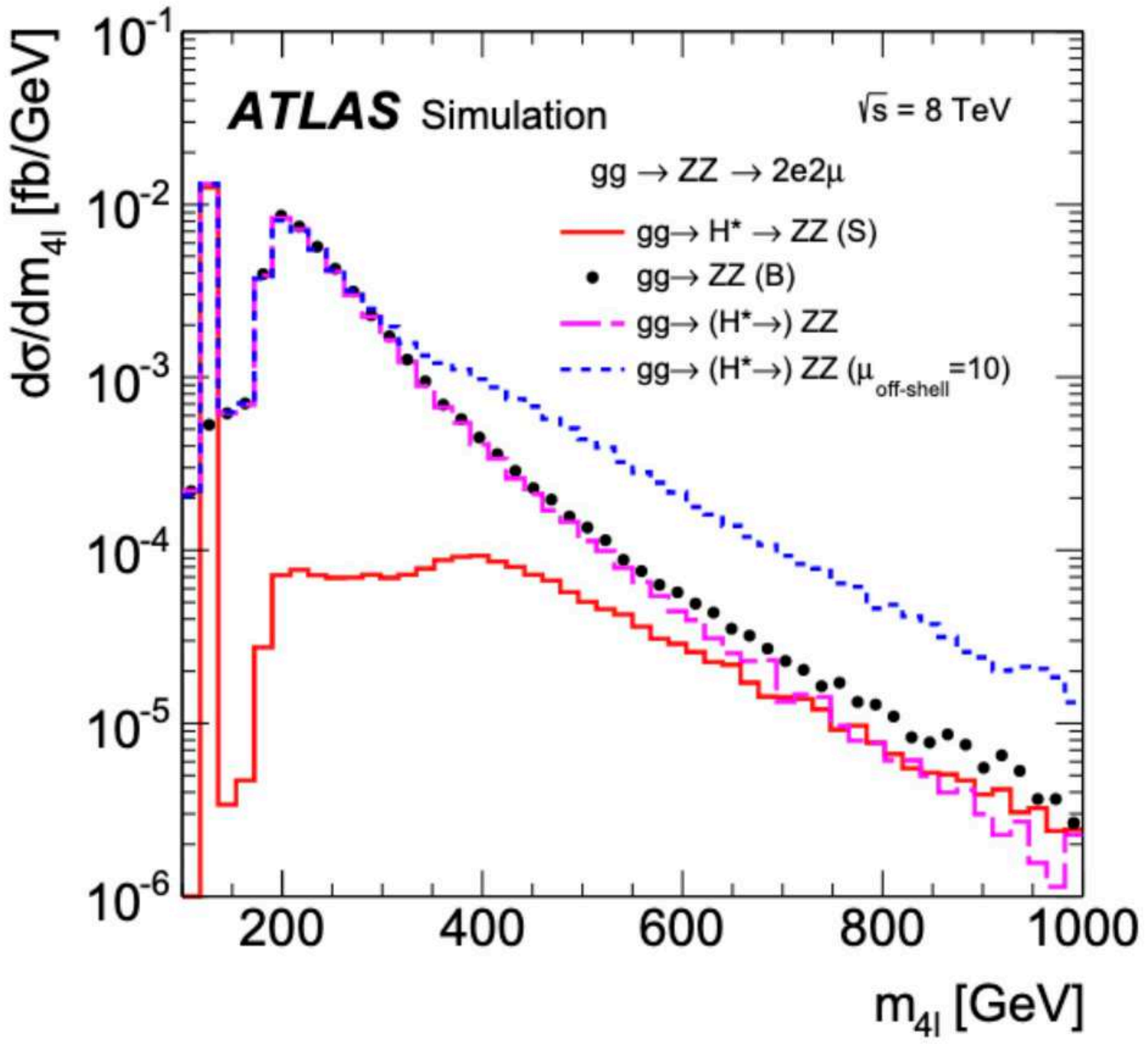
40 SIGMA SIGNALS AND WHERE TO FIND THEM

Bumps

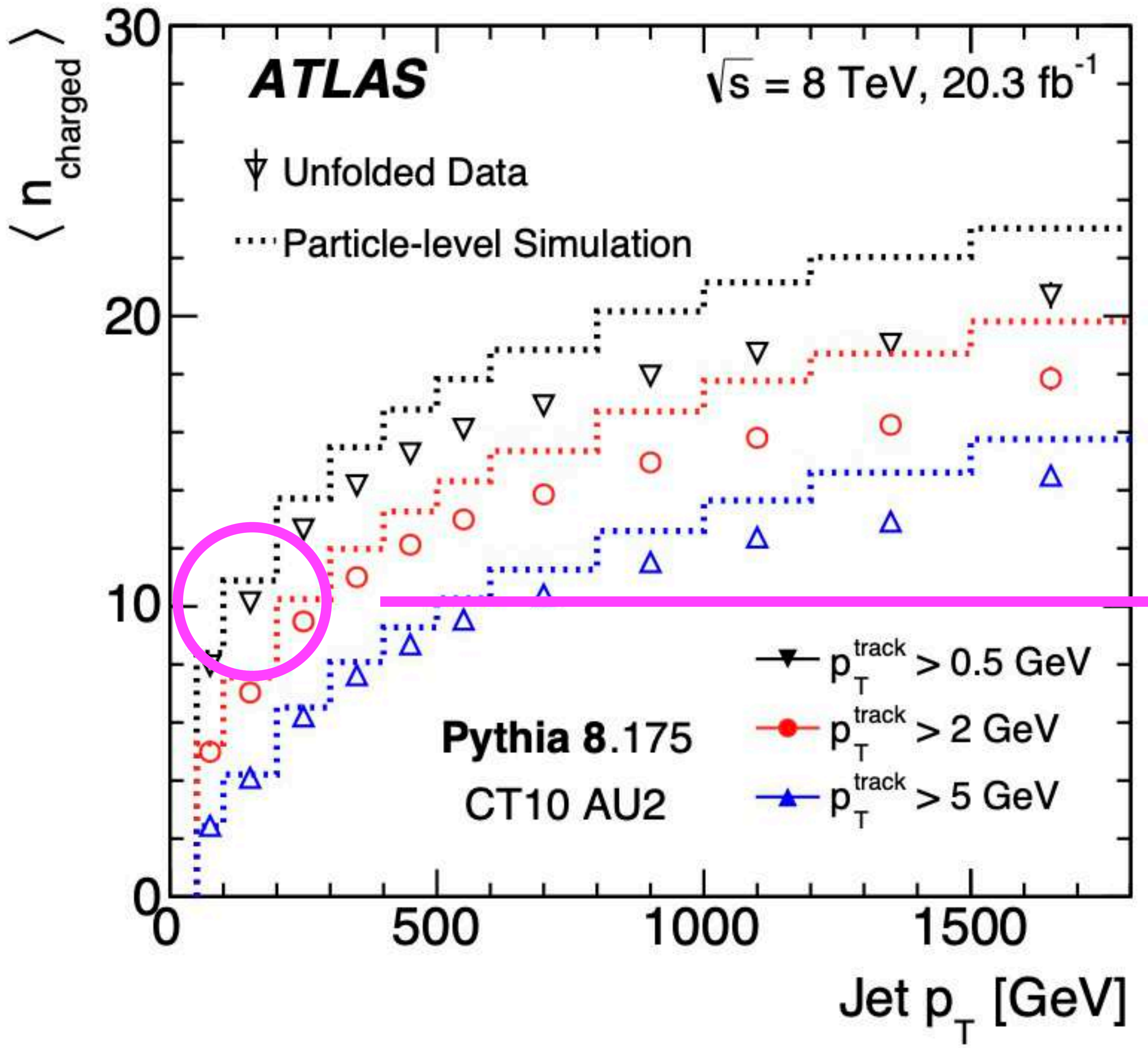


CMS Data (2.9 pb-1)
Fit
10% JES Uncertainty
QCD Pythia + CMS Simulation
Excited Quark
String
√s = 7 TeV
|η| < 2.5 & |Δη| < 1.3
S (1 TeV)
S (2 TeV)
q* (0.5 TeV)
q* (1.5 TeV)
dσ/dm (pb/GeV)
Dijet Mass (GeV)

Tails



ATLAS Simulation
√s = 8 TeV
gg → ZZ → 2e2μ
gg→ H* → ZZ (S)
gg→ ZZ (B)
gg→ (H*→) ZZ
gg→ (H*→) ZZ (μoff-shell=10)
dσ/dm4l [fb/GeV]
m4l [GeV]



Roughly ~ 10 charged tracks in a 100 GeV jet

Dijet event

~ 20 x 3 = **60**

kinematical variables

$$\sim N^{60}$$

Possible signals

$N > 1$

But unknown

PDF of the data

$$f(\{\vec{p}_i\}|\mathbf{w})$$

$$\sum_n a_n \left(\sum_i \vec{p}_i \cdot \vec{w}_i \right)^n$$

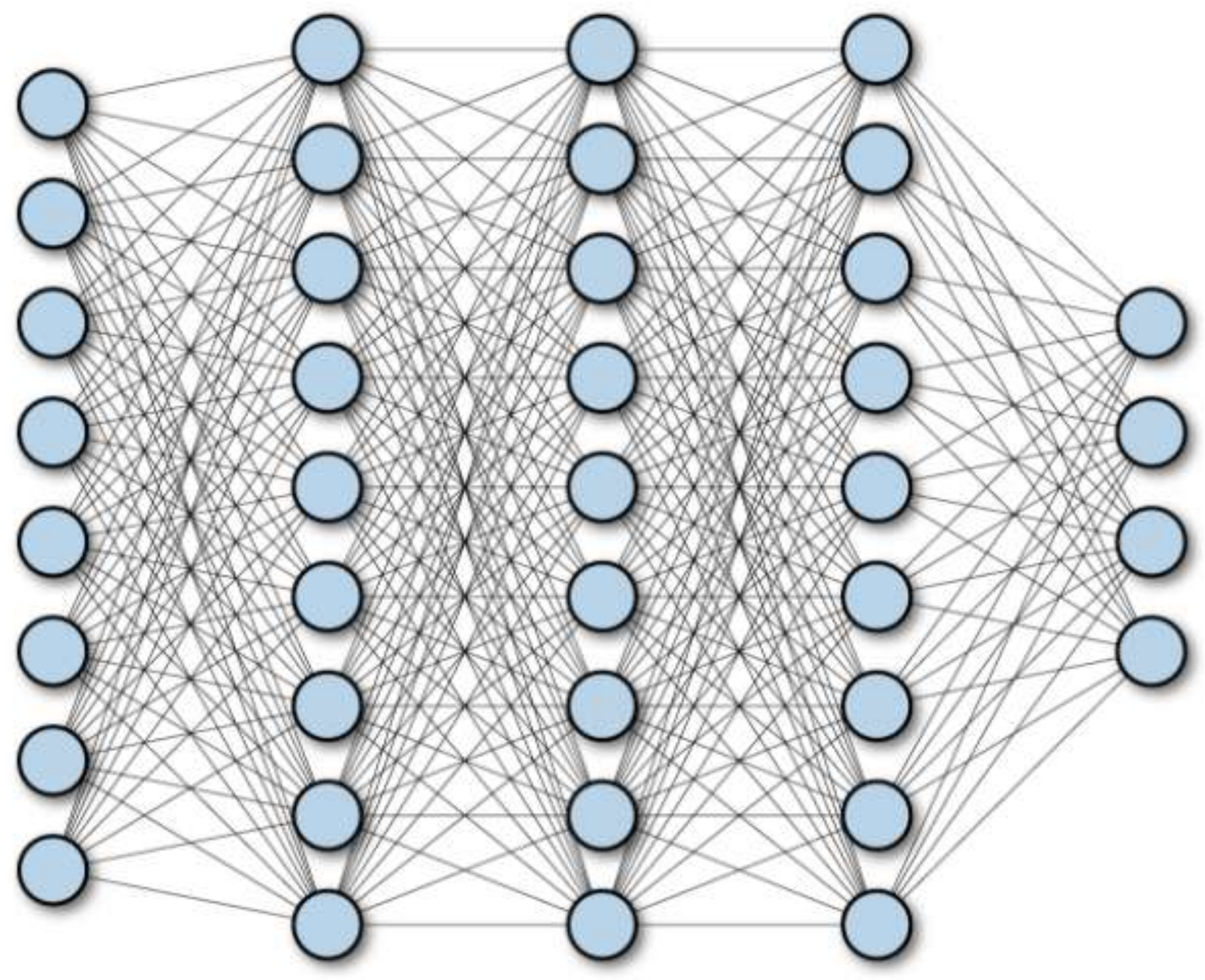
PDF of the data

$$f(\{\vec{p}_i\}|\mathbf{w})$$

$$\sum_n a_n \left(\sum_i \vec{p}_i \cdot \vec{w}_i \right)^n$$

PDF of the data

$$f(\{\vec{p}_i\}|\mathbf{w})$$



PDF of the data

$$f(\{\vec{p}_i\}|\mathbf{w})$$

The number of parameters **w** is proportional to D

PDF of the data

$$f(\{\vec{p}_i\}|\mathbf{w})$$

If the **w**'s can take N distinct values you have

$$\sim N^{\alpha D}$$

possible PDFs

$$\sim N^{60}$$

And this does not even include:

1) All particles from Pile Up

2) The growth at 13 TeV

3) All the hits in the tracker from tracks that don't look "normal"

$$\sim N^{10^4}$$

And this does not even include:

1) All particles from Pile Up

2) The growth at 13 TeV

3) All the hits in the tracker from tracks that don't look "normal"

THE DREAM

Raw data = Hits in tracker and muon system + energy deposits in the calorimeters

THE DREAM

Raw data → Dr. Black Box

THE DREAM

Raw data → Dr. Black Box → New Physics?

THE PROBLEMS

THE PROBLEMS

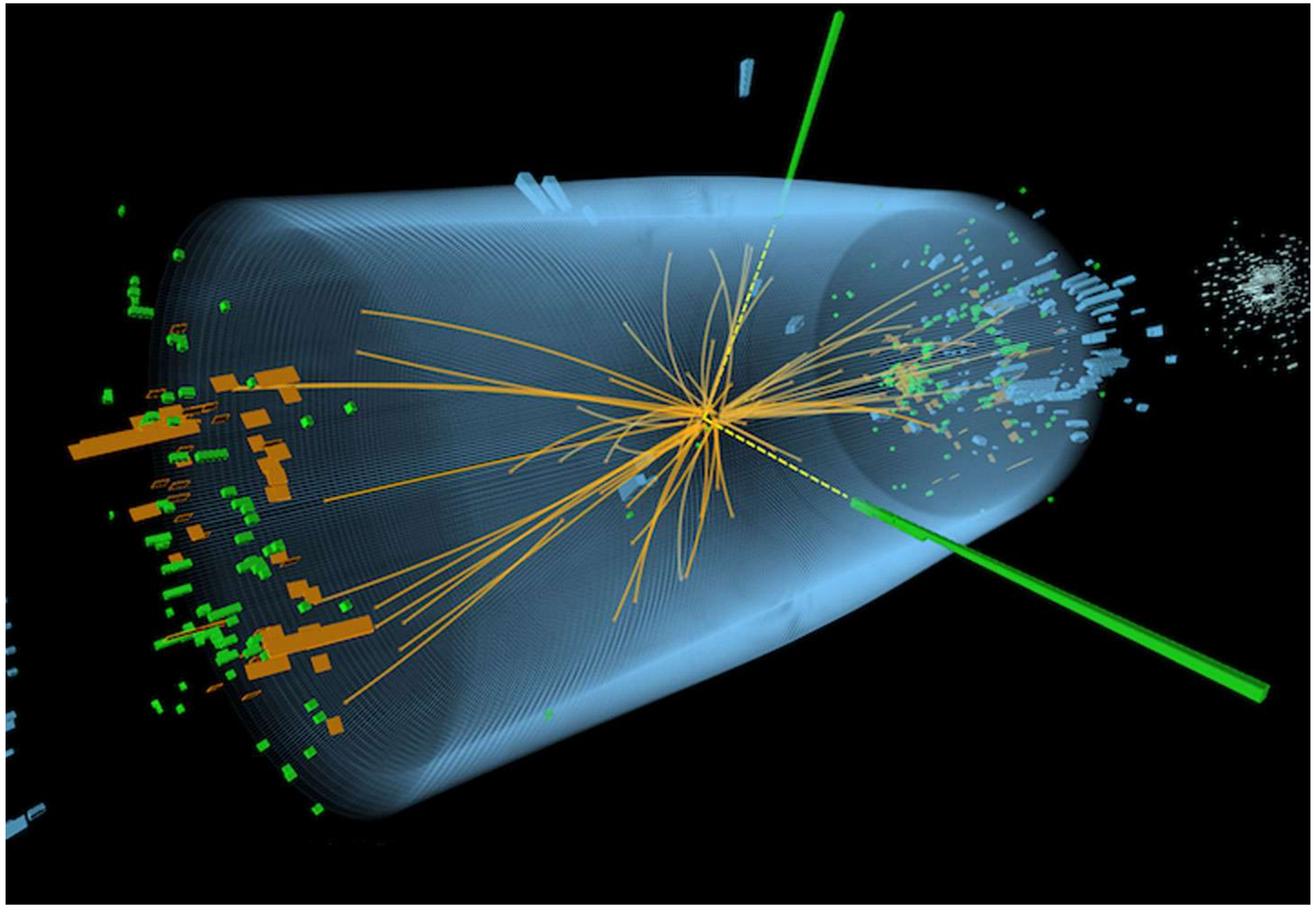
Raw data →

Old Physics →

Curse of dimensionality → New Physics?


BLESSINGS AND CURSES

STANDARD HEP SOLUTION



Take the two leading photons and compute the invariant mass

$$f(m_{\gamma\gamma}|\hat{\mathbf{w}})$$

No curse, but misses many signals

THE STATISTICAL PROBLEM

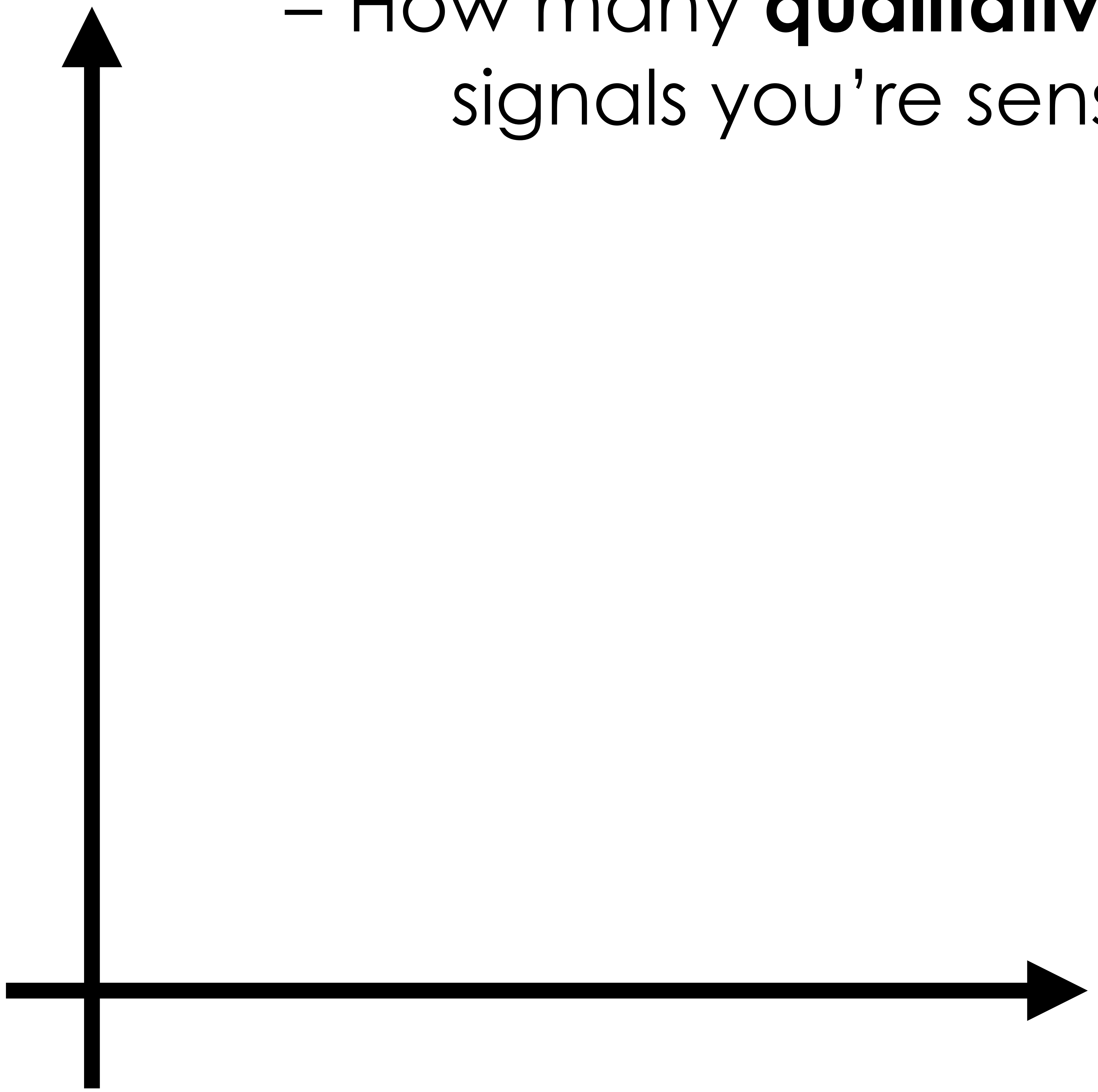
Goodness-of-fit in high-dimensional spaces

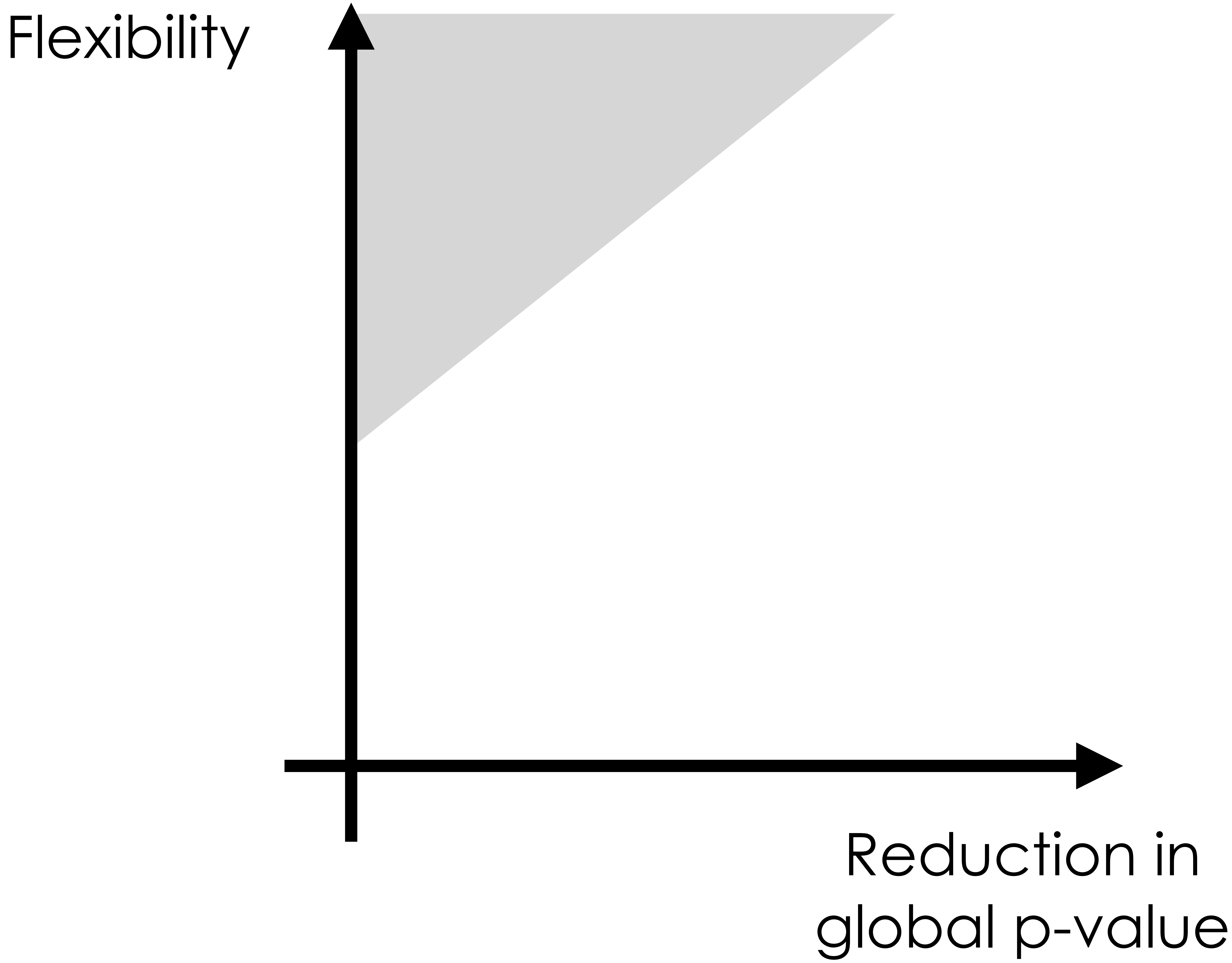
THE STATISTICAL PROBLEM

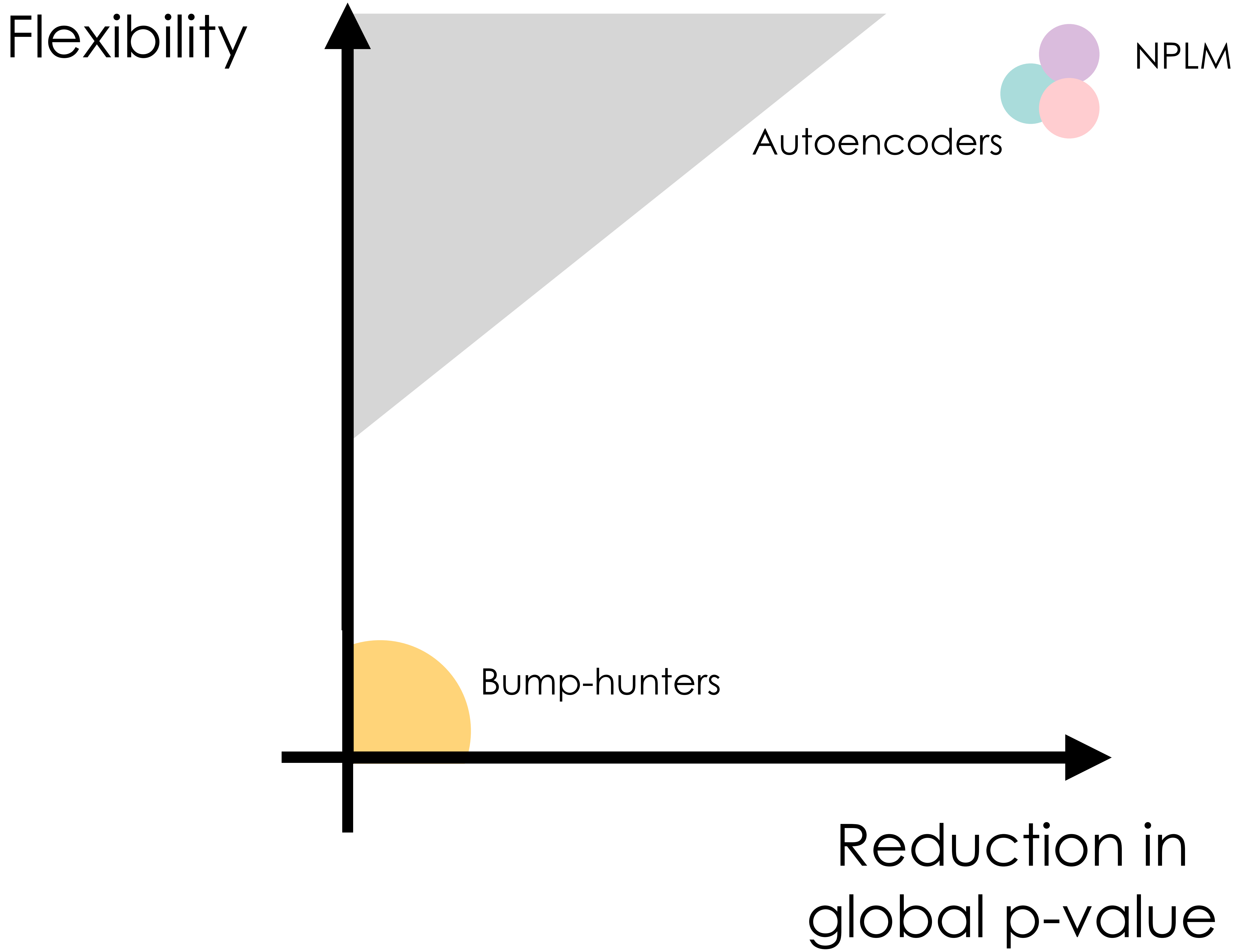
- **Kernel-based tests (MMD, kernel Stein discrepancies)**: measure a reproducing-kernel Hilbert space distance between empirical and null distributions.
- **Energy / distance-based tests (energy distance, distance covariance)**: use pairwise distances to detect distributional differences or dependence; in high dimensions standard Euclidean distances can concentrate and lose discriminatory power.
- **Projection-based tests**: reduce the problem to many univariate tests by projecting data (random or optimized projections) and combining p-values; effective when alternatives lie on low-dimensional subspaces.
- **Graph / nearest-neighbor and rank-based tests**: build graphs (k-NN, MST) or use ranks to compare local geometry between samples; often robust to some high-d effects but sensitive to metric choice.
- **Residual- or model-based tests (RP tests, parametric bootstrap)**: fit a high-dimensional model (e.g., Lasso), analyze residuals with flexible regressors or prediction proxies, and calibrate via parametric/bootstrap schemes; tailored to detect model misspecification (nonlinearity, heteroscedasticity, omitted variables).

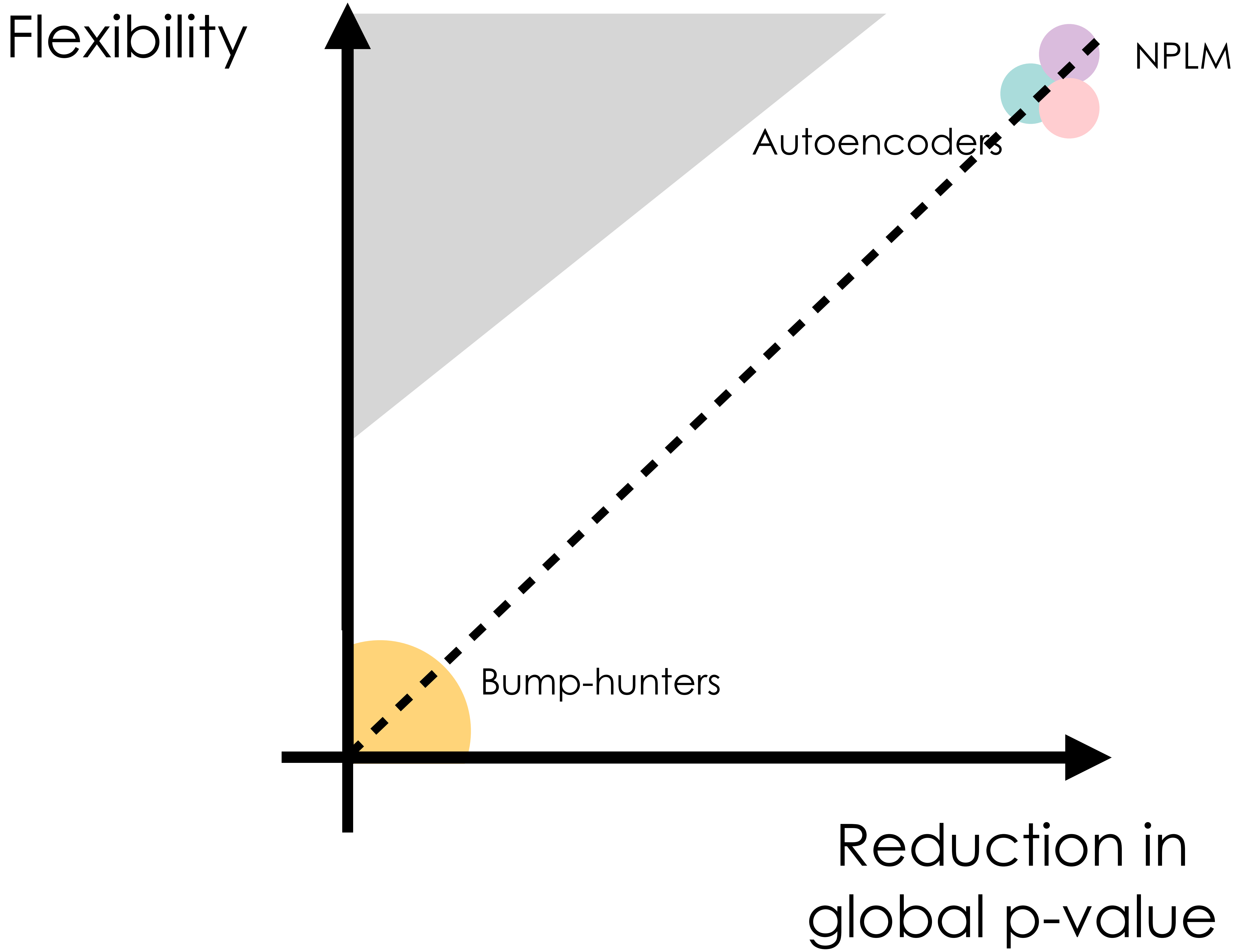
Flexibility

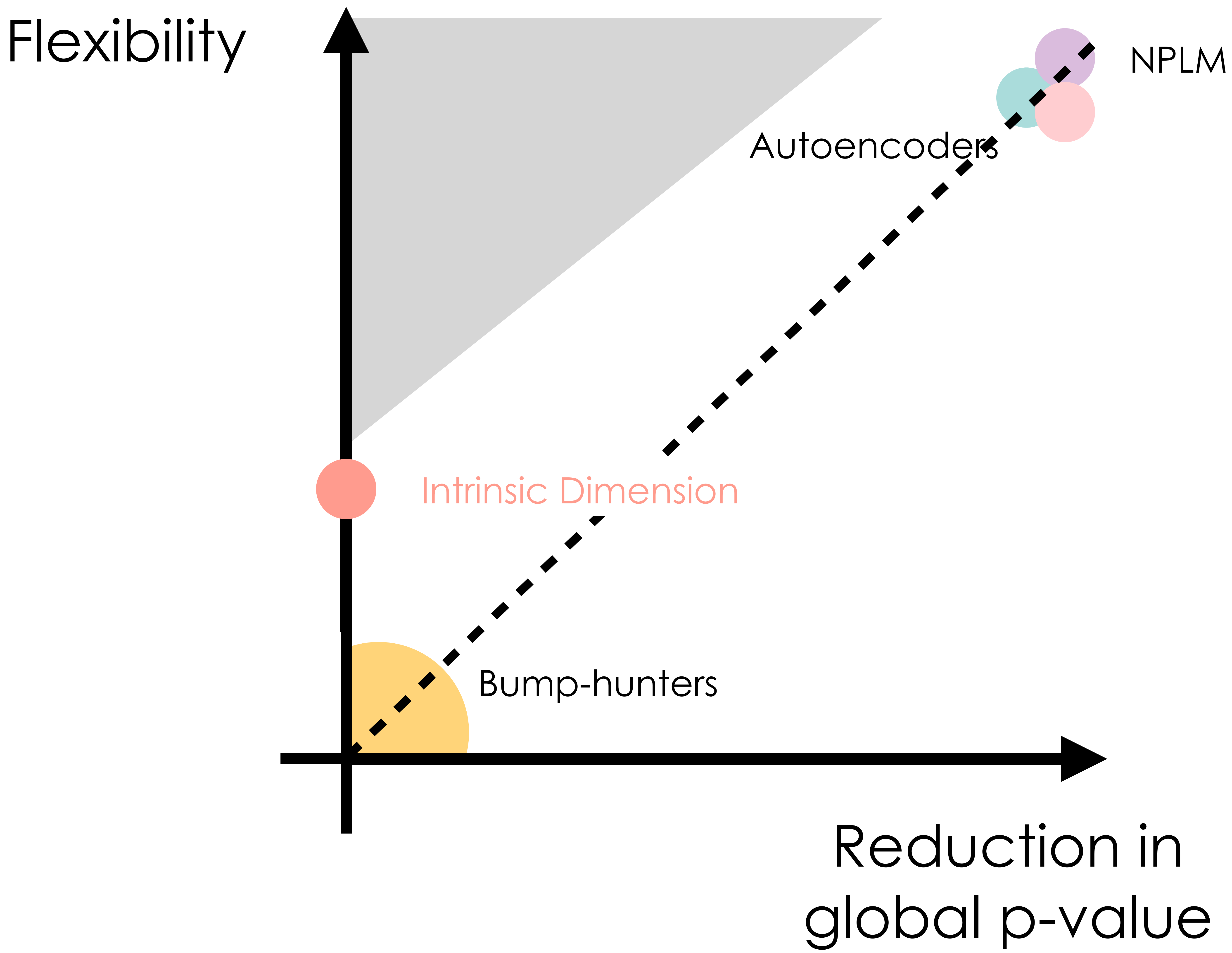
= How many **qualitatively different** signals you’re sensitive to




Flexibility
Reduction in global p-value

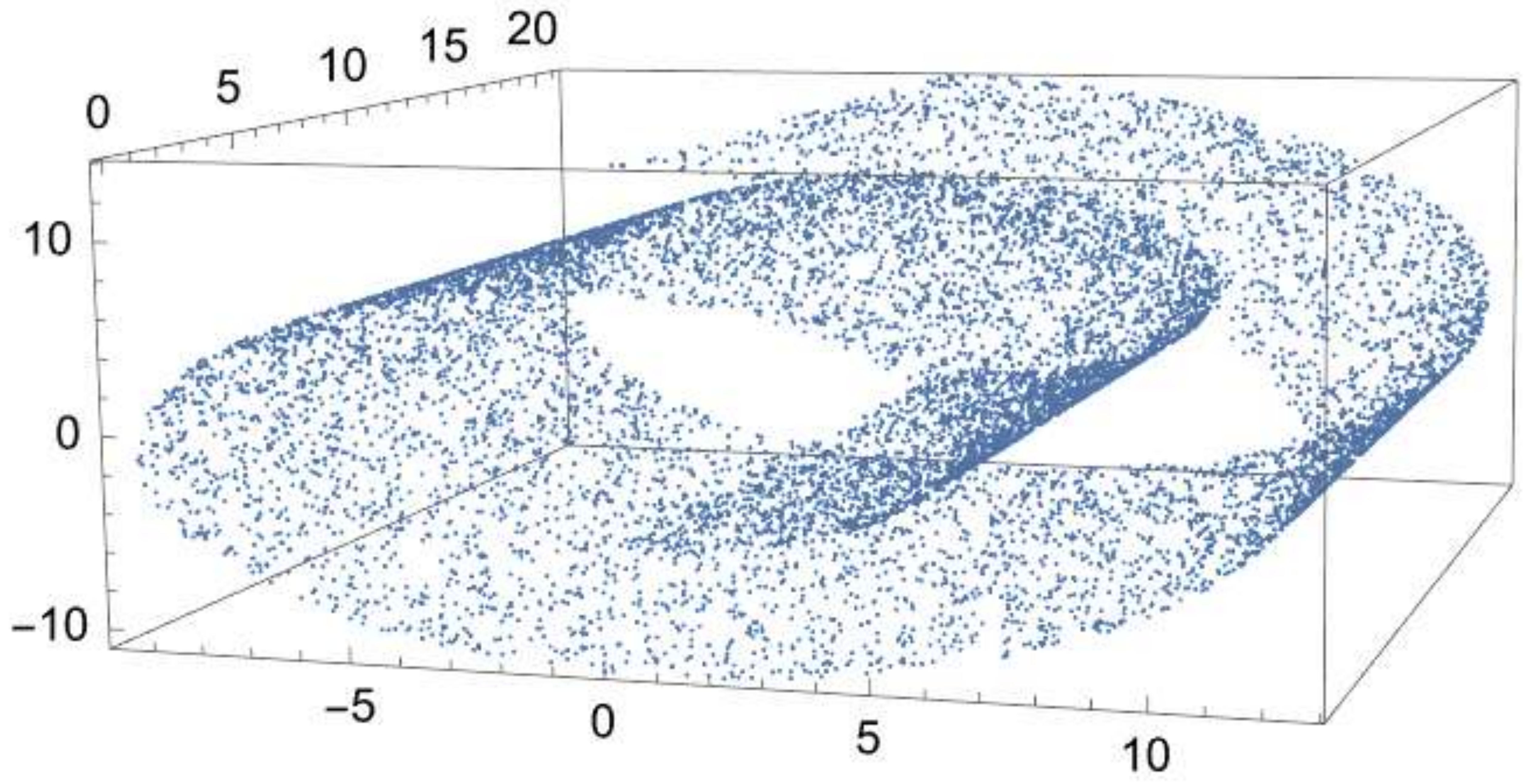

Flexibility
NPLM
Autoencoders
Bump-hunters
Reduction in global p-value


Flexibility
NPLM
Autoencoders
Bump-hunters
Reduction in
global p-value


Flexibility
NPLM
Autoencoders
Intrinsic Dimension
Bump-hunters
Reduction in global p-value

THE INTRINSIC DIMENSION OF COLLIDER EVENTS

INTRINSIC DIMENSION



Can you tell that D=2 just from the data?

INTRINSIC DIMENSION

$$r_i = \text{Distance between reference point and its i-th nearest neighbor}$$

INTRINSIC DIMENSION

$$\mu_{ij} \equiv \frac{r_j}{r_i}$$

INTRINSIC DIMENSION

$$\mu_{ij} \equiv \frac{r_j}{r_i} \longrightarrow$$

Under mild assumptions

the PDF can be derived analytically

INTRINSIC DIMENSION

$$\mu_{ij} \equiv \frac{r_j}{r_i} \longrightarrow f(\mu_{ij}) = \frac{D(\mu_{ij}^D - 1)^{j-i-1}\mu_{ij}^{-D(j-1)-1}}{B(j-i,i)}$$

INTRINSIC DIMENSION

1. Compute μ_{ij} for every point in your dataset

2. Use all values of μ_{ij} to compute its PDF

3. Fit for the intrinsic dimension

WHY?

$$r_i =$$

Earth Mover's Distance

[Komiske, Metodiev, Thaler '19]

As you vary i and j you are exploring non-trivial features of the kinematics

THE INTRINSIC DIMENSION FOR NEW PHYSICS

[RTD, A. GLIOTI, G. RIGO, A. VALENTI, arXiv:2511.20760]

1. Flexibility
2. Robustness to Showering and Detector Effects
3. Robustness to Systematics
4. No Curse of Dimensionality

SIGNAL 1

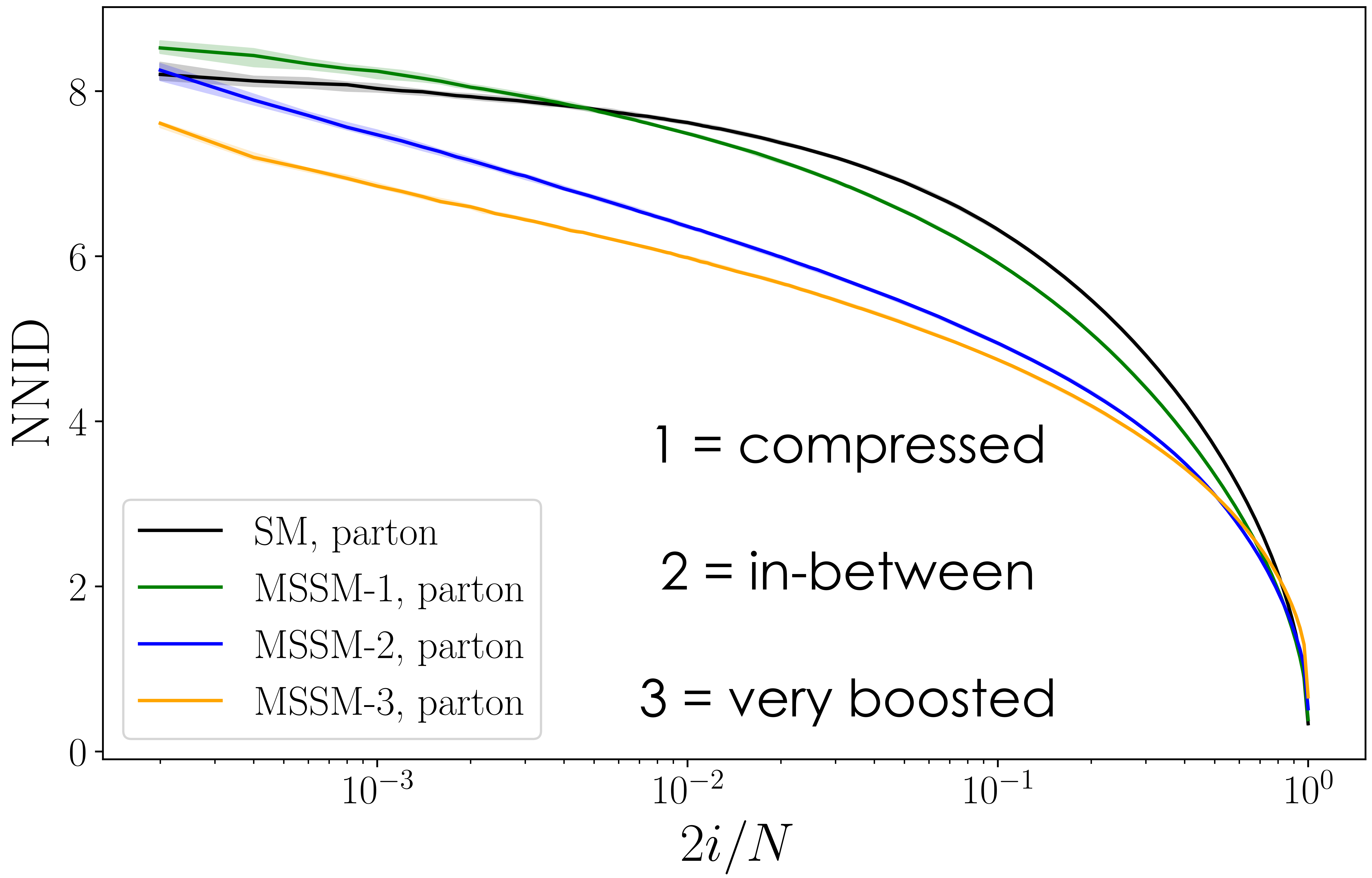
SM: $$pp \to W^+ Z \to \mu^+ \mu^- e^+ \nu_e$$

SIGNAL: $$pp \to \chi^+ \chi_2^0 \to W^+ Z \chi^0 \chi^0 \to \mu^+ \mu^- e^+ \nu_e \chi^0 \chi^0$$

Note: We never use any information about MET

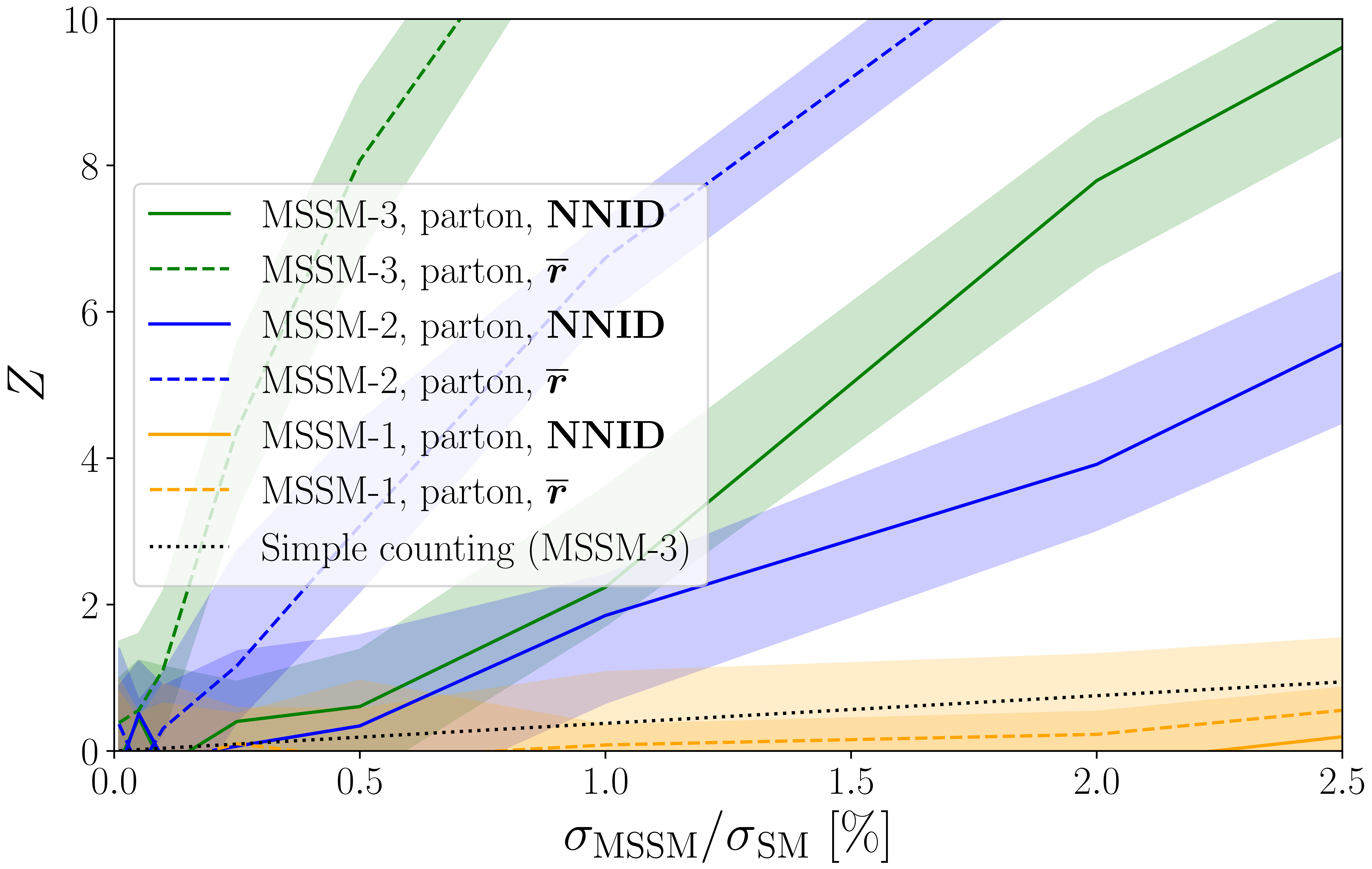
Note: We use agnostic cuts that penalize our signals

SIGNAL 1

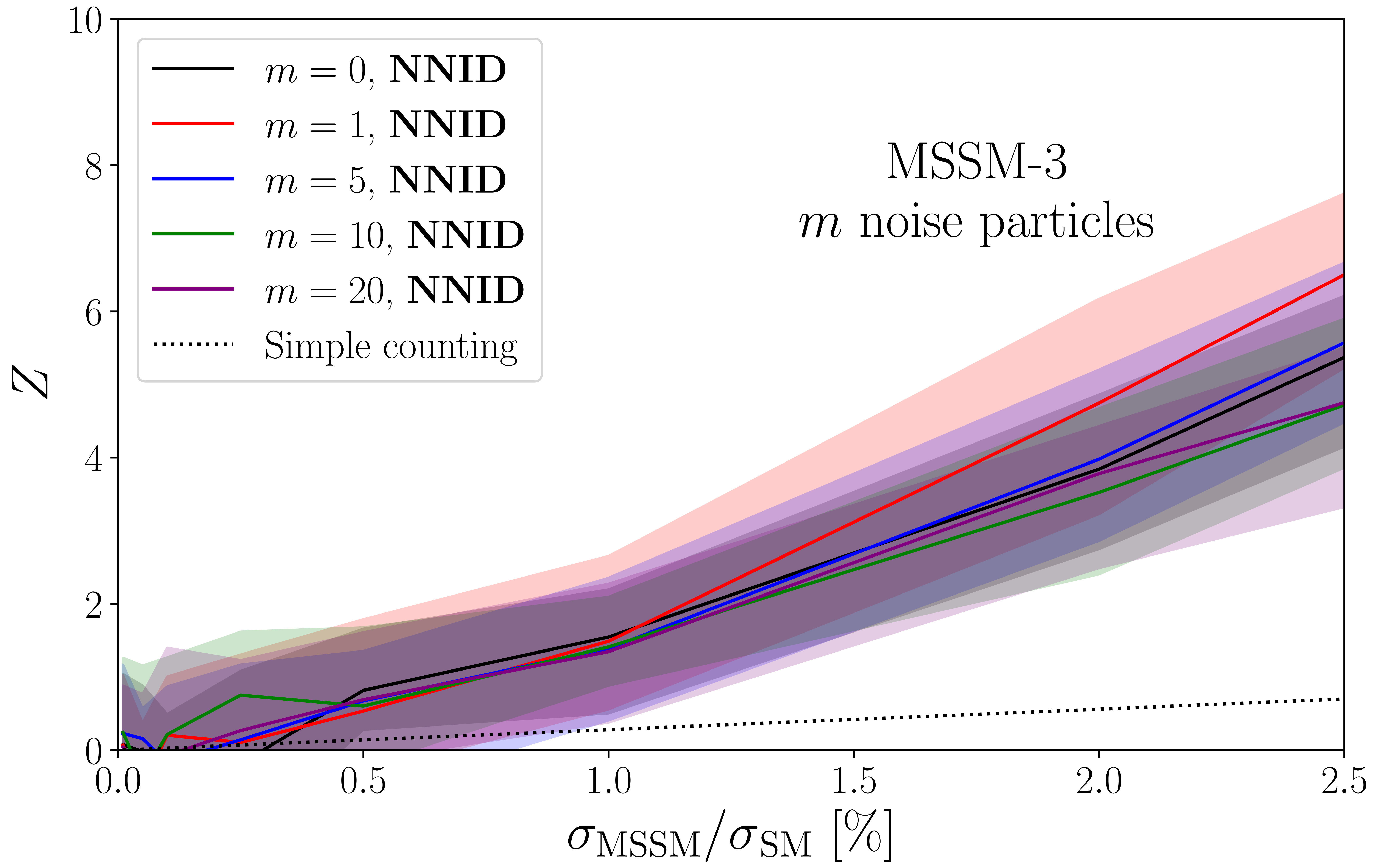


NNID
8
6
4
2
0
10^{-3}
10^{-2}
10^{-1}
10^{0}
$2i/N$
SM, parton
MSSM-1, parton
MSSM-2, parton
MSSM-3, parton
1 = compressed
2 = in-between
3 = very boosted

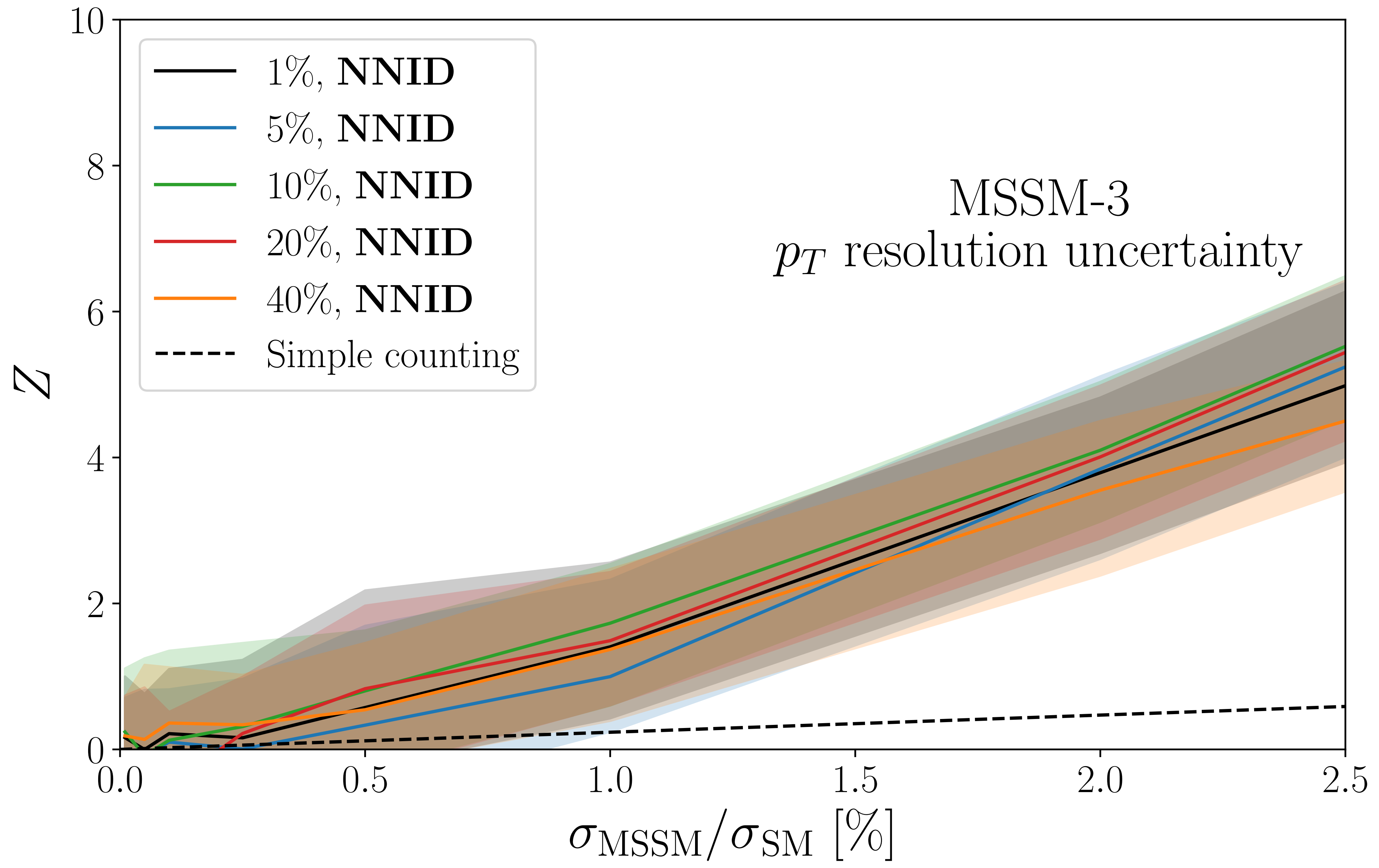
SIGNAL 1 — FLEXIBILITY

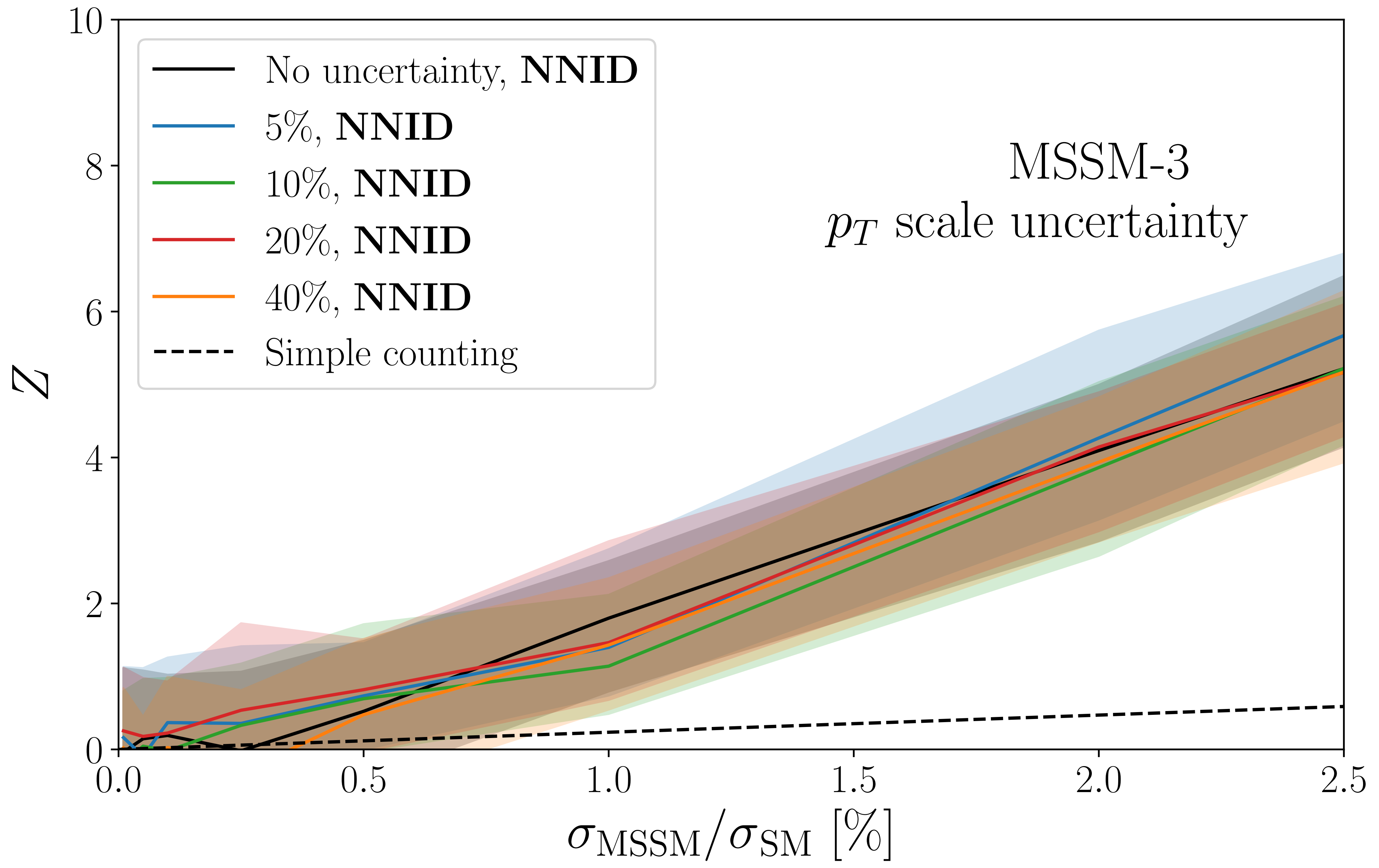


SIGNAL 1 — NO CURSE

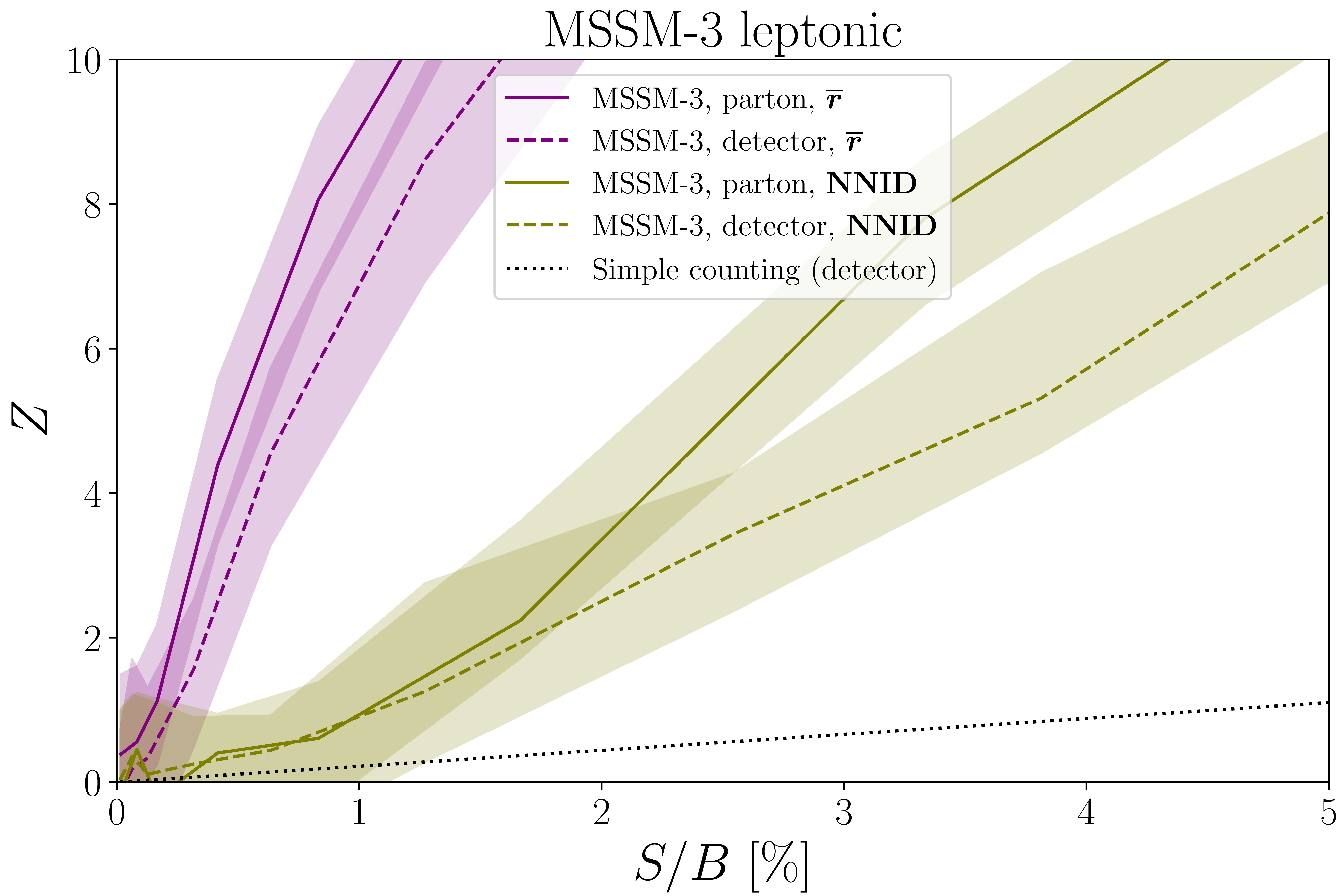


SIGNAL 1 — NO SYSTEMATICS

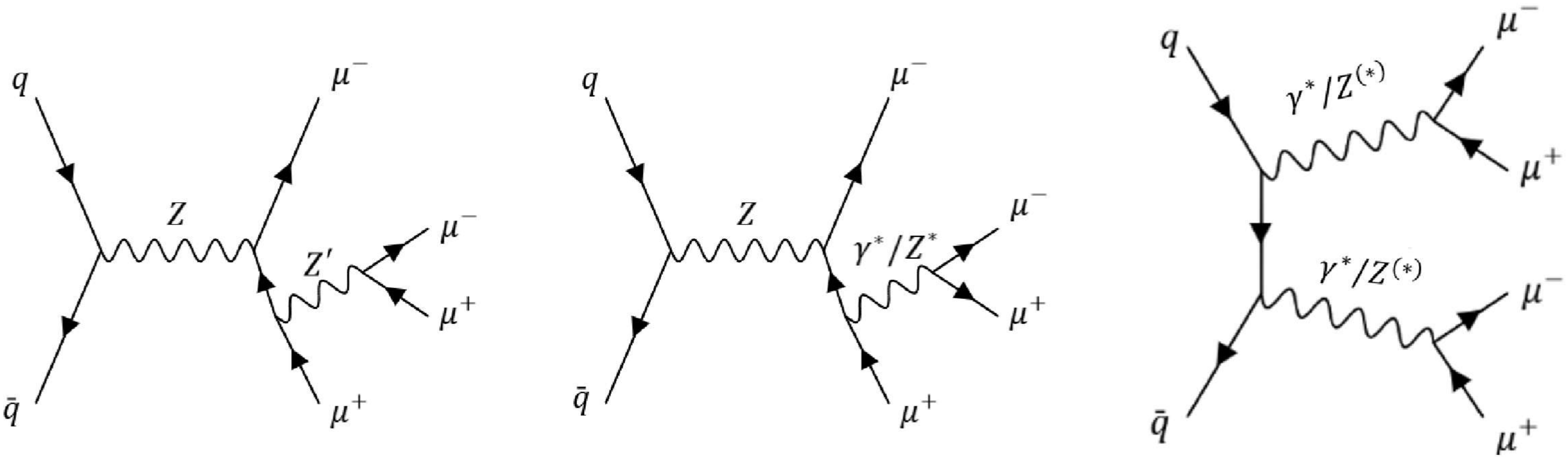




SIGNAL 1 — DETECTOR EFFECTS



SIGNAL 2



q
$\bar{q}$
Z
μ^-
μ^+
Z'
μ^-
μ^+
q
$\bar{q}$
Z
μ^-
μ^+
γ^*/Z^*
μ^-
μ^+
q
$\bar{q}$
$\gamma^*/Z^{(*)}$
μ^-
μ^+
$\gamma^*/Z^{(*)}$
μ^-
μ^+

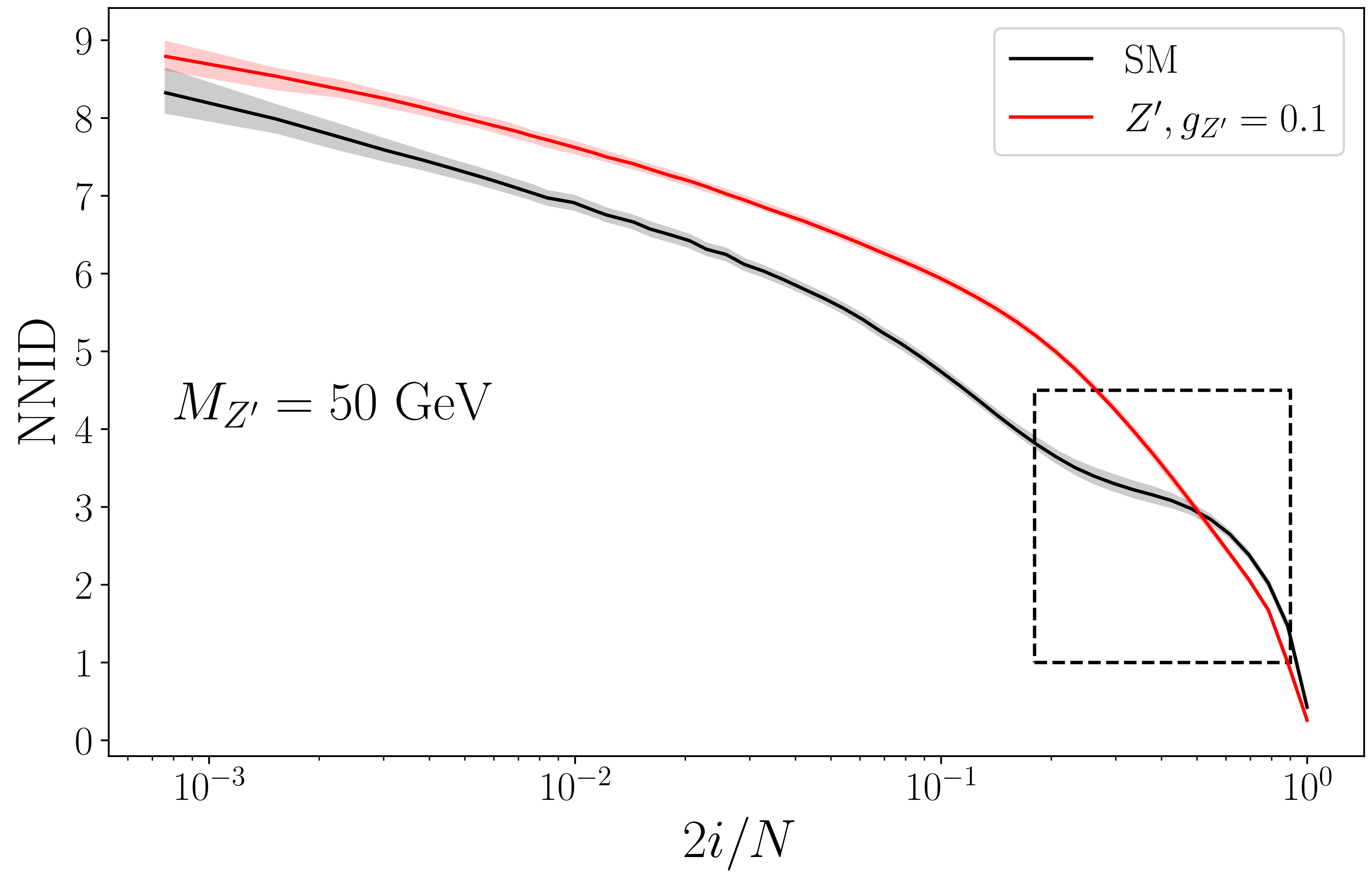
(a) (b) (c)

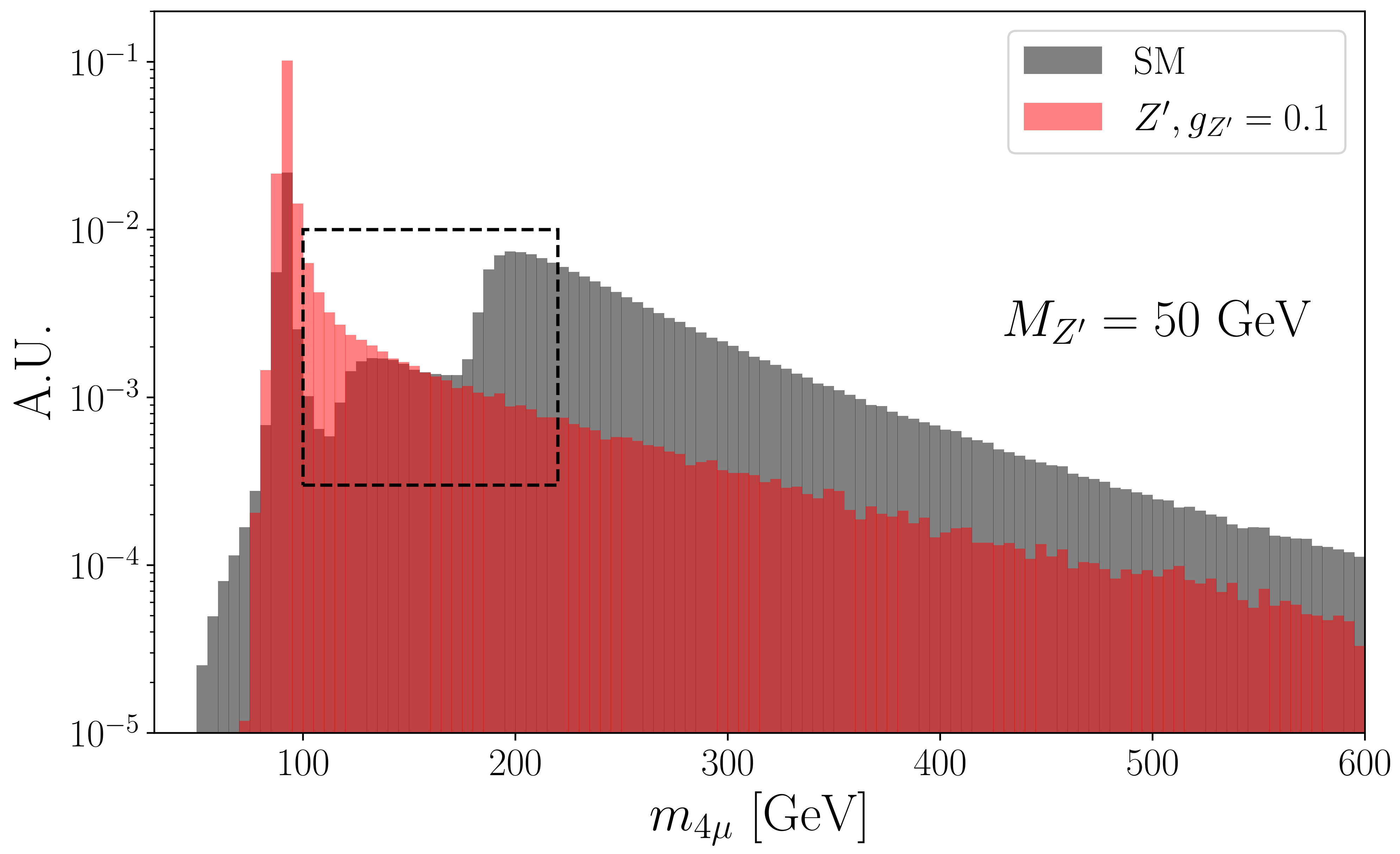
SIGNAL 2

SM: $$pp \to ZZ \to 4\mu$$

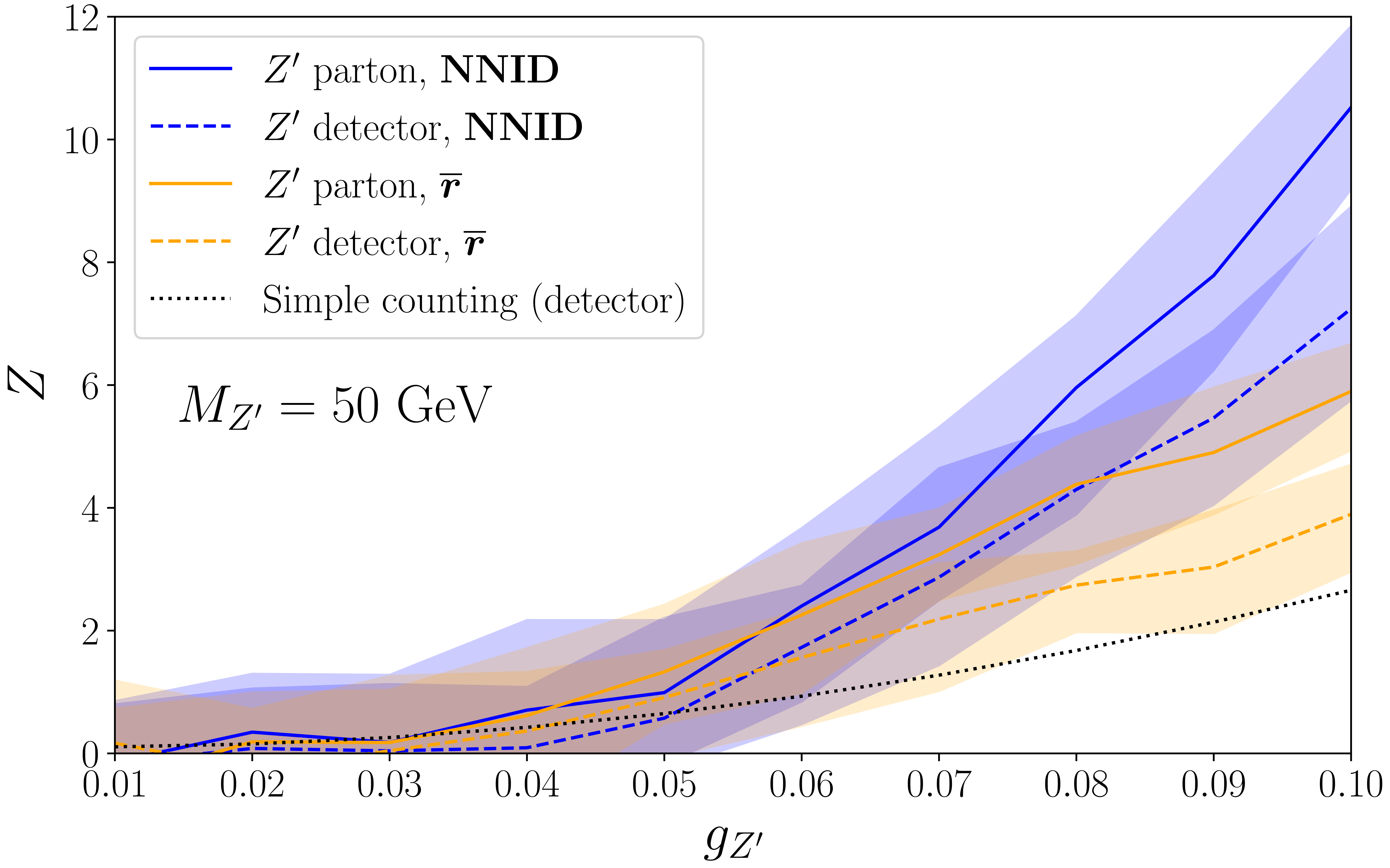
SIGNAL: $$pp \to Z \to \mu^+\mu^- Z_{L_\mu - L_\tau} \to 4\mu$$

SIGNAL 2





SIGNAL 2 — FLEXIBILITY



SIGNAL 3

$$pp \to WZ \to \mu^+\mu^- jj$$

~ 30-40 particles per event

D ~ 100

SIGNAL 3

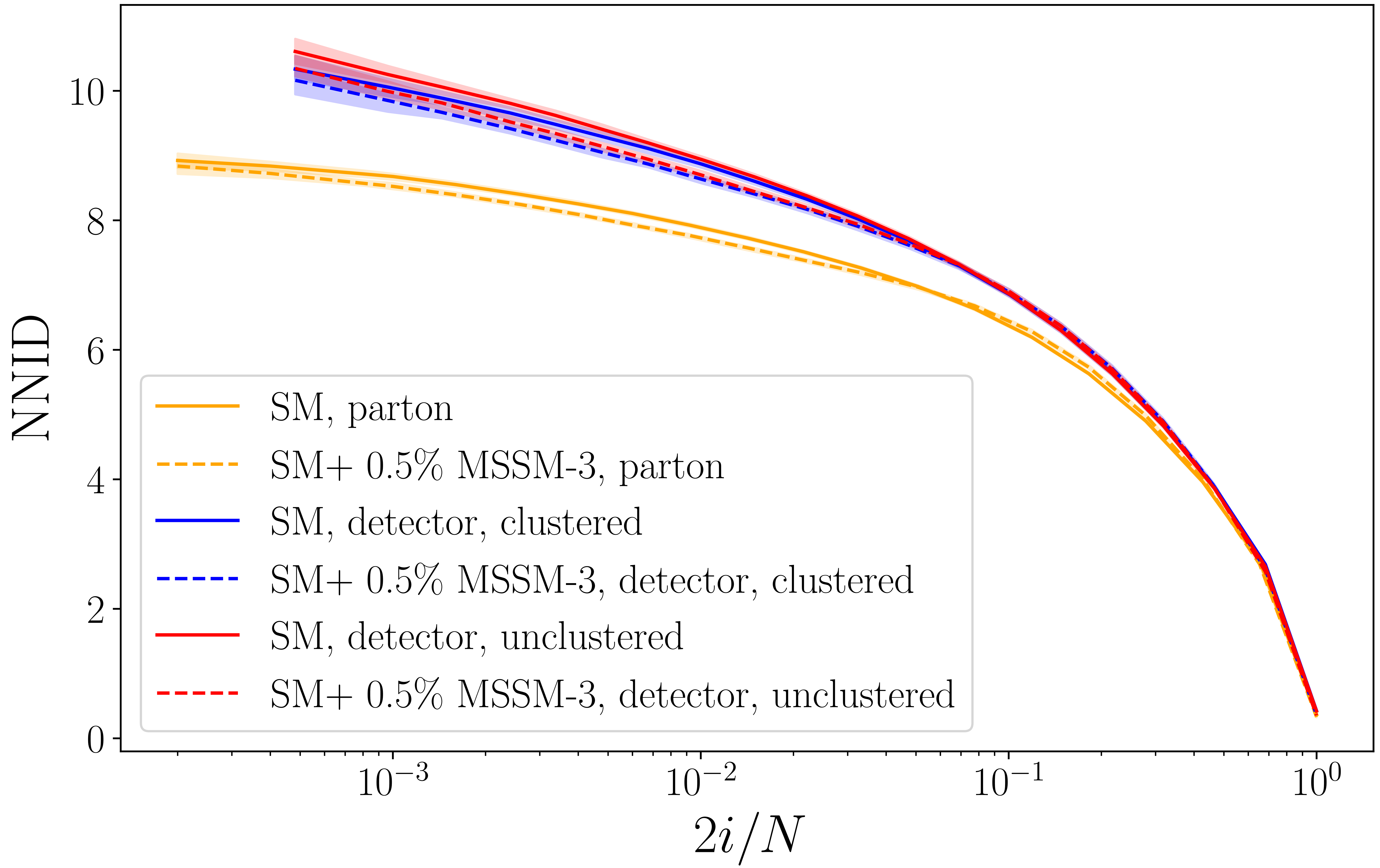
SM: $$pp \to W^+ Z \to \mu^+ \mu^- jj$$

SIGNAL: $$pp \to \chi^+ \chi_2^0 \to W^+ Z \chi^0 \chi^0 \to \mu^+ \mu^- jj \chi^0 \chi^0$$

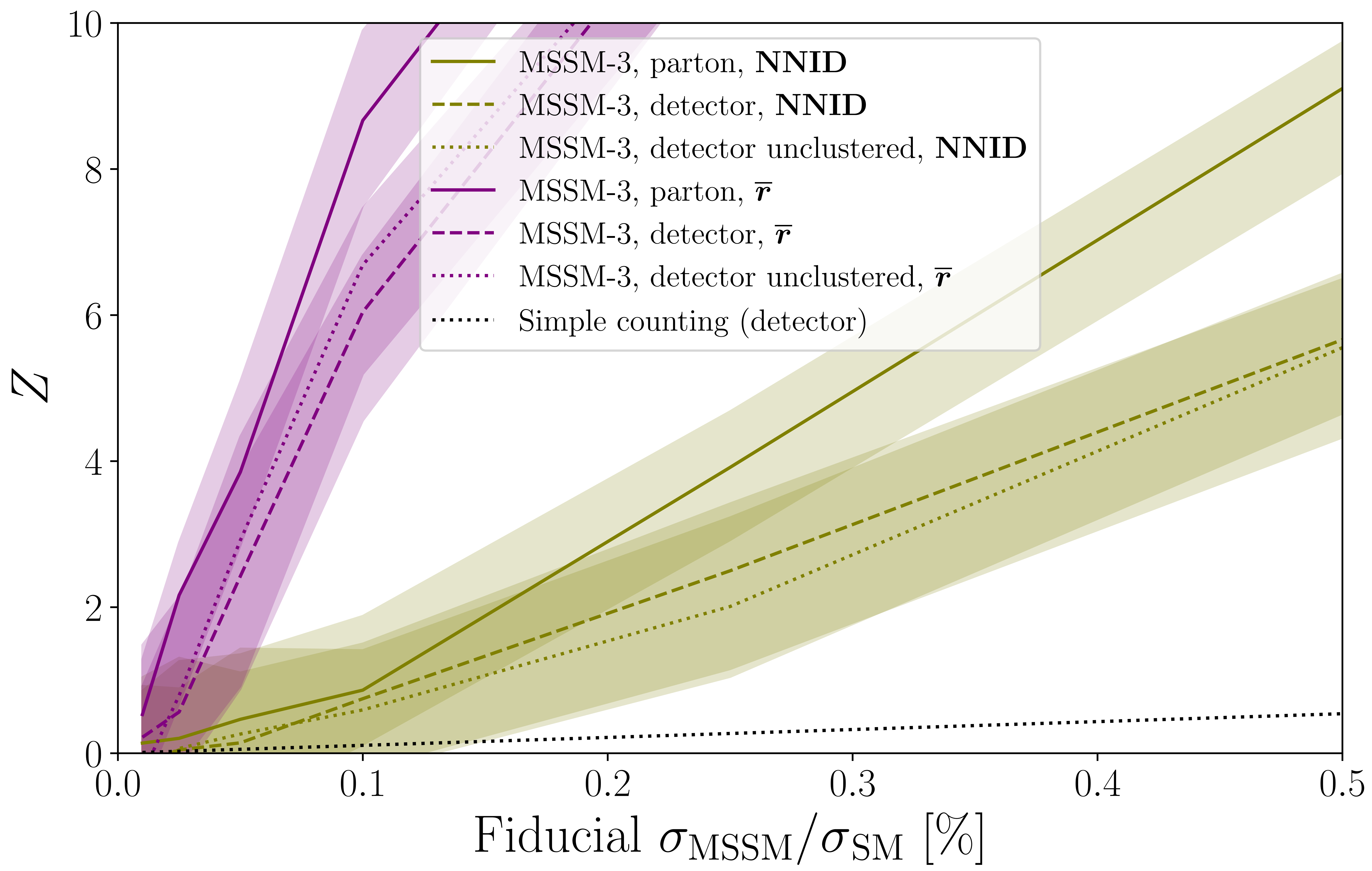
Note: We never use any information about MET

Note: We use agnostic cuts that penalize our signals

SIGNAL 3


NNID
10
8
6
4
2
0
10^{-3}
10^{-2}
10^{-1}
10^{0}
$2i/N$
SM, parton
SM+ 0.5% MSSM-3, parton
SM, detector, clustered
SM+ 0.5% MSSM-3, detector, clustered
SM, detector, unclustered
SM+ 0.5% MSSM-3, detector, unclustered

SIGNAL 3 — FLEXIBILITY



SO FAR WE HAVE IGNORED ONE PROBLEM

Raw data → Dr. Black Box → New Physics?

Old Physics → Dr. Black Box

OLD PHYSICS

1. Use it on final states where the MC has been tested to death

2. Test for time variations in LHC data (DQM)

3. Test the exact and approximate symmetries of the Standard Model (CP, Lepton flavor, Lorentz, ...)

OLD PHYSICS

SAMPLE 1

All events with 1 electron

SAMPLE 2

All events with 1 muon

CONCLUSION

Think globally, act locally

LOCALITY FOR NEW PHYSICS

[RTD, A. GLIOTI, J. THALER, A. VALENTI, in preparation]

BACKUP

WHY?

1. First thing to do before any analysis

2. First thing to do before building any autoencoder

WHY?

3. As you vary i and j you are exploring non-trivial features of the kinematics

WHY?

3. As you vary i and j you are exploring non-trivial features of the kinematics

4. What is the dimension of a jet?

Many more details and comparisons with other notions of dimension in

[RTD, A. GLIOTI, G. RIGO, A. VALENTI, arXiv:2511.20760]

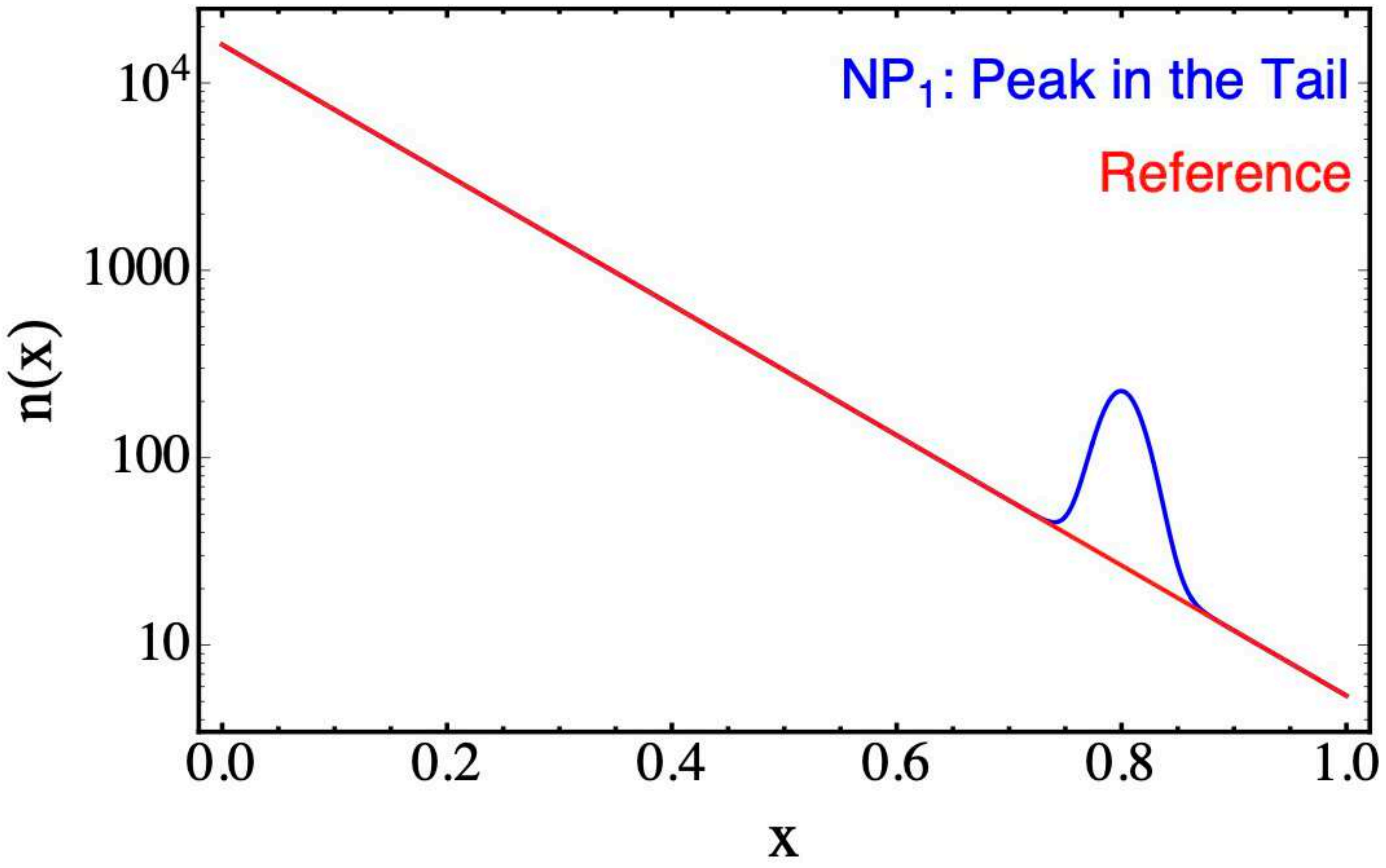
EXAMPLES — CWOLA Bumphunter

[J.H. Collins, K. Howe and B. Nachman '18]

[J.H. Collins, K. Howe and B. Nachman '19]

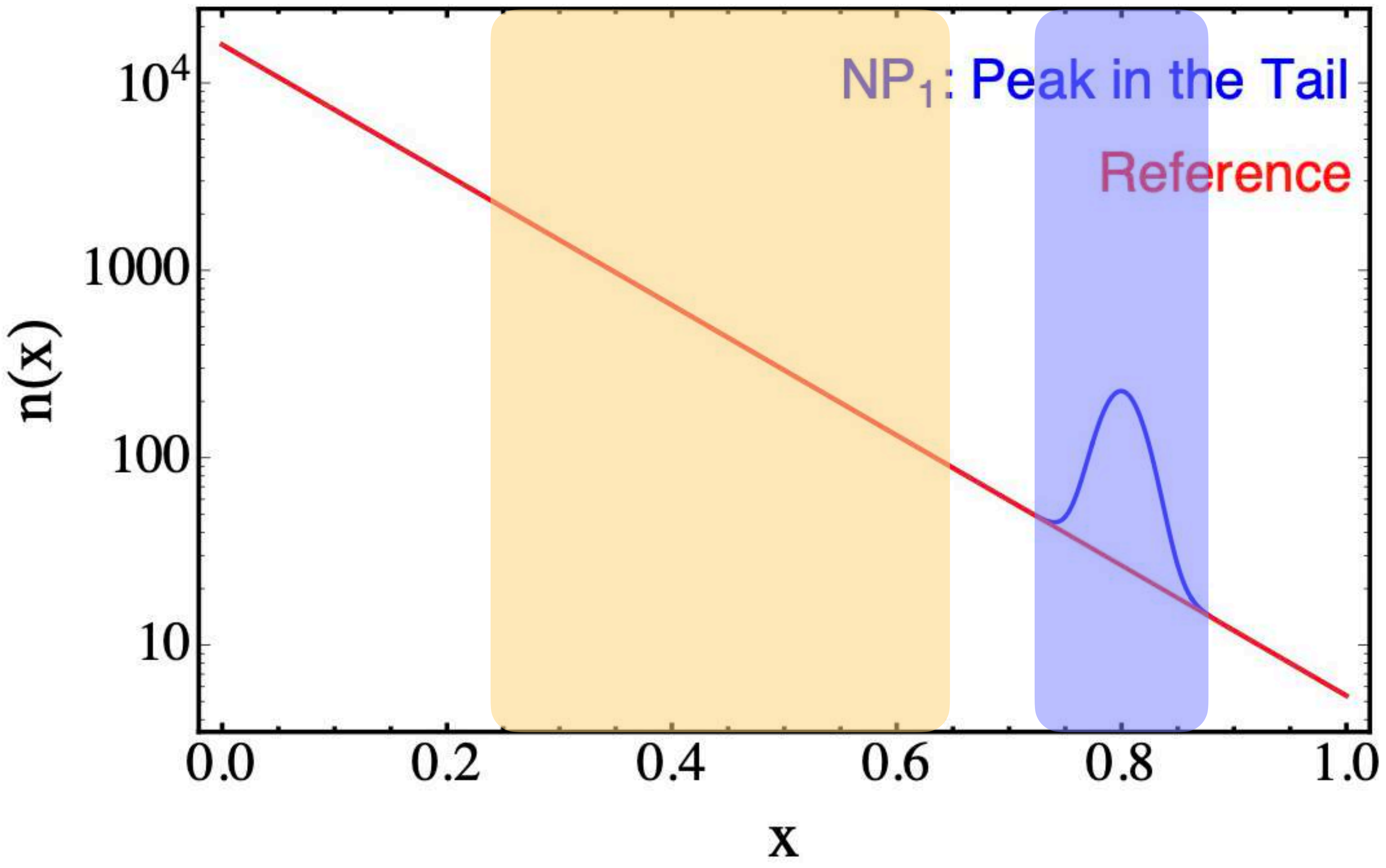
[K. Benkendorfer, L.L. Pottier and B. Nachman '20]

1. Choose an "Invariant mass" variable

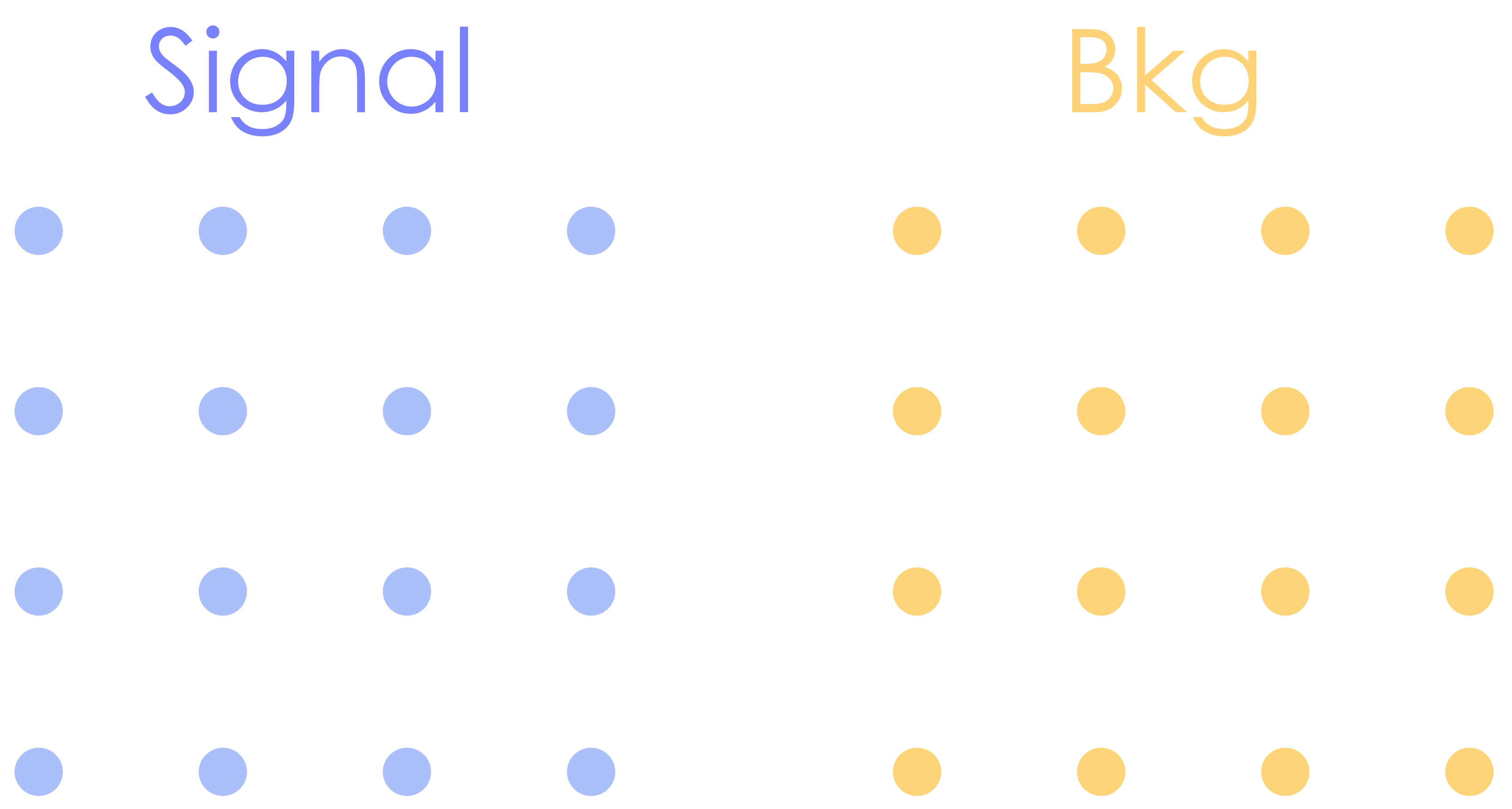


EXAMPLES — CWOLA Bumphunter

1. Choose an "Invariant mass" variable

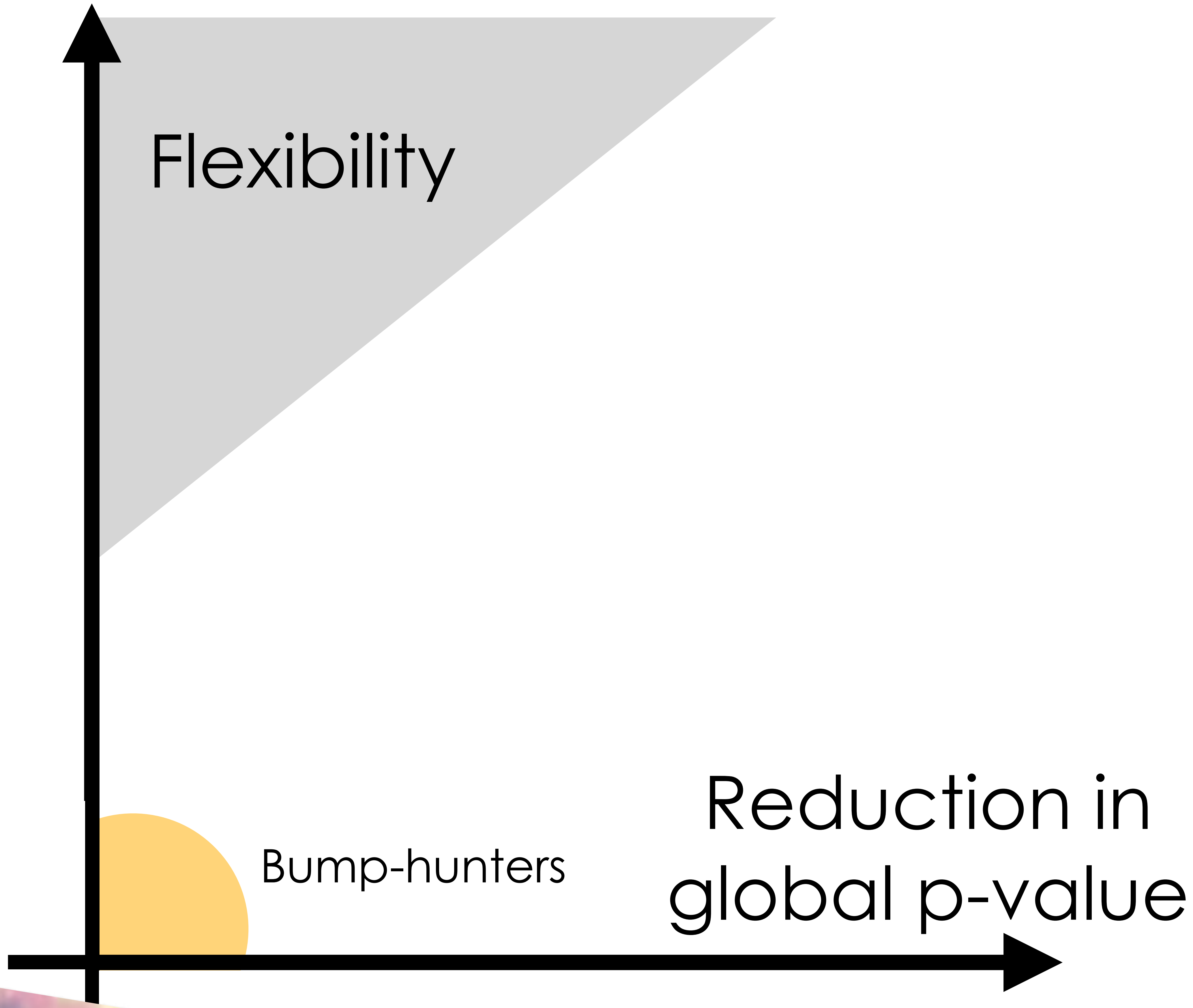


2. Get Signal and Background

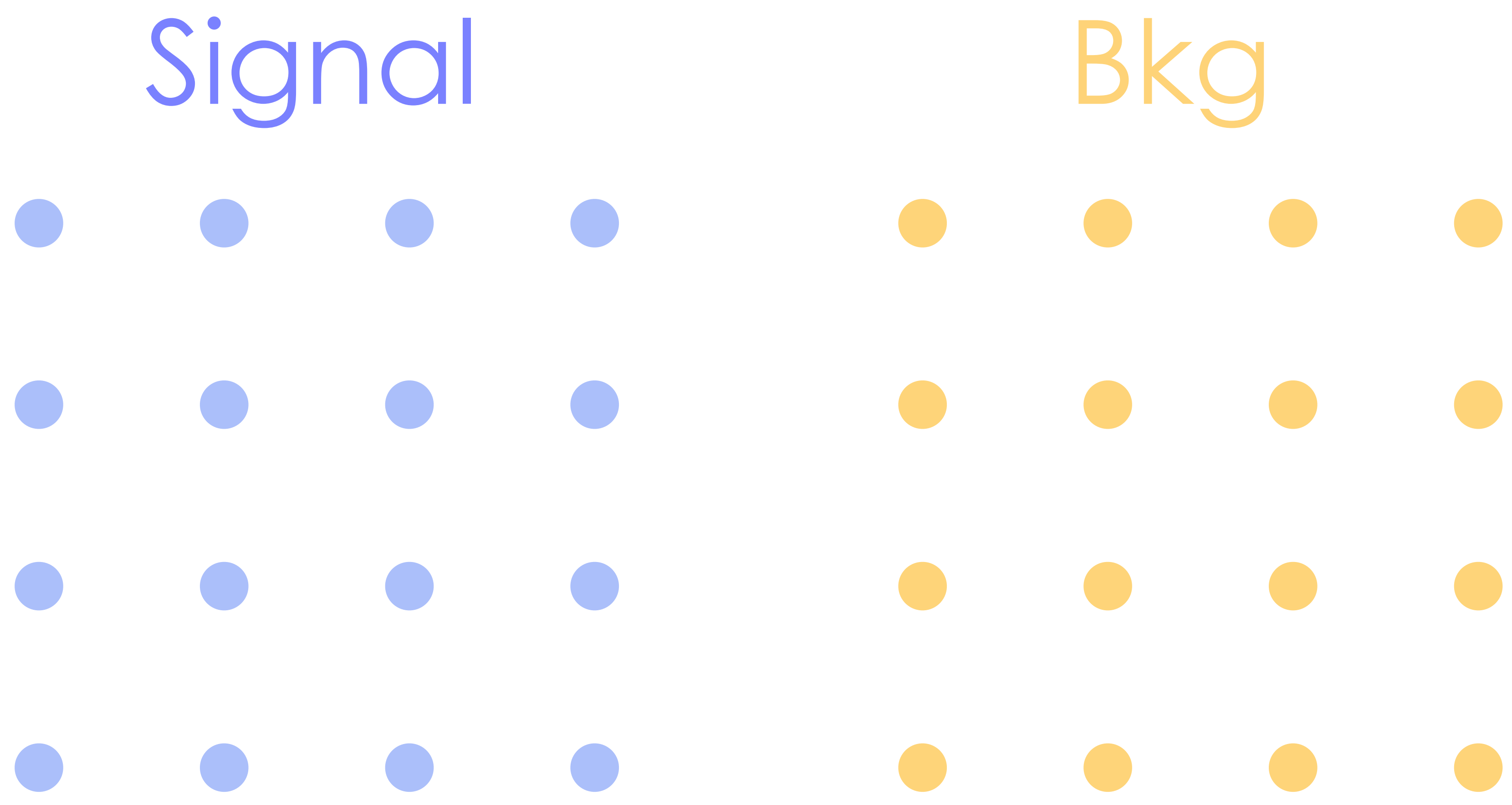


EXAMPLES — CWOLA Bumphunter

3. Optimize over the other variables using a classifier



2. Get Signal and Background



EXAMPLES — NPLM

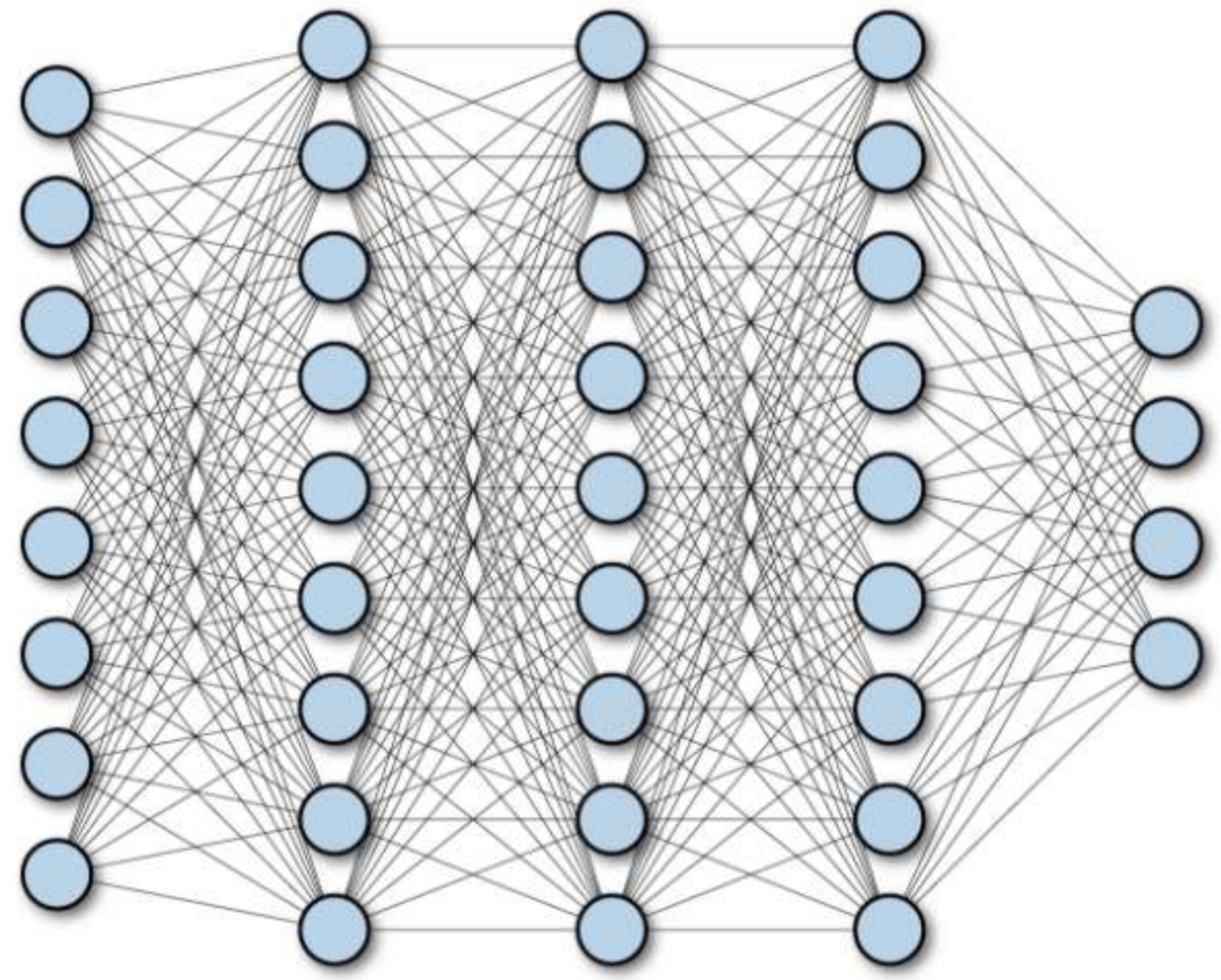
[RTD, A. Wulzer '18]

[RTD, G. Grosso, M. Pierini, A. Wulzer, M. Zanetti '19]

[RTD, G. Grosso, M. Pierini, A. Wulzer, M. Zanetti '21]

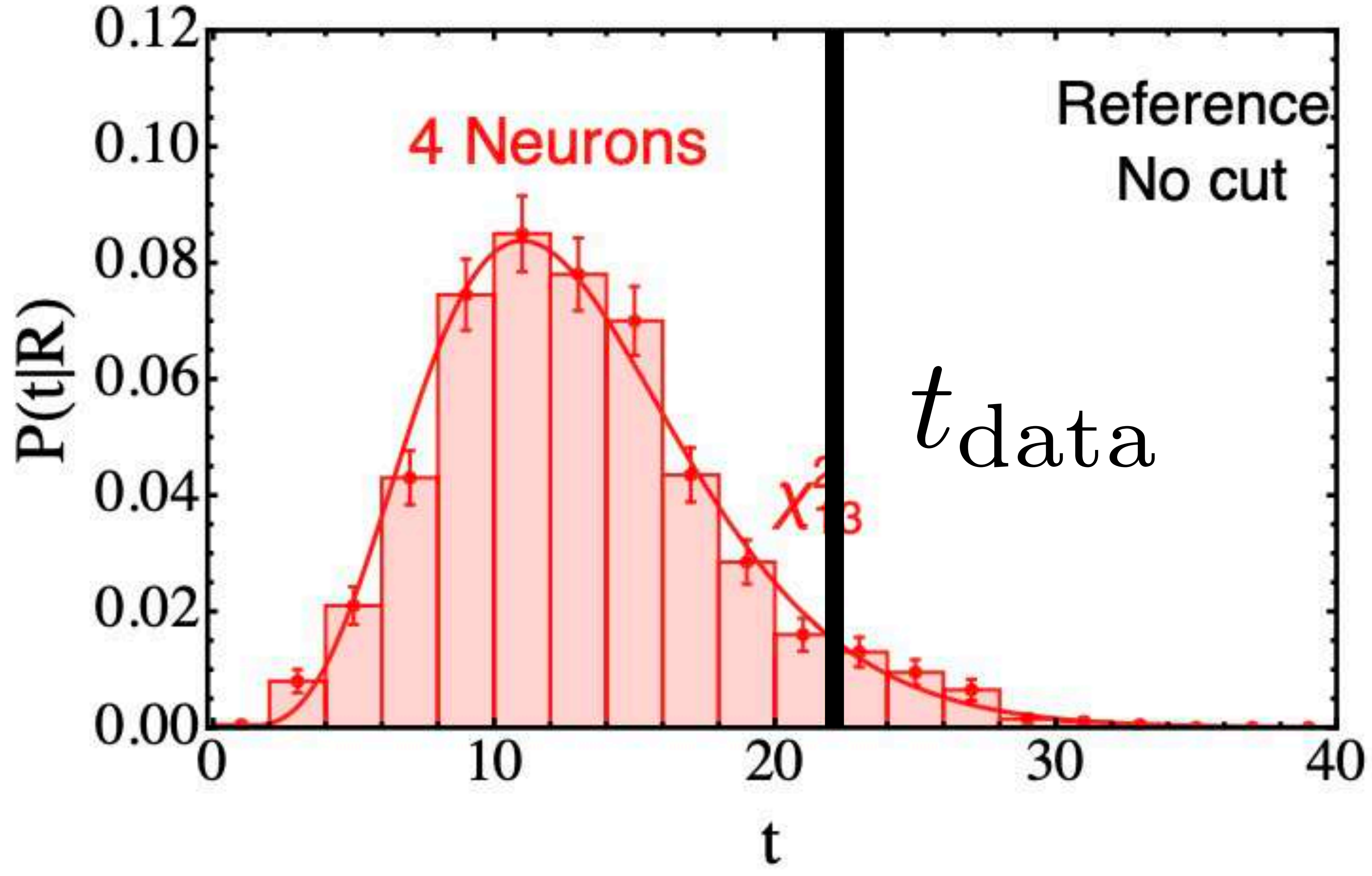
PDF of the data

$$f(\{\vec{p}_i\}|\mathbf{w})$$



EXAMPLES — NPLM

$$t(\mathcal{D}) = \sum_{p \in \mathcal{D}} \log \frac{f(p|\hat{\mathbf{w}})}{f(p|\mathrm{SM})}$$



EXAMPLES — NPLM

$$t(\mathcal{D}) = \sum_{p \in \mathcal{D}} \log \frac{f(p|\hat{\mathbf{w}})}{f(p|\mathrm{SM})}$$

Tunes Flexibility vs Curse

$$||\mathbf{w}|| \leq w_{\max}$$

EXAMPLES — Autoencoders

[M. Farina, Y. Nakai and D. Shih ’18]
[T. Heimel, G. Kasieczka, T. Plehn and J.M. Thompson ’18]
[T. Finke, M. Kr¨amer, A. Morandini, A. M¨uck and I. Oleksiyuk ’21]
[CMS ’25]

Same principle of NPLM, but

$$f\left(\{\vec{p}_i\}\middle|\mathbf{w}\right) = \text{pdf of the SM}$$

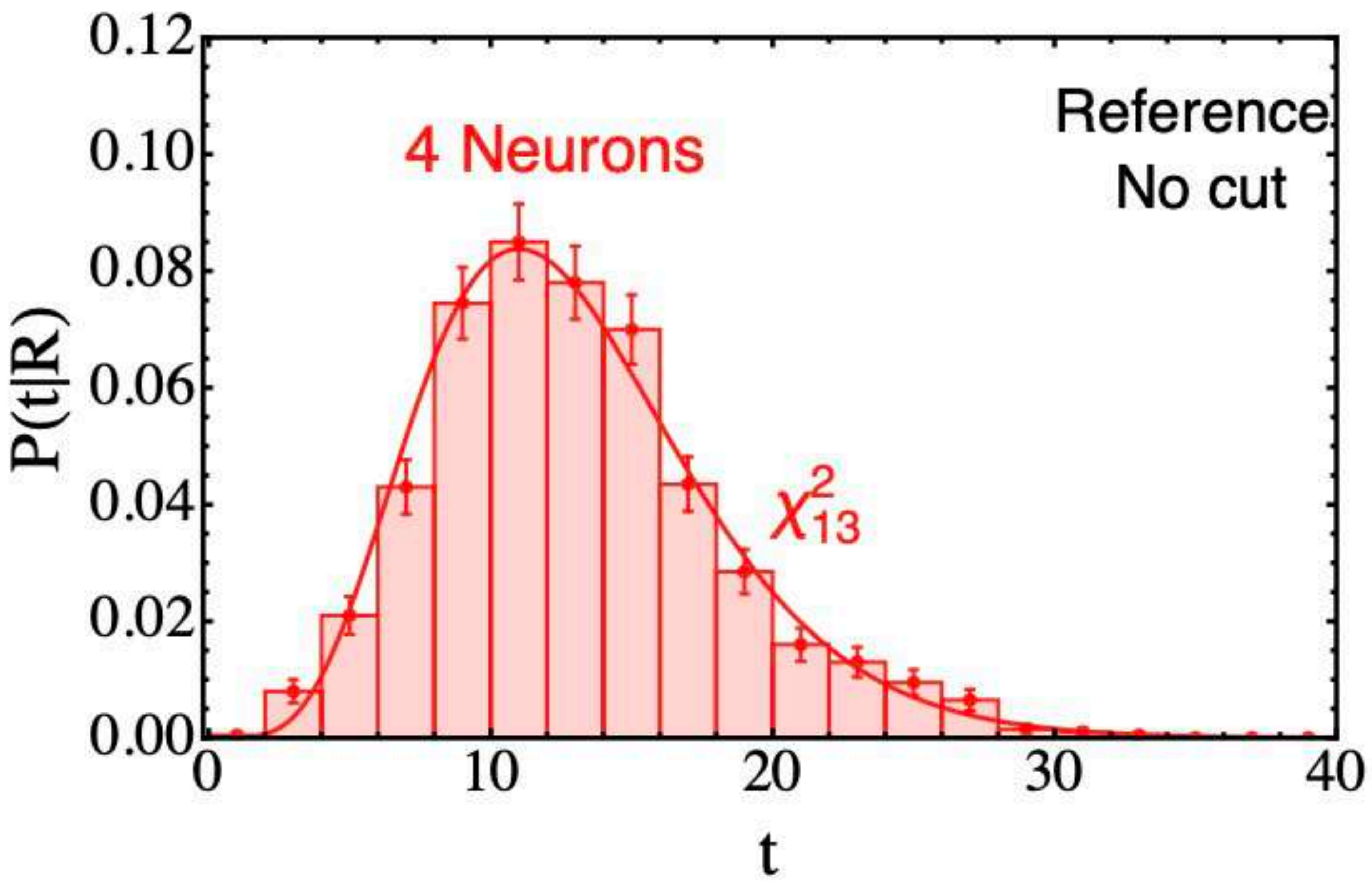
CURSE OF DIMENSIONALITY

PDF of the data vs SM prediction

$$f(\{\vec{p}_i\}|\mathbf{w}) \qquad \text{vs} \qquad f(\{\vec{p}_i\}|\mathrm{SM})$$

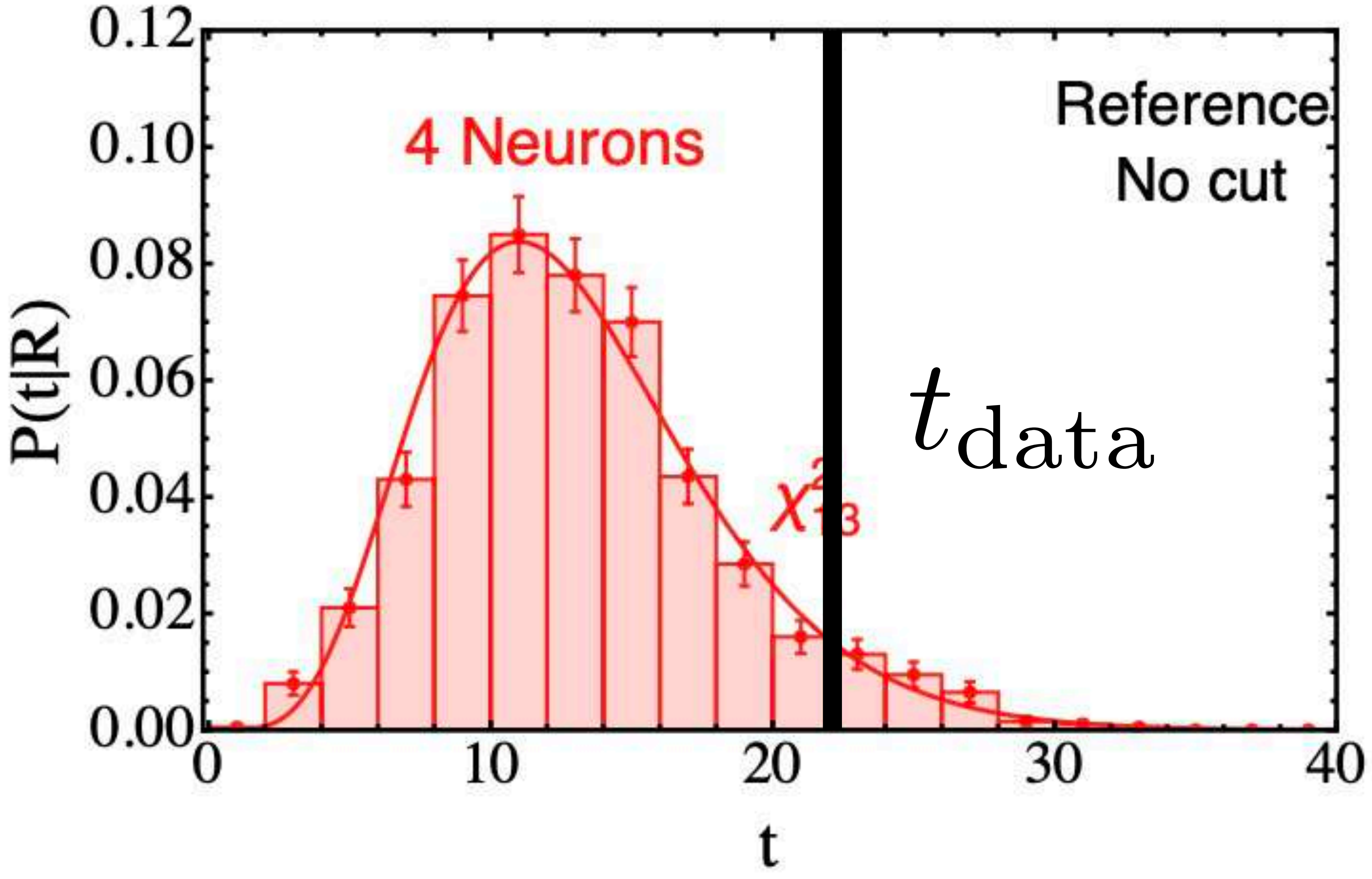
CURSE OF DIMENSIONALITY

$$t(\mathcal{D}) = \sum_{p \in \mathcal{D}} \log \frac{f(p|\hat{\mathbf{w}})}{f(p|\mathrm{SM})}$$



CURSE OF DIMENSIONALITY

$$t(\mathcal{D}) = \sum_{p \in \mathcal{D}} \log \frac{f(p|\hat{\mathbf{w}})}{f(p|\mathrm{SM})}$$



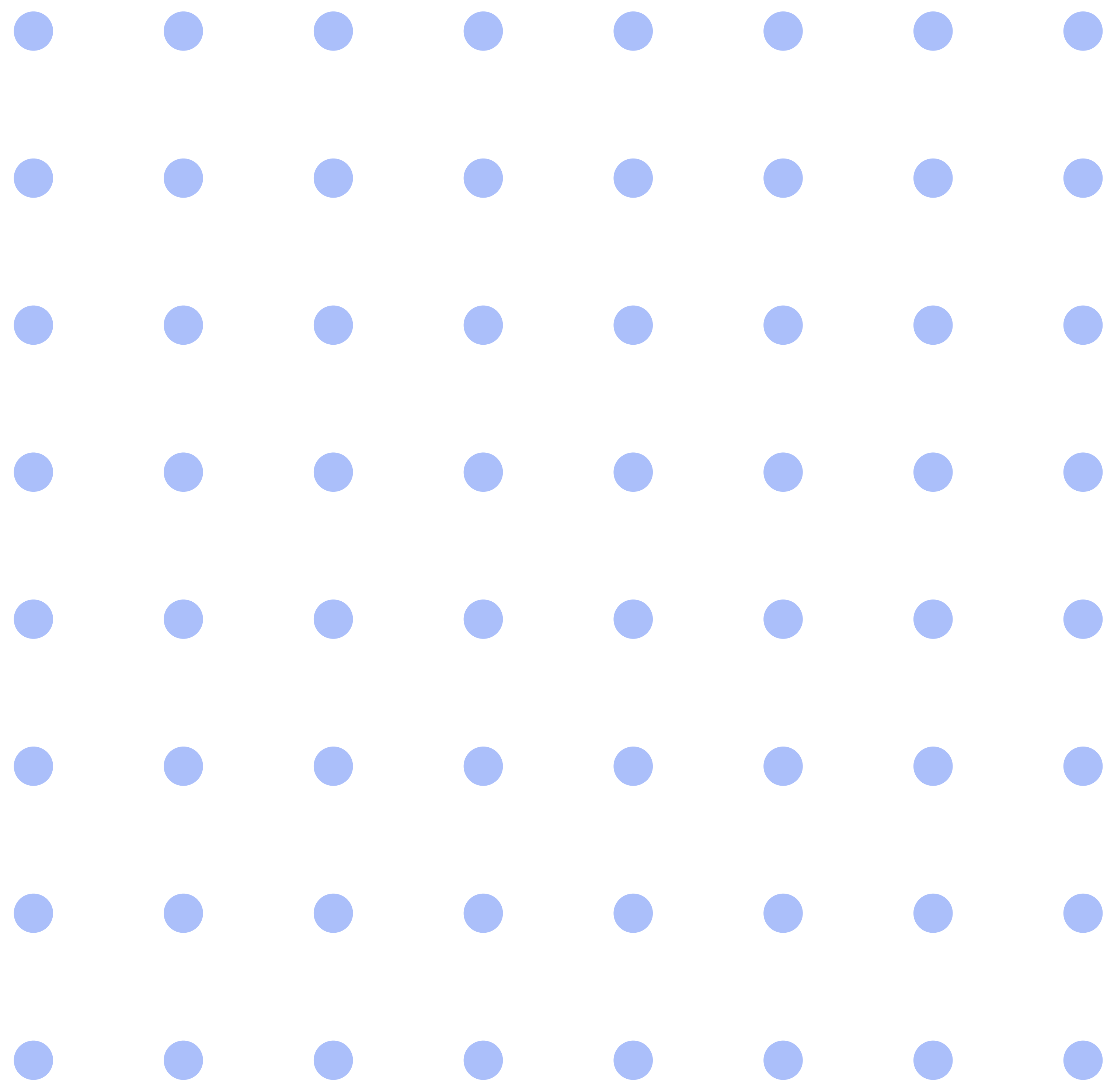
CURSE OF DIMENSIONALITY

PDF of the data

$$f(\{\vec{p}_i\}|\mathbf{w})$$

1D

2D

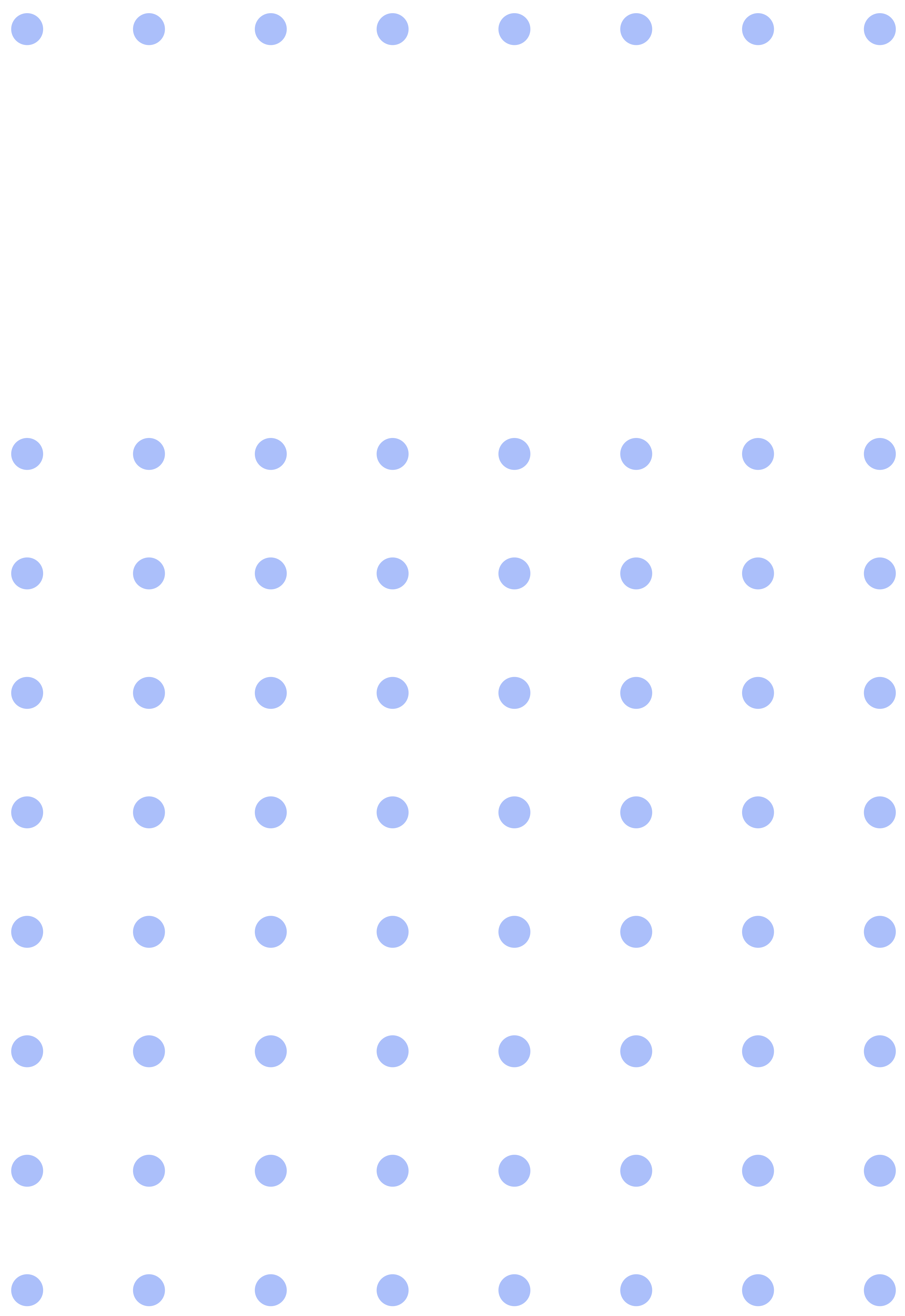


CURSE OF DIMENSIONALITY

$$N_{\text{data}} = \left(\frac{1}{\delta p}\right)^D$$

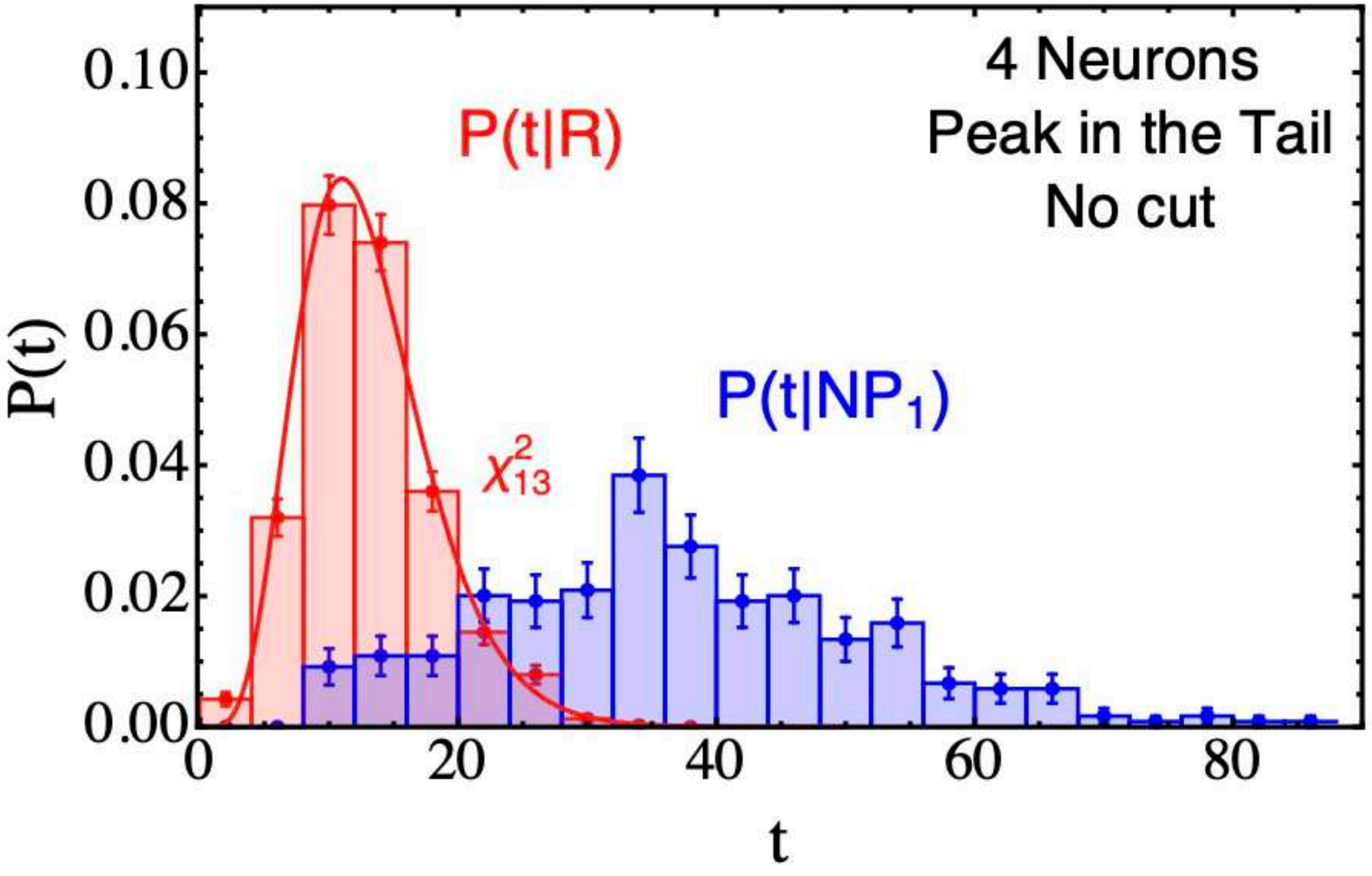
1D

2D

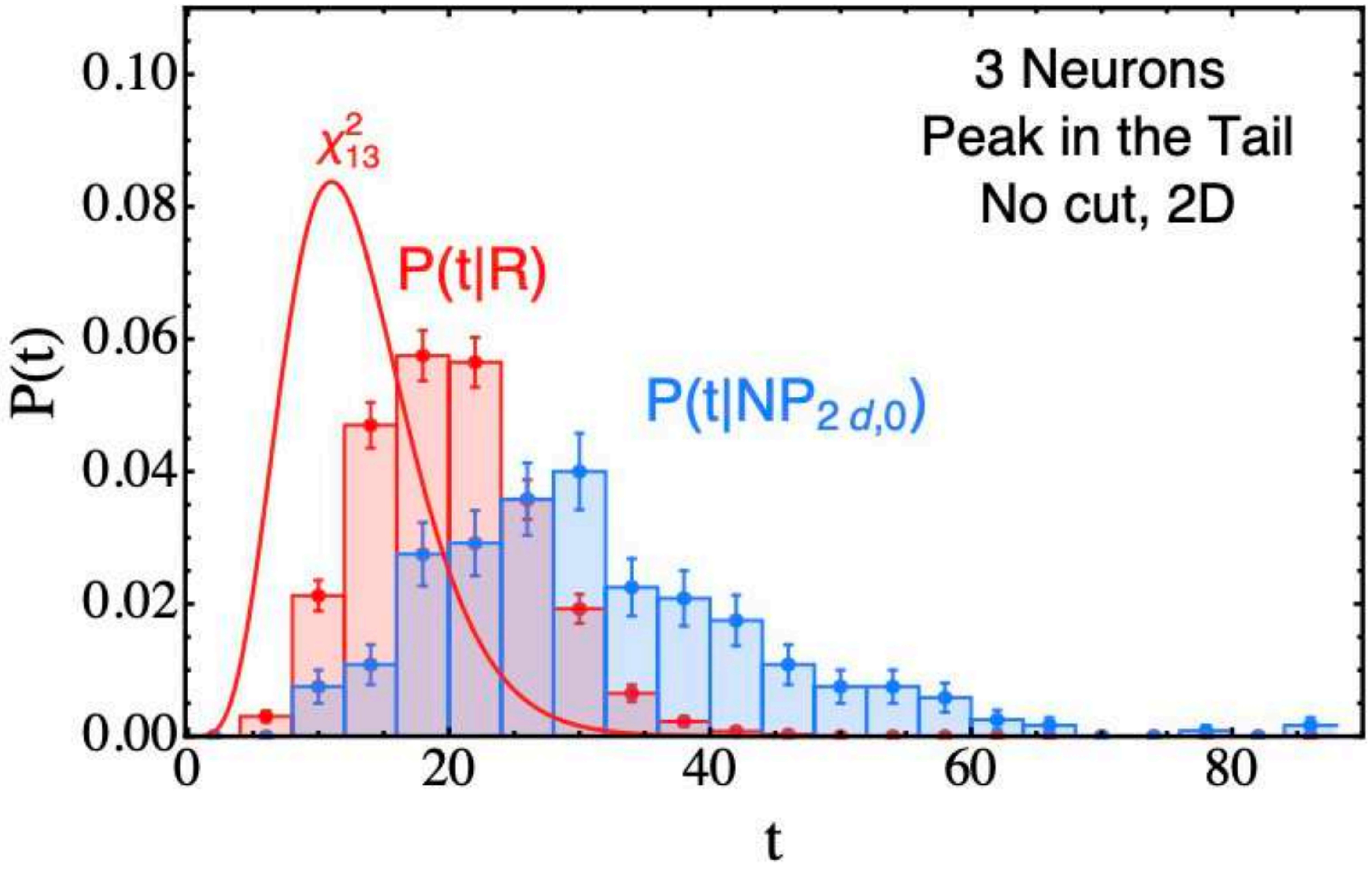


CURSE OF DIMENSIONALITY

1D



2D



CURSE OF DIMENSIONALITY

Number of observables

$$D \gg 1$$

Number of pdfs that can fit the data

$$N^D$$

Significance of observation

$$p_{\text{global}} = p_{\text{local}} \times N^D$$

$$r_i = \text{Earth Mover's Distance}$$

[Komiske, Metodiev, Thaler '19]

EMD

[D. Ba, A. S. Dogra, R. Gambhir, A. Tasissa, J. Thaler '23]

1. Finite when evaluated on atomic measures
2. Positive and close
3. Symmetric
4. Weakly continuous (IRC safe)
5. Faithfully lifts the ground metric (euclidean distance on the calorimeter)

EMD

$$\mathrm{EMD}(\mathcal{E}_A, \mathcal{E}_B) = \min_{\{\xi_{km}\}} \sum_{k,m} \xi_{km} \frac{\theta_{km}}{R} + \left| \sum_{k \in \mathcal{E}_A} E_k - \sum_{m \in \mathcal{E}_B} E_m \right|$$

$$\theta_{km} \equiv \sqrt{(y_k - y_m)^2 + (\phi_k - \phi_m)^2}$$

$$\xi_{km} \geq 0\,, \quad \sum_{k \in \mathcal{E}_A} \xi_{km} \leq E_m\,, \quad \sum_{m \in \mathcal{E}_B} \xi_{km} \leq E_k, \quad \sum_{k,m} \xi_{km} \leq E_{\min}$$

DISTANCE BETWEEN DATASETS

[Komiske, Metodiev, Thaler '19]

$$\Sigma\mathrm{MD}(\mathcal{D}_A, \mathcal{D}_B) = \min_{\{\zeta_{km}\}} \sum_{k,m} \zeta_{km} \frac{\mathrm{EMD}(\mathcal{E}_k, \mathcal{E}_m)}{R} + \left| \sum_{k \in \mathcal{D}_A} \sigma_k - \sum_{m \in \mathcal{D}_B} \sigma_m \right|$$

$$\Sigma\mathrm{MD}(\mathcal{D}_A, \mathcal{D}_B) = 0 \leftrightarrow \mathcal{D}_A = \mathcal{D}_B$$

Can we use it as a test statistic?

$$\zeta_{km} \leftrightarrow \mathbf{w}$$

$$\Sigma\mathrm{MD}(\mathcal{D}_A, \mathcal{D}_B)$$

Wasserstein distance

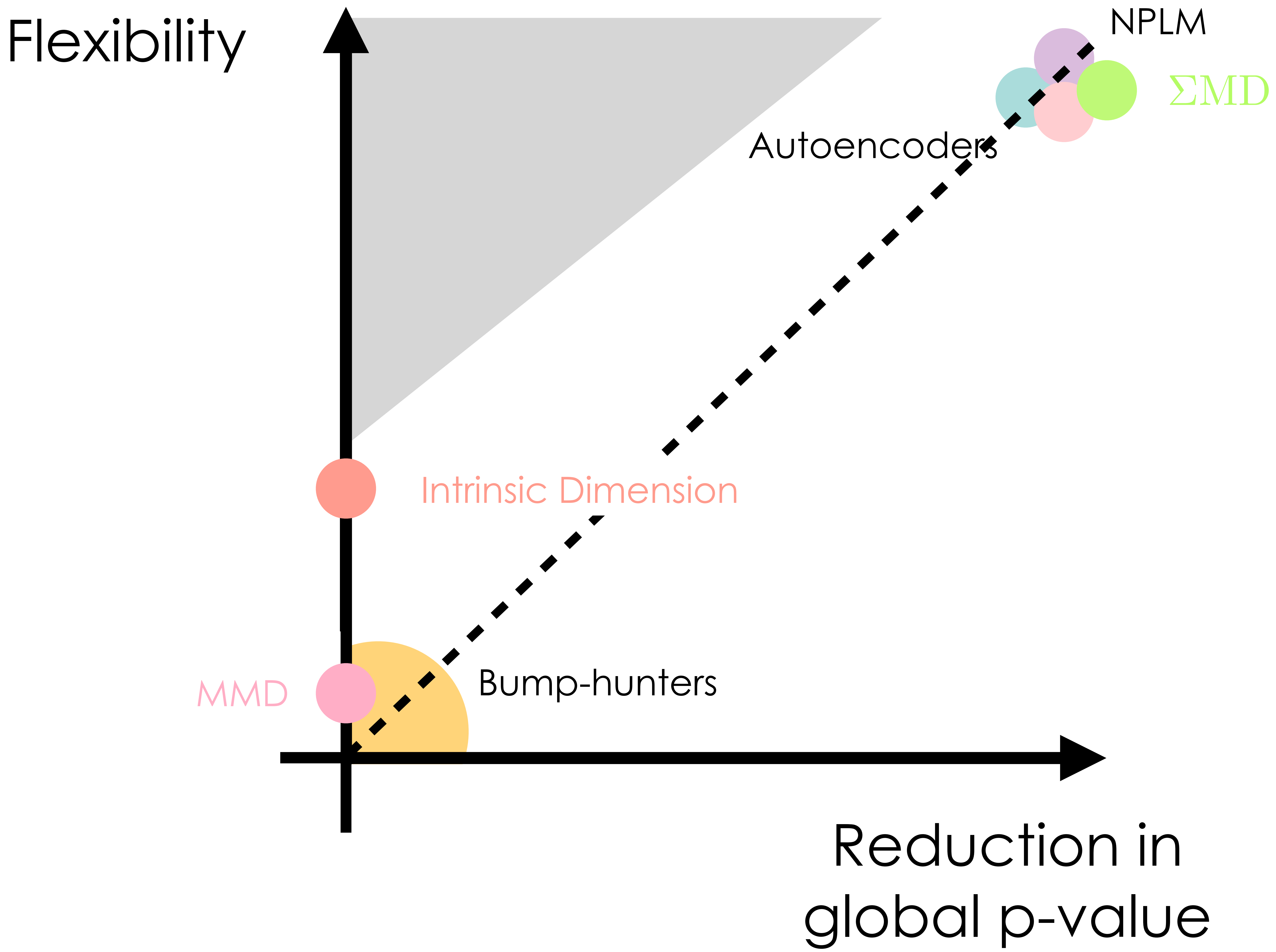
[Ramdas, Garcia Trillos, Cuturi '15]

$$W_1(\mathcal{D}_A, \mathcal{D}_B | \zeta) - \frac{1}{\lambda} \sum_{km} \zeta_{km} \log \zeta_{km}$$

$$\lambda \to 0$$

$$\mathrm{MMD}(\mathcal{D}_A, \mathcal{D}_B) = -\frac{1}{2} \sum_{k \in \mathcal{D}_A} \sum_{m \in \mathcal{D}_B} \mathrm{EMD}(\mathcal{E}_k, \mathcal{E}_m)(\sigma_k - \sigma'_k)(\sigma_m - \sigma'_m)$$

$$\lambda \leftrightarrow w_{\text{max}}$$


Flexibility
NPLM
ΣMD
Autoencoders
Intrinsic Dimension
MMD
Bump-hunters
Reduction in global p-value

Data

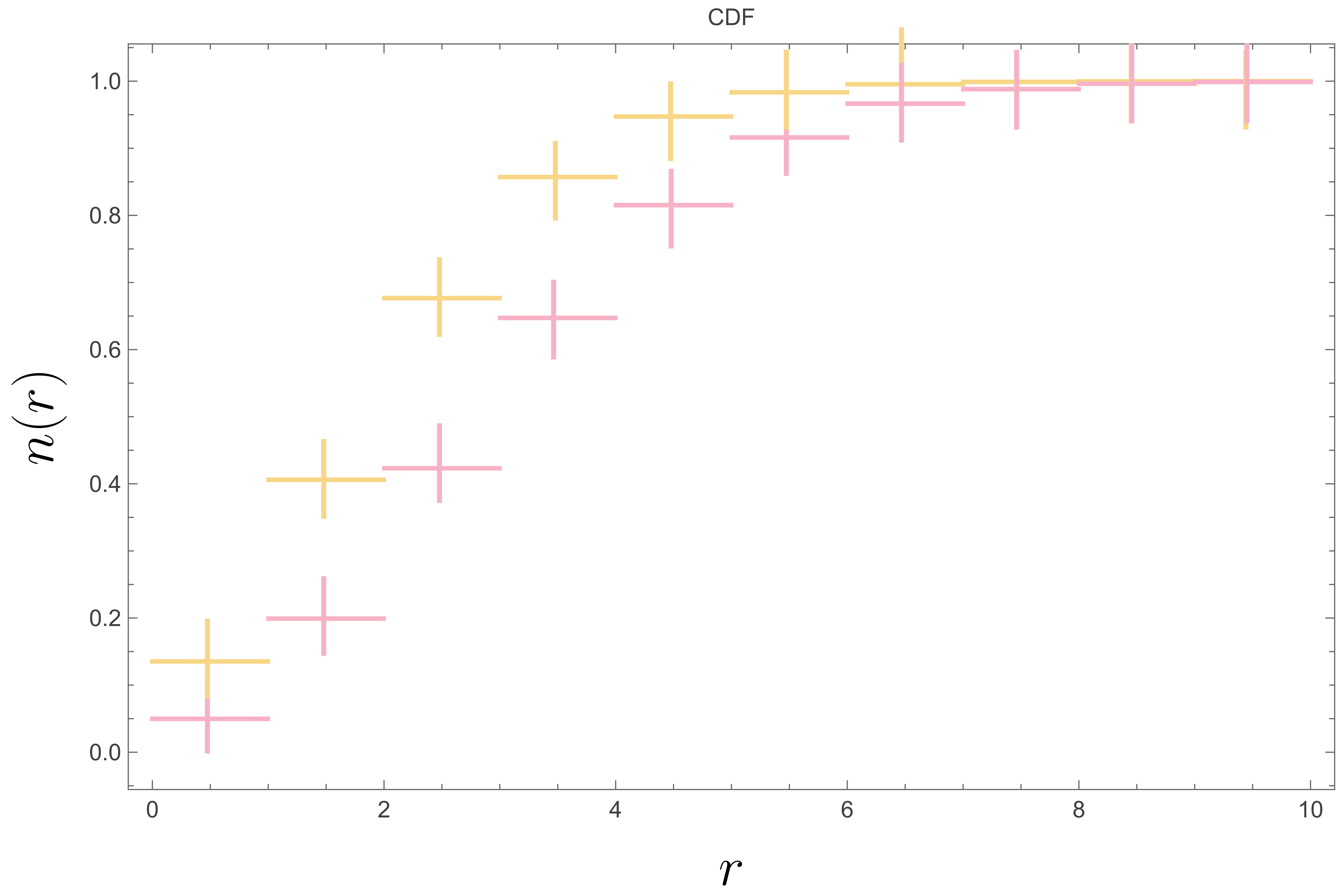
SM


Beacon St
Boston Common
Esplanade
Charles River
CHURCH GREEN

$$n_d(r)$$

$$n_d(r)$$

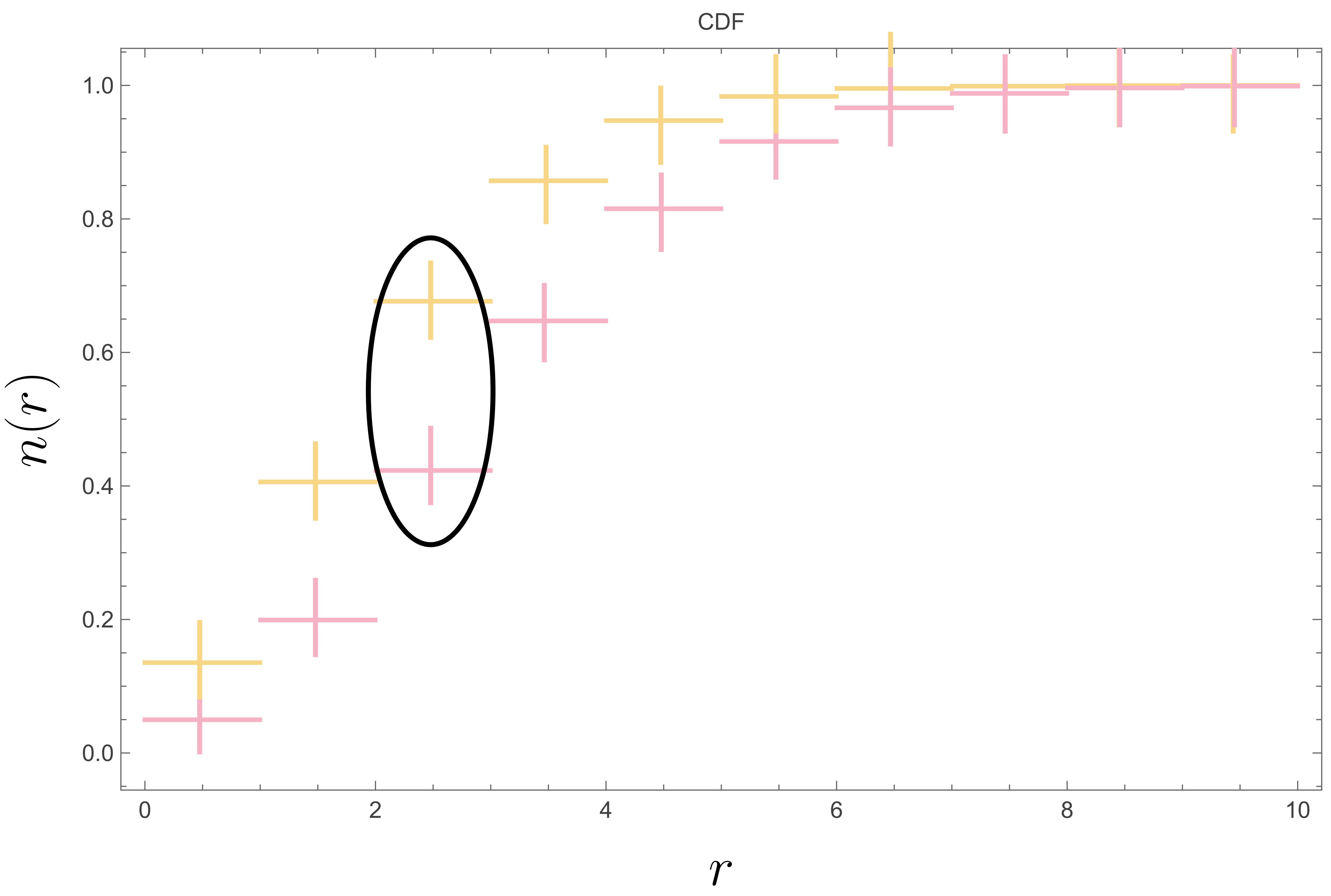
$$n_{\mathrm{SM}}(r)$$



$n_d(r)$

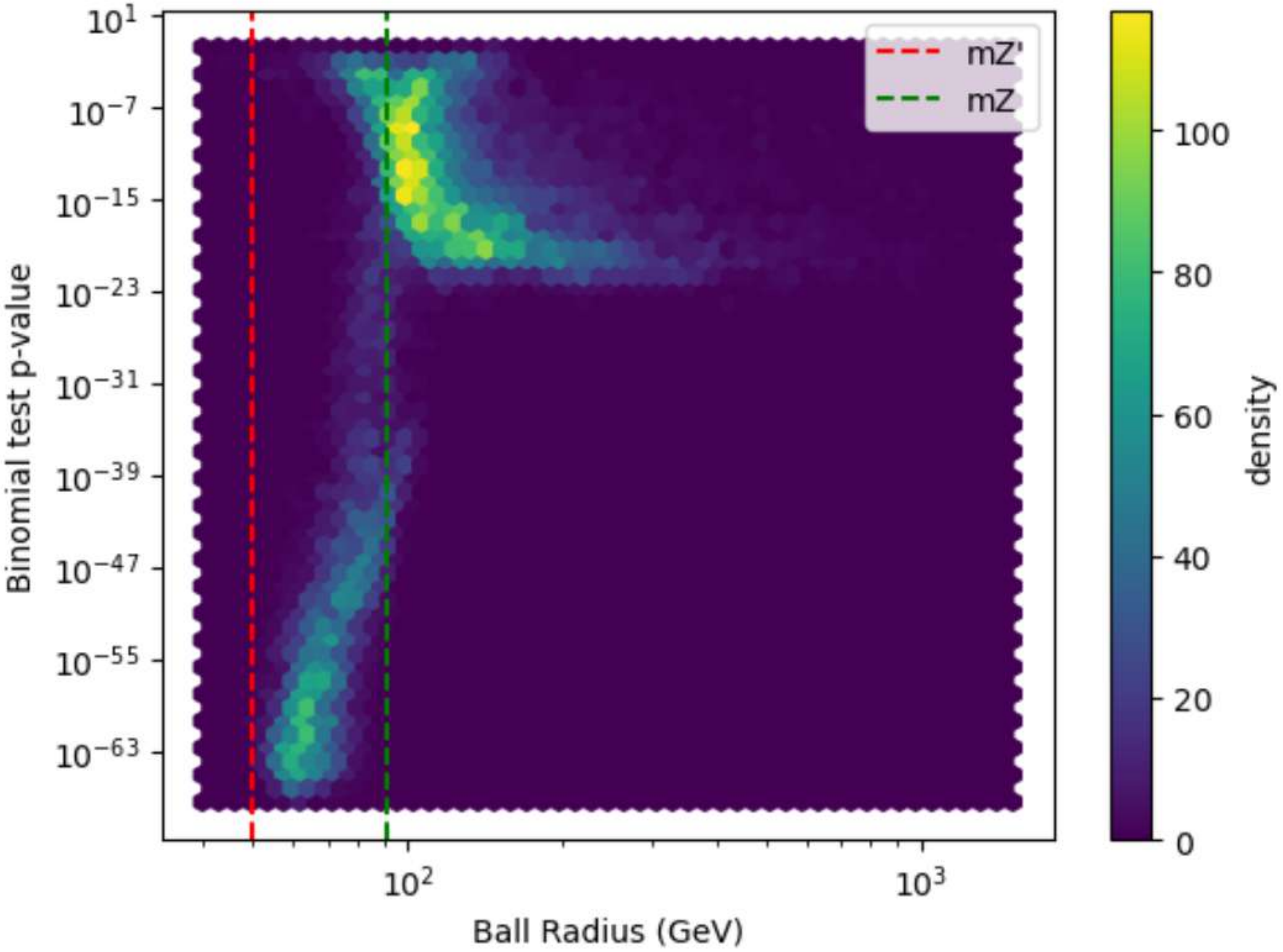
$n_{\mathrm{SM}}(r)$

TEST STATISTIC



$$\max_{r,d} p(r,d)$$

SOME VERY PRELIMINARY RESULTS



CONCLUSION

Think globally, act locally

MSSM Benchmarks

$$m_{\chi_1^0} = 150 \text{ GeV and } m_{\chi_2^0} = m_{\chi_1^+}, \text{ with } m_{\chi_2^0,\chi_1^+} = 250, 500, 1000 \text{ GeV}$$

EXAMPLES — Dimensionality reduction

[S. Bright-Thonney, C. Reissel, G. Grosso, N. Woodward, K. Govorkova, A. Novak et al. '25]

[G. Grosso, S.S.R. Hindupur, T. Fel, S. Bright-Thonney, P. Harris and D. Ba '25]

$$m_{\chi_1^0} = 150 \text{ GeV and } m_{\chi_2^0} = m_{\chi_1^+}, \text{ with } m_{\chi_2^0,\chi_1^+} = 250, 500, 1000 \text{ GeV}$$