Exploiting the discovery potential of the LHC data using the Data Directed Paradigm

Shikma Bressler

Bump hunt DDP - S. Volkovitch, F. DeVito Halevi, SB [2107.11573]
Symmetry DDP - M. Birman, B. Nachman, R. Sebbah, G. Sela, O. Turetz, SB [2203.07529]

ICHEP 2022 BOLOGNA

מכון ויצמן למדע
WEIZMANN INSTITUTE OF SCIENCE

Most searches conducted following the blind analysis paradigm

- Signal region selection motivated by
  - Theoretical considerations –
    highly motivated models are the first one to be tested
  - Final states (di-X resonance searches)
  - Topologies

Most searches conducted following the blind analysis paradigm

- Advantage
  - Smaller chance for biassing the results
  - Best sensitivity for pre-defined signals
- Disadvantages
  - Thousands of person years are spent studying region of the data that turn out to have nothing interesting in them
  - Large portion of the data is not fully exploited

Example I - Resonances

arXiv:1907.06659

	e	μ	τ	q/g	b	t	γ	Z/W	H	BSM → $SM_1 \times SM_1$: q/g	γ/π^0's	b	…	BSM → $SM_1 \times SM_2$: tZ/H	bH	…	BSM → complex: $\tau qq'$	eqq'	$\mu qq'$	…
e	[37,38]	[39,40]	[39]	∅	∅	∅	[41]	[42]	∅	∅	∅	∅		∅	∅	∅	∅	[43,44]	∅	
μ		[37,38]	[39]	∅	∅	∅	[41]	[42]	∅	∅	∅	∅		∅	∅	∅	∅	∅	[43,44]	
τ			[45,46]	∅	[47]	∅	∅	∅	∅	∅	∅	∅		∅	∅	∅	[48,49]	∅	∅	
q/g				[29,30,50,51]	[52]	∅	[53,54]	[55]	∅	∅	∅	∅		∅	∅	∅	∅	∅	∅	
b					[29,52,56]	[57]	[54]	[58]	[59]	∅	∅	∅		[60]	∅	∅	∅	∅	∅	
t						[61]	∅	[62]	[63]	∅	∅	∅		[64]	[60]	∅	∅	∅	∅	
γ							[65,66]	[67–69]	[68,70]	∅	∅	∅		∅	∅	∅	∅	∅	∅	
Z/W								[71]	[71]	∅	∅	∅		∅	∅	∅	∅	∅	∅	
H									[72,73]	[74]	∅	∅		∅	∅	∅	∅	∅	∅	
BSM → $SM_1 \times SM_1$: q/g										∅	∅	∅		∅	∅	∅	∅	∅	∅	
γ/π^0's											[75]	∅		∅	∅	∅	∅	∅	∅	
b												[76,77]		∅	∅	∅	∅	∅	∅	
⋮																				
⋮																				

Mostly inclusive searches – so the data should be exploited far beyond what is seen in this table

Example II – LU and LFC

- LU and LFC in the SM give rise to symmetry between e's, μ's and τ's
  - Up to Yukawa interactions and phase space effect
- The discovery of an e/μ asymmetry == New physics
  - Strong motivation to search for such asymmetries
- In practice, most of the data is not yet explored

Tested symmetry		inclusive	Object multiplicity						Decays			Topologies		
			τ	j	bj	γ	K	...	Z	W	...	VBF	$t\bar{t}$	...
LU - U(3)	e/μ									[31-34]				
	$ee/\mu\mu$ OS						[24]		[29,30]					
	$ee/\mu\mu$ SS													
LFC - $U(1)^3$	$e\mu/\mu e$ OS	[27,28]										[27,28]		
	$e\mu/\mu e$ SS													
	$e^+\mu^-/e^-\mu^+$	[35]		[35]										

Table is a WIP

Example II – LU and LFC

- LU and LFC in the SM give rise to symmetry between e's, μ's and τ's
  - Up to Yukawa interactions and phase space effect
- The discovery of an e/μ asymmetry == New physics
  - Strong motivation to search for such asymmetries
- In practice, most of the data is not yet explored

Higgs LFV decays in ATLAS

R_K anomaly

Table is a WIP

Tested symmetry		inclusive	Object multiplicity						Decays			Topologies		
			τ	j	bj	γ	K	...	Z	W	...	VBF	$t\bar{t}$	...
LU - U(3)	e/μ $ee/\mu\mu$ OS $ee/\mu\mu$ SS						[24]		[29,30]	[31-34]				
LFC - $U(1)^3$	$e\mu/\mu e$ OS $e\mu/\mu e$ SS	[27,28]										[27,28]		
	$e^+\mu^-/e^-\mu^+$	[35]		[35]										

Searches based on the data

- Event-based anomaly detection
  - Autoencoders etc.
- Semi-supervised approaches

The LHC Olympics 2020

A Community Challenge for Anomaly Detection in High Energy Physics

Gregor Kasieczka (ed),[1] Benjamin Nachman (ed),[2,3] David Shih (ed),[4] Oz Amram,[5] Anders Andreassen,[6] Kees Benkendorfer,[2,7] Blaz Bortolato,[8] Gustaaf Brooijmans,[9] Florencia Canelli,[10] Jack H. Collins,[11] Biwei Dai,[12] Felipe F. De Freitas,[13] Barry M. Dillon,[8,14] Ioan-Mihail Dinu,[5] Zhongtian Dong,[15] Julien Donini,[16] Javier Duarte,[17] D. A. Faroughy[10] Julia Gonski,[9] Philip Harris,[18] Alan Kahn,[9] Jernej F. Kamenik,[8,19] Charanjit K. Khosa,[20,30] Patrick Komiske,[21] Luc Le Pottier,[2,22] Pablo Martín-Ramiro,[2,23] Andrej Matevc,[8,19] Eric Metodiev,[21] Vinicius Mikuni,[10] Inês Ochoa,[24] Sang Eon Park,[18] Maurizio Pierini,[25] Dylan Rankin,[18] Veronica Sanz,[20,26] Nilai Sarda,[27] Uroš Seljak,[2,3,12] Aleks Smolkovic,[8] George Stein,[2,12] Cristina Mantilla Suarez,[5] Manuel Szewc,[28] Jesse Thaler,[21] Steven Tsan,[17] Silviu-Marian Udrescu,[18] Louis Vaslin,[16] Jean-Roch Vlimant,[29] Daniel Williams,[9] Mikaeel Yunus[18]

Searches based on the data

- Event-based anomaly detection
  - Autoencoders etc.
- Semi-supervised approaches

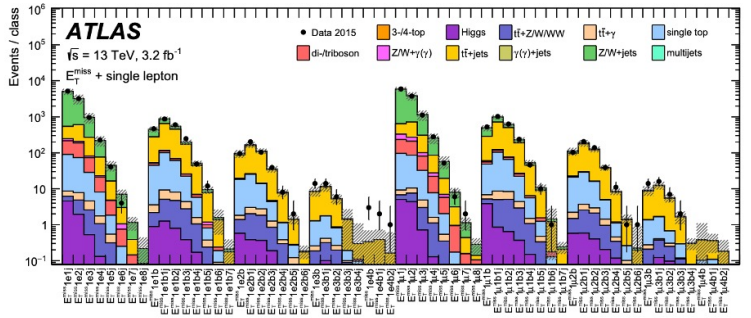
- Generic data/mc comparison
  - HERA → D0 & CDF → CMS & ATLAS
  - Sensitivity to MC mismodeling and statistics

arXiv:1807.07447

- Data is divided to mutually exclusive classes according to object multiplicity e.g. $1\mu, 3e3j, \ldots 10jMET, \ldots.$ etc.

Object	Label	p_{T} (min) [GeV]	Pseudorapidity
Isolated electron	e	25	$\lvert\eta\rvert < 1.37$ or $1.52 < \lvert\eta\rvert < 2.47$
Isolated muon	μ	25	$\lvert\eta\rvert < 2.7$
Isolated photon	γ	50	$\lvert\eta\rvert < 1.37$ or $1.52 < \lvert\eta\rvert < 2.37$
b-tagged jet	b	60	$\lvert\eta\rvert < 2.5$
Light (non-b-tagged) jet	j	60	$\lvert\eta\rvert < 2.8$
Missing transverse momentum	$E_{\mathrm{T}}^{\mathrm{miss}}$	200	

- A total of 704 event classes with at least 1 data event and/or 0.1 SM event prediction



Searches based on the data

- Event-based anomaly detection
  - Autoencoders etc.
- Semi-supervised approaches

- Generic data/mc comparison
- The data directed paradigm
  - Identify a property of the SM and look for regions exhibiting significant deviation from this property
    No MC is needed.

The Data Directed Paradigm

- Our proposal relies solely on the data
  - Not limited by MC
- Based on two key ingredients
  - A theoretically well established property of the SM based on which deviations from the SM predictions can be searched for
  - An efficient tool that allows rapid scanning of many final states in search for such a deviation
- Complementary to ML-based method developed in an attempt to enhance the Signal/Background ratio

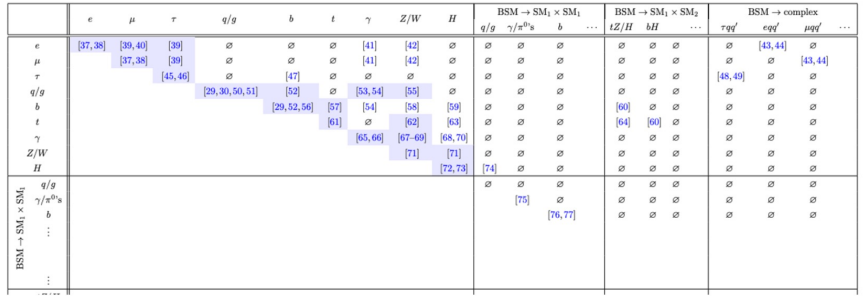


Table is a WIP

Tested symmetry		inclusive	Object multiplicity τ	j	bj	γ	K	...	Decays Z	W	...	Topologies VBF	$t\bar{t}$	...
LU - U(3)	e/μ $ee/\mu\mu$ OS $ee/\mu\mu$ SS						[24]		[29,30]	[31-34]				
LFC - $U(1)^3$	$e\mu/\mu e$ OS $e\mu/\mu e$ SS	[27,28]										[27,28]		
CP	$e^+\mu^-/e^-\mu^+$	[35]		[35]										

The Data Directed Paradigm

- Our proposal relies solely on the data
  - Not limited by MC
- Based on two key ingredients
  - A theoretically well established property of the SM based on which deviations from the SM predictions can be searched for
  - **An efficient tool that allows rapid scanning of many final states in search for such a deviation**
- Complementary to ML-based method developed in an attempt to enhance the Signal/Background ratio

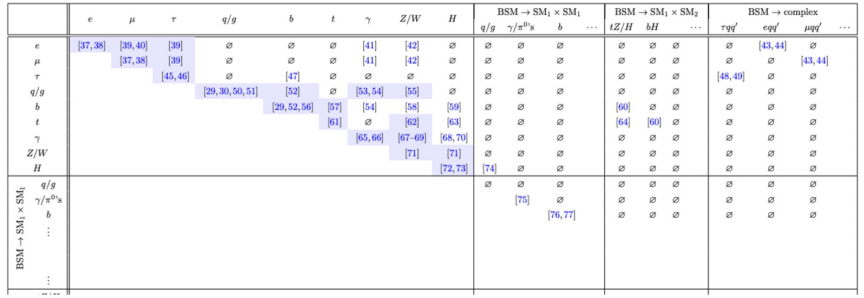


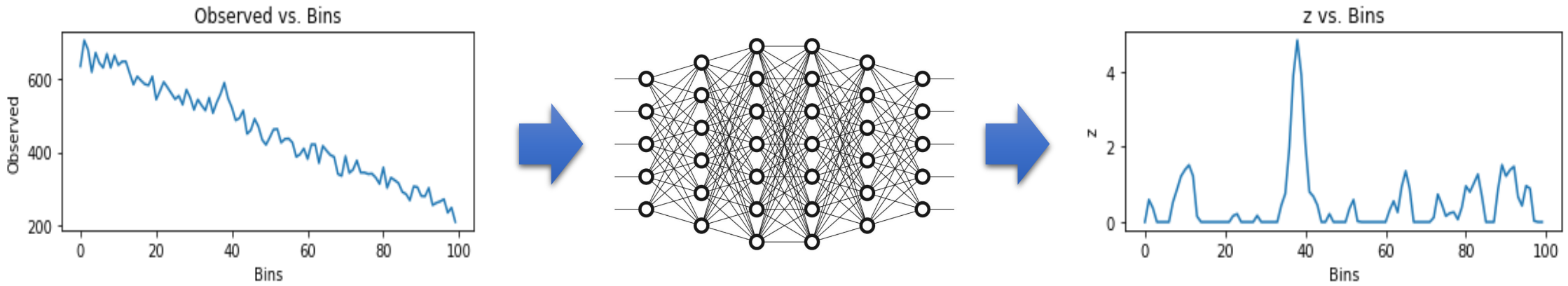
Table is a WIP

Tested symmetry		inclusive	Object multiplicity τ	j	bj	γ	K	...	Decays Z	W	...	Topologies VBF	$t\bar{t}$	...
LU - U(3)	e/μ $ee/\mu\mu$ OS $ee/\mu\mu$ SS						[24]		[29,30]	[31-34]				
LFC - $U(1)^3$	$e\mu/\mu e$ OS $e\mu/\mu e$ SS	[27,28]										[27,28]		
CP	$e^+\mu^-/e^-\mu^+$	[35]		[35]										

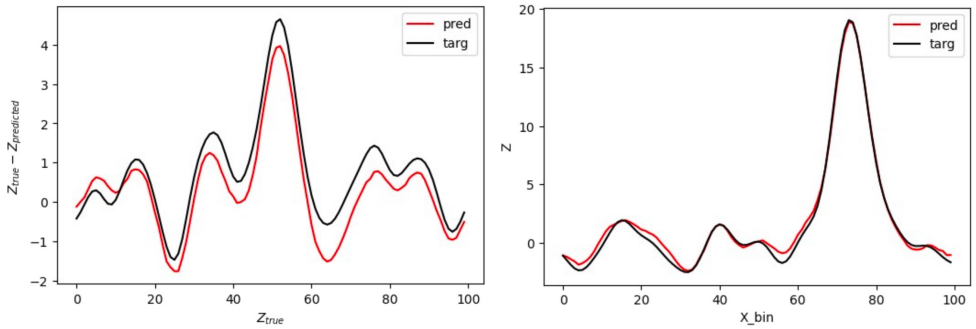
Example I - resonances

S. Volkovitch, F. DeVito Halevi, SB [2107.11573]

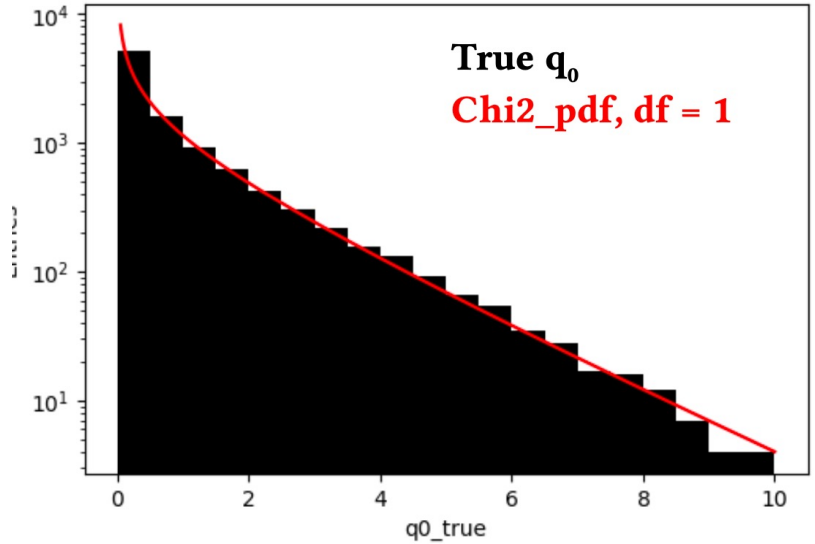
- The SM property:
  - In absence of resonances –
    most invariant mass distributions are smoothly falling
- The tool:
  - An NN that maps invariant mass distributions to significances (q0)

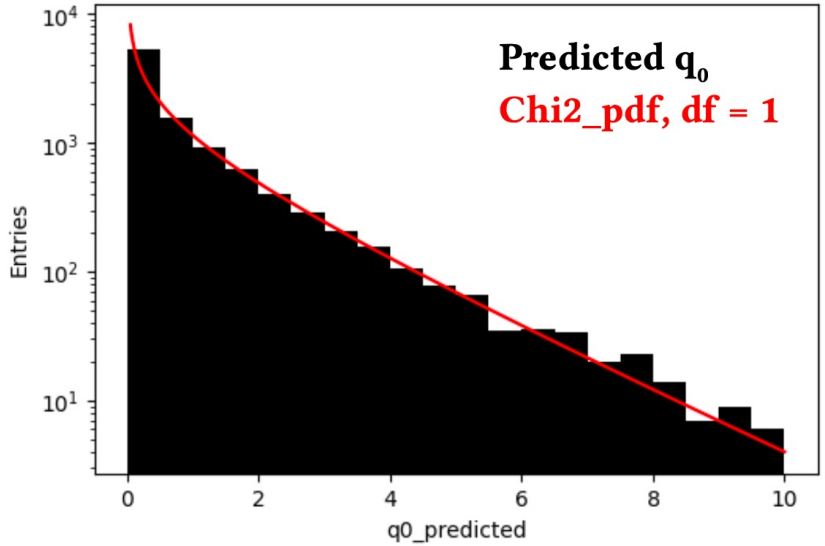


Example I - resonances



- Looking at 1 mass bin with no signal injected
- This is the standard background-only scenario
- Using profile-likelihood tests statistics - $q_0 = Z^2$ should follow the χ^2 distributions
- This is also the case for the NN prediction





Example I – resonances - Performance

- Significance predicted in almost no time
  Precision almost as good as that expected in blind analysis
  - The latter exploits the PLR test statistics and rely on exact knowledge of the signal and background shape

  *concept proved on synthetic data

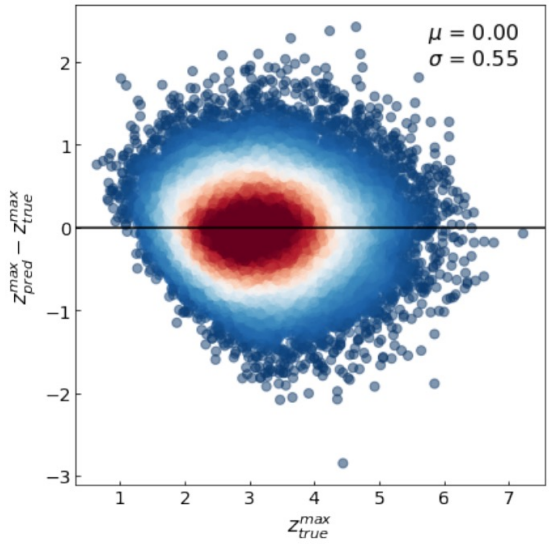


Fig. 2 The difference between $z^{\max}_{\text{pred}}$ and $z^{\max}_{\text{true}}$ as a function of $z^{\max}_{\text{true}}$. Dense regions are shown in red (roughly corresponding to the 1σ region), while sparse regions are shown in blue.

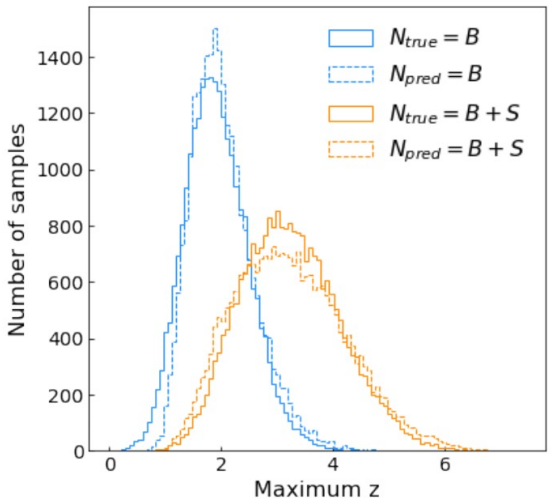


Fig. 3 The distribution of $z^{\max}_{\text{true}}$ (solid line) and $z^{\max}_{\text{pred}}$ (dashed line) for samples with no signal added (blue) and for samples with a 3σ significance signal added (orange)

Example II - symmetries

M. Birman, B. Nachman, R. Sebbah, G. Sela, O. Turetz, SB [2203.07529]

- The SM property:
  - Any exact or approximate symmetry of the SM: e/μ, CP, forward backward, …
- The tool:
  - Symmetries allows splitting the data into two mutually exclusive datasets
    - Under the symmetry assumption – they originate from the same underline distributions
  - N_σ test statistics that identifies rapidly asymmetries between two datasets –
    in this case the datasets are projected to histograms

$$\mathrm{N}_\sigma(B,A) = \frac{1}{\sqrt{M}} \sum_{i=1}^{M} \frac{B_i - A_i}{\sqrt{A_i + B_i}}$$

Example II - symmetries

- An “ideal” analysis
  - Background shape is perfectly known
  - Signal shape and resolution are perfectly know
  - 0 uncertainties
  - Sensitivity calculate with profile likelihood test statistics

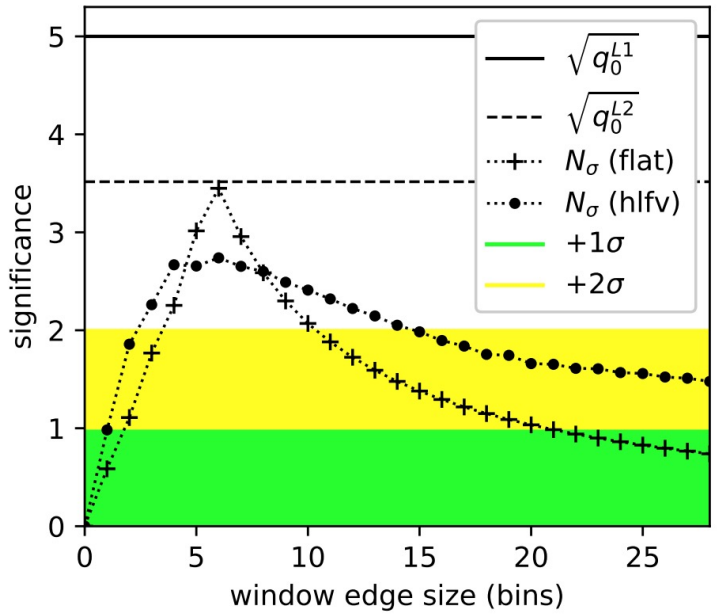
$$Z = \sqrt{q_0^{L1}}$$

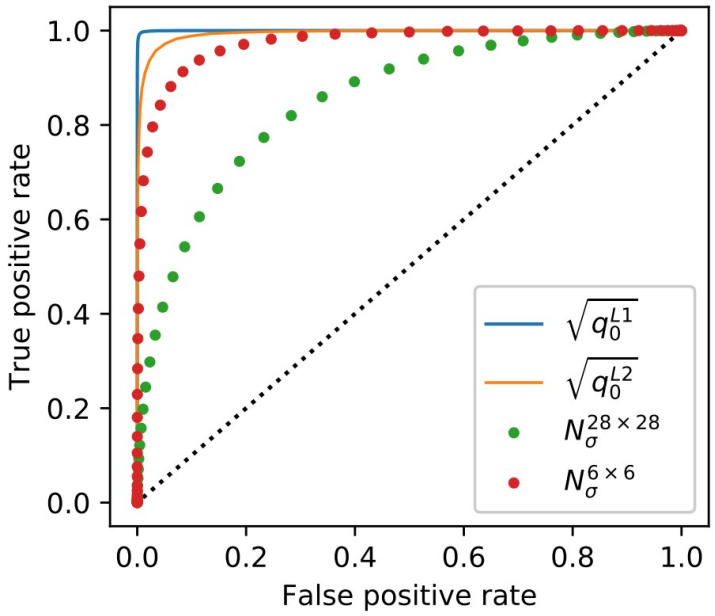
- Analysis relaying on symmetry considerations
  - Data is split into two mutually exclusive sample – one serves as a background estimate to the other
  - Background model from the data based on symmetry assumption
    - Systematic uncertainty due to available statistics in the “other” sample
  - Signal shape and resolution are perfectly know
  - Sensitivity calculate with profile likelihood test statistics

$$Z = \sqrt{q_0^{L2}}$$

Example II – symmetries - Performance

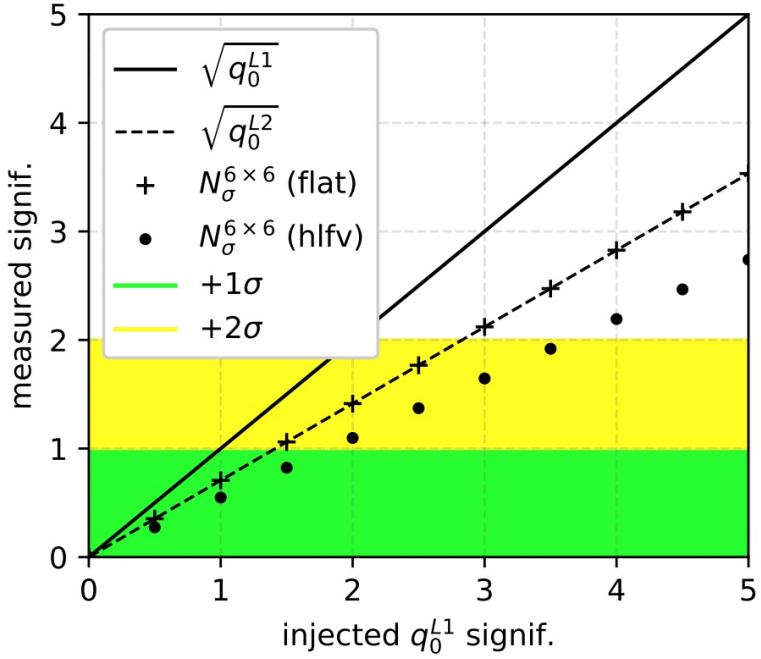
- The N_σ calculation is rapid
- Can scan with no time as many matrices as one want
- Can scan sub-matrices
- Sensitivity is restored fast

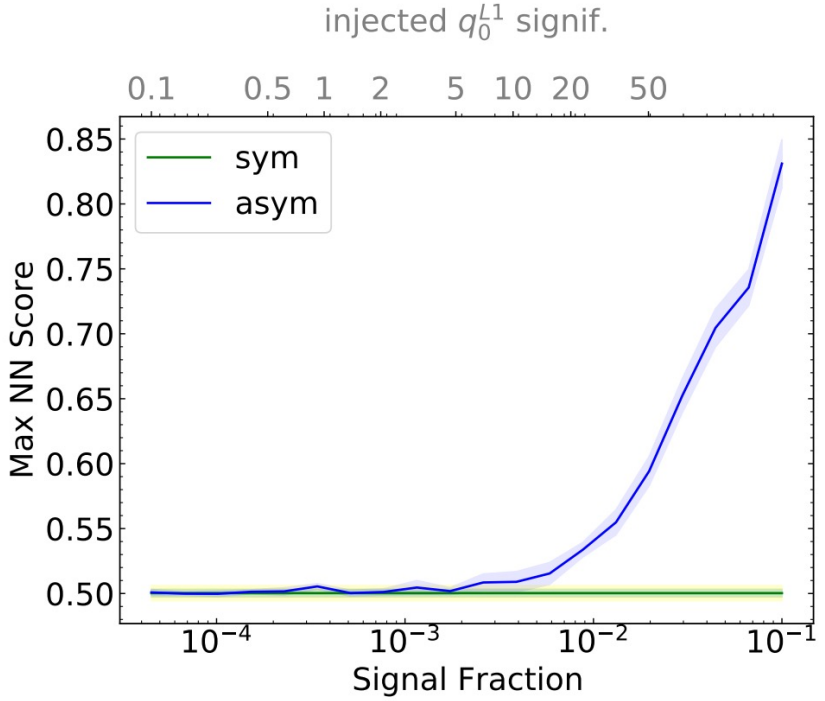




Semi-supervised training vs the DDP

- First attempt to identify asymmetries using ML techniques
  - Demonstrate the potential of such methods but, so far, does not perform as good as the N_σ test





Summary

- Thousands of person hours invested so far in search for BSM physics
- Resulted in an impressive set of bounds on many BSM models
- No hints for BSM physics
- The data is far from being fully exploited –
  New physics could easily be hidden in the already collected data
- Complementary search paradigms should be exploited
- The Data Directed Paradigm is one such possibility
  - Allows scanning rapidly many different final states and many different selection and mark those that are potentially interesting
- Concept demonstrated with two different properties of the SM
- ATLAS searches are slowly ramping up