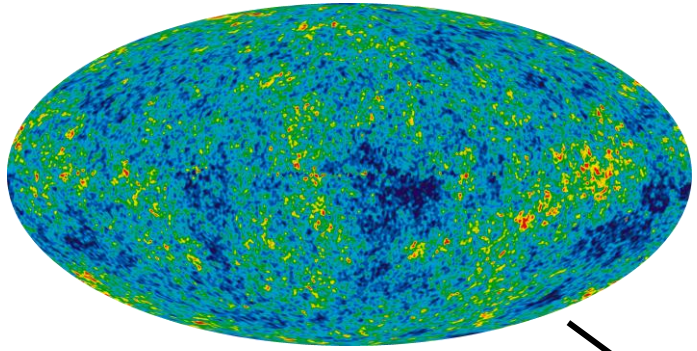
CP-Analyses with Symbolic Regression

[H. Bahl, E. Fuchs, M. Menen, T. Plehn '25]

Quantum Universe Attract Workshop

Hamburg, 24.11.2025

Marco Menen



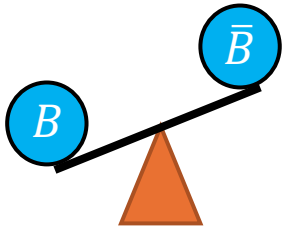
D/H

^{3}He/H ^{7}Li/H

Baryon asymmetry of the Universe (BAU)

[Planck '18]
[S. Navas et al. '24]

$$\eta = \frac{n_B - n_{\bar{B}}}{s} \approx 8.7 \cdot 10^{-11}$$



Sakharov conditions for successful baryogenesis:

[A. D. Sakharov '67]

1. Departure from thermal equilibrium
2. CP (and C) violation
3. Baryon number violation

Not sufficient amounts in SM

⇩

BSM Higgs sectors

The Higgs Characterization Model:

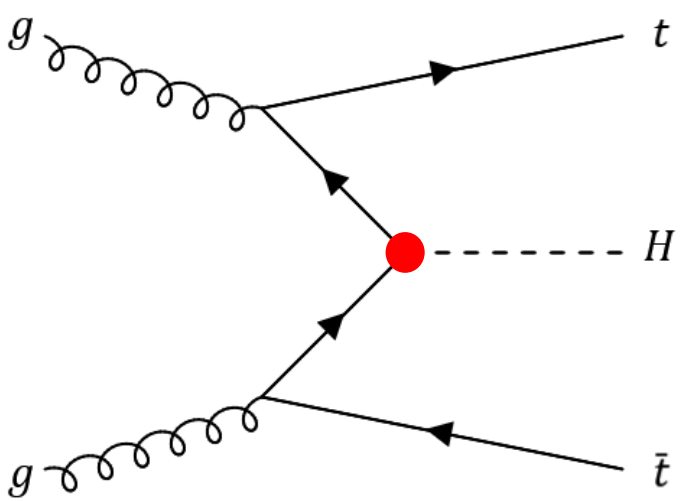
[P. Artoisenet et al. '13]

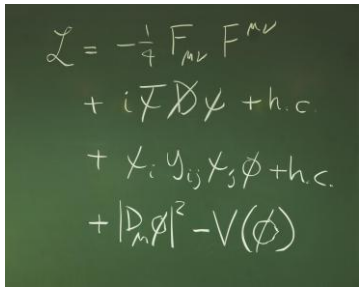
Modification of SM top-Yukawa coupling via CP-violating phase

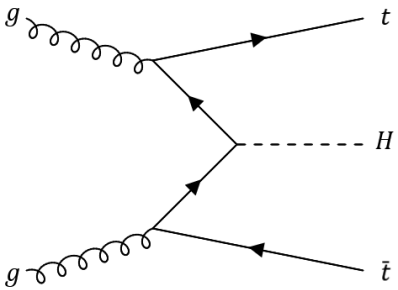
Total rate modifier g_t, CP-mixing angle α_t

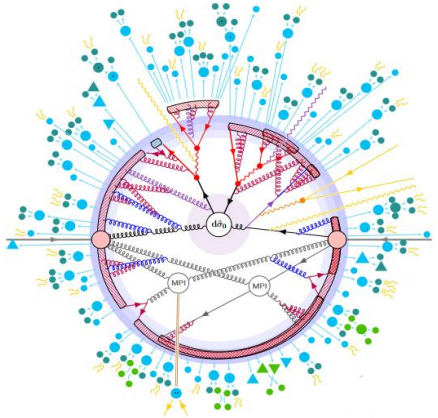
$$\mathcal{L}^{\mathrm{HCM}}_{t\bar{t}H} \supset \frac{y_t}{\sqrt{2}} t \; g_t \left(\cos\alpha_t + i\gamma_5 \sin\alpha_t\right) \bar{t} H$$

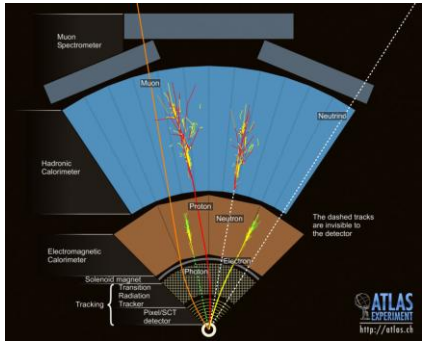
SM: $g_t = 1 \qquad \alpha_t = 0$











Analysis

$$f\left(x_{\text{parton}}|x_{\text{reco}}\right) = ?$$

Symmetries imprinted

Fast evaluation

Easily reusable

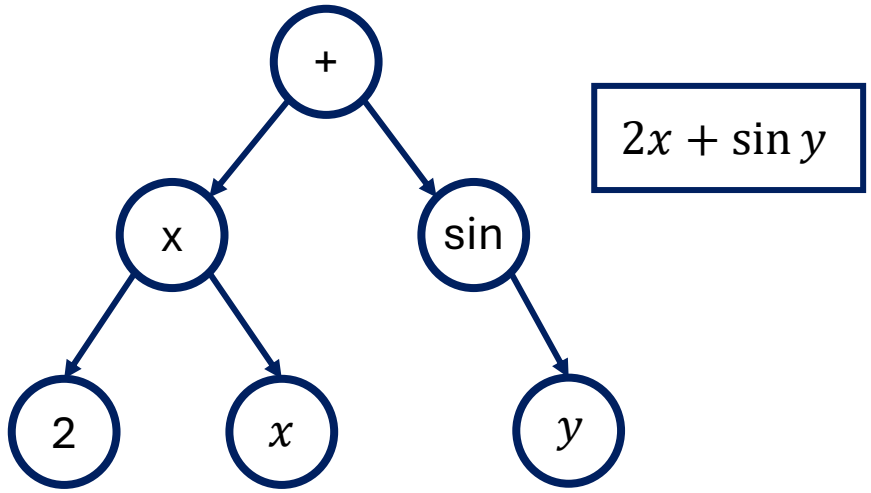
Maximal interpretability

Symbolic Regression: Equation discovery from data

PySR (Genetic algorithm)

[M. Cranmer '23]

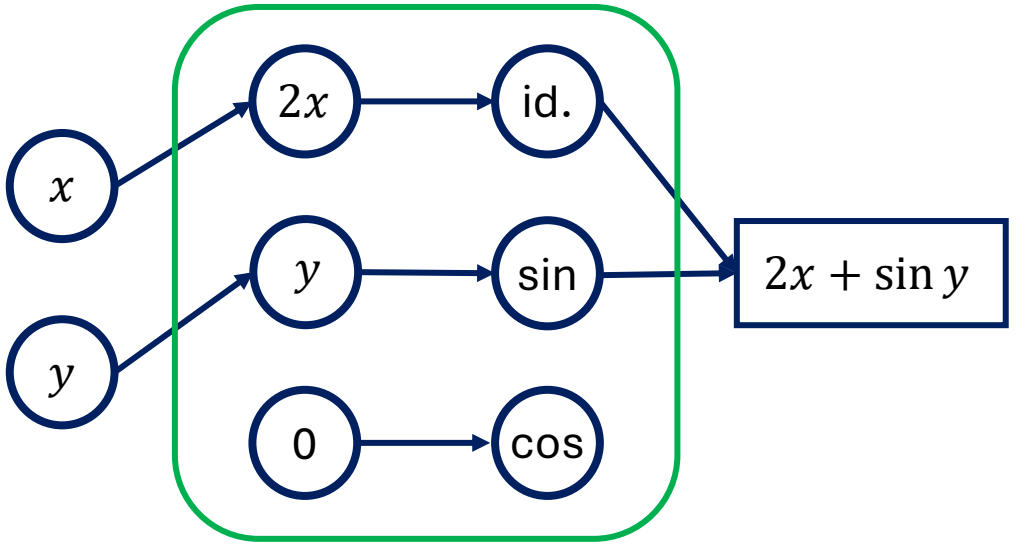
- Easy equation → complex equation
- Based on binary tree
- “Survival of the fittest”



SymbolNet (Deep neural network)

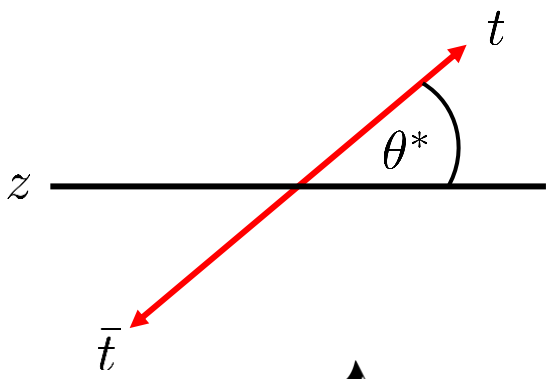
[Tsoi, Loncar, Dasu, Harris '24]

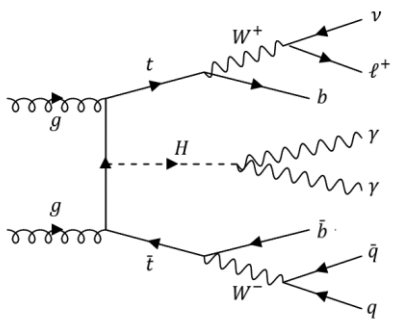
- Complex equation → easy equation
- Based on symbolic layers
- Backpropagation

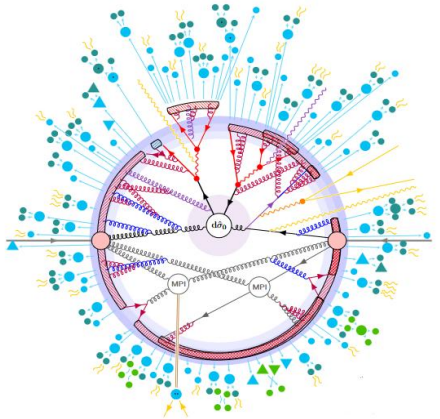


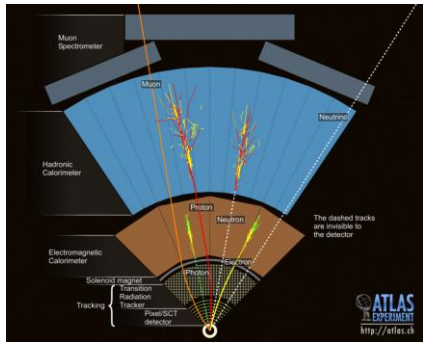
Goal: Reconstruct Collins-Soper angle for $t\bar{t}H$

[J. C. Collins, D. E. Soper '77]









Scenario 1 • • • Scenario 6

- Define benchmark scenarios with increasing complexity
- Scenario 1: $t\bar{t}$-frame variables, full neutrino information, no shower or detector effects
- Scenario 6: lab-frame variables, only E_T^{miss} information, an additional QCD jet, variables are smeared according to detector resolution

Scenario 1:

Equation for $\cos\theta^*$ is known:

$$\cos\theta^* = \frac{\sum p_{z,i}^{t\bar{t}}}{\sum \|p_i^{t\bar{t}}\|}$$

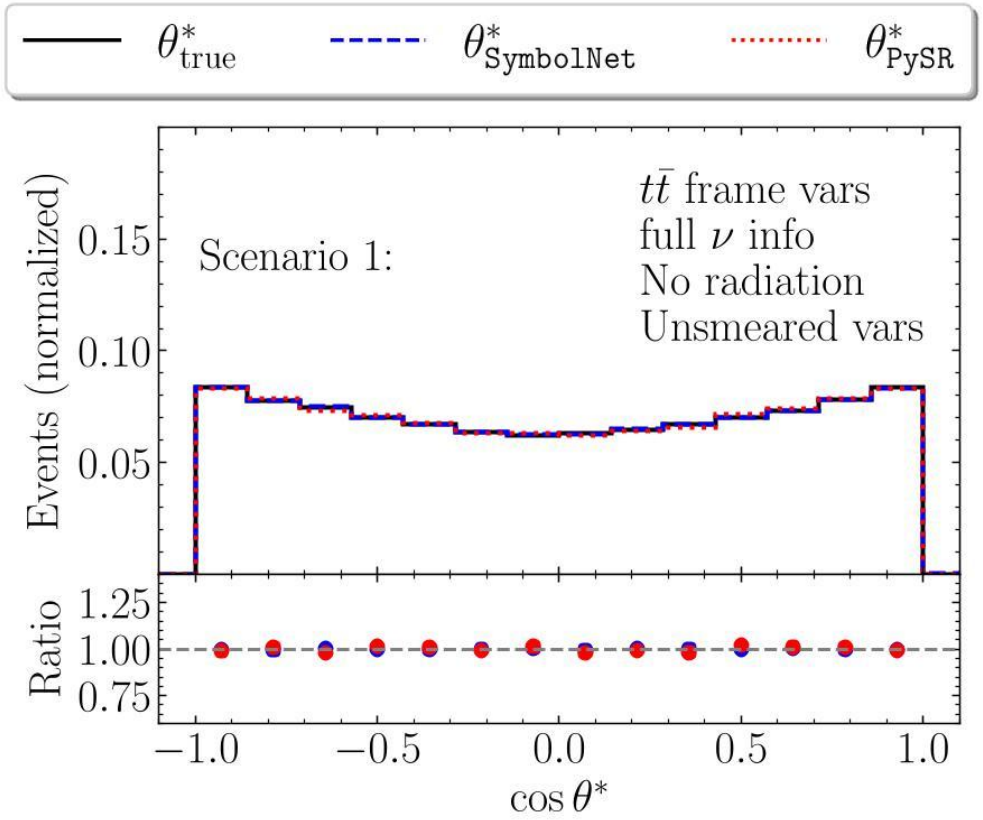
Both methods reconstruct it well

$$\cos\theta^*_{\mathrm{PySR}} = \frac{p_{z,b}+p_{z,\bar{l}}+p_{z,\nu}}{\|p_b+p_{\bar{l}}+p_\nu\|_3}$$

$$\cos\theta^*_{\mathrm{SymbolNet}} = \frac{p_{z,b}+p_{z,\bar{l}}+p_{z,\nu}+p_{z,\bar{b}}+p_{z,q}+p_{z,\bar{q}}}{\|p_b+p_{\bar{l}}+p_\nu+p_{z,\bar{b}}+p_{z,q}+p_{z,\bar{q}}\|_3}$$

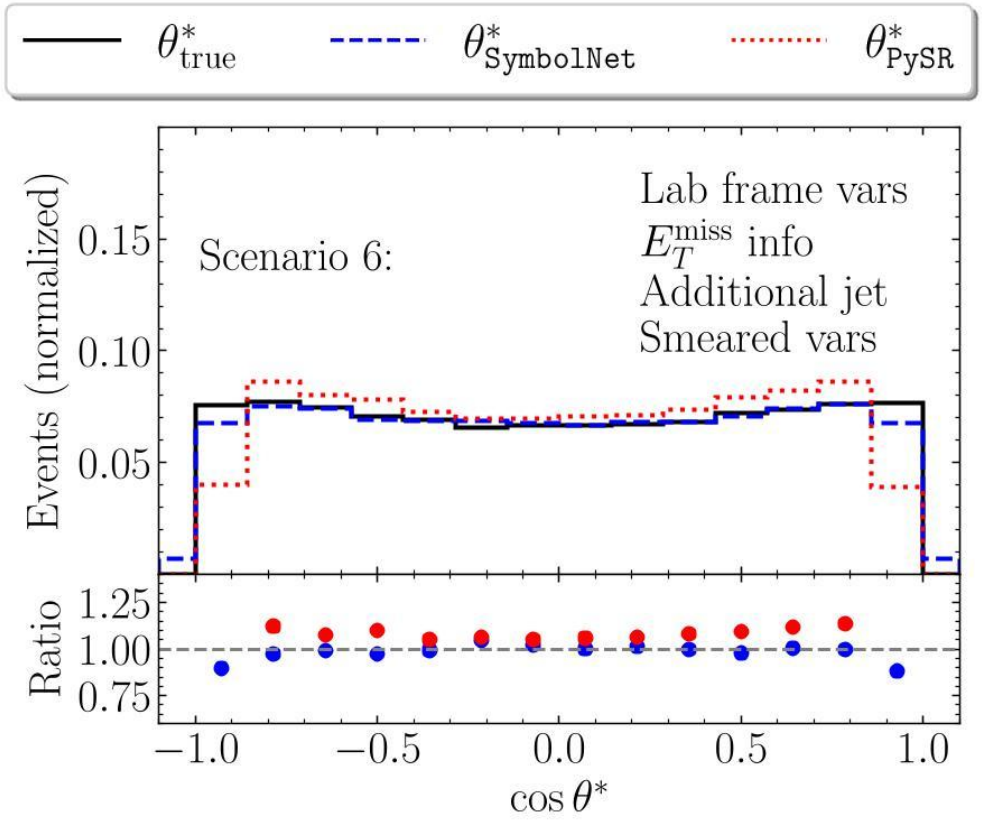
Leptonically decaying top

Hadronically decaying top



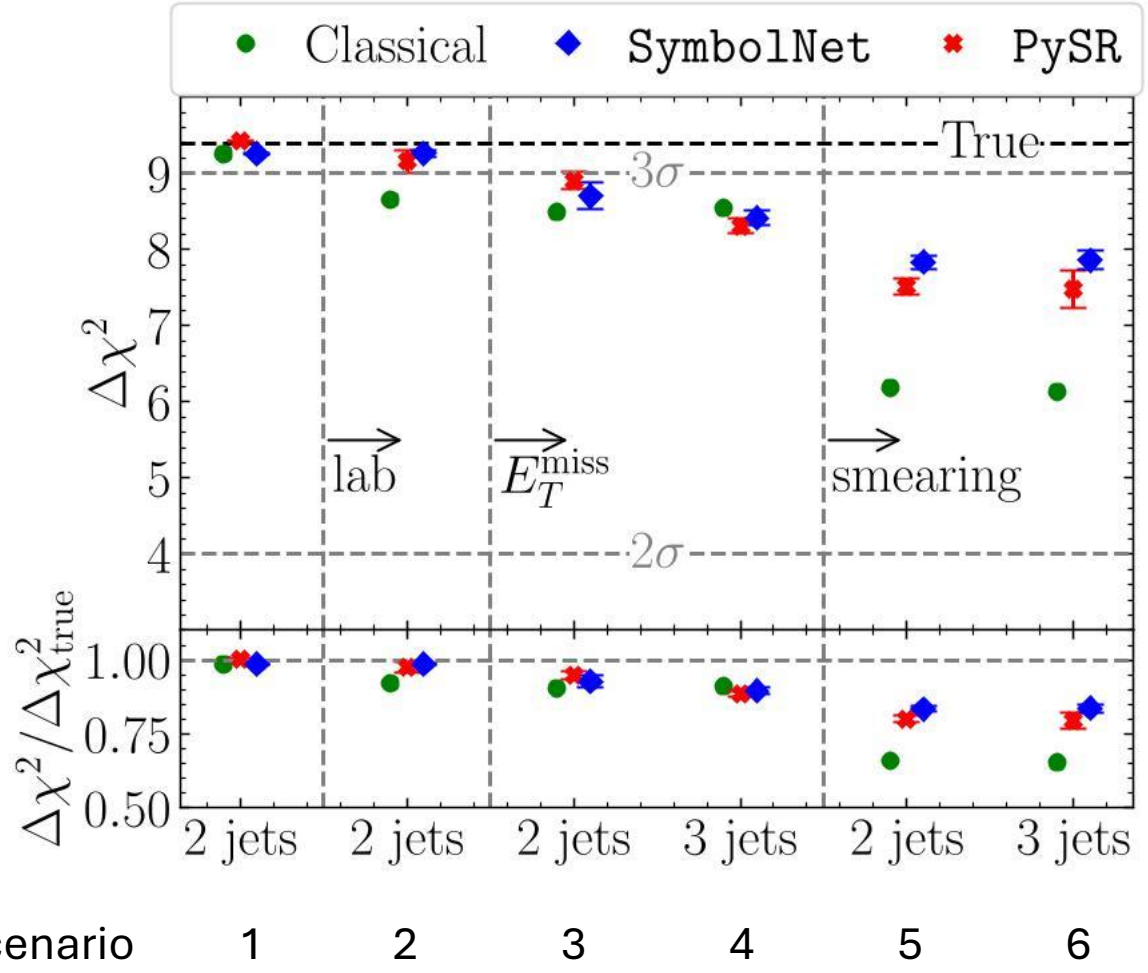
Scenario 6:

- Full information not available at detector level
- Center of the $\cos\theta^*$ is reconstructed well by both algorithms
- SymbolNet performs slightly better than PySR
- But the SymbolNet equation is also much more complicated



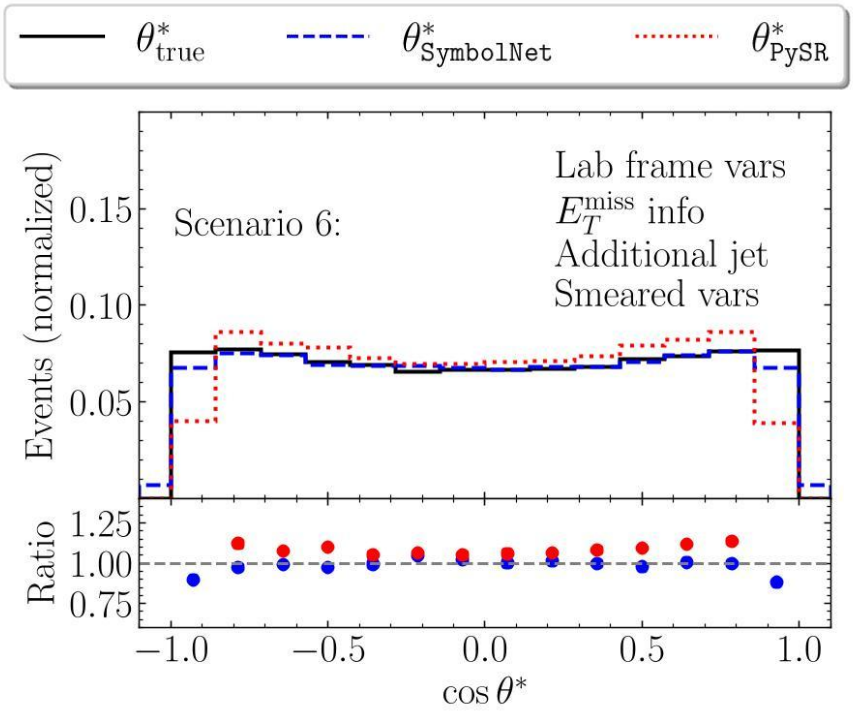
Significances to exclude $\alpha_t = 45°$ with a SM measurement at $\mathcal{L} = 300\text{fb}^{-1}$

- „Classical“ method: Reconstruct top quark via mass requirements
- SR approaches outperform classical reconstruction
- SymbolNet performs slightly better in complex scenarios
- Around 80% of CP information is retained by SR



Conclusions / Outlook

- Can find an equation to reconstruct a parton-level variable using detector-level information
- Retain around 80% of CP information
- SR algorithms outperform classical reconstruction method



⇒ Many potential applications, also outside of physics

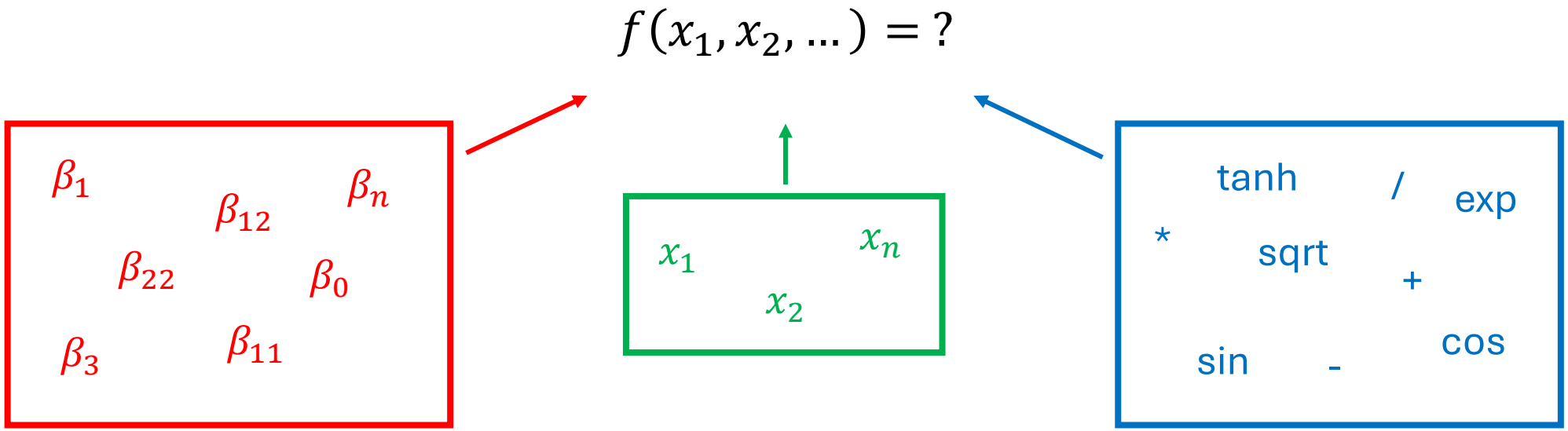
⇒ Further improvements might be possible from adding scientific knowledge

[Shojaee, Meidani, Gupta, Farimani, Reddy '24]

Backup

Default regression: Provide equation, estimate parameters

Symbolic regression: Estimate equation from parameters, variables, operators



General tradeoff: Goodness of fit ⟷ complexity of equation

Very (NP) hard task ⇒ Tools needed

PySR (Genetic algorithm):

[M. Cranmer '23]

Updates to populations via inner loop:

1. Select a tree to mutate determined via its fitness

2. Select a mutation to apply:

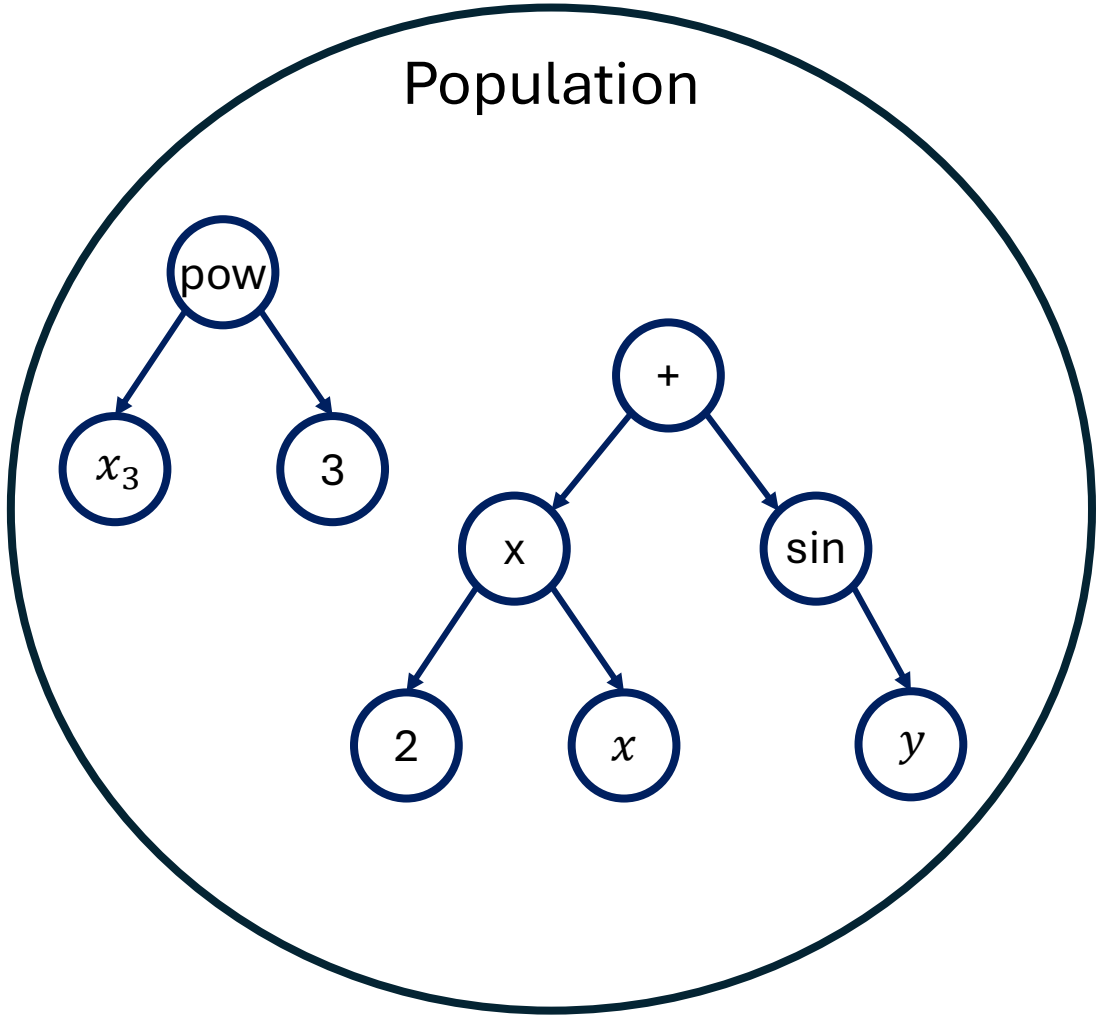
   Operator mutation

   Crossover

   Simplification

   Constant optimization

3. Replace the least fit member of the population by the mutated tree



PySR (Genetic algorithm):

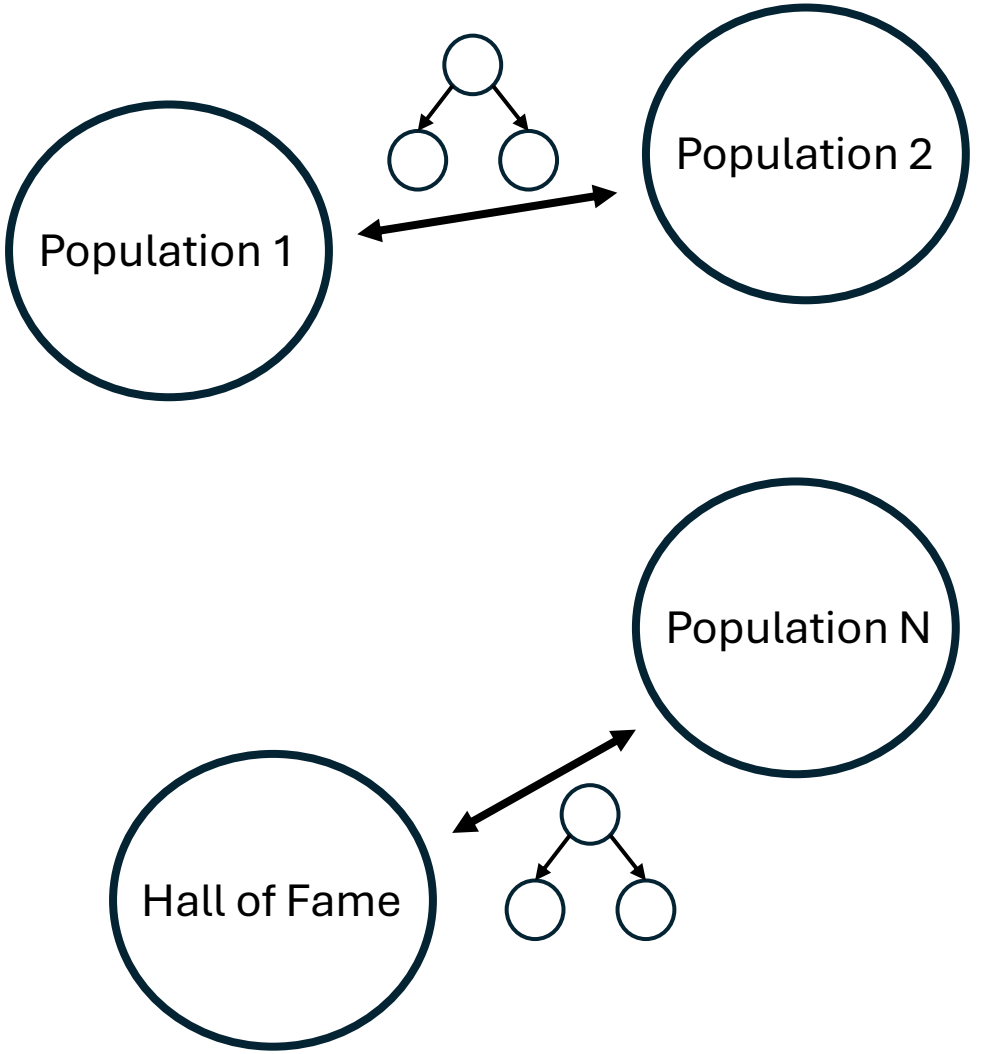
[M. Cranmer '23]

Migration of trees via outer loop:

1. Multiple individual populations evolve independently

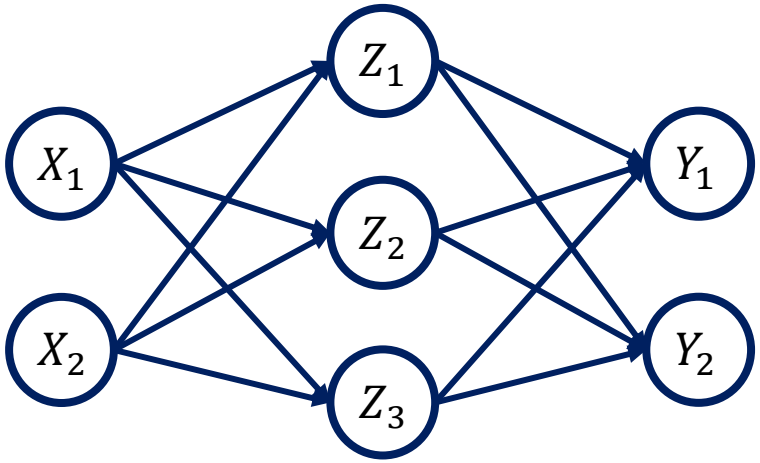
2. After a fixed number of mutations, trees migrate between populations

3. The trees with the highest fitness are stored in the hall of fame



SymbolNet (DNN):

[Tsoi, Loncar, Dasu, Harris '24]

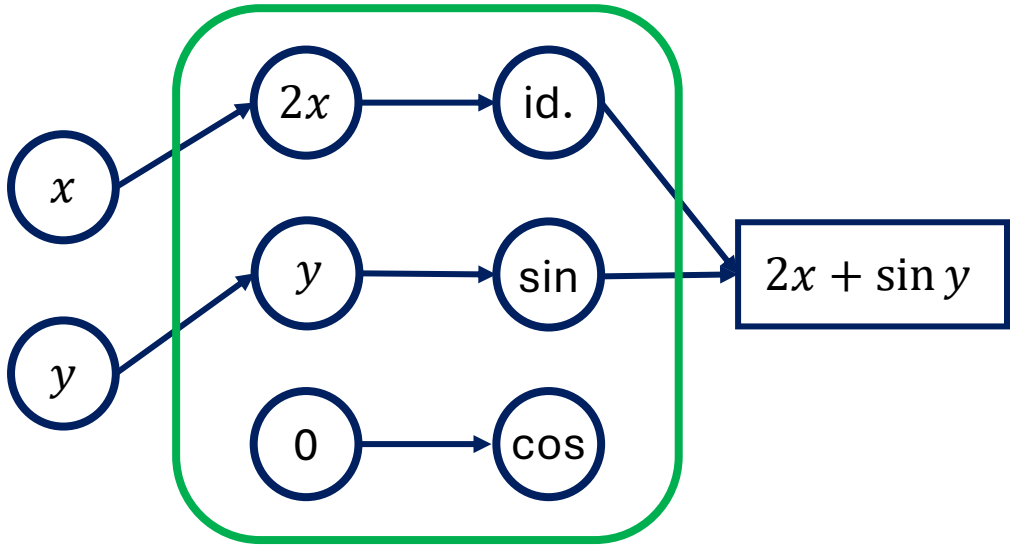


Symbolic layers:

- Linear combination of input nodes
- Unary and binary operators are applied

Updates via backpropagation:

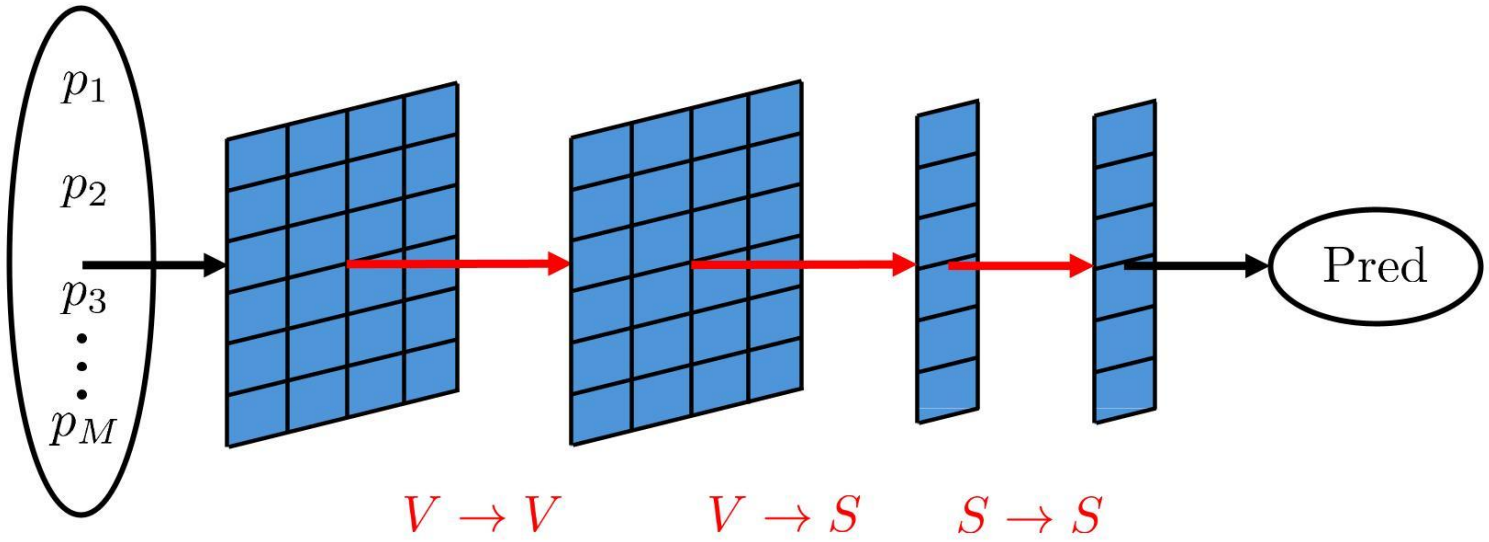
- Usual tuning of weights
- Removal of connections between nodes and operators
- Hard equation → easy equation



Vectorized SymbolNet:

[Bahl, Fuchs, Menen, Plehn '25]

Modification of SymbolNet to process 4-vectors:

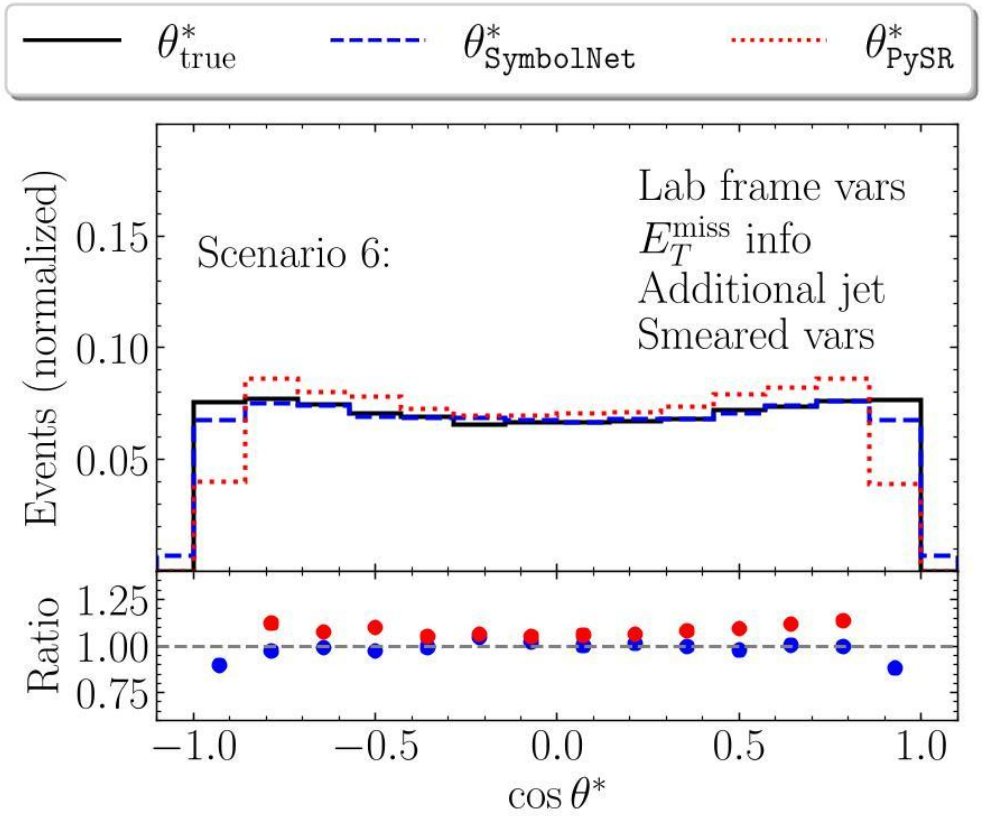


$V \to V$: Operators that preserve vector dimension (addition, Lorentz boost, ...)
$V \to S$: Operators that collapse vector dimension (norm, scalar product, ...)
$S \to S$: Default scalar operators

Scenario 6:

$$\cos\theta^*_{\texttt{PySR}} = \sin\left[\left(1.114p_{z,b} + 2.143p_{z,\bar{l}} - 0.858p_{z,\bar{b}} - 0.426p_{z,q} - 1.088p_{z,\bar{q}}\right)\Big/ \left(E_T^{\text{miss}}E_q + E_b + E_{\bar{b}} + 1.85E_{\bar{l}} + E_{\bar{q}} - \left(p_{x,b} + p_{x,\bar{l}}\right)\left(p_{x,\bar{b}} + p_{x,q} + p_{x,\bar{q}}\right) - 0.863 - \frac{0.205p_{z,q}}{\sqrt{E_q}}\right)\right]$$

$$\begin{aligned}\cos\theta^*_{\texttt{SymbolNet}} =& 0.971\Big(\text{boost}\Big[-1.521p_b - 3.066p_{\bar{l}} + 1.421p_{\bar{b}} + 2.218p_q + 2.043p_{\bar{q}}\Big| \\ & - 2.845p_b - 4.559p_{\bar{l}} - 2.575p_{\bar{b}} - 2.753p_q - 2.676p_{\bar{q}}\Big] \\ & - 0.256\tanh\left(0.646p_{\bar{b}} - 0.634p_{\bar{l}} + 0.648p_q\right)\Big)\Big|_z \Big/ \\ & \Big(\Big\|\text{boost}\Big[-1.521p_b - 3.066p_{\bar{l}} + 1.421p_{\bar{b}} + 2.218p_q + 2.043p_{\bar{q}}\Big| \\ & - 2.845p_b - 4.559p_{\bar{l}} - 2.575p_{\bar{b}} - 2.753p_q - 2.676p_{\bar{q}}\Big]\Big\|_3 \\ & + 0.774\Big\langle 0.071\,\text{boost}\Big[-1.521p_b - 3.066p_{\bar{l}} + 1.421p_{\bar{b}} + 2.218p_q + 2.043p_{\bar{q}}\Big| \\ & - 2.845p_b - 4.559p_{\bar{l}} - 2.575p_{\bar{b}} - 2.753p_q - 2.676p_{\bar{q}}\Big] \\ & + \tanh\left(0.646p_{\bar{b}} - 0.634p_{\bar{l}} + 0.648p_q\right)\times \\ & \tanh\left(0.646p_{\bar{b}} - 0.634p_{\bar{l}} + 0.648p_q\right)\Big\rangle_3 - 0.223\Big)\end{aligned}$$



Significances to exclude different α_t with a SM measurement at $\mathcal{L} = 300\text{fb}^{-1}$

- Total exclusion power heavily decreases towards $\alpha_t = 0°$
- Relative amount of CP information compared to parton-level drops for small α_t
- SR approaches outperform classical reconstruction for all α_t

Scenario 6: