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]
Symmetrized NPLM – SB, I. Savoray, Y. Zurgil [2401.09530]

AI SSAI
Anomaly Detection Workshop

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 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 $\to$ SM$_1$ $\times$ SM$_1$				BSM $\to$ SM$_1$ $\times$ SM$_2$			BSM $\to$ complex			
										q/g	γ/π^0's	b	$\cdots$	tZ/H	bH	$\cdots$	$\tau qq'$	eqq'	$\mu qq'$	$\cdots$
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 $\to$ SM$_1$ $\times$ SM$_1$: q/g										∅	∅	∅		∅	∅	∅	∅	∅	∅	
γ/π^0's											[75]	∅		∅	∅	∅	∅	∅	∅	
b												[76, 77]		∅	∅	∅	∅	∅	∅	
$\vdots$																				
$\vdots$																				

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

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									[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]										

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 measurements

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									[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]										

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 to enhance the Signal/Background ratio

	e	μ	τ	q/g	b	t	γ	Z/W	H
e	[37,38]	[39,40]	[39]	∅	∅	∅	[41]	[42]	∅
μ		[37,38]	[39]	∅	∅	∅	[41]	[42]	∅
τ			[45,46]	∅	[47]	∅	∅	∅	∅
q/g				[29,30,50,51]	[52]	∅	[53,54]	[55]	∅
b					[29,52,56]	[57]	[54]	[58]	[59]
t						[61]	∅	[62]	[63]
γ							[65,66]	[67–69]	[68,70]
Z/W								[71]	[71]
H									[72,73]

	BSM → SM₁ × SM₁ q/g	γ/π^0's	b	$\cdots$	BSM → SM₁ × SM₂ tZ/H	bH		$\cdots$	BSM → complex $\tau qq'$	eqq'	$\mu qq'$	$\cdots$
e	∅	∅	∅		∅	∅	∅		∅	[43,44]	∅	
μ	∅	∅	∅		∅	∅	∅		∅	∅	[43,44]	
τ	∅	∅	∅		∅	∅	∅		[48,49]	∅	∅	
q/g	∅	∅	∅		∅	∅	∅		∅	∅	∅	
b	∅	∅	∅		[60]	∅	∅		∅	∅	∅	
t	∅	∅	∅		[64]	[60]	∅		∅	∅	∅	
γ	∅	∅	∅		∅	∅	∅		∅	∅	∅	
Z/W	∅	∅	∅		∅	∅	∅		∅	∅	∅	
H	[74]	∅	∅		∅	∅	∅		∅	∅	∅	
BSM → SM₁ × SM₁ q/g	∅	∅	∅		∅	∅	∅		∅	∅	∅	
γ/π^0's		[75]	∅		∅	∅	∅		∅	∅	∅	
b			[76,77]		∅	∅	∅		∅	∅	∅	
⋮												
⋮												

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 to enhance the Signal/Background ratio

										$\mathrm{BSM} \to \mathrm{SM}_1 \times \mathrm{SM}_1$				$\mathrm{BSM} \to \mathrm{SM}_1 \times \mathrm{SM}_2$			$\mathrm{BSM} \to$ complex			
	e	μ	τ	q/g	b	t	γ	Z/W	H	q/g	γ/π^0's	b	$\cdots$	tZ/H	bH	$\cdots$	$\tau qq'$	eqq'	$\mu qq'$	$\cdots$
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]	∅	∅		∅	∅	∅	∅	∅	∅	
$\mathrm{BSM} \to \mathrm{SM}_1 \times \mathrm{SM}_1$: q/g										∅	∅	∅		∅	∅	∅	∅	∅	∅	
γ/π^0's											[75]	∅		∅	∅	∅	∅	∅	∅	
b												[76,77]		∅	∅	∅	∅	∅	∅	
⋮																				
⋮																				

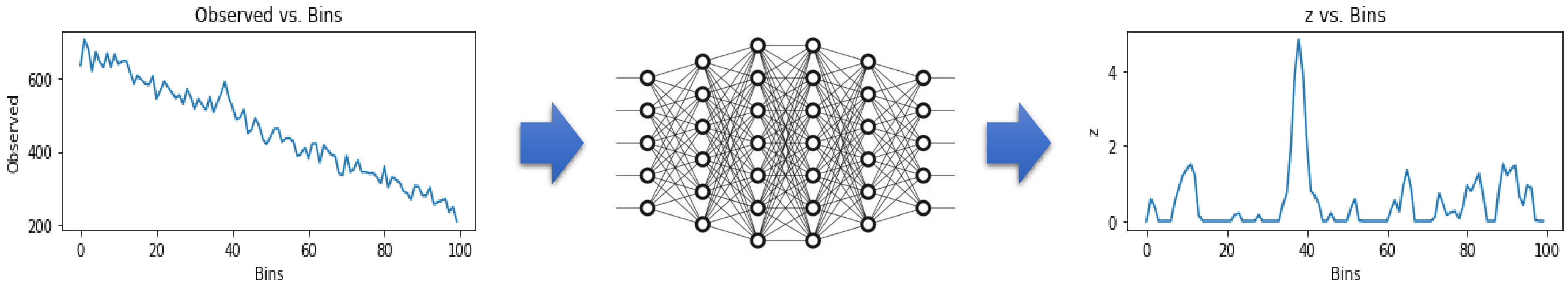
			Object multiplicity						Decays			Topologies		
Tested symmetry		inclusive	τ	j	bj	γ	K	...	Z	W	...	VBF	$t\bar{t}$	...
LU - U(3)	[illegible] $ee/\mu\mu$ OS $ee/\mu\mu$ SS						[24]		[29,30]	[31-34]				
LFC - $U(1)^3$	$e\mu/\mu e$ OS $e\mu/\mu e$ SS	[27,28]										[27,28]		
CP	$e^+\mu^-/e^-\mu^+$	[35]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

Table is a WIP

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 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: $\underline{e/\mu}$, CP, forward backward, …
- Tools I and II:
  - Symmetries allows splitting the data into two mutually exclusive datasets
    - Under the symmetry assumption – they originate from the same underline distributions

1) N_σ test statistics that identifies rapidly asymmetries between two datasets – in this case the datasets are projected to histograms
2) Cwola-like weakly supervised training of a classifier

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

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

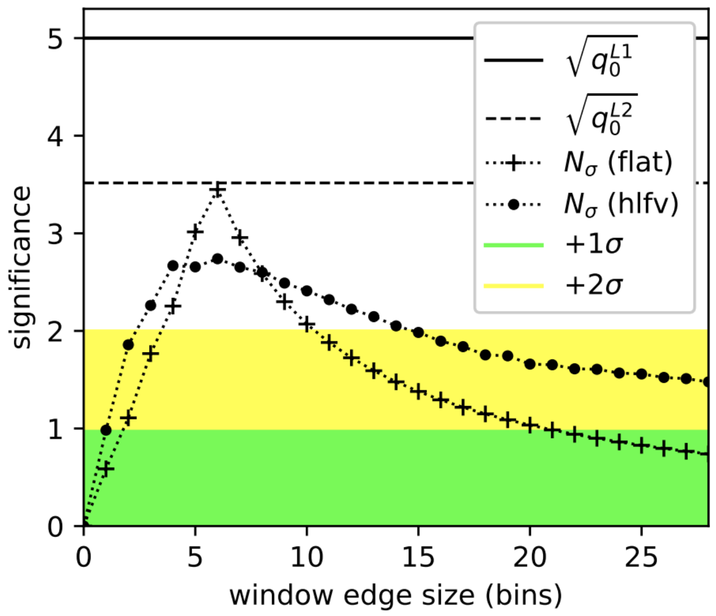
$$Z = \sqrt{q_0^{L1}}$$

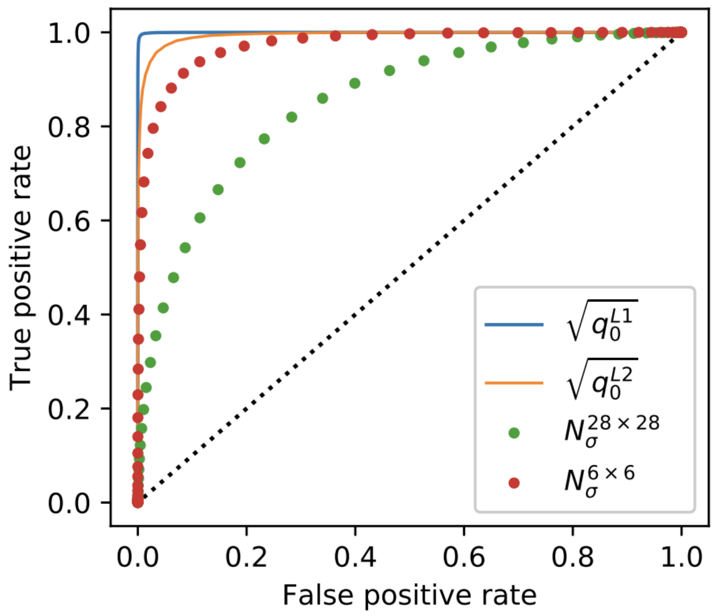
- Analysis relaying on symmetry considerations
  - Data is split into two mutually exclusive sample – one serves as a background estimate to the other
  - Background model from the data based on symmetry assumption
    - Systematic uncertainty due to available statistics in the "other" sample
  - Signal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

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

Example II – symmetries – tool I

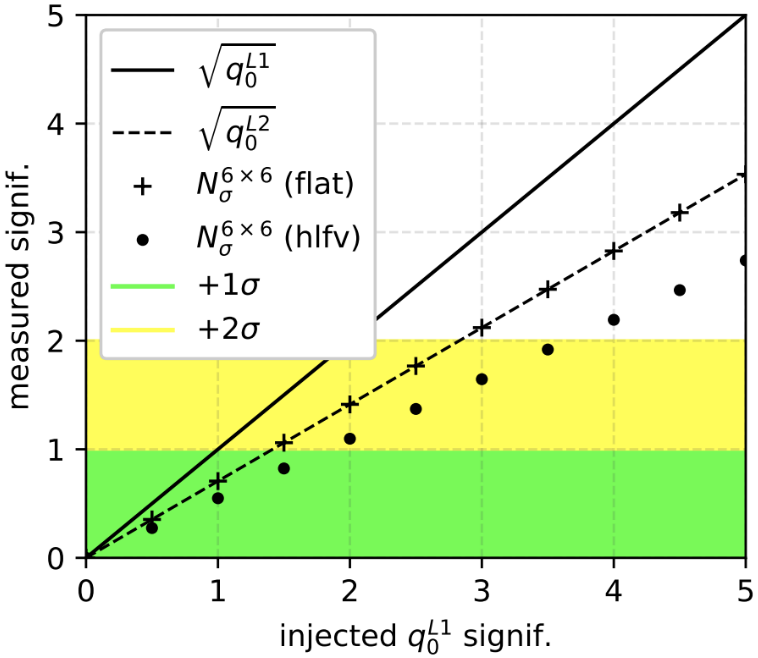
- The N_σ calculation is rapid
- Can scan with no time as many matrices as one want
- Can scan sub-matrices
- Sensitivity is restored fast

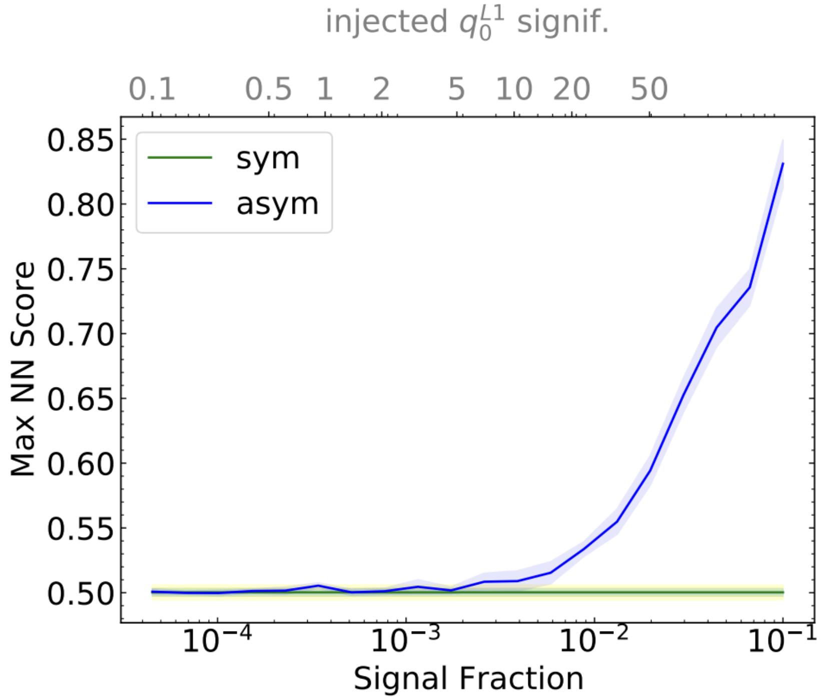




Example II – symmetries – tool II

- 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





Tool III - Symmetrized NPLM

- Learning New Physics from Machine (NPLM) R.T. D'Agnolo, A. Wulzer [1806.02350]
  - For a dataset A and a reference dataset B much larger than A
  - Train a classifier with minus log likelihood ratio as loss function
    - Weakly-supervised
  - Minimize the loss → Maximizing the likelihood ratio → optimized test statistics
  - Loss pdf under background only data approaches $\chi^2_{n_{dof}}$
  - n_{dof} determined by the number of free parameters in the model → number of free parameters in the NN
- NN machinery is exploited for optimal fitting procedure
  - Flexible set of functions
  - Efficient minimization procedure

Symmetrized NPLM

Original NPLM

R.T. D'Agnolo, A. Wulzer [1806.02350]

- The NULL hypothesis: $\mathcal{H}_0 :$ $n_{\mathbf{A}}\left(x|\mathcal{H}_0\right) = \frac{N_{\mathbf{A}}}{N_{\mathbf{B}}} n_{\mathbf{B}}\left(x,\nu\right)$
- The alternative hypothesis: $\mathcal{H}_1 :$ $n_{\mathbf{A}}\left(x|\mathcal{H}_1\right) = \frac{N_{\mathbf{A}}}{N_{\mathbf{B}}} e^{f(x,\mu)} n_{\mathbf{B}}(x,\nu)$
- The test statistics: $t = 2\log\left(\frac{\max_{\nu,\mu}\left(\mathcal{L}\left(\mathcal{H}_1|\mathbf{A}\right)\right)}{\max_{\nu}\left(\mathcal{L}\left(\mathcal{H}_0|\mathbf{A}\right)\right)}\right)$
- The extended likelihood function: $\mathcal{L}\left(\mathcal{H}|\mathbf{A}\right) = \frac{e^{-N_{\mathbf{A}}(\mathcal{H})}}{\tilde{N}_{\mathbf{A}}!} \prod_{x\in\mathbf{A}} n_{\mathbf{A}}(x|\mathcal{H})$
  n is the number density function
- Then $t = 2\left(\hat{N}_{\mathbf{A}}\left(\mathcal{H}_0\right) - \hat{N}_{\mathbf{A}}\left(\mathcal{H}_1\right) + \log\left(\prod_{x\in\mathbf{A}} \frac{\hat{n}_{\mathbf{A}}(x|\mathcal{H}_1)}{\hat{n}_{\mathbf{A}}(x|\mathcal{H}_0)}\right)\right)$
- Parametrize $\hat{n}_{\mathbf{A}}(x|\mathcal{H}_1) = e^{\hat{f}(x)} \hat{n}_{\mathbf{A}}(x|\mathcal{H}_0)$
- To obtain $t = 2\left(-\int\left(e^{\hat{f}(x)} - 1\right)\hat{n}_{\mathbf{A}}(x|\mathcal{H}_0)dx + \sum_{x\in\mathbf{A}} \hat{f}\left(x\right)\right)$

Symmetrized NPLM

Original NPLM

R.T. D'Agnolo, A. Wulzer [1806.02350]

- For B much larger than A the integral on the number density function can be replaced with a summation

$$t = t_{\mathbf{B}}\left(\mathbf{A}\right) \equiv -2\left(\frac{\hat{N}_{\mathbf{A}}\left(\mathcal{H}_0\right)}{\tilde{N}_{\mathbf{B}}}\sum_{x\in\mathbf{B}}\left(e^{\hat{f}(x)}-1\right)-\sum_{x\in\mathbf{A}}\hat{f}\left(x\right)\right)$$

- The goal is to find a parametrization to f that minimizes t
- A useful parametrization of f is obtained via the output of a NN
  - Highly expressive, continuous and smooth
- For a fully connected NN with N_{neu} neurons

$$f\left(x\right) = b_{\mathrm{out}} + \sum_{\alpha=1}^{N_{\mathrm{neu}}} w_{\mathrm{out}}^{\alpha}\sigma\left(w_{\alpha}x + b_{\alpha}\right)$$

- The NN is trained to find the w's and b's that minimizes the loss
- When applied to the test statistics – obtained it's score
- For a given background only pdf – gives also the significance

Symmetrized NPLM

Original NPLM - challenges

SB, I. Savoray, Y. Zurgil [2401.09530]

- B is not necessarily larger than A – in searches for symmetry breaking this is by far not the case
- The asymptotic $\chi^2_{n_{dof}}$ is not respected for other A/B ratios
  - Due to mis-modeling of the NULL hypothesis which ignores statistical uncertainty in B
- There is a need to constrain the weights of the NN to obtain the $\chi^2_{n_{dof}}$ distribution due to overfitting
- Points existing in A but not in B results in divergence of f

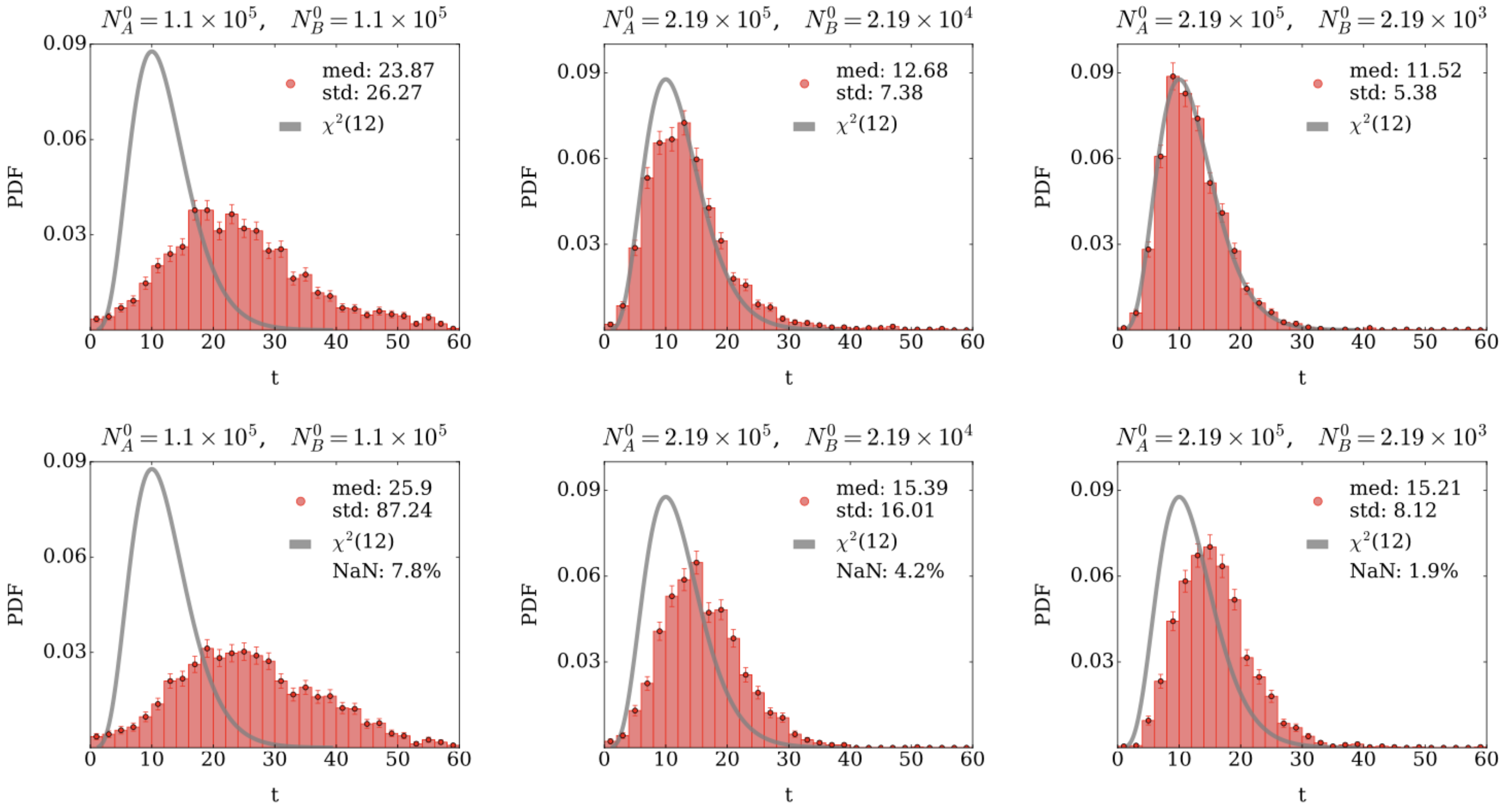
Symmetrized NPLM

Original NPLM - challenges

SB, I. Savoray, Y. Zurgil [2401.09530]

With WC

Without WC



Symmetrized NPLM

The symmetrized formalism

SB, I. Savoray, Y. Zurgil [2401.09530]

- Treat A and B on equal footing
  - The NULL hypothesis – A and B are drawn from the same pdf
  - The alternative hypothesis A and B are drawn from different pdfs
- Construct likelihood test to simultaneously fit A and B

$$t = 2\log\left(\frac{\max_{\mu,\nu}\left(\mathcal{L}\left(\mathcal{H}_1|\mathbf{A},\mathbf{B}\right)\right)}{\max_{\nu}\left(\mathcal{L}\left(\mathcal{H}_0|\mathbf{A},\mathbf{B}\right)\right)}\right) = 2\log\left(\frac{\max_{\mu,\nu}\left(\mathcal{L}\left(\mathcal{H}_1|\mathbf{A}\right)\mathcal{L}\left(\mathcal{H}_1|\mathbf{B}\right)\right)}{\max_{\nu}\left(\mathcal{L}\left(\mathcal{H}_0|\mathbf{A}\right)\mathcal{L}\left(\mathcal{H}_0|\mathbf{B}\right)\right)}\right)$$

$$t = 2\log\left(\frac{\max_{p_\mathbf{A},p_\mathbf{B}}\left(\mathcal{L}\left(N_\mathbf{A},p_\mathbf{A}\left(x\right)|\mathbf{A}\right)\mathcal{L}\left(N_\mathbf{B},p_\mathbf{B}\left(x\right)|\mathbf{B}\right)\right)}{\max_{p_0}\left(\mathcal{L}\left(N_\mathbf{A},p_0\left(x\right)|\mathbf{A}\right)\mathcal{L}\left(N_\mathbf{B},p_0\left(x\right)|\mathbf{B}\right)\right)}\right)$$

$$t_{N_\mathbf{B}\gg N_\mathbf{A}} \to 2\log\left(\frac{\max_{p_\mathbf{A}}\left(\mathcal{L}\left(N_\mathbf{A},p_\mathbf{A}\left(x\right)|\mathbf{A}\right)\right)}{\mathcal{L}\left(N_\mathbf{A},\hat{p}_\mathbf{B}\left(x\right)|\mathbf{A}\right)}\right)$$

Symmetrized NPLM

The symmetrized formalism

SB, I. Savoray, Y. Zurgil [2401.09530]

- Parametrize:

$$\mathcal{H}_0 : \quad \begin{aligned} n_{\mathbf{A}}(x) &= \frac{\tilde{N}_{\mathbf{A}}}{\int n_{\mathcal{R}}(x)\,dx} e^{h(x)} n_{\mathcal{R}}(x) \\ n_{\mathbf{B}}(x) &= \frac{\tilde{N}_{\mathbf{B}}}{\int n_{\mathcal{R}}(x)\,dx} e^{h(x)+r} n_{\mathcal{R}}(x)\ , \end{aligned}$$

$$\mathcal{H}_1 : \quad \begin{aligned} n_{\mathbf{A}}(x) &= \frac{\tilde{N}_{\mathbf{A}}}{\int n_{\mathcal{R}}(x)\,dx} e^{f(x)} n_{\mathcal{R}}(x) \\ n_{\mathbf{B}}(x) &= \frac{\tilde{N}_{\mathbf{B}}}{\int n_{\mathcal{R}}(x)\,dx} e^{g(x)} n_{\mathcal{R}}(x)\ , \end{aligned}$$

- h(x), f(x),g(x) obtained as output of the fit (NN) procedure

$$t = 2\log\left(\frac{\max\limits_{f(x),g(x)}\left(\mathcal{L}\left(\mathcal{H}_1|\mathbf{A},\mathbf{B}\right)\right)}{\max\limits_{h(x),r}\left(\mathcal{L}\left(\mathcal{H}_0|\mathbf{A},\mathbf{B}\right)\right)}\right)$$

Symmetrized NPLM

The symmetrized formalism

SB, I. Savoray, Y. Zurgil [2401.09530]

- The maximization of the NULL hypothesis is analytic:

$$-2\log\left(\max_{h(x),r}\left(\mathcal{L}\left(\mathcal{H}_0|\mathbf{A},\mathbf{B}\right)\right)\right)=2\sum_{x\in\mathbf{A},\mathbf{B}}\left[\frac{1}{\tilde{N}_\mathbf{A}+\tilde{N}_\mathbf{B}}\left(\tilde{N}_\mathbf{A}e^{\hat{h}(x)}+\tilde{N}_\mathbf{B}\left(e^{\hat{h}(x)+\hat{r}}-\hat{r}\right)\right)-\hat{h}\left(x\right)\right]$$

- With $\hat{h}\left(x\right)=0$ and $\hat{r}=0.$

- Perform to fits

$$t_{\mathbf{A}+\mathbf{B}}\left(\mathbf{A}\right)=-2\cdot\min_{f(x)}\left[-\frac{1}{\tilde{N}_\mathbf{A}+\tilde{N}_\mathbf{B}}\sum_{x\in\mathbf{A},\mathbf{B}}\tilde{N}_\mathbf{A}\left(e^{f(x)}-1\right)+\sum_{x\in\mathbf{A}}f\left(x\right)\right]$$

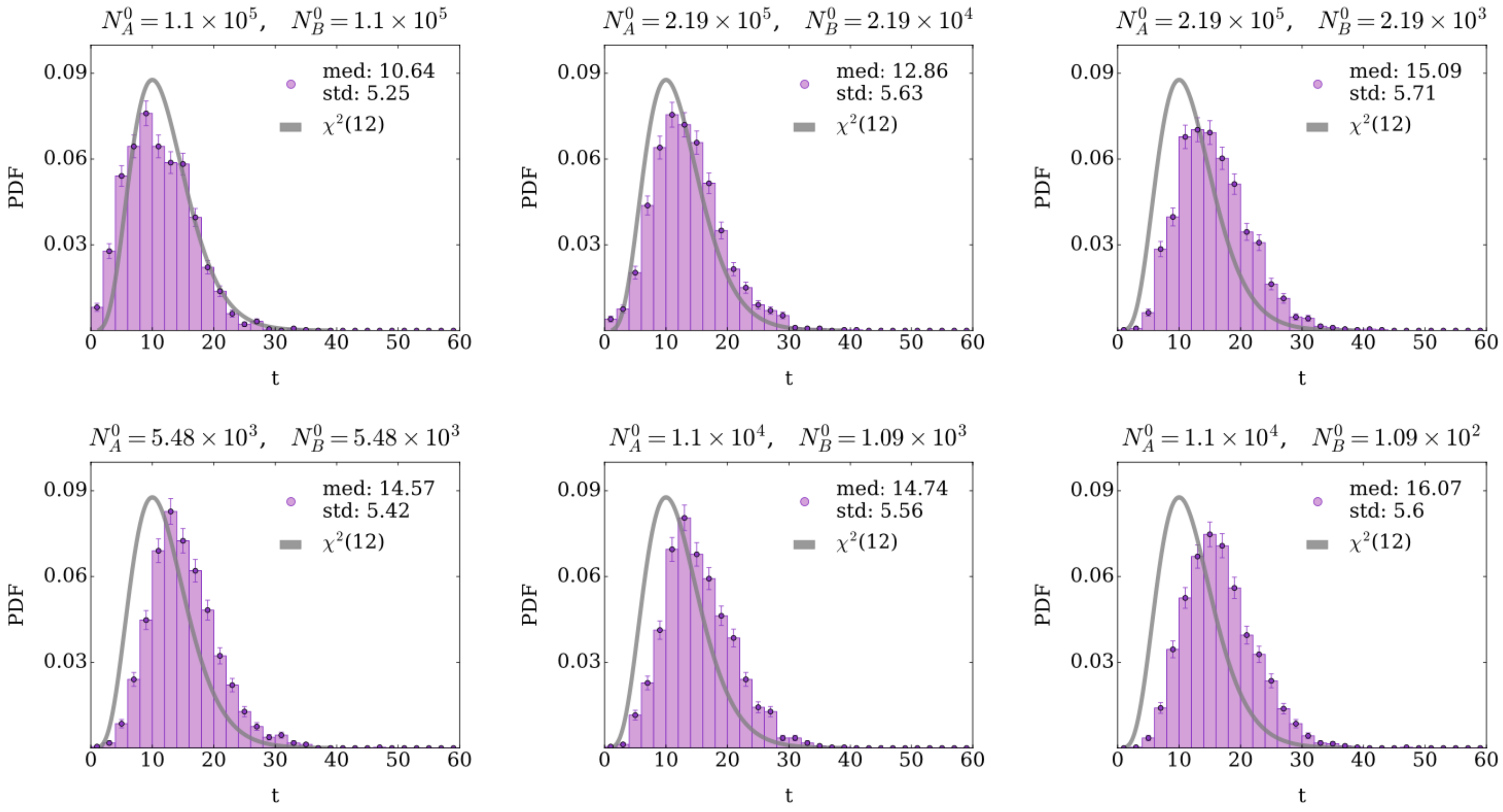
$$t_{\mathbf{A}+\mathbf{B}}\left(\mathbf{B}\right)=-2\cdot\min_{g(x)}\left[-\frac{1}{\tilde{N}_\mathbf{A}+\tilde{N}_\mathbf{B}}\sum_{x\in\mathbf{A},\mathbf{B}}\tilde{N}_\mathbf{B}\left(e^{g(x)}-1\right)+\sum_{x\in\mathbf{B}}g\left(x\right)\right]$$

Symmetrized NPLM

The symmetrized formalism

SB, I. Savoray, Y. Zurgil [2401.09530]

- No WC
- No divergences
- Better agreement with $\chi^2_{n_{dof}}$

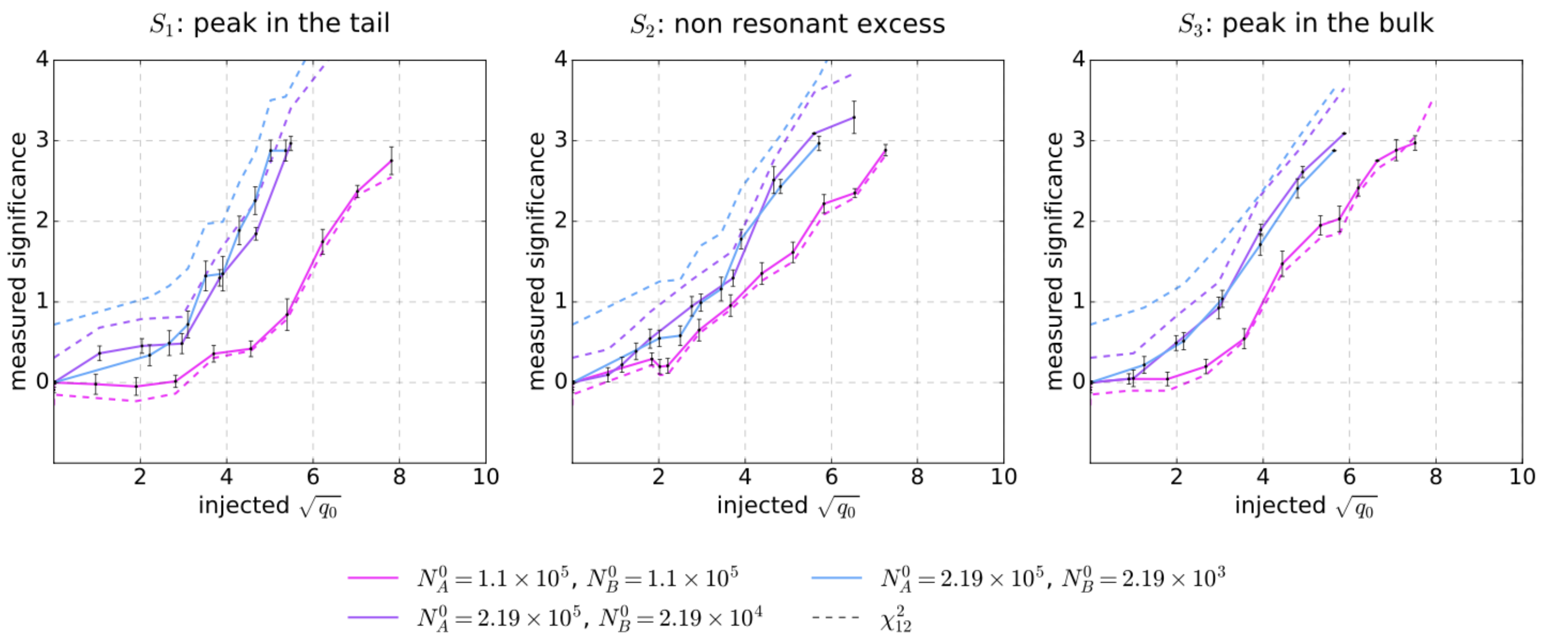


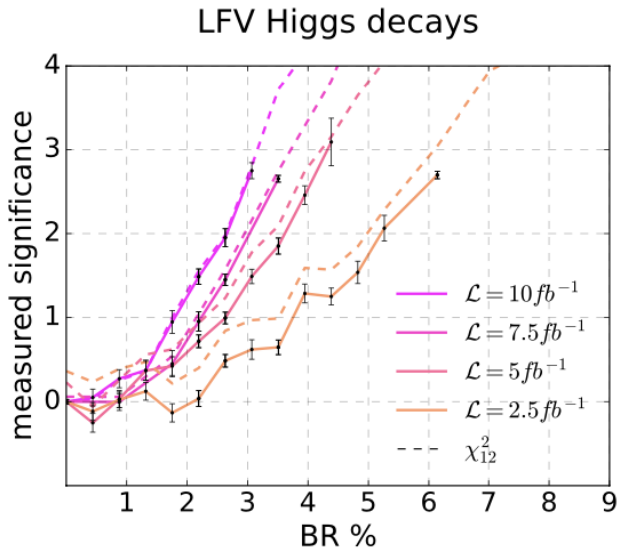
Symmetrized NPLM

The symmetrized formalism

SB, I. Savoray, Y. Zurgil [2401.09530]

- Good sensitivity for all resonant signals
- Good sensitivity for LFV Higgs decays



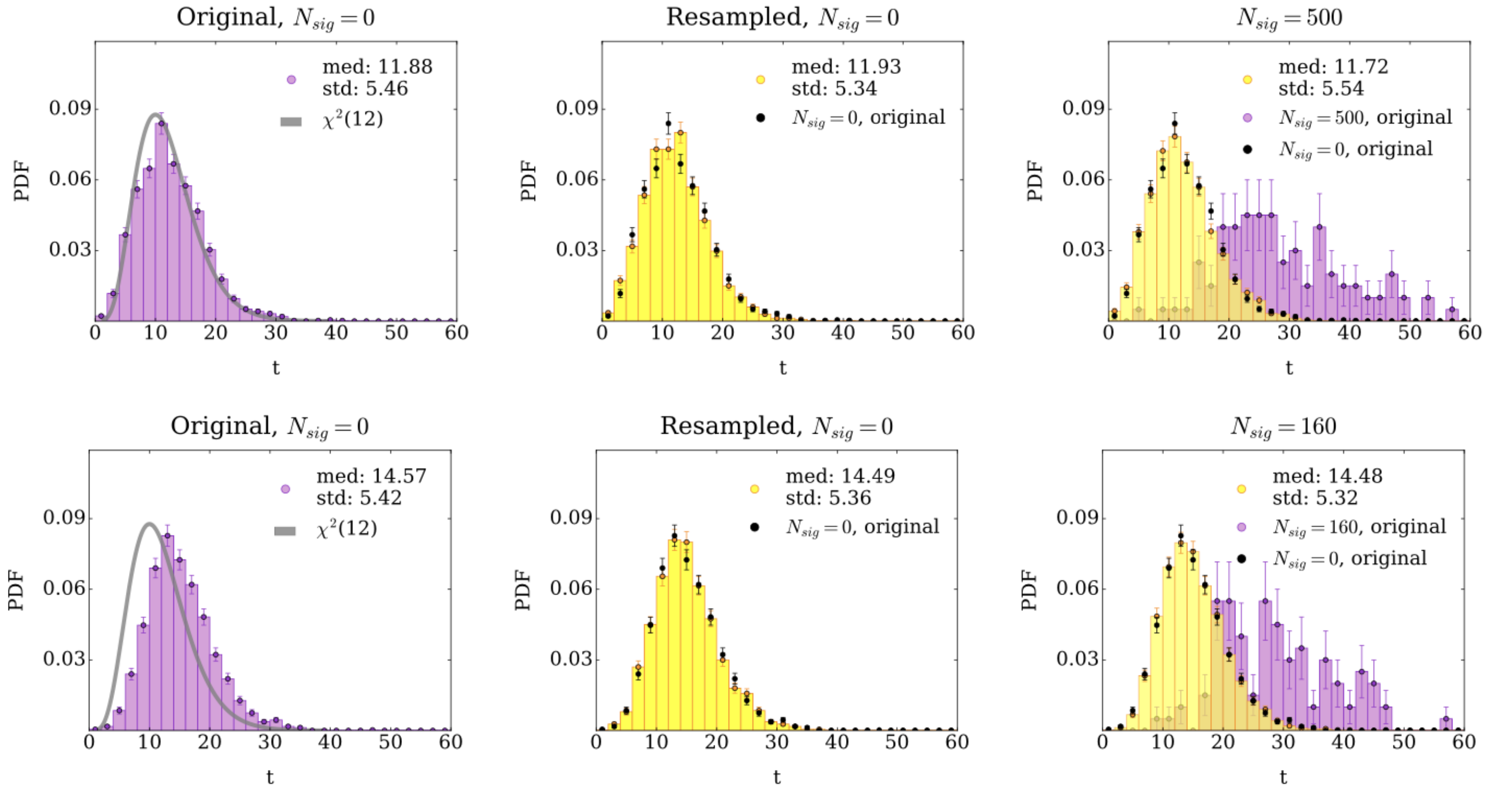


Symmetrized NPLM

The symmetrized formalism

SB, I. Savoray, Y. Zurgil [2401.09530]

- If needed, the background only pdf can also be generated from permutations
- Works well for both scenarios where the χ^2 approximation works well or not



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