



Symbolic Extraction of Non-Perturbative TMD from Drell-Yan Data

Cole Granger • Cristiano Fanelli

William & Mary — School of Computing, Data Sciences and Physics

with the MAP Collaboration — Bacchetta, Bertone, Bissolotti, Cerutti, Radici, Rodini, Rossi



CAREER



WILLIAM
& MARY
CHARTERED 1693

DATA SCIENCE

The result, up front

An analytic non-perturbative TMD extracted directly from data

A neural network fit to global Drell–Yan data, distilled by symbolic regression into a compact closed-form expression.

9

free numerical constants

1.040

χ^2/ndf over 482 data
points

31

total expression
complexity

Comparable to the parent neural-network fit ($\chi^2/\text{ndf} = 0.941$) – at a tiny fraction of the parameters.

A challenging problem

TMDs, the non-perturbative problem, and a tension between flexibility and interpretability



Fitting f_{NP} : flexibility vs. interpretability

Traditional parametrizations

Gaussian / exponential forms with low-dimensional kinematic dependence. Practical and widely used — but they impose functional assumptions that can bias the extraction.

Neural-network fits

Flexible; they reduce model dependence. But the fit is high-dimensional and hard to interpret or compare directly with traditional analyses.

Can we combine the flexibility of a network with the interpretability of a formula?

Symbolic regression as a bridge

1

FIT

Train a factorized neural network directly on global Drell–Yan data at N³LL accuracy.

2

DISTILL

Apply symbolic regression (PySR) to each network component to obtain compact analytic expressions.

3

SELECT

Choose the final formula from the Pareto front balancing experimental χ^2 and complexity

What is distillation?

This is a technique in ML that refers to using a larger, higher complexity model as a teacher for a smaller, lower complexity model to save compute.

Symbolic Regression in Hadron Physics and Beyond

SR has already recovered interpretable structures across several areas:

- GPDs from lattice + phenomenology (Dotson et al., EXCLAIM Coll.);
- Precision LHC pheno with rediscovery of known QED angular distributions (Bendavid et al)
- Fragmentation forms from SIDIS (Makke et al.)

positions SR — paired with their new ECC method — as a practical, interpretable tool for extracting physical laws, and a step toward SR-assisted nucleon tomography

A. Dotson et al., "Generalized Parton Distributions from Symbolic Regression," arXiv:2504.13289 [hep-ph] (2025).

M. Morales-Alvarado, D. Conde, J. Bendavid, V. Sanz, M. Ubiali, "Symbolic Regression for Precision LHC Physics," NeurIPS 2024 (ML & Physical Sciences), arXiv:2412.07839 [hep-ph].

J. Bendavid, D. Conde, M. Morales-Alvarado, V. Sanz, M. Ubiali, "Angular Coefficients from Interpretable Machine Learning with Symbolic Regression," JHEP 02 (2026) 081 [arXiv:2508.00989 [hep-ph]].

N. Makke, S. Chawla, "Inferring Interpretable Models of Fragmentation Functions using Symbolic Regression," Mach. Learn. Sci. Tech. 6, 025003 (2025) [arXiv:2501.07123 [hep-ph]].

Symbolic Regression in Hadron Physics and Beyond

SR has already recovered interpretable structures across several areas:

- GPDs from lattice + phenomenology (Dotson et al., EXCLAIM Coll.);
- Precision LHC pheno with rediscovery of known QED angular distributions (Bendavid et al)
- Fragmentation forms from SIDIS (Makke et al.)

A. Dotson et al., "Generalized Parton Distributions from Symbolic Regression," arXiv:2504.13289 [hep-ph] (2025).

M. Morales-Alvarado, D. Conde, J. Bendavid, V. Sanz, M. Ubiali, "Symbolic Regression for Precision LHC Physics," NeurIPS 2024 (ML & Physical Sciences), arXiv:2412.07839 [hep-ph].

J. Bendavid, D. Conde, M. Morales-Alvarado, V. Sanz, M. Ubiali, "Angular Coefficients from Interpretable Machine Learning with Symbolic Regression," JHEP 02 (2026) 081 [arXiv:2508.00989 [hep-ph]].

N. Makke, S. Chawla, "Inferring Interpretable Models of Fragmentation Functions using Symbolic Regression," Mach. Learn. Sci. Tech. 6, 025003 (2025) [arXiv:2501.07123 [hep-ph]].

SR has not been used to extract non-perturbative TMD directly from data

*Other techniques

- Z. Kang, AI-assisted modeling and Bayesian inference of unpolarized quark transverse momentum distributions from Drell-Yan data (2026)

Method

From TMD factorization to a factorized network to a distilled formula



A factorized, distillable ansatz

- f_{NP} is parametrized as

$$f_{\text{NP}}(x, b_T, \zeta) = \exp \left[-g_2^2(b_T) b_T^2 \ln(\zeta/Q_0^2) \right] \cdot R(x, b_T)$$

where

$$R(x, b_T) = \exp \left[h_x(x) h_b(b_T) + h_{xb}(x, b_T) \right]$$

- **Why factorize?** Symbolic regression scales poorly with dimensionality — a full 2D fit yields expressions too complex to distill (no candidate reached $\chi^2/\text{Ndat} < 1.3$).
- Decompose into two 1D pieces $h_x(x)$, $h_b(b_T)$ plus a cross term $h_{xb}(x, b_T)$ that captures x - b_T correlation.
- Boundary conditions $h_b(0)=0$, $h_{xb}(x,0)=0$ are enforced by construction $\Rightarrow$ **$R(x,0)=1$ exactly.**

Keeping h_{xb} “honest”

- L1 sparsity prior
- Capacity cap (4 neurons)
- Subset-balanced χ^2

A small network, trained on the observable

- Each component (h_x, h_b, g_2, h_{xb}) is a shallow MLP with only 4 or 8 neurons – **$O(100)$ parameters total**.
- Trained directly against the experimental χ^2 via a precomputed **convolution grid**: quadrature weights, momentum fractions and scales stored per data bin.

The NNs learn only the non-perturbative function. All perturbative QCD ingredients are frozen into a precomputed NangaParbat convolution skeleton, allowing efficient training with PyTorch while preserving the full MAP theory calculation. [Phys. Rev. Lett. 135, 021904 (2025)]

- A balanced loss across 59 experiment-specific subsets stops low-statistics sets from being sacrificed.
- Overfitting guards: tiny network, Pareto-knee selection, and symbolic compression.

$O(100) \rightarrow 9$

network parameters distilled to
analytic constants

... while the global χ^2 changes from
0.941 (NN) to 1.040 (SR)

Total Loss

$$\mathcal{L} = \mathcal{L}_{\chi^2}^{\text{bal}} + \lambda_{L_1} \|\theta_{h_{xb}}\|_1 + \lambda_{L_2} \|\theta_{\text{shape}}\|^2$$

loss term that pushes every subset χ^2 proxy value towards global mean

$$\mathcal{L}_{\chi^2}^{\text{bal}} = \frac{1}{N_{\text{exp}}} \sum_s w_s \frac{\chi_s^2}{N_s} + \lambda_{\text{bal}} \frac{1}{N_{\text{exp}}} \sum_s \left(\frac{\chi_s^2}{N_s} - \langle \chi^2 / N \rangle \right)^2$$

L2 weight decay on the parameters of the remaining shape networks (h_x, h_b, g_2), discouraging arbitrarily large weights without strongly constraining the fit.

L1 sparsity penalty applied exclusively to the hxb network parameters

The global Drell-Yan dataset

Source	Experiments	Ndat
Fermilab	E288, E605, E772	233
Tevatron	CDF, D0	71
RHIC	STAR	7
LHC	ATLAS, CMS, LHCb	171
Total	—	482

482 kinematic bins • 9 experiments • 59 subsets

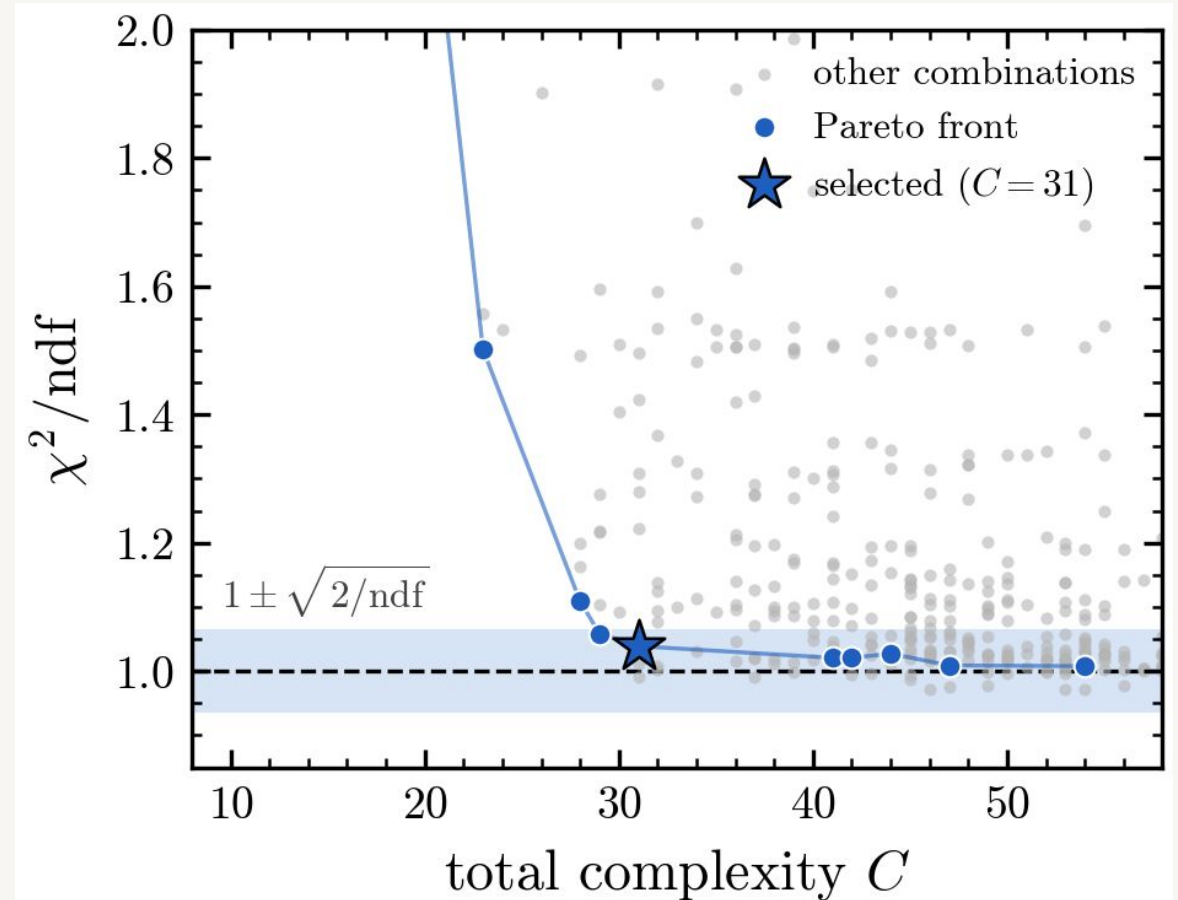
Uniform selections:

- $q_T/Q < 0.2$ — ensure TMD factorization applies
- Υ -resonance veto (9–11 GeV)

Discovering the formula: PySR + Pareto selection

- PySR evolves populations of candidate expressions (crossover, mutation, simplification), returning a **Pareto front**: complexity vs. accuracy.
- SR loss is MSE vs a sample grid from the NN
- SR is restricted to $b_T < 2.5 \text{ GeV}^{-1}$ — beyond which TMD evolution leaves the data essentially unable to constrain f_{NP}
- Per-component candidates are combined; each full f_{NP} is evaluated on the real data.

Selection: minimum complexity with reduced χ^2 compatible with 1, worst experimental subset $\chi^2 < 3$ for points on the Pareto front → **complexity 31**.



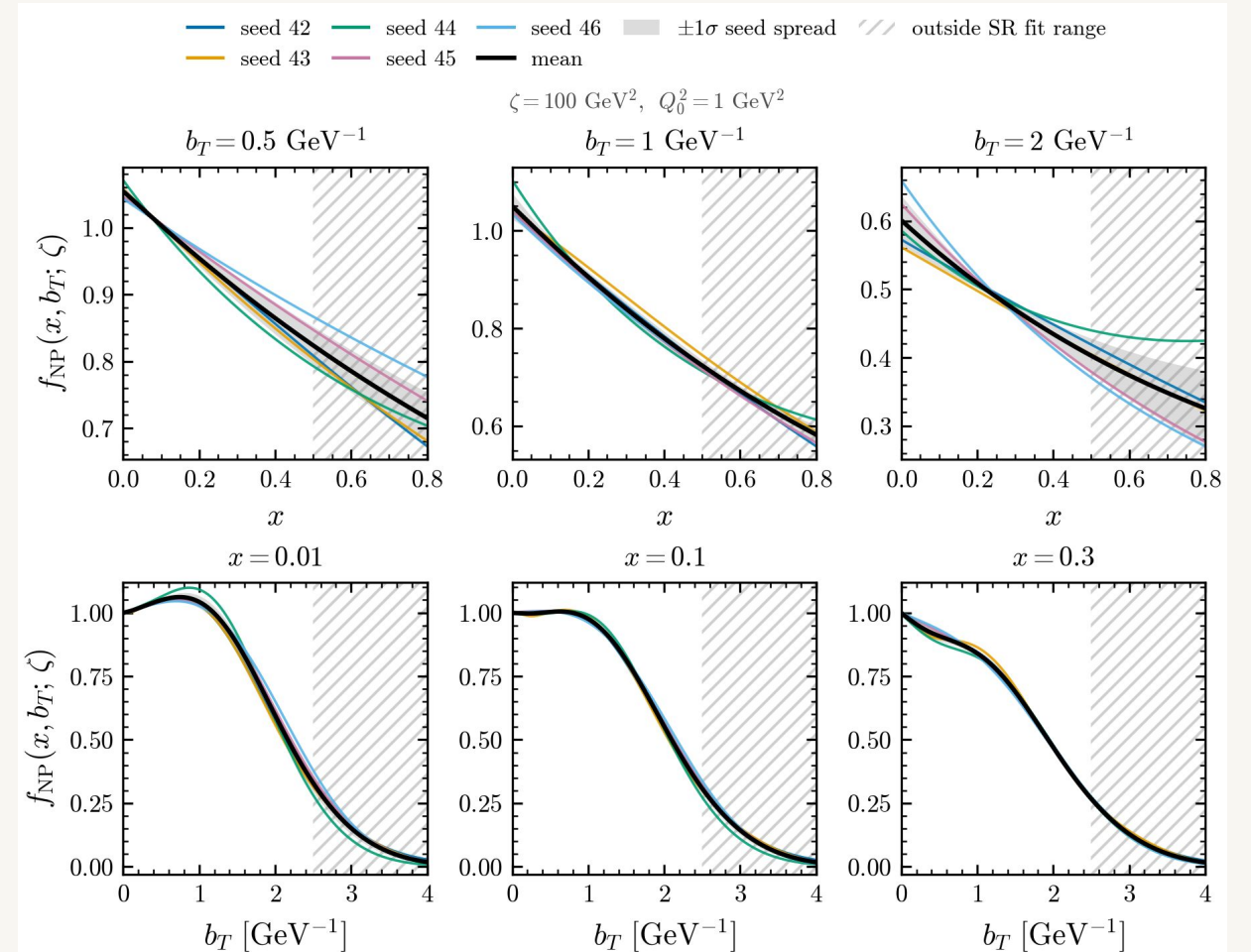
Results

The formula, its fit quality, and what it means



Robustness across network seeds

- The pipeline was run over multiple independent NN seeds (different random initializations).
- Narrow spread $\Rightarrow$ the recovered functional form is stable; wide spread $\Rightarrow$ genuinely distinct solutions of comparable quality.
- The band widens at large b_T ($b_T \gtrsim 4 \text{ GeV}^{-1}$): there the χ^2 gradient is Sudakov-suppressed, so the network is poorly constrained.



The SR-driven formula

$$f_{NP}(\mathbf{x}, b_T, \zeta) = \exp[-g_2^2(b_T) b_T^2 \ln(\zeta/Q_0^2)] \cdot \exp[h_x(\mathbf{x}) h_b(b_T) + h_{xb}(\mathbf{x}, b_T)]$$

9 constants • complexity 31 • $\chi^2/\text{ndf} = 1.040$

$$\mathbf{h}_x(\mathbf{x}) \quad h_x(x) = -0.854 - 1.45 x$$

$$\mathbf{h}_b(b_T) \quad h_b(b_T) = \frac{b_T}{s} (0.919 - 0.368 \sqrt{b_T/s})$$

$$\mathbf{h}_{xb}(\mathbf{x}, b_T) \quad h_{xb}(x, b_T) = \frac{b_T}{s} (0.766 - 0.0567 (b_T/s - x)^2)$$

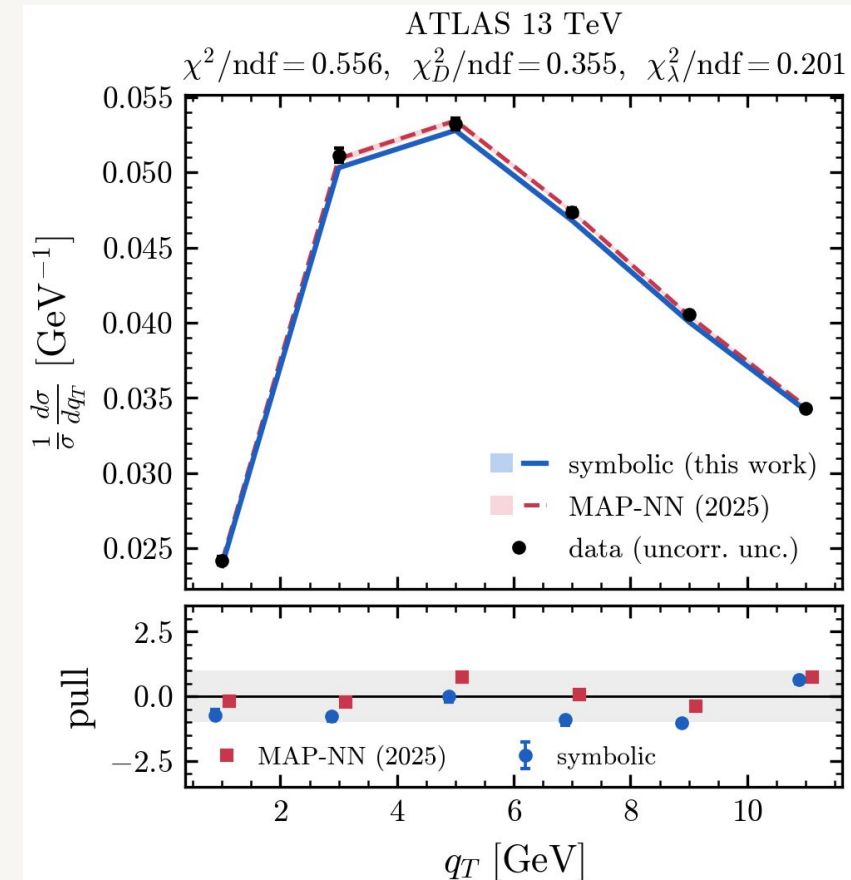
$$\mathbf{g}_2(b_T) \quad g_2(b_T) = 0.169 + 0.0328 \sqrt{b_T/s + 0.636}$$

Here s is a scale factor of 1 GeV^{-1} to ensure all components are dimensionless

The formula describes the data

Global $\chi^2/\text{ndf} = 1.040 \pm 0.065$ (473 dof) — compatible with 1, and close to the parent NN (0.941).

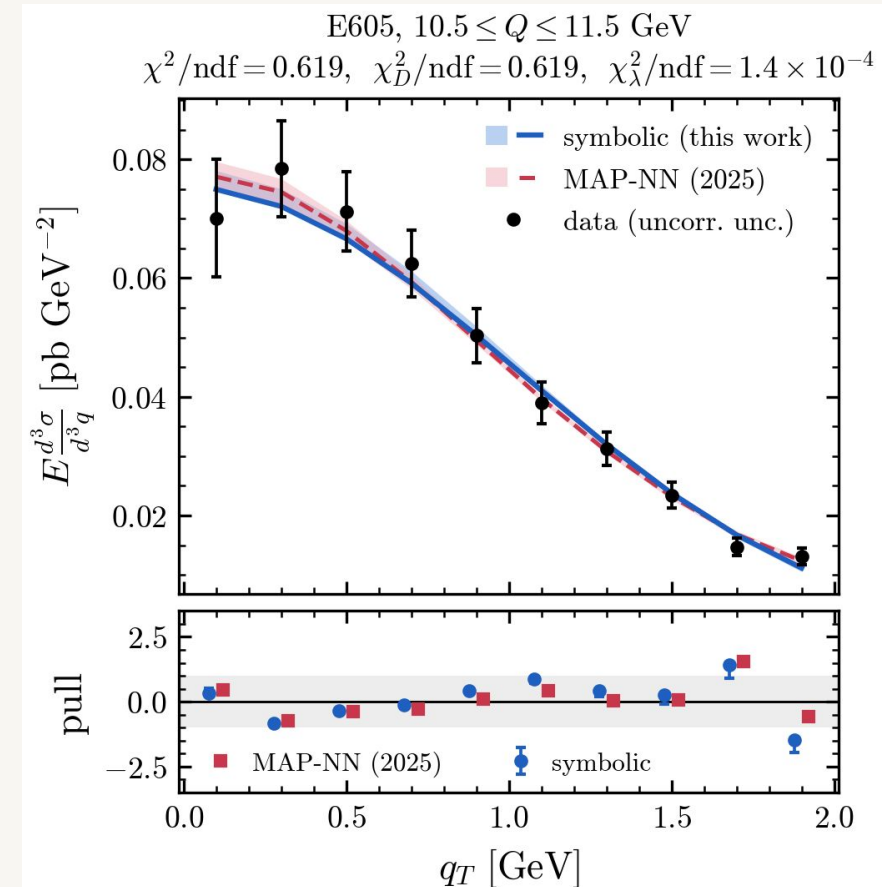
Experiment	Ndat	NN	Symbolic
Fermilab	233	1.057	1.179
RHIC	7	1.070	1.101
Tevatron	71	0.803	0.864
LHCb	21	1.083	1.074
CMS	78	0.420	0.429
ATLAS	72	1.213	1.274
Total	482	0.941	1.040



The formula describes the data

Good fit overall; tensions in low- Q fixed-target and central-rapidity collider bins

Experiment	Ndat	NN	Symbolic
Fermilab	233	1.057	1.179
RHIC	7	1.070	1.101
Tevatron	71	0.803	0.864
LHCb	21	1.083	1.074
CMS	78	0.420	0.429
ATLAS	72	1.213	1.274
Total	482	0.941	1.040



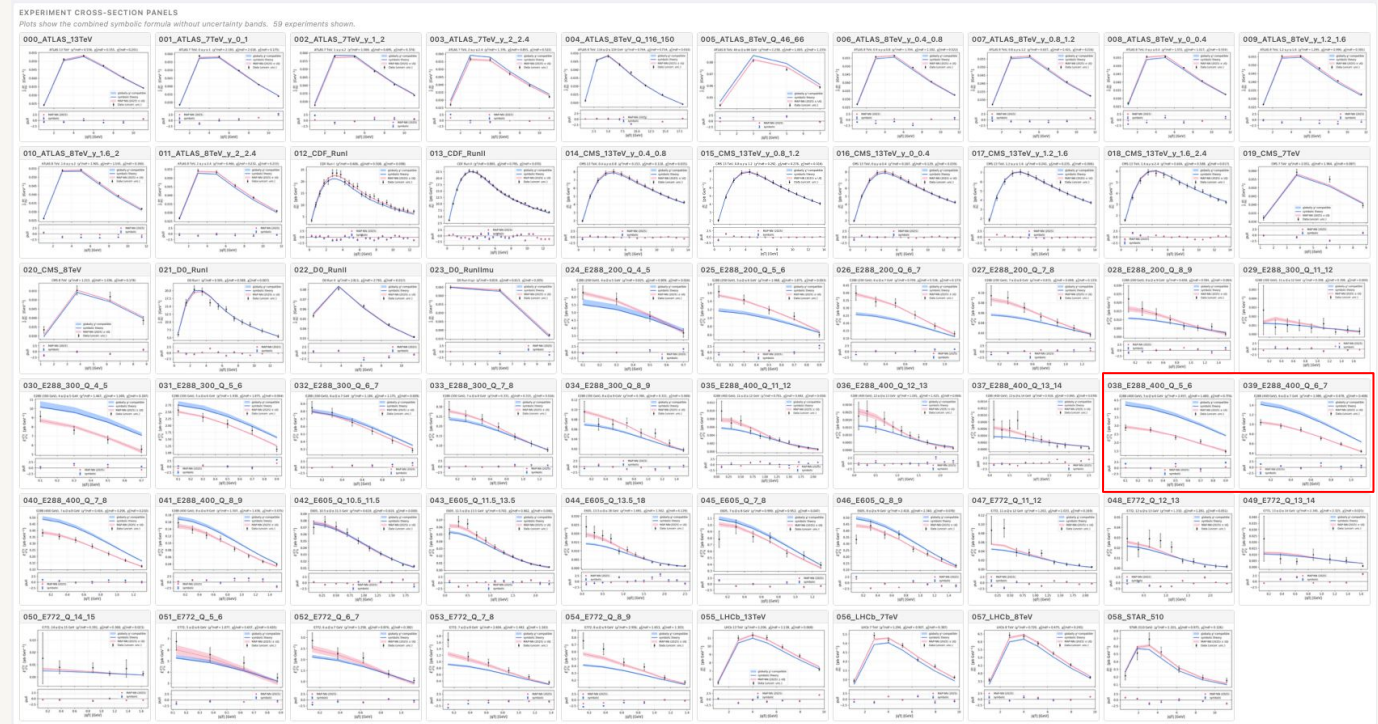
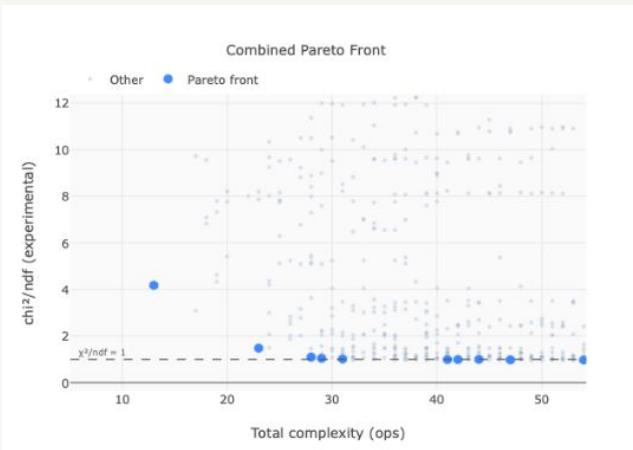
The “Pareto Explorer”

https://wmdatophys.github.io/Results/Symbolic_TMD_PDF/pareto_explorer_global_maxexp_front.html

Choose a point in the Pareto front

→ Display fit to experimental cross-sections

Tension in low-Q
fixed target



Tensions concentrate in two regions: low-Q fixed-target (E772, E605) and central-rapidity collider bins (ATLAS 7 TeV $|\eta| < 1$, D0 Run II). All other subsets fall below $\chi^2 \approx 2.5$.

What the components mean

$h_x(x)$

A simple linear function, negative across all x — a smooth modulation of the b_T -space suppression amplitude.

$h_b(b_T)$

A $\sqrt{b_T}$ -corrected linear ramp $\times b_T$; vanishes at $b_T = 0$, peaks ≈ 0.85 near $b_T \approx 2.77 \text{ GeV}^{-1}$.

$h_{xb}(x, b_T)$

Contributes at leading order. At $(x, b_T) = (0.1, 1)$: $h_{xb} = +0.72$ vs. backbone $h_x * h_b = -0.55$ — the cross term partially cancels the factorized suppression. Its leading piece ($0.766 b_T$) is separable; the genuine correlation is the $(b_T - x)^2$ modulation.

$g_2(b_T)$

Its small- b_T value ($\approx 0.20 \text{ GeV}$; replicas give 0.205 ± 0.034) matches prior global fits. The mild rise ($0.195 \rightarrow 0.227$) is not resolved by the data.

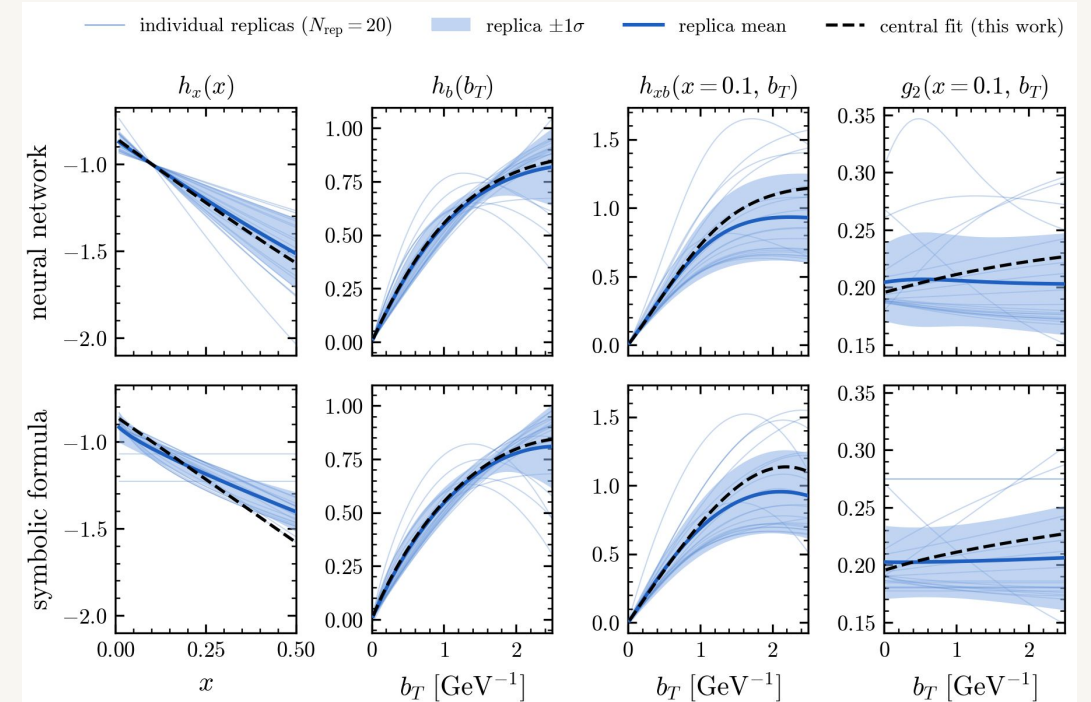
Data support a mild x - b_T correlation on top of a separable backbone.

Uncertainty Quantification

- Each constant suggested by PySR is treated as a free parameter
- We first vary these parameters independently to find a minimum and maximum value with which the global χ^2 is compatible with 1, and the worst experimental χ^2 is < 3
- From this hypercube, we sample parameters jointly, and keep samples with the same criteria
- This captures constant ranges consistent with global $\chi^2 \approx 1$. Experimental fluctuations are propagated separately through MC replicas (next slide); theory uncertainties are not yet included.

MC Replica Fits

- 20 pseudo-data replicas from the **full experimental covariance**, including correlated systematics: $d \rightarrow d + Lz$, $C = LL^T$
- Each replica: retrain the network, rerun PySR, pick the formula closest to complexity 31
- h_x and h_b are stable to **7%** and **15%** over the data-constrained region
- h_{xb} is **non-zero and positive in 20/20 replicas**
- **14 of the 20** replica formulas keep an explicit x - b_T coupling



Replica results

- $h_{xb}(0.1, 1 \text{ GeV}^{-1}) = 0.69 \pm 0.21$ (reported fit: 0.73)
- $g_2(b_T \rightarrow 0) = 0.205 \pm 0.034 \text{ GeV}$
- The b_T -slope of g_2 is **not resolved**: 7 replicas rise and 13 fall

Is h_{xb} imposed by the regularization?

Refits of the reported configuration, one setting changed per run

Change	NN χ^2/N	Worst subset	$h_{xb} / h_x h_b$	Formula C	Formula χ^2/ndf
Paper	0.94	2.89	+0.73/-0.56	31	1.040
$\lambda_{L1} = 10^{-5}$	0.93	2.93	+0.74/-0.55	35	1.005
$\lambda_{L1} = 10^{-3}$	0.94	2.89	+0.73/-0.56	47	0.989
h_{xb} width 2	0.91	2.98	+0.57/-0.33	32	1.011
h_{xb} width 8	0.91	2.95	+0.84/-0.58	28	1.053
$w_{E772} = 2$	0.93	2.95	+0.75/-0.56	33	1.006
$w_{E772} = 1$	0.94	2.98	+0.73/-0.56	56	0.978

The L1 penalty is inert. At $\lambda_{L1} = 10^{-4}$ it adds about 0.0007 to a loss of order 1.

Width moves the split, not the fit. It shifts how the exponent is divided between h_{xb} and $h_x h_b$. The global χ^2 is unchanged.

The E772 weight only speeds convergence. At $w = 2$ a similar modulation survives: $b(0.769 - 0.064(b - \sqrt{x})^2)$

Paper: $\lambda_{L1} = 10^{-4}$, 4 neurons, $w_{E772} = 3$. $h_{xb} / h_x h_b$ is evaluated at $x = 0.1$, $b_T = 1 \text{ GeV}^{-1}$. NN χ^2/N and worst subset χ^2/N_s are from the paper (Table III).

The cross term makes the formula possible

Remove h_{xb}

The network fit degrades: $\chi^2/N = 1.26$, worst subset 4.97 (vs 0.94 and 2.89 with h_{xb})

No compact formula can be found: the best is $\chi^2/\text{ndf} \geq 1.27$

Constant g_2 (as in MAP 2025)

The network fit is poorer: $\chi^2/N = 1.11$, worst subset 3.34 (vs 0.94 and 2.89)

No compact formula can be found: the best is $\chi^2/\text{ndf} \geq 1.14$, from enumerating all 5,567 combinations

Without b_T -dependence in g_2 , h_b has to absorb structure that PySR can only reproduce with high-order polynomials

h_{xb} is required for a compact formula that fits the data. A b_T -dependent g_2 improves the fit, but its slope is not statistically significant.

Limitations & what's next

Caveats

- Result is conditioned on the factorized enveloping ansatz; alternatives may give different components at similar fit quality.
- How the exponent is split between h_{xb} and $h_x h_b$ depends on network width and training seed. The fit quality and the sum $h_x h_b + h_{xb}$ do not.
- The b_T -dependence of g_2 is chosen because it makes the fit distillable. The data do not resolve its slope.
- f_{NP} is flavor-universal
- Constrained only for $b_T < 2.5 \text{ GeV}^{-1}$; the h_b node (crossing 0) at $b_T \approx 6.2 \text{ GeV}^{-1}$ is an extrapolation.

What's next

- Flavor dependence with semi-inclusive DIS data.
- Extension to polarized TMDs.

Conclusions - Data-Driven Symbolic Extraction

- A flexible neural-network representation, distilled by symbolic regression, yields a compact analytic parametrization without prescribing its component functions.
- The extracted non-perturbative TMD achieves a $\chi^2/\text{ndf} = 1.040$ with only 9 constants, close to its parent network fit (0.941)
- A cross term is required to make the fit distillable. The data fix $h_x h_b + h_{xb}$ but not how it splits; the non-separable $(b_T - x)^2$ part stays in h_{xb} in every split.

Interactive Pareto explorer on GitHub · Supported by NSF CAREER #2443510
wmdataphys.github.io

