Symbolic Regression

Tilman Plehn

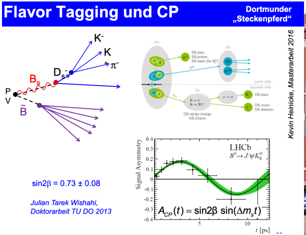
Universität Heidelberg

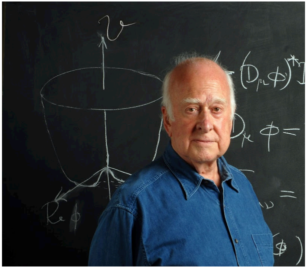
ErUM-Data, Wiehl, September 2022

Modern LHC physics

Classic motivation

- dark matter
- baryogenesis
- Higgs VEV





Modern LHC physics

Classic motivation

- dark matter
- baryogenesis
- Higgs VEV

LHC physics

- fundamental questions
- huge data set
- complete uncertainty control
- first-principle precision simulations

Modern LHC physics

Classic motivation

- dark matter
- baryogenesis
- Higgs VEV

LHC physics

- fundamental questions
- huge data set
- complete uncertainty control
- first-principle precision simulations

Traditional methods

- discover in rates
- unveil little black holes
- find supersymmetry
- travel extra dimensions
- measure couplings

Modern LHC physics

Classic motivation

- dark matter
- baryogenesis
- Higgs VEV

LHC physics

- fundamental questions
- huge data set
- complete uncertainty control
- first-principle precision simulations

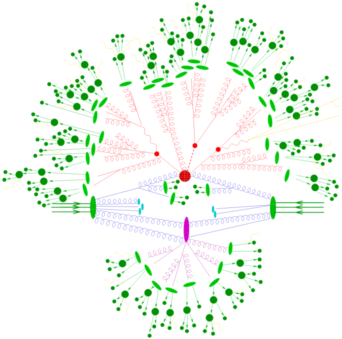
Traditional methods

- discover in rates
- unveil little black holes
- find supersymmetry
- travel extra dimensions
- measure couplings

First-principle simulations

- start with Lagrangian
- calculate scattering using QFT
- simulate events
- simulate detectors

$\rightarrow$ LHC events in virtual worlds



Modern LHC physics

Classic motivation

- dark matter
- baryogenesis
- Higgs VEV

LHC physics

- fundamental questions
- huge data set
- complete uncertainty control
- first-principle precision simulations

Traditional methods

- discover in rates
- unveil little black holes
- find supersymmetry
- travel extra dimensions
- measure couplings

First-principle simulations

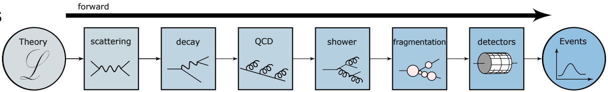
- start with Lagrangian
- calculate scattering using QFT
- simulate events
- simulate detectors

$\rightarrow$ LHC events in virtual worlds

Simulation-based inference

- compare simulations and data
- analyze data systematically [SMEFT]
- understand LHC dataset [SM or BSM]
- publish useable results

$\rightarrow$ With a little help from data science...



Why Formulas

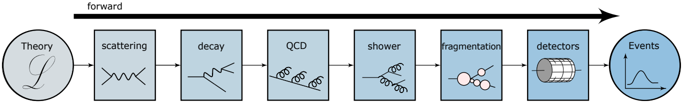
Modern LHC physics — all numerics

- Lagrangian defining the relevant parameters through formula

$$\mathcal{L} = \overline{Q}_L i \not{D} Q_L + \overline{Q}_R i \not{D} Q_R - \frac{1}{4} F_{\mu\nu} F^{\mu\nu} - \mu^2 |\phi|^2 - \lambda |\phi|^4 - \frac{f_{\phi,2}}{3\Lambda^2} |\phi|^6$$

- extract Feynman rules
- compute and square transition amplitudes
- add parton-shower gluon radiation
- simulate hadronization/fragmentation and detector response

$\to$ Nothing to look at and understand



Why Formulas

Modern LHC physics — all numerics

- Lagrangian defining the relevant parameters through formula

$$\mathcal{L} = \overline{Q}_L i\not{D} Q_L + \overline{Q}_R i\not{D} Q_R - \frac{1}{4} F_{\mu\nu} F^{\mu\nu} - \mu^2 |\phi|^2 - \lambda |\phi|^4 - \frac{f_{\phi,2}}{3\Lambda^2} |\phi|^6$$

- extract Feynman rules
- compute and square transition amplitudes
- add parton-shower gluon radiation
- simulate hadronization/fragmentation and detector response

$\to$ Nothing to look at and understand

Benefit of formulas

- recognizable content

$$\dot{N}(t) = -\lambda N(t) \qquad \text{decay law}$$

$$E = \frac{m}{2}\dot{x}^2 + \frac{k}{2}x^2 \qquad \text{harmonic oscillator}$$

$$E = mc^2 \qquad \text{something with Einstein}$$

- symmetry properties $\quad t = p_{T,1} p_{T,2} \sin(2\Delta\phi)$ $\quad$ independent of p_z
- Taylor series $\quad \sin\phi = \phi + \mathcal{O}(\phi^2)$

$\to$ Way to understand physics

Why Formulas

Modern LHC physics — all numerics

- Lagrangian defining the relevant parameters through formula

$$\mathcal{L} = \overline{Q}_L i \not{D} Q_L + \overline{Q}_R i \not{D} Q_R - \frac{1}{4} F_{\mu\nu} F^{\mu\nu} - \mu^2 |\phi|^2 - \lambda |\phi|^4 - \frac{f_{\phi,2}}{3\Lambda^2} |\phi|^6$$

- extract Feynman rules
- compute and square transition amplitudes
- add parton-shower gluon radiation
- simulate hadronization/fragmentation and detector response

$\to$ Nothing to look at and understand

Formulas as ML-models

- neural networks best interpolation [NNPDF]
- interpolation vs extrapolation

  planetary movements
  time series in cancer research
  weather forecast
  background modelling with one/two sideband(s)
  LHC-simulation of kinematics tails

$\to$ feature-based networks useless

$\to$ model-based ML implicit bias, formulas, differential equations

AI-Feynman

Properties of physics formulas [Udrescu & Tegmark]

- units limiting allowed structures [$A(p^2=0, m=0)=0$]
- Taylor low-order polynomials everywhere
- smoothness nature is smooth and differentiable
- symmetry translation, rotation, scaling...
- compositionality $f(x,y,z) = f_1(x,y)\, f_2(y,z)$
- separability $f(x,y,z) = f_1(x)\, f_2(y,z)$

$\to$ Basis for extracting physics formulas from data?

AI-Feynman

Properties of physics formulas [Udrescu & Tegmark]

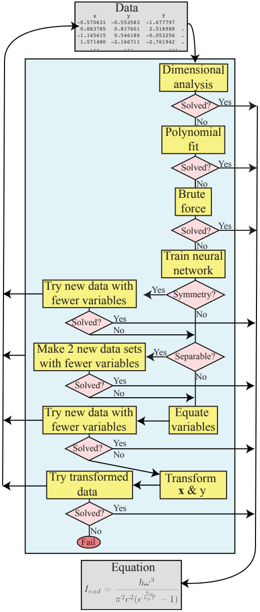
- units limiting allowed structures $[A(p^2 = 0,\ m = 0) = 0]$
- Taylor low-order polynomials everywhere
- smoothness nature is smooth and differentiable
- symmetry translation, rotation, scaling...
- compositionality $f(x,y,z) = f_1(x,y)\ f_2(y,z)$
- separability $f(x,y,z) = f_1(x)\ f_2(y,z)$

$\to$ Basis for extracting physics formulas from data?

Algorithm

- start with numerical dataset $f(x,y)$
- Brute force the standard algorithm
- represent formula by 1D string [pocket calculator]
- objects $x\ \pi\ \dots$
  single argument $\sqrt{\ }\ \exp\ \log\ \sin\ \dots$
  two arguments $+\ -\ *\ /$
- loss function balancing complexity and precision

$$L = \log \mathtt{rank} + \lambda \log \max\left(1, \frac{\mathrm{RMS}}{10^{-15}}\right)$$



AI-Feynman

Properties of physics formulas [Udrescu & Tegmark]

- units limiting allowed structures $[A(p^2 = 0, m = 0) = 0]$
- Taylor low-order polynomials everywhere
- smoothness nature is smooth and differentiable
- symmetry translation, rotation, scaling...
- compositionality $f(x, y, z) = f_1(x, y)\ f_2(y, z)$
- separability $f(x, y, z) = f_1(x)\ f_2(y, z)$

$\to$ Basis for extracting physics formulas from data?

Benchmarking and naming

Feynman	Equation	Time [s]	Methods	Data
I.6.20	$f = e^{-\frac{\theta^2}{2\sigma^2}} / \sqrt{2\pi\sigma^2}$	2992	ev, bf-log	10^2
I.9.18	$F = \frac{Gm_1 m_2}{(x_2-x_1)^2+(y_2-y_1)^2+(z_2-z_1)^2}$	5975	all	10^6
I.10.7	$m = \frac{m_0}{\sqrt{1-\frac{v^2}{c^2}}}$	14	unit, bf	10
I.11.19	$A = x_1 y_1 + x_2 y_2 + x_3 y_3$	184	unit, pf	10^2
I.34.10	$\omega = \frac{\omega_0}{1-v/c}$	13	unit, bf	10
I.34.27	$E = \hbar\omega$	8	unit	10
I.40.1	$n = n_0 e^{-\frac{mgx}{k_b T}}$	20	unit, bf	10

AI-Feynman

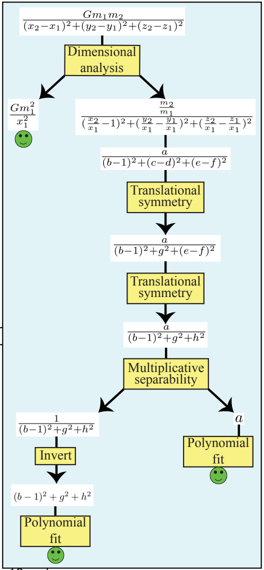
Properties of physics formulas [Udrescu & Tegmark]

- units — limiting allowed structures $[A(p^2 = 0, m = 0) = 0]$
- Taylor — low-order polynomials everywhere
- smoothness — nature is smooth and differentiable
- symmetry — translation, rotation, scaling...
- compositionality — $f(x, y, z) = f_1(x, y)\; f_2(y, z)$
- separability — $f(x, y, z) = f_1(x)\; f_2(y, z)$

$\to$ Basis for extracting physics formulas from data?

Benchmarking and naming

Feynman	Equation	Time [s]	Methods	
I.6.20	$f = e^{-\frac{\theta^2}{2\sigma^2}} / \sqrt{2\pi\sigma^2}$	2992	ev, bf-log	
I.9.18	$F = \frac{G m_1 m_2}{(x_2-x_1)^2+(y_2-y_1)^2+(z_2-z_1)^2}$	5975	all	
I.10.7	$m = \frac{m_0}{\sqrt{1-\frac{v^2}{c^2}}}$	14	unit, bf	
I.11.19	$A = x_1 y_1 + x_2 y_2 + x_3 y_3$	184	unit, pf	
I.34.10	$\omega = \frac{\omega_0}{1-v/c}$	13	unit, bf	
I.34.27	$E = \hbar\omega$	8	unit	
I.40.1	$n = n_0 e^{-\frac{mgx}{k_b T}}$	20	unit, bf	10



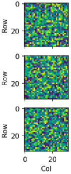
PySR

PySR alternative approach [Miles Cranmer]

- motivation: explainable AI
- modeling language of physics: formulas

$\rightarrow$ combine networks and formulas [slides from Miles's talk at Hammers & Nails 2022]

Empirical fit:

Kepler's third law	Planck's law	Neural Network Weights
$P^2 \propto a^3$	$B = \frac{2h\nu^3}{c^2}\left(\exp\left(\frac{h\nu}{k_\mathrm{B}T}\right) - 1\right)^{-1}$	
$\downarrow$	$\downarrow$ (Partially)	$\downarrow$
Newton's law of gravitation, to explain it	Quantum mechanics, to explain it	???

Networks and formulas

Formulas vs networks [Cranmer, Cranmer, etal]

- network to extract actual degrees of freedom
- networks fast to evaluate
- access to derivatives

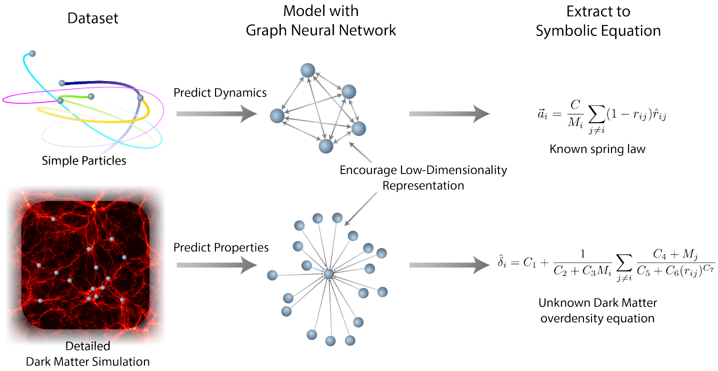
$\rightarrow$ Formulas through networks

Networks and formulas

Formulas vs networks [Cranmer, Cranmer, etal]

- network to extract actual degrees of freedom
- networks fast to evaluate
- access to derivatives

$\to$ Formulas through networks



Formula encoding

Analytic formulas for LHC observables [Brehmer, Butter, TP, Soybelman]

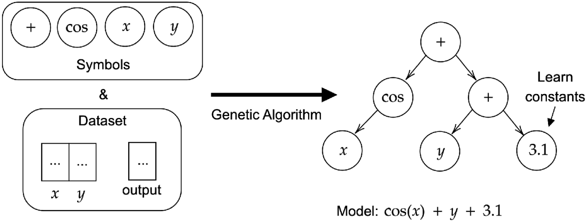
- function $t(x|\theta)$ approximated by tree
- order-one phase space parameters $\quad x_p = p_T/m_H, \Delta\eta, \Delta\phi$ [node]
- operators $\quad \sin x, x^2, x^3, x+y, x-y, x*y, x/y$ [node]

$\to$ figures of merit [complexity = number of nodes]

$$\text{MSE} = \frac{1}{n}\sum_{i=1}^{n}\Big[g_i(x) - t(x,z|\theta)\Big]^2$$

$$\text{score} \approx \text{MSE} + \text{parsimony} \cdot \text{complexity}$$

- Symbolic regression finds analytic expressions to fit a dataset.
- Pioneering work by Langley et al., 1980s; Koza et al., 1990s; Lipson et al., 2000s



Optimization

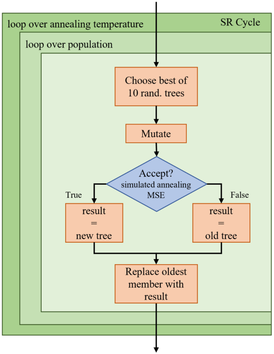
Simulated annealing

- combine trees to populations
- mutate trees   exchange, add, delete nodes
- acceptance probability

$$p = \exp\left(-\frac{\text{score}_\text{new} - \text{score}_\text{old}}{\alpha T \ \text{score}_\text{old}}\right)$$

- added: proper fit of pre-factors

→ Hall of Fame: best equation per complexity



Optimization

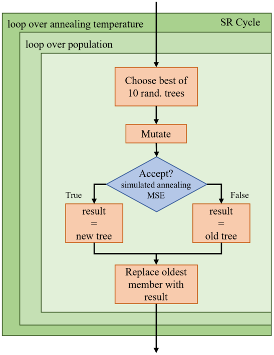
Simulated annealing

- combine trees to populations
- mutate trees exchange, add, delete nodes
- acceptance probability

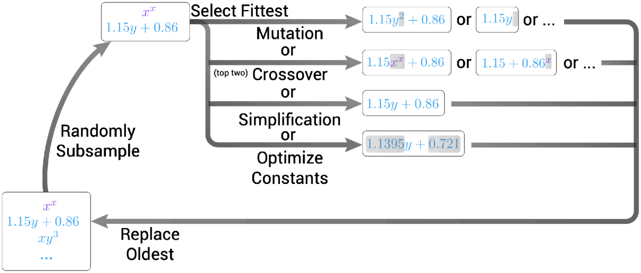
$$p = \exp\left(- \frac{\text{score}_\text{new} - \text{score}_\text{old}}{\alpha\, T\ \text{score}_\text{old}} \right)$$

- added: proper fit of pre-factors

$\to$ Hall of Fame: best equation per complexity



Miles' example



Orbital mechanics

Force law from orbital mechanics [Lemos, Jeffrey, Cranmer, Ho, Battaglia]

- data 3- years of solar system [sun, planets, big moons]
- objects scalar property per body, call it 31 masses
- graph network interaction as 465 edges, summed
- loss mean weighted error
- PySR interpret edges as force
- post-processing re-train masses and force law

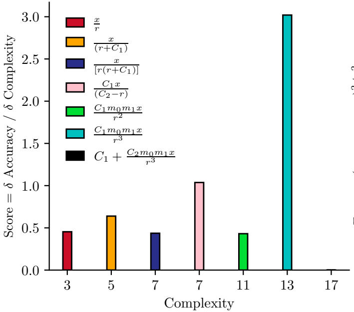
$\rightarrow$ Gravity it is...

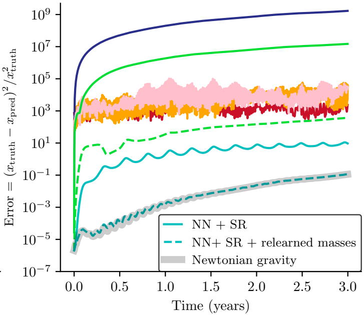
Orbital mechanics

Force law from orbital mechanics [Lemos, Jeffrey, Cranmer, Ho, Battaglia]

- data 3- years of solar system [sun, planets, big moons]
- objects scalar property per body, call it 31 masses
- graph network interaction as 465 edges, summed
- loss mean weighted error
- PySR interpret edges as force
- post-processing re-train masses and force law

→ Gravity it is...





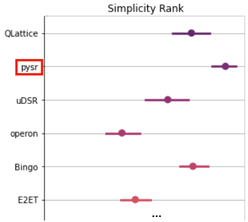
PySR competition

Standard in ML: conference challenges

- Kaggle tracking, Higgs
- ML4Jets top tagging, anomaly seraches, autoencoder, detector simulation

How well does this work?

- GECCO 2022 symbolic regression competition - equations track
- PySR won second place overall, but ***by far*** finds the simplest solutions:
- (First place is a proprietary software from a startup, so PySR is technically the best overall open-source symbolic regression!)



PySR competition

Standard in ML: conference challenges

- Kaggle — tracking, Higgs
- ML4Jets — top tagging, anomaly seraches, autoencoder, detector simulation

MilesCranmer / PySR Public

High-Performance Symbolic Regression in Python

Apache-2.0 license

761 stars 81 forks

- Open-source, free forever
- Extensible Python API compatible with scikit-learn
- Can be distributed over 1000-core clusters
- Custom operators, losses, constraints

Optimal observables

Measure model parameter θ optimally

- single-event likelihood [from Monte Carlo]

$$p(x|\theta) = \frac{1}{\sigma_\text{tot}(\theta)} \frac{d^d\sigma(x|\theta)}{dx^d}$$

- expanded locally in θ, define score [just taylor log]

$$\log \frac{p(x|\theta)}{p(x|\theta_0)} \approx (\theta - \theta_0)\, \nabla_\theta \log p(x|\theta)\Big|_{\theta_0} \equiv (\theta - \theta_0)\; t(x|\theta_0) \equiv (\theta - \theta_0)\; \mathcal{O}^\text{opt}(x)$$

- parton level, as used in ATLAS

$$p(x|\theta) \approx |\mathcal{M}|_0^2 + \theta |\mathcal{M}|_\text{int}^2 \qquad \Rightarrow \qquad t(x|\theta_0) \sim \frac{|\mathcal{M}|_\text{int}^2}{|\mathcal{M}|_0^2} \; ,$$

$\rightarrow$ Easy at parton level, LEP physics...

Optimal observables

Measure model parameter θ optimally

- single-event likelihood [from Monte Carlo]

$$p(x|\theta) = \frac{1}{\sigma_\text{tot}(\theta)} \frac{d^d\sigma(x|\theta)}{dx^d}$$

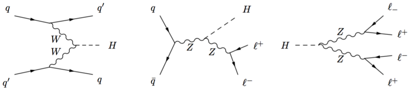
- expanded locally in θ, define score [just taylor log]

$$\log \frac{p(x|\theta)}{p(x|\theta_0)} \approx (\theta - \theta_0)\, \nabla_\theta \log p(x|\theta)\Big|_{\theta_0} \equiv (\theta - \theta_0)\, t(x|\theta_0) \equiv (\theta - \theta_0)\, \mathcal{O}^\text{opt}(x)$$

- parton level, as used in ATLAS

$$p(x|\theta) \approx |\mathcal{M}|_0^2 + \theta |\mathcal{M}|_\text{int}^2 \qquad \Rightarrow \qquad t(x|\theta_0) \sim \frac{|\mathcal{M}|_\text{int}^2}{|\mathcal{M}|_0^2} \,,$$

$\rightarrow$ Easy at parton level, LEP physics...



Discrete symmetry

- CPV at dimension-6
- unique CP-observable [C-even, P-odd, $\hat{T}$-odd]

$$t \propto \epsilon_{\mu\nu\rho\sigma}\; k_1^\mu\; k_2^\nu\; q_1^\rho\; q_2^\sigma\; \text{sign}\left[(k_1 - k_2)\cdot(q_1 - q_2)\right] \xrightarrow{\text{lab frame}} \sin\Delta\phi_{jj}$$

$\rightarrow$ Computable, modulo prefactor from D6-operator

Optimal observables after detector

Computing score using MadMiner

- likelihood ratio at detector level

$$\log \frac{p(x_d|\theta)}{p(x_d|\theta_0)} = \log \frac{\int dx_p \; T(x_d|x_p) \; p(x_p|\theta)}{\int dx_p \; T(x_d|x_p) \; p(x_p|\theta_0)}$$

- minimization problem for

$$F(x_d) = \int dx_p \; \left| g(x_d, x_p) - \hat{g}(x_d) \right|^2 \; T(x_d|x_p) \; p(x_p|\theta)$$

  smart choice

$$g(x_d, x_p) = \frac{p(x_p|\theta)}{p(x_p|\theta_0)} \qquad \Rightarrow \qquad \hat{g}_*(x_d) = \frac{p(x_d|\theta)}{p(x_d|\theta_0)}$$

- same for unobservable phase-space directions [joint score $t(x, z|\theta)$]

$\rightarrow$ Minimization means ML, function as NN

Optimal observables after detector

Computing score using MadMiner

- likelihood ratio at detector level

$$\log \frac{p(x_d|\theta)}{p(x_d|\theta_0)} = \log \frac{\int dx_p \; T(x_d|x_p) \; p(x_p|\theta)}{\int dx_p \; T(x_d|x_p) \; p(x_p|\theta_0)}$$

- minimization problem for

$$F(x_d) = \int dx_p \;\; |g(x_d, x_p) - \hat{g}(x_d)|^2 \;\; T(x_d|x_p) \; p(x_p|\theta)$$

smart choice

$$g(x_d, x_p) = \frac{p(x_p|\theta)}{p(x_p|\theta_0)} \qquad \Rightarrow \qquad \hat{g}_*(x_d) = \frac{p(x_d|\theta)}{p(x_d|\theta_0)}$$

- same for unobservable phase-space directions [joint score $t(x, z|\theta)$]

$\to$ Minimization means ML, function as NN

Going back to formulas [Brehmer, Butter, TP, Soybelman]

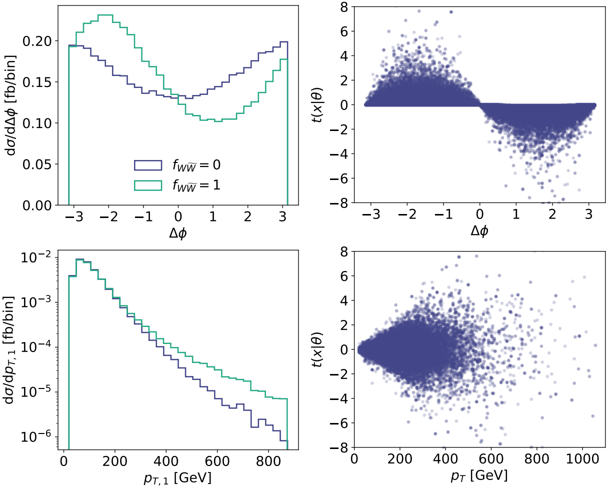
- detector-level score from MadMiner
- parton-level score analytically
- good enough formula for controlled use?

$\to$ Symbolic regression

Around Standard Model

Score around Standard Model

- shift in distributions, reflected in score [parton level]
  CP-effect in $\Delta\phi_{jj}$
  D6-effect in $p_{T,j}$



Around Standard Model

Score around Standard Model

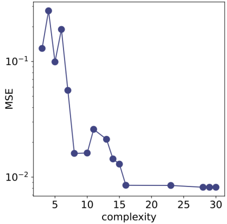
- shift in distributions, reflected in score [parton level]
  CP-effect in $\Delta\phi_{jj}$
  D6-effect in $p_{T,j}$
- best 4-parameter formula including $\Delta\eta$ [without/with detector]

$$t = -x_{p,1}\,\left(x_{p,2} + c\right)\,(a - b\Delta\eta)\,\sin(\Delta\phi + d)$$

with $a = 1.086(11)$ $b = 0.10241(19)$ $c = 0.24165(8)$ $d = 0.00662(32)$
$a = 0.926(2)$ $b = 0.08387(35)$ $c = 0.3542(20)$ $d = 0.00911(67)$

$\to$ Mostly expected formula

compl	dof	function	MSE
3	1	$a\,\Delta\phi$	$1.30 \cdot 10^{-1}$
4	1	$\sin(a\Delta\phi)$	$2.75 \cdot 10^{-1}$
5	1	$a\Delta\phi x_{p,1}$	$9.93 \cdot 10^{-2}$
6	1	$-x_{p,1}\sin(\Delta\phi + a)$	$1.90 \cdot 10^{-1}$
7	1	$(-x_{p,1} - a)\sin(\sin(\Delta\phi))$	$5.63 \cdot 10^{-2}$
8	1	$(a - x_{p,1})x_{p,2}\sin(\Delta\phi)$	$1.61 \cdot 10^{-2}$
14	2	$x_{p,1}(a\Delta\phi - \sin(\sin(\Delta\phi)))(x_{p,2} + b)$	$1.44 \cdot 10^{-2}$
15	3	$-(x_{p,2}(a\Delta\eta^2 + x_{p,1}) + b)\sin(\Delta\phi + c)$	$1.30 \cdot 10^{-2}$
16	4	$-x_{p,1}(a - b\Delta\eta)(x_{p,2} + c)\sin(\Delta\phi + d)$	$8.50 \cdot 10^{-3}$
28	7	$(x_{p,2} + a)(bx_{p,1}(c - \Delta\phi) - x_{p,1}(d\Delta\eta + ex_{p,2} + f)\sin(\Delta\phi + g))$	$8.18 \cdot 10^{-3}$



Away from Standard Model

Score away from Standard Model

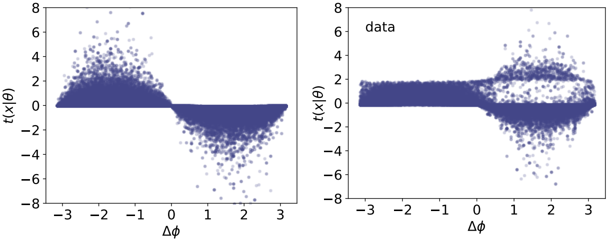
- scaling beyond linearization

$$p(x|\theta) = |\mathcal{M}|_0^2 + \theta|\mathcal{M}|_{\text{int}}^2 + \theta^2|\mathcal{M}|_{\text{quad}}^2$$

- saturating score

	$\theta \ll 1$	$\theta \gtrsim 1$
approximation	leading term $\dfrac{|\mathcal{M}|_{\text{int}}^2}{|\mathcal{M}|_0^2} + \dfrac{1}{|\mathcal{M}|_0^2}\left(2|\mathcal{M}|_{\text{quad}}^2 - \dfrac{|\mathcal{M}|_{\text{int}}^4}{|\mathcal{M}|_0^2}\right)\theta$	quadratic term $\dfrac{2}{\theta}$
scaling	mostly constant	decreasing with θ

- combination of different regimes



Away from Standard Model

Score away from Standard Model

- scaling beyond linearization

$$p(x|\theta) = |\mathcal{M}|_0^2 + \theta|\mathcal{M}|_\text{int}^2 + \theta^2|\mathcal{M}|_\text{quad}^2$$

- saturating score

	$\theta \ll 1$	$\theta \gtrsim 1$
approximation	leading term $\frac{|\mathcal{M}|_\text{int}^2}{|\mathcal{M}|_0^2} + \frac{1}{|\mathcal{M}|_0^2}\left(2|\mathcal{M}|_\text{quad}^2 - \frac{|\mathcal{M}|_\text{int}^4}{|\mathcal{M}|_0^2}\right)\theta$	quadratic term $\frac{2}{\theta}$
scaling	mostly constant	decreasing with θ

- combination of different regimes
- regression including division [rational function, complexity 31]

$$t(x_{p,\times}, s_\phi, \Delta\eta | f_{W\widetilde{W}} = 1) = \frac{a' x_{p,\times}(e' s_\phi^2 x_{p,\times} - s_\phi \Delta\eta - f')}{(b' x_{p,\times} + s_\phi - g')(e' s_\phi^2 x_{p,\times} - s_\phi\Delta\eta - f') - c' s_\phi^2 - d'}$$

with $a' = 0.75$ $b' = 0.38$ $c' = 4.2$ $d' = 4.6$ $e' = 1.1$ $f' = 0.26$ $g' = 0.21$

$\to$ Optimal observables more complex

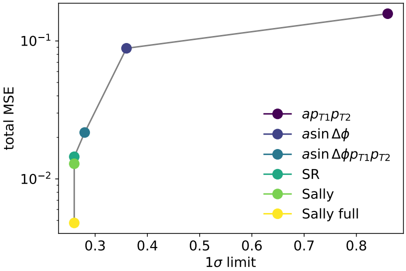
cmpl	dof	function	MSE
3	1	$ax_{p,\times}$	0.124
12	2	$ax_{p,\times}/(x_{p,\times}/\Delta\eta + \Delta\eta + b)$	0.116
15	2	$(s_\phi + a)(-s_\phi + x_{p,\times} - b)/(-s_\phi + x_{p,\times} + \Delta\eta/x_{p,\times})$	0.054
26	4	$a/(b - (s_\phi - c - d/(s_\phi^2 - s_\phi\Delta\eta - s_\phi/x_{p,\times} + ex_{p,\times}^2))/x_{p,\times})$	0.048
31	7	$a/(b - (s_\phi + (cs_\phi^2 - d)/(es_\phi^2 x_{p,\times}^2 - s_\phi\Delta\eta + f) - g)/x_{p,\times})$	0.039

Expectation for analysis

So what does the formula buy us?

- MSE for score:

  very wrong formula
  wrong formula
  right formula
  MadMiner



Expectation for analysis

So what does the formula buy us?

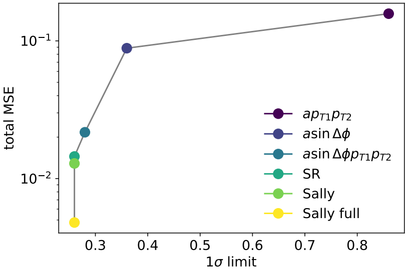
- MSE for score:

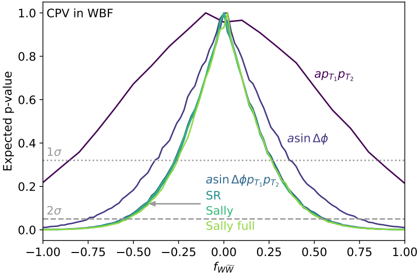
  very wrong formula
  wrong formula
  right formula
  MadMiner

- expected limits:

  very wrong formula
  wrong formula
  right formula $\approx$ MadMiner

$\rightarrow$ Statistically limited for Run 2





ML for LHC Theory

ML-applications in LHC analysis and theory

- just another numerical tool for a numerical field
- driven by money from data industry, medical research
- goals are...

  ...improve established tasks
  ...develop new tools for straightforward tasks
  ...come up with new ideas, now possible
- example   recovering formulas from complex observables

$\rightarrow$ Opportunity for young people to make a difference!