Machine learning for high-energy physics: from theory to discovery

Marco Letizia

Machine Learning Genoa Center and INFN, Sezione di Genova

Outline

Machine learning and data-driven modelling

Maximise discovery potential: anomalies

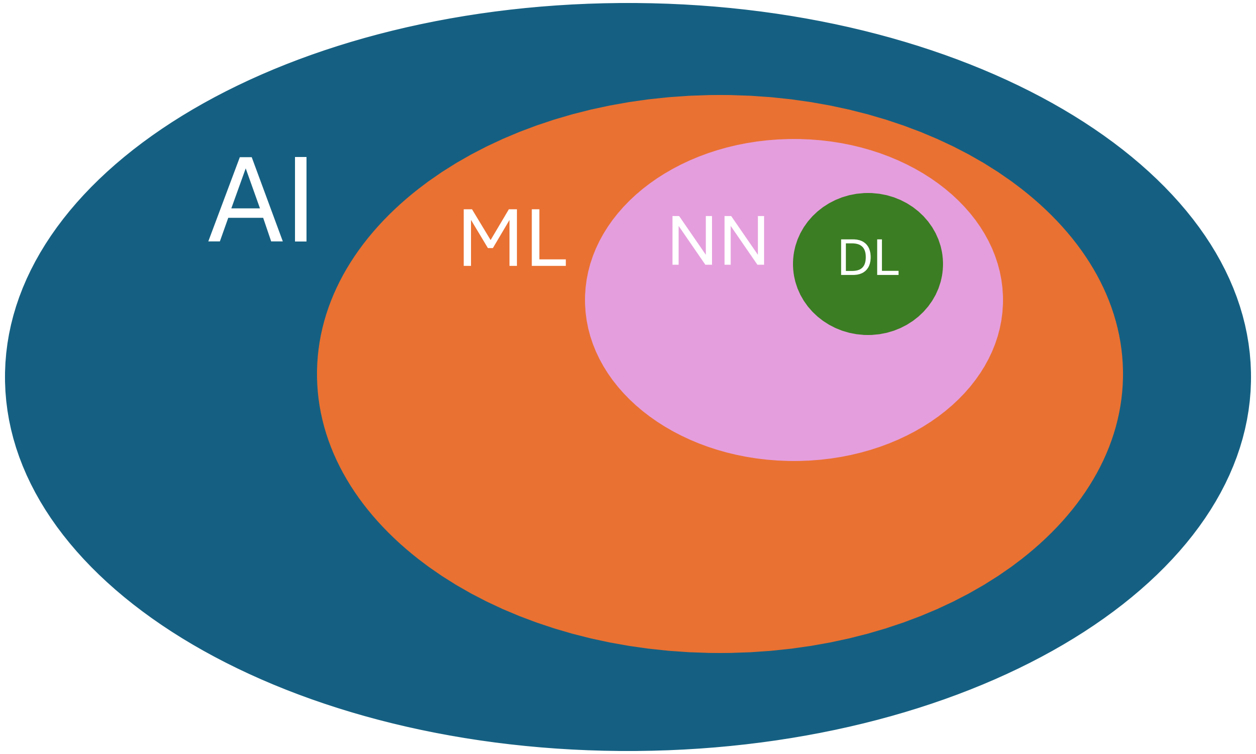
Trust in ML

Machine learning

Design algorithms that can perform tasks without being explicitly programmed.

BUT

is it just *glorified curve fitting*?

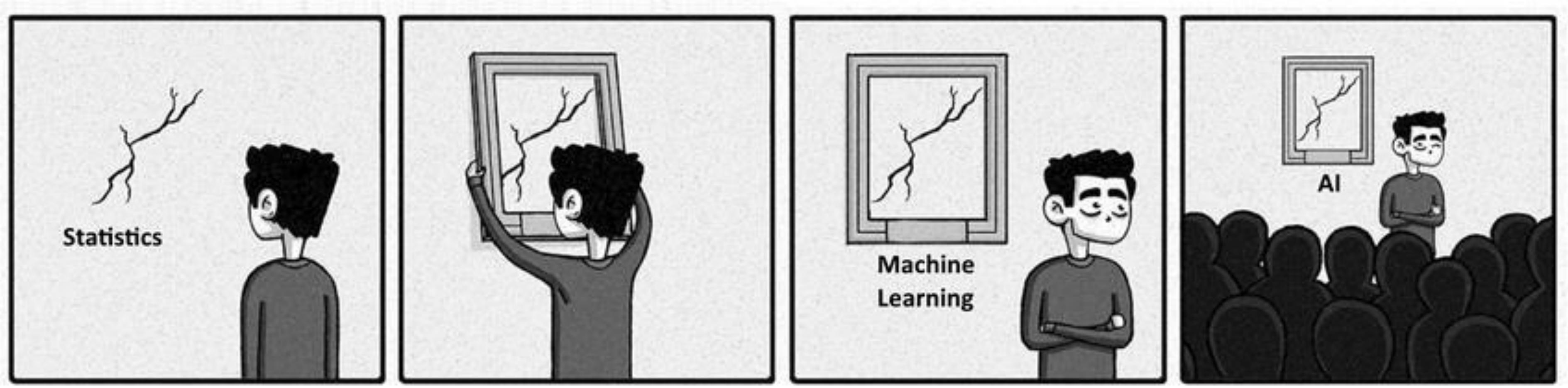


Machine learning

Des

BU

is i



original by sandserif comics

Machine learning

(supervised)

$$(x_i, y_i)_{i=1}^{n} \rightsquigarrow f: \mathcal{X} \to \mathcal{Y},$$

$\mathcal{X} \subseteq \mathbb{R}^d,$
$\mathcal{Y} \subseteq \mathbb{R}^d, \{0,1\}$

$f_w,\qquad w \in \mathbb{R}^p,\qquad p \gg n$ model

$$\min_w \frac{1}{n} \sum_{i=1}^{n} \left(y_i - f_w(x_i)\right)^2 \approx 0$$ fit

Not good if we want to generalise!

Machine learning

(supervised)

$\mathbb{R}^d,$
$\mathbb{R}^d, \{0,1\}$

$f_w, \quad w \in$

$$\min_w \frac{1}{n} \sum_{i=1}^{n} (y_i$$

With four parameters I can fit an elephant, and with five I can make him wiggle his trunk.

John von Neumann

Not good if we want to generalise!

Machine learning

(supervised)

$$\hat{w}_\lambda = \arg\min_w \frac{3}{n} \sum_{i=1}^{n/3} \left(y_i - f_w(x_i)\right)^2 + \lambda \|w\|^2$$

fit

$n/3$

$$\hat{\lambda} = \arg\min_\lambda \frac{3}{n} \sum_{i=n/3+1}^{2n/3} \left(y_i - f_{\hat{w}_\lambda}(x_i)\right)^2$$

validate

$n/3$

$$\frac{3}{n} \sum_{i=2n/3+1}^{n} \left(y_i - f_{\hat{w}_{\hat{\lambda}}}(x_i)\right)^2$$

test

$n/3$

Machine learning

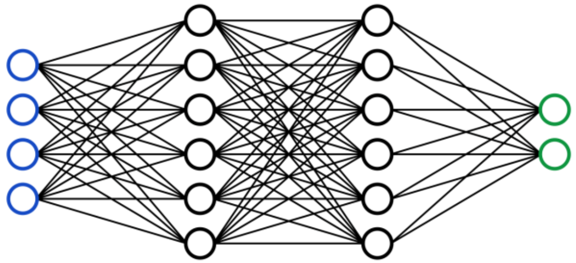
Classic vs data-driven modeling

- Paradigm shift: modeling by data-driven algorithms
  - Potentially lose explainability and a mechanistic/reductionist view.
- Careful pipeline needed!
- Algorithm design
  - can be physics-informed!

Machine learning

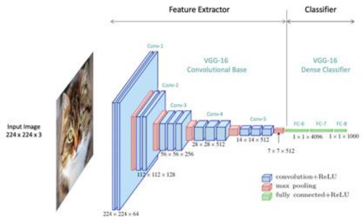
- Expressive ML models: decision trees, kernel methods, neural networks, ...

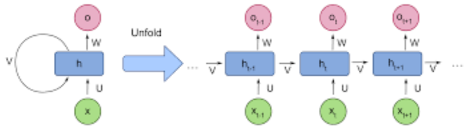
- (Deep) neural networks advantage is feature extraction in high-dimensions and in modeling high-level correlations.

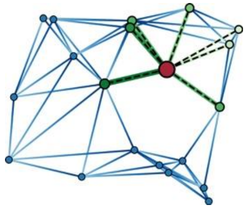


- Structured/unstructured data and architectures:

inductive bias: images – CNN, time series – RNN, graphs – GNN.

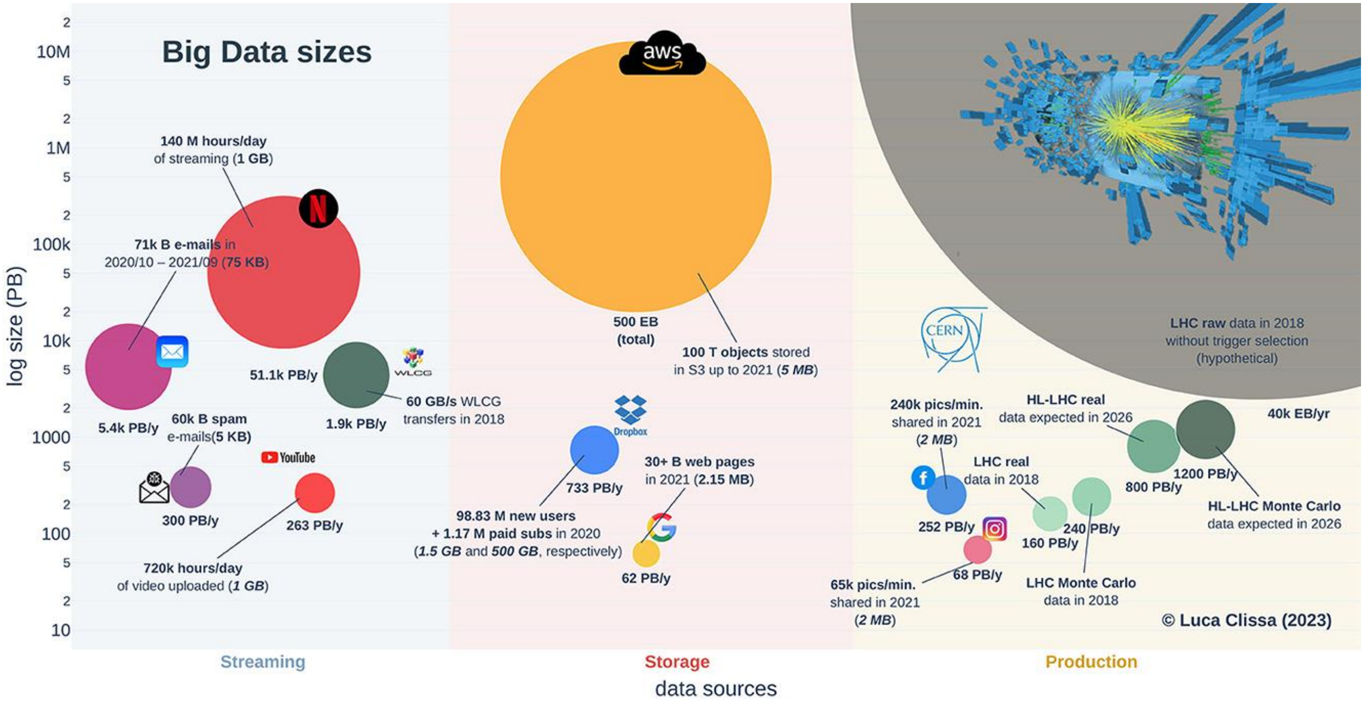




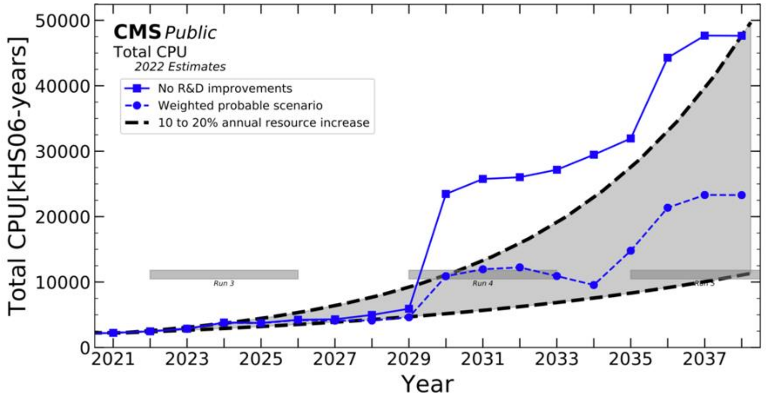


Machine learning in HEP

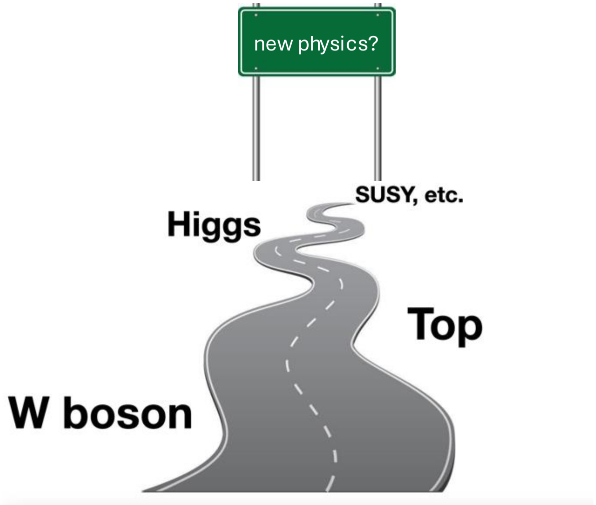
High-energy physics is a great playground!



Clissa et al, Frontiers in Big Data (2023)



CMS Phase-2 Computing Model: Update Document (2022)



Outline

Machine learning and data-driven modelling

Maximise discovery potential: anomalies

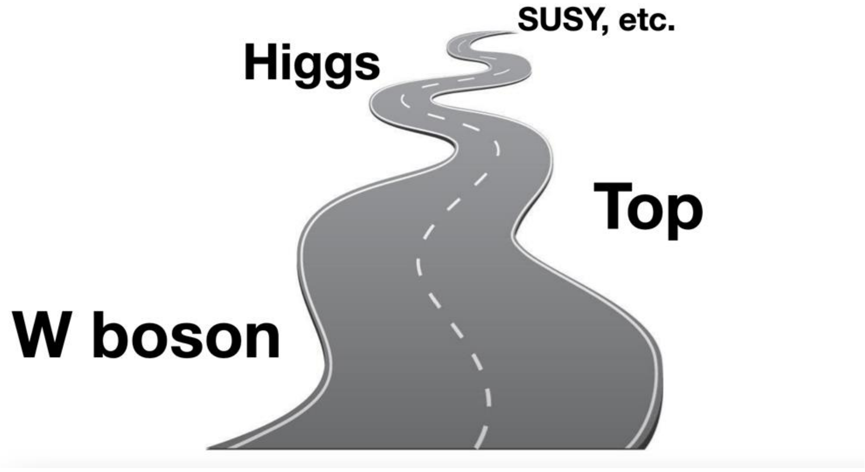
Should we care about interpretability

Anomalies

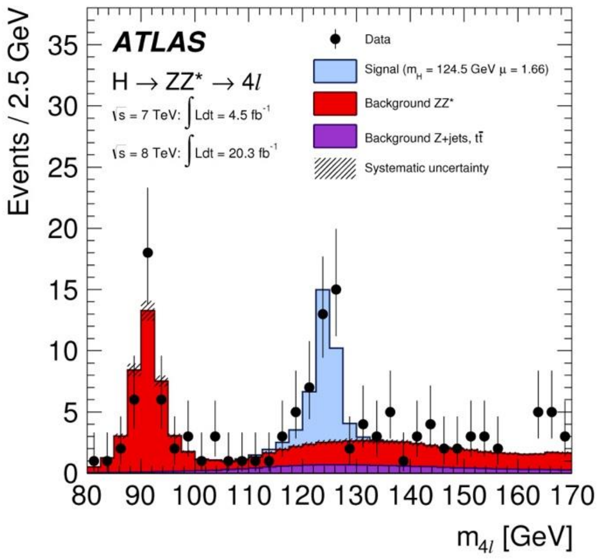
Traditionally strong theory prior

→ likelihood-ratio hypothesis testing (Neyman-Pearson)

$$t_i(\mathcal{D}) = 2\log\frac{\mathcal{L}(\mathcal{D}|NP_i)}{\mathcal{L}(\mathcal{D}|bkg)}$$



[Sketch: A. Wulzer]



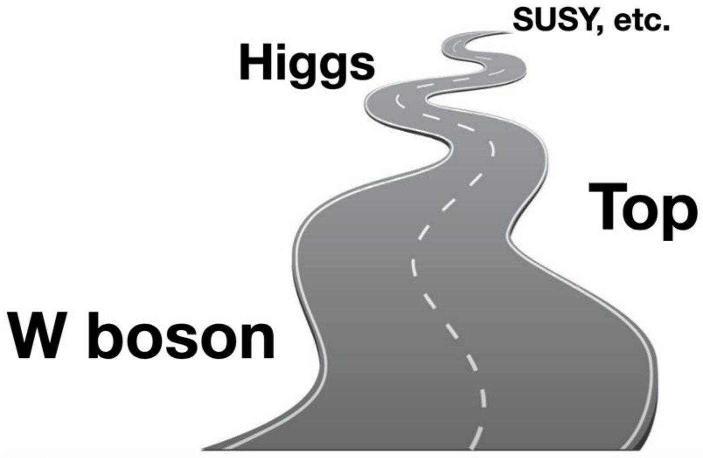
Anomalies

Traditionally strong theory prior

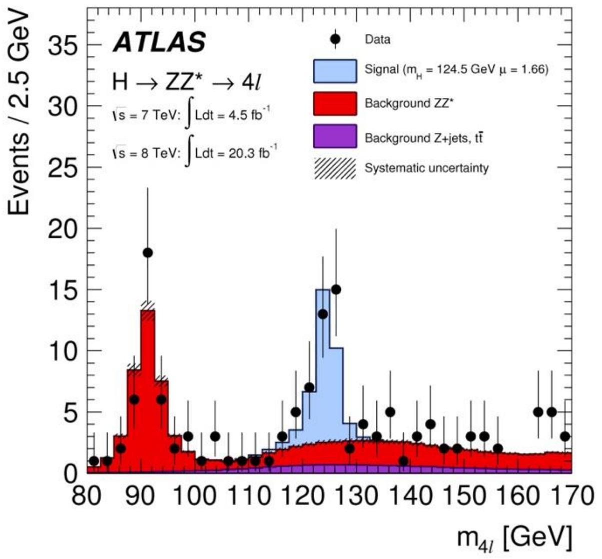
→ likelihood-ratio hypothesis testing (Neyman-Pearson)

$$t_i(\mathcal{D}) = 2\log\frac{\mathcal{L}(\mathcal{D}|NP_i)}{\mathcal{L}(\mathcal{D}|bkg)}$$

make it "any NP"?

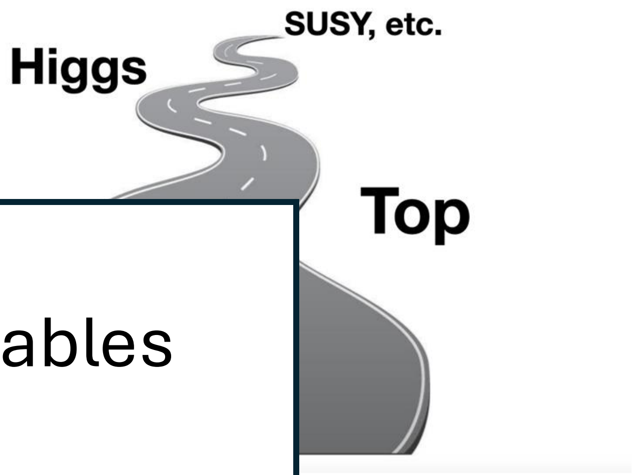


[Sketch: A. Wulzer]



Anomalies

Traditionally strong theory prior



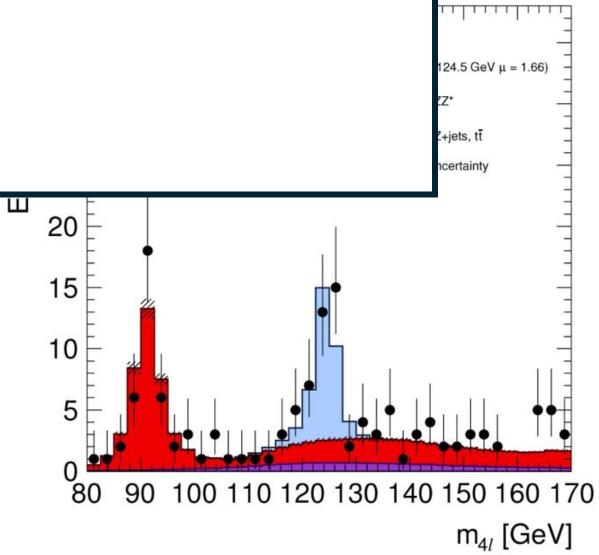
→ likeli

(Ney

Infinite number of possibilities and observables

Machine learning to maximise discovery potential
towards model-independence

$t_i(\mathcal{D})$ = ... $\mathcal{L}(\mathcal{D}|bkg)$

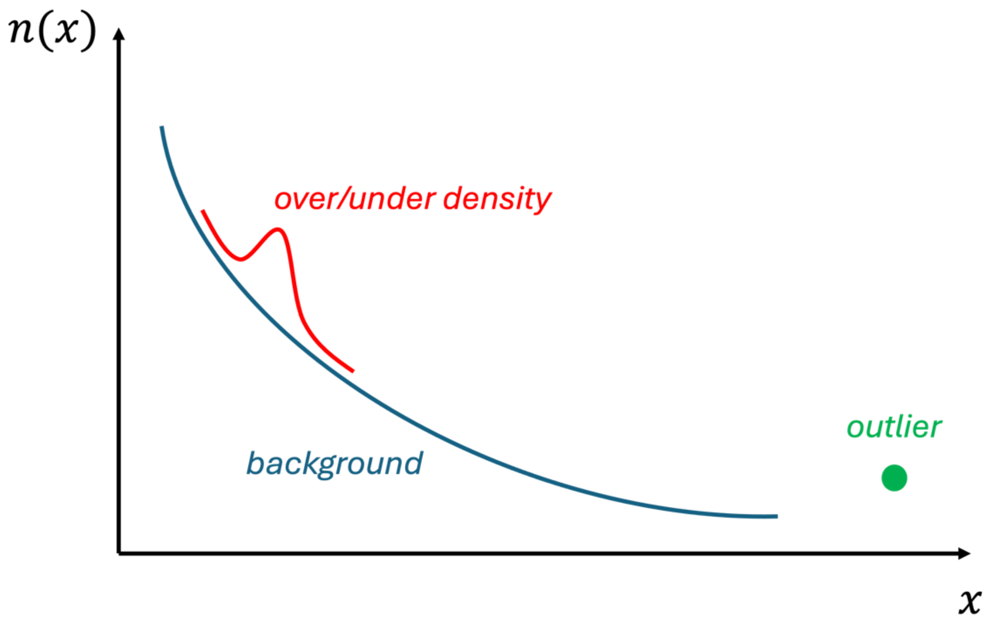


Anomalies

Machine learning to maximise discovery potential

→ anomaly detection

Anomalies are patterns in data that do not conform to a well-defined notion of normal behavior.
“Anomaly detection: A survey”, Chandola, Banerjee, Kumar 2010

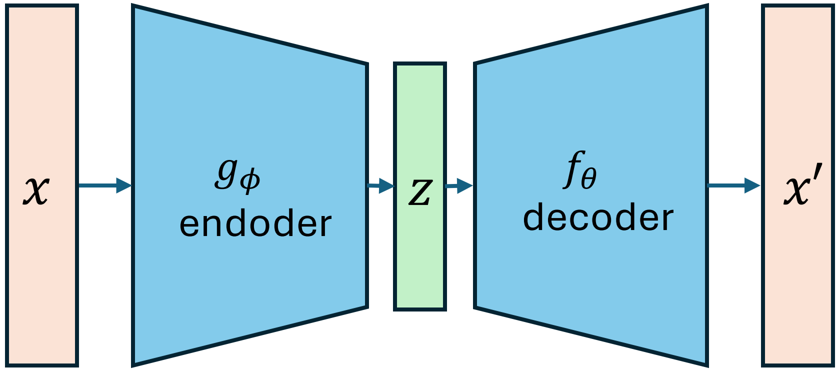


Hard problem:

- Agnostic method
- Multivariate
- Unbinned
- Large scale
- Rare/hidden anomalies
- Uncertainties in bkg
- Statistically robust

Anomalies

Autoencoders for *outlier detection*



$$x, x' \in \mathbb{R}^D$$

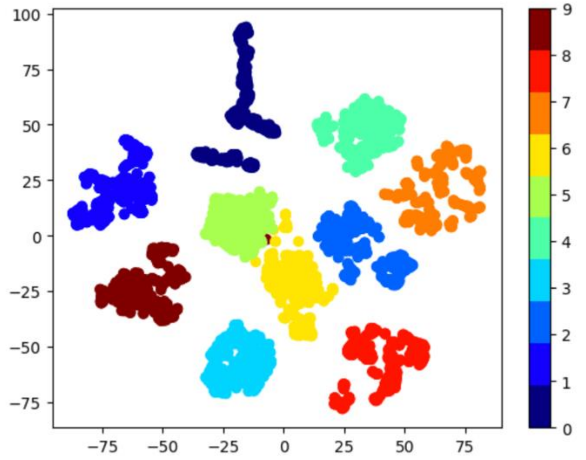
$$z \in \mathbb{R}^d, \quad d \ll D$$

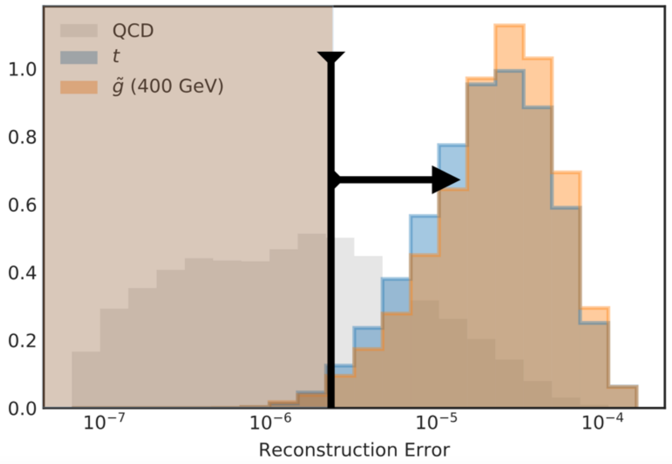
$$x' = f_\theta(z) = f_\theta\big(g_\phi(x)\big) \approx x$$

$$L = \frac{1}{n}\sum_i (x_i - x_i')^2$$

[Real-time anomaly detection at L1 on FPGA]

AXOL1TL



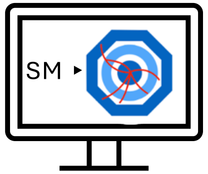


Anomalies

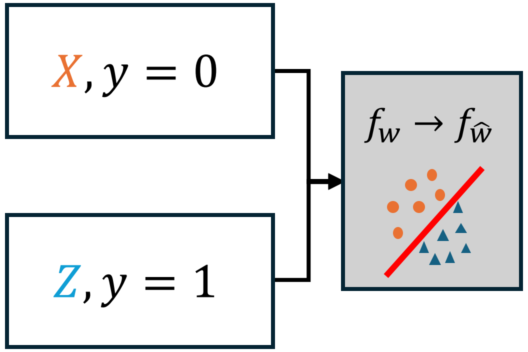
Classifier-based two-sample tests

$$p_{SM} = p_{data}?$$

Train a classifier to separate background from measurements



$$X = \{x_1, \dots, x_n\} \sim p_{SM}, \qquad Z = \{z_1, \dots, z_m\} \sim p_{data}$$

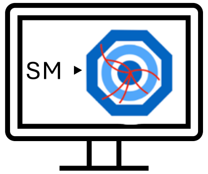


Baker, Cousins (1984), Friedman (2003), Lopez-Paz, Oquab (2017),
Metodiev, Nachman, Thaler (2017), D'Agnolo, Wulzer (2018),
ML, Losapio, Rando, Grosso, Wulzer, Pierini, Zanetti, Rosasco (2022)

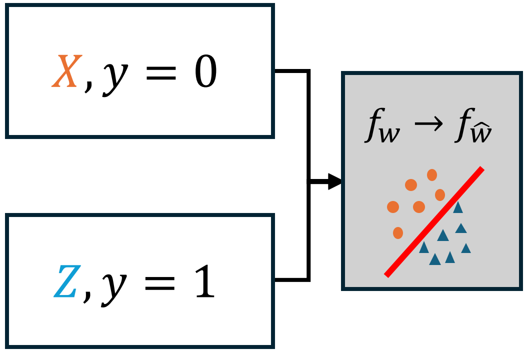
Anomalies

Classifier-based two-sample tests $\quad p_{SM} = p_{data}?$

Train a classifier to separate background from measurements

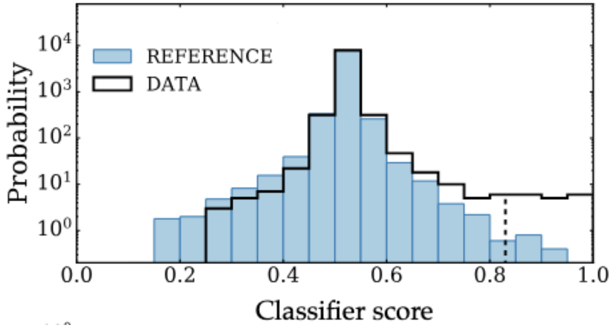


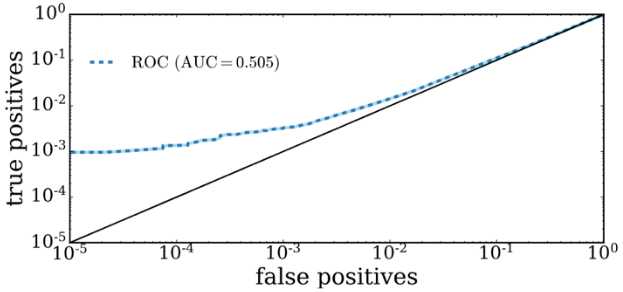
$X = \{x_1, \dots, x_n\} \sim p_{SM}, \qquad Z = \{z_1, \dots, z_m\} \sim p_{data}$



perform a test on the classifier output:

accuracy, AUC, KS, χ^2, ...



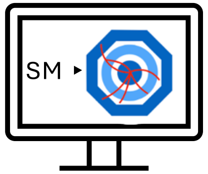


Baker, Cousins (1984), Friedman (2003), Lopez-Paz, Oquab (2017),
Metodiev, Nachman, Thaler (2017), D'Agnolo, Wulzer (2018),
ML, Losapio, Rando, Grosso, Wulzer, Pierini, Zanetti, Rosasco (2022),...

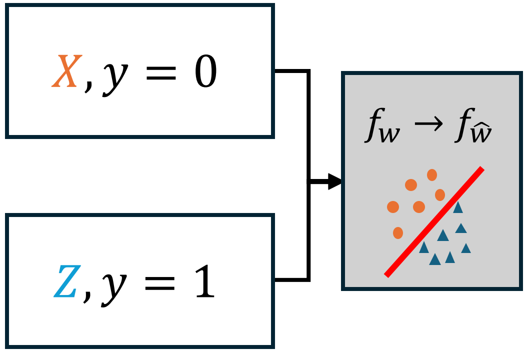
Anomalies

Classifier-based two-sample tests $\quad p_{SM} = p_{data}?$

Train a classifier to separate background from measurements



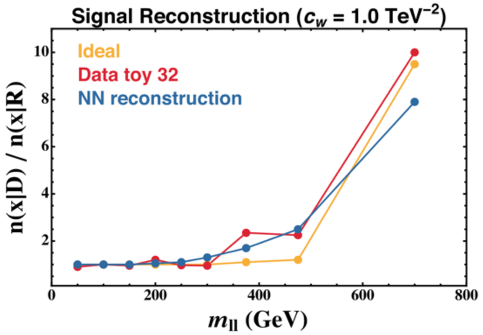
$X = \{x_1, \dots, x_n\} \sim p_{SM}, \qquad Z = \{z_1, \dots, z_m\} \sim p_{data}$

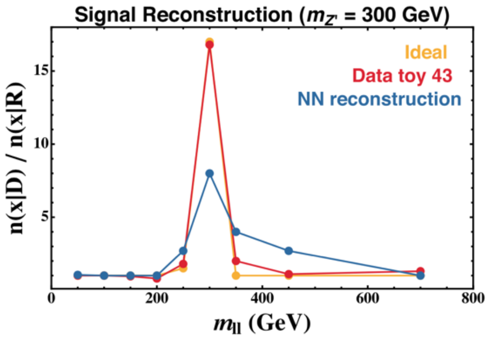


data-driven Neyman-Pearson testing:

"likelihood-ratio trick"

$$f_{\hat{w}} \approx \log \frac{p_{data}}{p_{SM}} \quad \to \quad t_{\hat{w}}(\mathcal{D}) = 2 \log \prod_{x \in \mathcal{D}} f_{\hat{w}}(x)$$





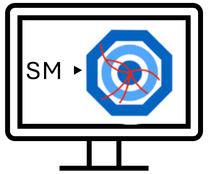
Baker, Cousins (1984), Friedman (2003), Lopez-Paz, Oquab (2017), Metodiev, Nachman, Thaler (2017), D'Agnolo, Wulzer (2018), ML, Losapio, Rando, Grosso, Wulzer, Pierini, Zanetti, Rosasco (2022),...

Anomalies

Classifier-based two-sample tests

$$p_{SM} = p_{data}?$$

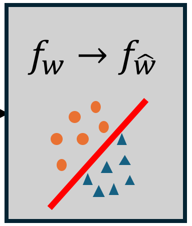
Train a classifier to separate background from measurements



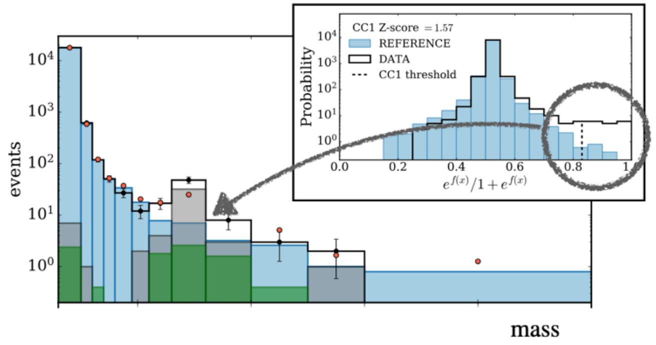
$X = \{x_1, \ldots, x_n\} \sim p_{SM},$ $Z = \{z_1, \ldots, z_m\} \sim p_{data}$

$X, y = 0$

$Z, y = 1$



enhance signal hypotheses (e.g. bump hunts)



[Baker, Cousins (1984)](), [Friedman (2003)](), [Lopez-Paz, Oquab (2017)](),
[Metodiev, Nachman, Thaler (2017)](), [D'Agnolo, Wulzer (2018)](),
[ML, Losapio, Rando, Grosso, Wulzer, Pierini, Zanetti, Rosasco (2022)](),…

Anomalies

To establish significance we need to calibrate

the SM is good $\Longrightarrow$ accuracy = 55% $\rightarrow$ Is it significant? Estimate null hypothesis

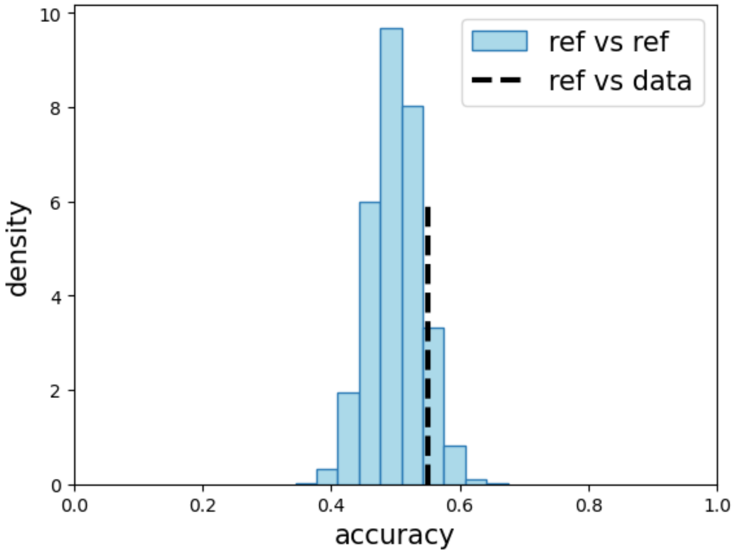
(permutation, bootstrap,…)

Leverage simulations:

SM vs SM

$$x_1, \ldots, x_m \sim p_{\mathrm{SM}}$$
$$x_1, \ldots, x_m \sim p_{\mathrm{SM}}$$

…



$$p_{\mathrm{value}} = \int_{t_{obs}}^{\infty} dt\, p(t)$$
$$Z = \Phi^{-1}(1 - p_{\mathrm{value}})$$

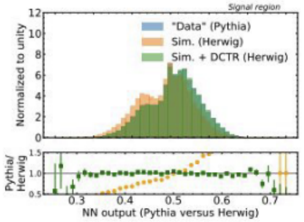
D'Agnolo, Wulzer (2018),
ML, Losapio, Rando, Grosso, Wulzer, Pierini, Zanetti, Rosasco (2022)
Grosso, ML, Pierini, Wulzer (2023)

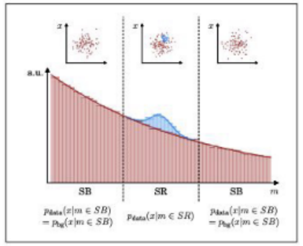
Anomalies

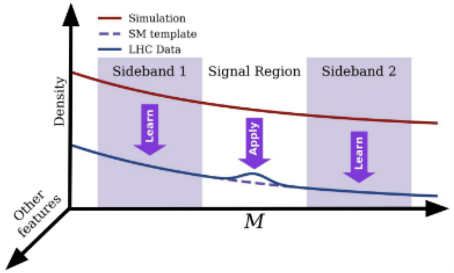
Examples of enhanced "traditional" model-independence:

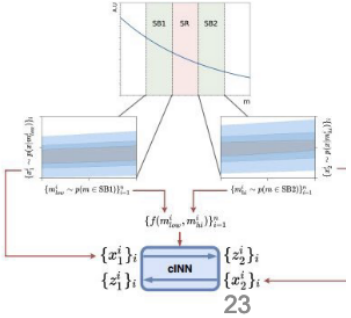
learning high-dimensional bkg templates* for bump hunts

In-situ BG estimate	Learn from simulation	Learn from data (SB)
Modeling the likelihood ratio	SALAD	CATHODE* [*see also LaCATHODE & ANODE]
Morphing the features	FETA	CURTAINs & Flow4Flows









[*Fidelity of simulation alone insufficient]

[source: T. Golling, Corfu2024 Workshop on Future Accelerators]

Anomalies

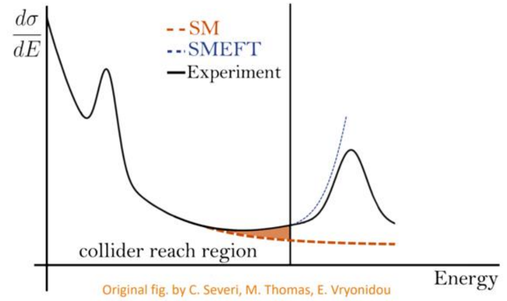
Examples of enhanced “traditional” model-independence:

Effective field theories

[see David and Claudia’s talks]

$$\mathcal{L}_{SMEFT} = \mathcal{L}_{SM}^{(d=4)} + \sum_i \frac{C_i^{(5)}}{\Lambda} O_i^{(5)} + \sum_i \frac{C_i^{(6)}}{\Lambda^2} O_i^{(6)} + \dots$$

[Source: David’s talk]



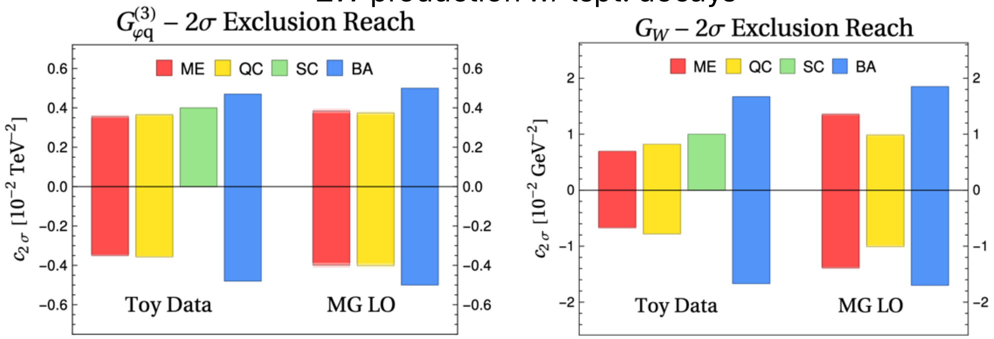
[Source: Victor’s talk]

Unbinned methodologies for new physics searches in EFT*

$$d\sigma_0(x;c) = d\sigma_1(x)\left\{[1 + c\,\alpha(x)]^2 + [c\,\beta(x)]^2\right\}$$

$$L[\mathrm{n}_\alpha(\cdot), \mathrm{n}_\beta(\cdot)] = \sum_{c_i \in \mathcal{C}} \left\{ \sum_{\mathrm{e}\in \mathrm{S}_0(c_i)} w_\mathrm{e}[f(x_\mathrm{e}, c_i)]^2 + \sum_{\mathrm{e}\in \mathrm{S}_1(c_i)} w_\mathrm{e}[1 - f(x_\mathrm{e}, c_i)]^2 \right\}$$

ZW production w/ lept. decays



Buchmuller, Wyler (1985)

Grzadkowski, Iskrzyński, Misiak, Rosiek (2017)

...

*Chen, Glioti, Panico, Wulzer (2020)

Chatterjee, Frohner, Lechner, Schöfbeck, Schwarz (2022)

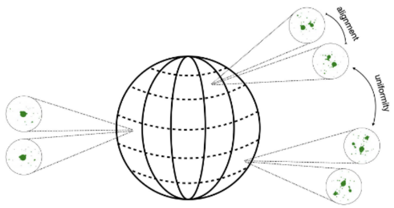
Ambrosio, Hoeve, Madigan, Rojo, Sanz (2022)

...

Anomalies

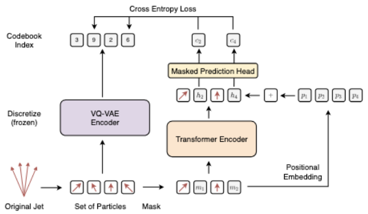
Foundation models

Symmetry Augmentation



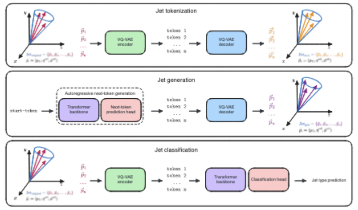
[Dillon, Kasieczka, Olischlager, Plehn, Sorrenson, Vogel, SciPost 2021]

Masked Particle Modeling



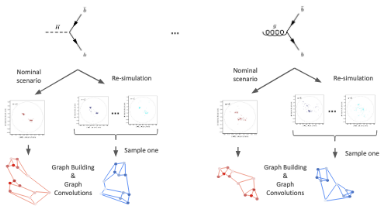
[Heinrich, Golling, Kagan, Klein, Leigh, Osadchy, Raine, arXiv 2024]

Next Token Prediction



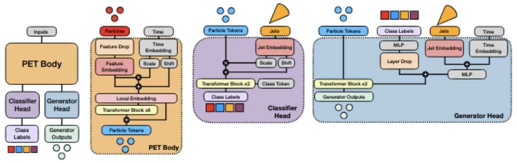
[Birk, Hallin, Kasieczka, arXiv 2024]

Re-Simulation Similarity

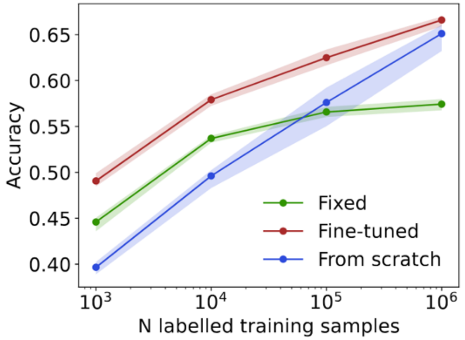


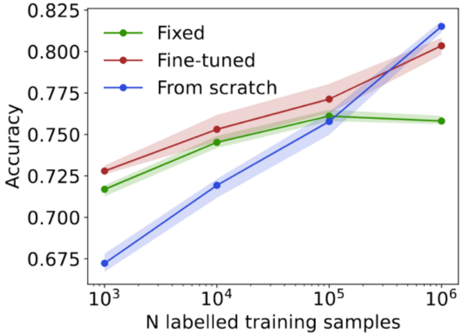
[Harris, Kagan, Krupa, Maier, Woodward, arXiv 2024]

Multi-Category Classification



[Mikuni, Nachman, arXiv 2024]





[source: J. Thaler, PHYSTAT - Stats meets ML, London 2024]

Could boost high-d, low-n

Are they anomaly-preserving?

Outline

Machine learning and data-driven modelling

Maximise discovery potential: anomalies

Should we care about interpretability

Interpretability

Warm, fuzzy feeling that you understand what your NN is doing.

J. Thaler (PHYSTAT workshop - Stat meets ML, London 2024)

Other characterisations might be more useful

- Explainability
- Robustness
- Accuracy
- Trustworthiness
- Uncertainty quantification
- ...

Interpretability

Systematic uncertainties are crucial for deployment

Grosso, D'Agnolo, Wulzer, Zanetti, Pierini (2021)

$$t(\mathcal{D},\,\mathcal{A}) = 2\log\left[\frac{\max_{\mathbf{w},\,\boldsymbol{\nu}} \mathcal{L}(\mathrm{H}_{\mathbf{w},\,\boldsymbol{\nu}}|\mathcal{D})\mathcal{L}(\boldsymbol{\nu}|\mathcal{A})}{\max_{\boldsymbol{\nu}} \mathcal{L}(\mathrm{R}_{\boldsymbol{\nu}}|\mathcal{D})\mathcal{L}(\boldsymbol{\nu}|\mathcal{A})}\right]\cdot\frac{\mathcal{L}(\mathrm{R}_{\mathbf{0}}|\mathcal{D})\mathcal{L}(\mathbf{0}|\mathcal{A})}{\mathcal{L}(\mathrm{R}_{\mathbf{0}}|\mathcal{D})\mathcal{L}(\mathbf{0}|\mathcal{A})}$$

$$= \boxed{\tau(\mathcal{D},\,\mathcal{A})} - \boxed{\Delta(\mathcal{D},\,\mathcal{A})}$$

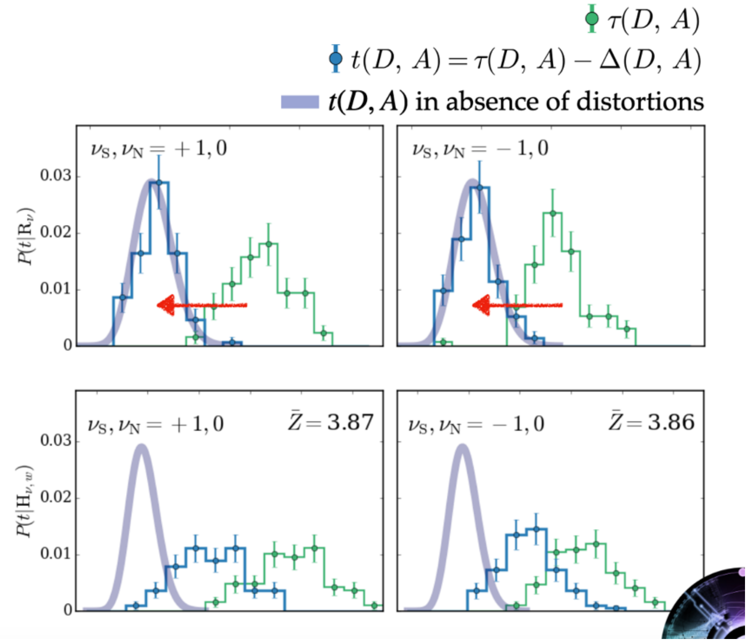
Tau term:

$$\tau(\mathcal{D},\,\mathcal{A}) = 2\max_{\mathbf{w},\,\boldsymbol{\nu}}\log\left[\frac{\mathcal{L}(\mathrm{H}_{\mathbf{w}},\,\boldsymbol{\nu}|\mathcal{D})\,\mathcal{L}(\boldsymbol{\nu}|\mathcal{A})}{\mathcal{L}(\mathrm{R}_{\mathbf{0}}|\mathcal{D})\,\mathcal{L}(\mathbf{0}|\mathcal{A})}\right] = -2\min_{\mathbf{w},\boldsymbol{\nu}} L\left[f(x,\mathbf{w}),\; r(x;\boldsymbol{\nu})\right]$$

NN model

Delta term:

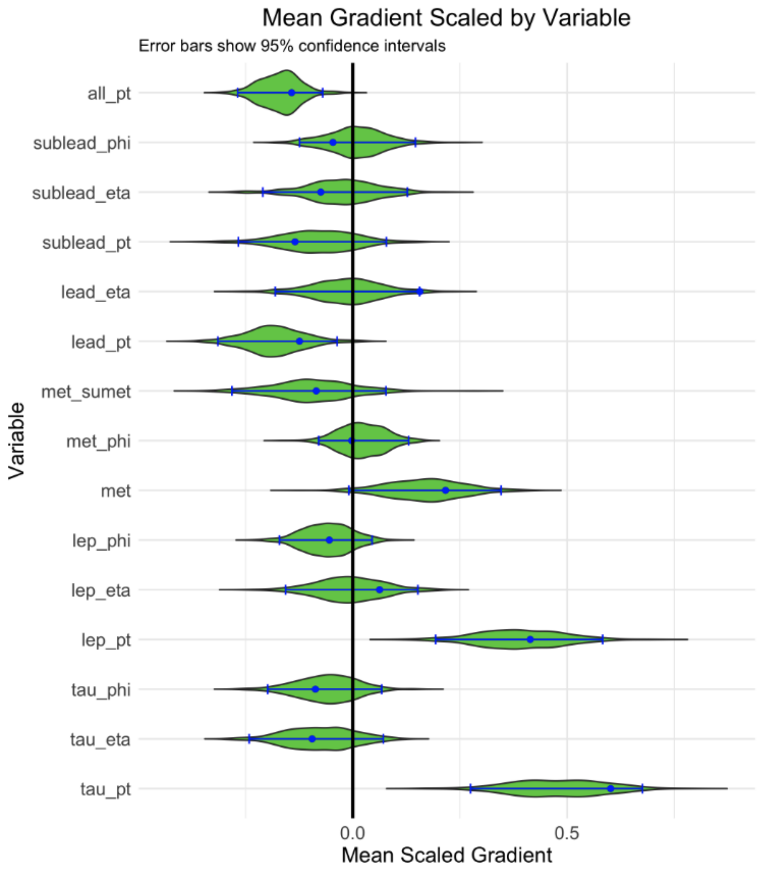
$$\Delta(\mathcal{D},\,\mathcal{A}) = 2\max_{\boldsymbol{\nu}}\log\left[\frac{\mathcal{L}(\mathrm{R}_{\boldsymbol{\nu}}|\mathcal{D})\,\mathcal{L}(\boldsymbol{\nu}|\mathcal{A})}{\mathcal{L}(\mathrm{R}_{\mathbf{0}}|\mathcal{D})\,\mathcal{L}(\mathbf{0}|\mathcal{A})}\right] = -2\min_{\boldsymbol{\nu}} L\left[r(x;\boldsymbol{\nu})\right]$$



[Gaia Grosso, Phystat London 2024]

Interpretability

Which features drive the decision?



Interpreting classifiers using active subspace methods

[Chakravarti, Kuusela, Lei, Wasserman (2021)]

Make sure they are physically relevant?

Can we get surpised?

Interpretability

Generative modeling a promising framework for fast simulations: normalizing flow and diffusion models

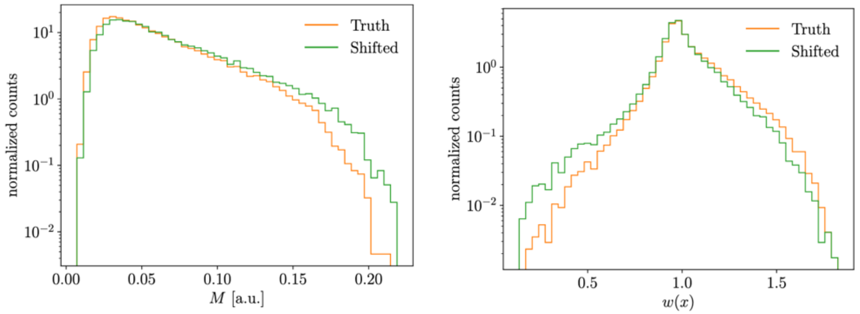
Robust evaluation is crucial for precision sciences

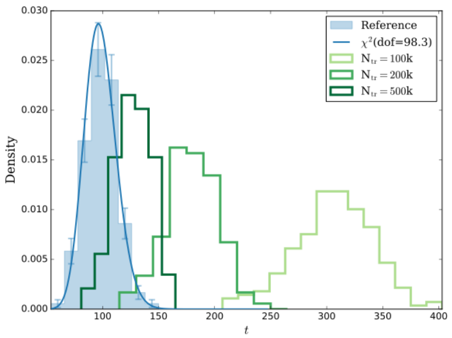
Test the tests

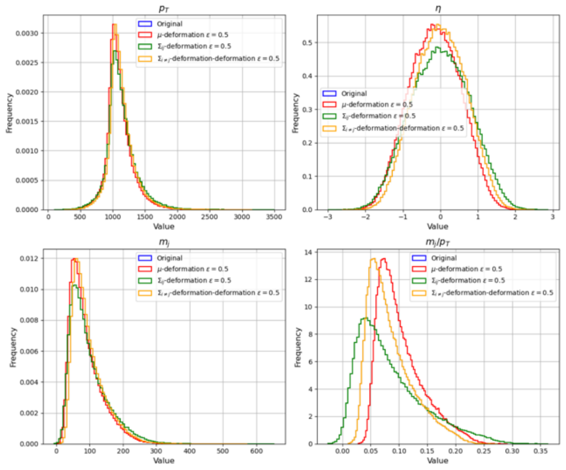
Refereeing the Referees: Evaluating Two-Sample Tests for Validating Generators in Precision Sciences

Samuele Grossi[a,b], Marco Letizia[b,c], and Riccardo Torre[a,b]

Again classifiers $p_{gen} = e^{f} p_{data}$







Scaled Jet features with $\mathbf{n = m = 5 \cdot 10^4}$

Statistic	μ-deformation $\epsilon_{95\%CL}$	$\epsilon_{99\%CL}$	t (s)	Σ_{ii}-deformation $\epsilon_{95\%CL}$	$\epsilon_{99\%CL}$	t (s)
t_{SW}	$0.01623^{+0.0045}_{-0.0069}$	$0.02098^{+0.0049}_{-0.0059}$	12410	$0.02089^{+0.0073}_{-0.008}$	$0.02834^{+0.0077}_{-0.0079}$	**1054**
$t_{\overline{KS}}$	$0.01585^{+0.0043}_{-0.0063}$	$0.01927^{+0.0043}_{-0.0056}$	17174	$\mathbf{0.02085^{+0.0064}_{-0.008}}$	$\mathbf{0.02567^{+0.006}_{-0.0075}}$	38871
t_{SKS}	$\mathbf{0.0113^{+0.0044}_{-0.005}}$	$\mathbf{0.0141^{+0.0037}_{-0.0045}}$	32620	$0.02254^{+0.0074}_{-0.0099}$	$0.02773^{+0.0073}_{-0.0089}$	28803
t_{FGD}	$0.02106^{+0.0062}_{-0.0079}$	$0.02659^{+0.0058}_{-0.0069}$	**11583**	$0.02133^{+0.0078}_{-0.0097}$	$0.02741^{+0.0071}_{-0.008}$	14254
t_{MMD}	$0.06739^{+0.013}_{-0.021}$	$0.08802^{+0.013}_{-0.011}$	46972	$0.0318^{+0.015}_{-0.0083}$	$0.04328^{+0.014}_{-0.012}$	28709
Statistic	$\Sigma_{i\neq j}$-deformation $\epsilon_{95\%CL}$	$\epsilon_{99\%CL}$	t (s)	pow_+-deformation $\epsilon_{95\%CL}$	$\epsilon^{pow_+}_{99\%CL}$	t (s)
t_{SW}	$0.0503^{+0.016}_{-0.019}$	$0.07052^{+0.015}_{-0.014}$	**1008**	$0.02465^{+0.011}_{-0.0081}$	$0.03314^{+0.0099}_{-0.0095}$	**1025**
$t_{\overline{KS}}$	$1.02009^{+0.0072}_{-0.001}$	$1.02812^{+0.003}_{-0.008}$	16410	$0.0232^{+0.0074}_{-0.011}$	$0.02698^{+0.01}_{-0.0092}$	35198
t_{SKS}	$0.06201^{+0.02}_{-0.029}$	$0.07573^{+0.02}_{-0.024}$	35383	$0.0402^{+0.015}_{-0.015}$	$0.04921^{+0.015}_{-0.015}$	47807
t_{FGD}	$\mathbf{0.00627^{+0.0016}_{-0.0018}}$	$\mathbf{0.00809^{+0.0015}_{-0.0018}}$	14008	$0.02237^{+0.013}_{-0.011}$	$0.0281^{+0.011}_{-0.0084}$	24967
t_{MMD}	$0.0794^{+0.039}_{-0.031}$	$0.112^{+0.031}_{-0.026}$	29620	$\mathbf{0.01898^{+0.012}_{-0.0094}}$	$\mathbf{0.02472^{+0.012}_{-0.0076}}$	66075
Statistic	pow_--deformation $\epsilon_{95\%CL}$	$\epsilon^{pow_-}_{99\%CL}$	t (s)	$\mathcal{N}$-deformation $\epsilon_{95\%CL}$	$\epsilon^{\mathcal{N}}_{99\%CL}$	t (s)
t_{SW}	$0.02527^{+0.011}_{-0.011}$	$0.03513^{+0.0084}_{-0.01}$	**993**	$0.11836^{+0.027}_{-0.028}$	$0.14062^{+0.018}_{-0.026}$	**910**
$t_{\overline{KS}}$	$\mathbf{0.02125^{+0.01}_{-0.0092}}$	$\mathbf{0.02649^{+0.0074}_{-0.009}}$	16472	$0.10579^{+0.014}_{-0.019}$	$0.11672^{+0.012}_{-0.016}$	31727
t_{SKS}	$0.03986^{+0.013}_{-0.017}$	$0.04873^{+0.013}_{-0.013}$	27407	$0.08577^{+0.024}_{-0.028}$	$0.10148^{+0.021}_{-0.026}$	25899
t_{FGD}	$0.02163^{+0.015}_{-0.0097}$	$0.02954^{+0.014}_{-0.0087}$	12892	$\mathbf{0.07833^{+0.0094}_{-0.019}}$	$\mathbf{0.08847^{+0.0084}_{-0.0069}}$	13246
t_{MMD}	$0.02133^{+0.013}_{-0.0086}$	$0.02924^{+0.011}_{-0.0081}$	68458	$0.26032^{+0.037}_{-0.057}$	$0.29897^{+0.028}_{-0.036}$	42149
Statistic	$\mathcal{U}$-deformation $\epsilon_{95\%CL}$	$\epsilon^{\mathcal{U}}_{99\%CL}$	t (s)	Timing t^{null} (s)		
t_{SW}	$0.20487^{+0.042}_{-0.048}$	$0.2434^{+0.032}_{-0.035}$	**877**	**123**		
$t_{\overline{KS}}$	$0.18018^{+0.024}_{-0.035}$	$0.19884^{+0.018}_{-0.027}$	25630	1913		
t_{SKS}	$0.14529^{+0.04}_{-0.056}$	$0.1719^{+0.035}_{-0.048}$	42277	4383		
t_{FGD}	$\mathbf{0.13545^{+0.014}_{-0.032}}$	$\mathbf{0.15299^{+0.015}_{-0.012}}$	12782	1787		
t_{MMD}	$0.45177^{+0.066}_{-0.091}$	$0.52083^{+0.05}_{-0.047}$	56078	3504		

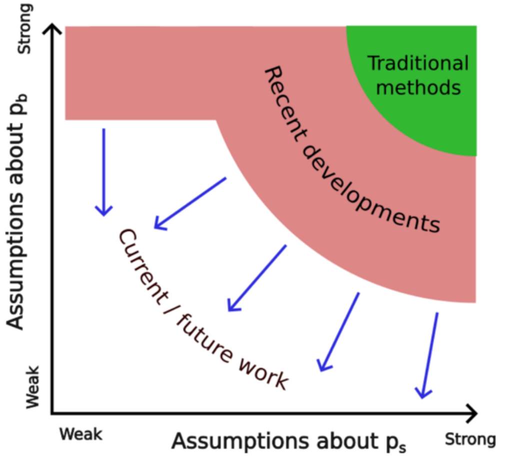
Kansal, Li, Duarte, Chernyavskaya, Pierini, Orzari, Tomei (2022)
Das, Favaro, Heimel, Krause, Plehn, Shih (2023)

Conclusions

- Machine learning can enable large scale model-independent searches:

  exploration AND exploitation

- Trust in ML: intepretability, robustness, uncertainty quantification,…



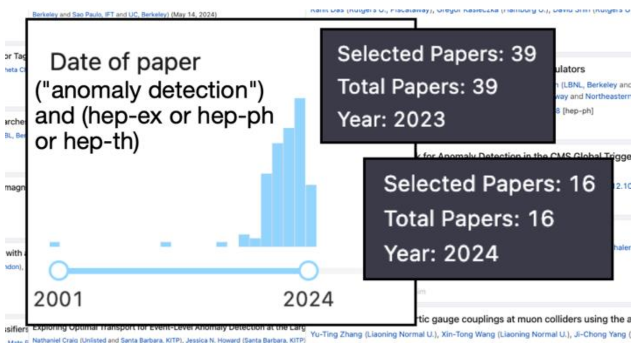
[source: M. Kuusela, PHYSTAT London 2024]

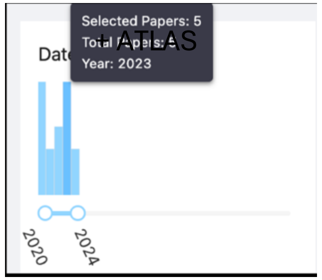
- Are foundation models robust beyond supervised tasks?

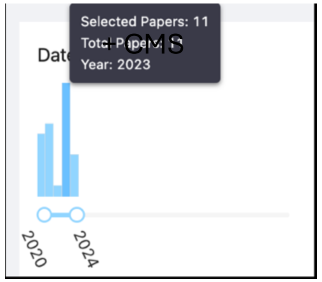
- Follow-up strategy after an anomalous detection?

- How to interpret signal-agnostic null results?

- Still a large gap between R&D and deployment







[source: T. Aarrestad, PHYSTAT London 2024]

Thank you