Sample Complexity of Scientific Discovery: PAC Learnability of Compositional Function Trees

Şuayp Talha Kocabay [1 2] **Talha Rüzgar Akkuş** [1] **Kerem Yalçın** [1]

Abstract

Scientific discovery via symbolic regression is often viewed as statistically and computationally intractable because the hypothesis space of expressions grows combinatorially with depth. This paper revisits the statistical side through the lens of PAC learning, focusing on compositional function trees built from a finite vocabulary of smooth operators (e.g., $\{+, \times, \sin, \exp\}$ and affine maps). We prove that the relevant generalization quantity—Rademacher complexity, hence the excess risk—does *not* necessarily blow up exponentially with the number of distinct symbolic structures, but is controlled by (i) the depth d and (ii) the Lipschitz constants of the base operators along the composed computation graph. Concretely, under mild Lipschitz conditions on operators and bounded affine leaves, a finite-union bound over a vocabulary of size $K{=}|\mathcal{H}_{\text{base}}|$ together with Maurer-type vector contraction yields $\mathfrak{R}_n(\mathcal{H}^d_{\text{comp}}) \leq (Kb\sqrt{2}\,L)^{d-1}\mathfrak{R}_n(\mathcal{H}^1_{\text{comp}})$ with arity bound b; corresponding high-probability risk bounds scale as $\mathcal{O}(L^d/\sqrt{n})$ when $K, b{=}O(1)$ and $\mathfrak{R}_n(\mathcal{H}^1_{\text{comp}}){=}O(n^{-1/2})$. We complement the theory with a modular codebase that trains differentiable operator trees (not MLPs) on synthetic "physics-like" targets of controlled depth and shows that the empirical generalization gap correlates positively with the predicted complexity term $(\widehat{L}^d)/\sqrt{n}$.

1. Introduction

Scientific machine learning frequently requires *interpretable* functional relationships from small datasets, not only accurate predictions. While deep neural networks can approximate complex functions, they typically do not yield concise symbolic laws, raising concerns in the sciences where interpretability, extrapolation, and falsifiability are central. Symbolic regression (SR) addresses this by searching over expressions built from a vocabulary of primitives, and has led to influential systems and benchmarks (Koza, 1992; Schmidt & Lipson, 2009a;b; Udrescu & Tegmark, 2020; Cranmer, 2023; La Cava et al., 2021).

However, SR is often described as "unlearnable" because the number of candidate expressions grows exponentially with depth, and finding the best expression is NP-hard in many formulations. This view conflates *computational* hardness with *statistical* learnability. PAC learning theory (Valiant, 1984) separates these issues: a hypothesis class can be PAC learnable (small sample complexity) even if optimizing over it is computationally challenging.

We formalize this separation for a compositional hypothesis space motivated by scientific discovery. Our core message is that statistical generalization of compositional operator trees is governed by compositional *stability* (Lipschitzness) rather than by the raw count of symbolic structures. Technically, we control generalization via Rademacher complexity (Bartlett & Mendelson, 2003; Mohri et al., 2018), extending classical contraction arguments (Ledoux & Talagrand, 1991) to tree-structured compositions.

Discrete search procedures incur computational costs proportional to the number of structures explored (Koza, 1992; Petersen et al., 2021), whereas uniform convergence controls how reliably empirical risk ranks hypotheses once attention is restricted to a class $\mathcal{H}^d_{\text{comp}}$ *and its continuous parameters*. In practice SR couples both: a learner proposes syntax trees and fits coefficients (often nonlinearly); our bounds clarify what statistical scaling one should expect from that nonlinear estimation step when stability holds across primitives.

Contributions.

- We define a compositional function-tree hypothesis space $\mathcal{H}^d_{\text{comp}}$ capturing symbolic expressions of depth d built from a base vocabulary $\mathcal{H}_{\text{base}}$.

[1]Independent Researcher [2]TÜBİTAK Science High School, Türkiye. Correspondence to: Şuayp Talha Kocabay <suayptalhakocabay.28@tubitakfenlisesi.k12.tr>, Talha Rüzgar Akkuş <talharuzgarakkus@gmail.com>, Kerem Yalçın <keremyalcin.28@tubitakfenlisesi.k12.tr>.

Accepted to the 2nd Workshop on Compositional Learning at ICML 2026, Seoul, South Korea. *Copyright 2026 by the author(s).*

- Under mild Lipschitz assumptions on base operators and bounded affine leaves (Theorem 3.2), we bound $\mathfrak{R}_n(\mathcal{H}^d_{\text{comp}})$ by $(Kb\sqrt{2}\,L)^{d-1}\mathfrak{R}_n(\mathcal{H}^1_{\text{comp}})$ for $K{=}|\mathcal{H}_{\text{base}}|$ (Equation (14) and Theorem 4.6), yielding PAC-style risk bounds scaling as $\mathcal{O}(L^d/\sqrt{n})$ when $K, b{=}O(1)$.

- We provide a clean PyTorch implementation of differentiable compositional trees and controlled-depth synthetic "physics" datasets, and empirically verify that the generalization gap tracks $(\widehat{L}^d)/\sqrt{n}$.

Theory versus empirical proxies. Our guarantees concern uniform convergence over $\mathcal{H}^d_{\text{comp}}$ and invoke population quantities (d, L) (localized when nonlinear operators such as $\exp$ appear). The experiments substitute d by the fixed structural depth of the fitted tree and L by a finite-batch estimator $\widehat{L}$. Thus alignment between observed gaps and $(\widehat{L}^d)/\sqrt{n}$ should be read as evidence that compositional stability predicts qualitative scaling rather than as identification of sharp numerical constants.

2. Related Work

Symbolic regression and scientific discovery. Classic SR is commonly instantiated via genetic programming (Koza, 1992) or program synthesis-style search, with modern toolchains emphasizing physics heuristics (Udrescu & Tegmark, 2020), large-scale benchmarking (La Cava et al., 2021), and differentiable relaxations (Cranmer et al., 2020; Cranmer, 2023). Recent work has also explored deep reinforcement learning and risk-seeking objectives for expression discovery (Petersen et al., 2021), as well as broader surveys of SR methodology and applications (Makke & Chawla, 2024). In parallel, sparse discovery methods such as SINDy (Brunton et al., 2016) target dynamical systems by searching over a pre-specified library with sparsity penalties. Our focus differs: we do not propose a new SR search algorithm; instead, we give a statistical learnability analysis of *compositional* operator classes and validate the predicted scaling empirically.

Generalization via complexity and stability. Rademacher complexity and related uniform convergence tools are standard in learning theory (Vapnik, 1998; Bartlett & Mendelson, 2003; Shalev-Shwartz & Ben-David, 2014; Mohri et al., 2018). For deep neural networks, stability and Lipschitz-like quantities (e.g., spectral norms) play central roles in generalization bounds (Bartlett et al., 2017). Our analysis is closest in spirit to compositional complexity controls (e.g., chain rules for Gaussian/Rademacher processes and vector contraction inequalities) (Ledoux & Talagrand, 1991; Talagrand, 2014; van der Vaart & Wellner, 1996; Maurer, 2016a;b), specialized here to operator trees commonly used in scientific modeling.

Depth versus width. Much neural-network theory emphasizes architectural depth jointly with width, norms, or margin quantities (Bartlett et al., 2017; Neyshabur et al., 2017). Here depth enters explicitly through repeated composition of a *fixed vocabulary* of interpretable operators and produces multiplicative Lipschitz accumulation along trees rather than implicit coupling through parameters. This distinction clarifies why symbolic/compositional models admit bounds stated directly in terms of (d, L) without appealing to implicit layers or infinite-width limits.

Refined contraction and localization. Recent refinements of Talagrand-type contraction for structured vector-valued maps sharpen dependence on norms and layer geometry beyond the textbook lemma (Truong, 2022). Parallel localization tools—covering-number-based local Rademacher complexities (Lei et al., 2016) and offset/local frameworks that yield estimator-aware excess-risk bounds (Kanade et al., 2024)—provide routes toward sharper or hypothesis-dependent guarantees than the uniform bounds emphasized here.

Discrete syntax graphs versus continuous estimation. Much SR methodology emphasizes discrete structure discovery combined with numerical tuning (Koza, 1992; Udrescu & Tegmark, 2020; Cranmer, 2023). From a learning-theoretic viewpoint it is helpful to separate (i) the outer discrete optimization problem—often akin to program synthesis (Petersen et al., 2021)—from (ii) the inner regression problem induced once an architecture family is fixed. Our complexity estimates speak primarily to (ii): depth enters compositionally through gradients/Lipschitz maps rather than through enumeration over exponentially many graphs.

3. Preliminaries

3.1. Learning setup

Let $\mathcal{X} \subset \mathbb{R}^p$ and $\mathcal{Y} \subset \mathbb{R}$. Given i.i.d. data $S = \{(x_i, y_i)\}_{i=1}^n$ drawn from an unknown distribution $\mathcal{D}$ on $\mathcal{X} \times \mathcal{Y}$, we consider squared loss $\ell(f(x), y) = (f(x) - y)^2$. For a hypothesis $f \in \mathcal{H}$, define true risk $R(f) = \mathbb{E}_{(x,y)\sim\mathcal{D}}[\ell(f(x), y)]$ and empirical risk $\widehat{R}_S(f) = \frac{1}{n}\sum_{i=1}^n \ell(f(x_i), y_i)$.

3.2. Trees, depth, and branching

To make the role of *depth* explicit, it is useful to describe a hypothesis $f \in \mathcal{H}^d_{\text{comp}}$ as a rooted operator tree whose internal nodes are operators from $\mathcal{H}_{\text{base}}$ and whose leaves

are either input coordinates or constants. Let b denote the maximum arity (branching factor) of operators (for our vocabulary, $b \leq 2$). We distinguish:

- **Depth** d: maximum number of operator applications along any root-to-leaf path.
- **Size** s: number of internal nodes (operators) in the tree.

While the number of distinct trees of depth d can grow super-exponentially in d, our bounds depend on d (and weakly on b) through stability, not enumeration.

3.3. Base operator class and compositional trees

Let $\mathcal{H}_{\text{base}}$ be a finite set of (possibly parameterized) operators acting on scalars or tuples of scalars. In this paper we focus on a small differentiable vocabulary typical of SR and physics-inspired modeling:

$$\mathcal{H}_{\text{base}} \subset \Big\{\text{add, mult, sin, cos, exp, affine}\Big\}. \quad (1)$$

We define a *compositional hypothesis space of depth* d recursively. For simplicity, assume a binary tree arity (unary operators can be modeled by taking one child to be a constant function).

Definition 3.1 (Compositional hypothesis space)**.** Let $\mathcal{H}^1_{\text{comp}}$ be the set of base (non-composed) hypotheses. For $d \geq 2$, define

$$\mathcal{H}^d_{\text{comp}} = \Big\{h \circ (g_1, g_2) \;:\; h \in \mathcal{H}_{\text{base}},\; g_1, g_2 \in \mathcal{H}^{d-1}_{\text{comp}}\Big\}. \quad (2)$$

Here (g_1, g_2) denotes the pair-valued map $x \mapsto (g_1(x), g_2(x))$ and h acts on $\mathbb{R}^2$ (or $\mathbb{R}$ in the unary case). The recursion induces a computation tree with depth d.

Parameters. Leaves labeled affine introduce slope/intercept parameters; compositions propagate gradients through operators exactly as in standard automatic differentiation.

Assumption 3.2 (Bounded affine leaves)**.** Assume $\mathcal{X}$ has bounded coordinates: $\|x\|_\infty \leq X_{\max}$ almost surely. Every affine primitive $x \mapsto \theta^\top x + \theta_0$ used at a leaf satisfies $\|\theta\|_2 \leq W_{\max}$ and $|\theta_0| \leq B_{\text{leaf}}$.

Under Assumption 3.2, leaves have uniformly bounded outputs on $\mathcal{X}$, and the depth-1 class $\mathcal{H}^1_{\text{comp}}$ is a finite union of bounded-parameter linear predictors; classical Rademacher bounds then yield $\mathfrak{R}_n(\mathcal{H}^1_{\text{comp}}) = \mathcal{O}(W_{\max} X_{\max} \sqrt{p/n})$ up to logarithmic factors (Bartlett & Mendelson, 2003; Mohri et al., 2018).

3.4. Local Lipschitz constants and bounded ranges

Since scientific data are bounded by measurement ranges, we work with *local* Lipschitz constants on the operator ranges induced by the data distribution and the hypothesis parameters. Formally, for a function $f : \mathcal{X} \to \mathbb{R}$, define its Lipschitz constant on $\mathcal{X}$ as

$$\text{Lip}_{\mathcal{X}}(f) := \sup_{x \neq x'} \frac{|f(x) - f(x')|}{\|x - x'\|_2}. \quad (3)$$

For operators like exp, $\text{Lip}_{\mathcal{X}}(\exp \circ f)$ is controlled by the range of $f(\mathcal{X})$ (cf. Theorem 4.3). This is standard in empirical process arguments that localize complexity to relevant regions (van der Vaart & Wellner, 1996; Koltchinskii & Panchenko, 2002).

3.5. Rademacher complexity

Given a sample $S = (x_i)_{i=1}^n$ and a real-valued function class $\mathcal{F}$, the empirical Rademacher complexity is

$$\mathfrak{R}_S(\mathcal{F}) = \mathbb{E}_\sigma\Big[\sup_{f \in \mathcal{F}} \frac{1}{n} \sum_{i=1}^n \sigma_i f(x_i)\Big], \quad (4)$$

where σ_i are i.i.d. Rademacher signs. We denote $\mathfrak{R}_n(\mathcal{F}) = \mathbb{E}_S[\mathfrak{R}_S(\mathcal{F})]$.

For a vector-valued class $\mathcal{G} \subset (\mathcal{X} \to \mathbb{R}^m)$, we use the *coordinatewise* empirical Rademacher complexity with *independent* signs per coordinate (Maurer, 2016b):

$$\mathfrak{R}^{(m)}_S(\mathcal{G}) = \mathbb{E}_\sigma\Big[\sup_{g \in \mathcal{G}} \frac{1}{n} \sum_{i=1}^n \sum_{j=1}^m \sigma_i^{(j)}\, g_j(x_i)\Big], \quad (5)$$

where $\sigma_i^{(1)}, \ldots, \sigma_i^{(m)}$ are independent Rademacher signs for each sample index i. When $m{=}2$ and $\mathcal{G} = \mathcal{F}_1 \times \mathcal{F}_2 = \{x \mapsto (f_1(x), f_2(x)) \;:\; f_1 \in \mathcal{F}_1, f_2 \in \mathcal{F}_2\}$, we write $\mathfrak{R}_S(\mathcal{F}_1 \times \mathcal{F}_2) := \mathfrak{R}^{(2)}_S(\mathcal{G})$.

Rademacher complexity yields uniform generalization bounds for Lipschitz losses (Bartlett & Mendelson, 2003; Mohri et al., 2018).

4. Main Theoretical Results

Our bounds formalize the intuition that *compositional trees generalize if each layer/operator is stable*. The key technical tool is a tree-structured contraction inequality, conceptually extending Ledoux–Talagrand contraction (Ledoux & Talagrand, 1991) and compositional Rademacher results (Maurer, 2016a) to our setting.

4.1. Lipschitz assumptions

Assumption 4.1 (Lipschitz operators)**.** Each base operator $h \in \mathcal{H}_{\text{base}}$ is (locally) L_h-Lipschitz on the relevant range

of its inputs with respect to the Euclidean norm. That is, for all admissible inputs u, v, $|h(u) - h(v)| \le L_h \|u - v\|_2$. Let $L = \max_{h \in \mathcal{H}_{\text{base}}} L_h$.

Remark 4.2. For operators such as exp, global Lipschitzness fails on $\mathbb{R}$. In scientific discovery, inputs are typically bounded by measurement ranges; our analysis is *local* to the operator ranges induced by the data distribution and the learned parameters. The experiments estimate $\widehat{L}$ data-dependently on the fitted map rather than certifying a sharp uniform Lipschitz constant over all of $\mathcal{H}^d_{\text{comp}}$. When L is measured relative to ranges induced by a specific hypothesis, contraction inequalities read hypothesis-dependent; uniform guarantees require either an envelope on admissible ranges (here enforced jointly by bounded outputs $|f| \le B$, bounded covariates, and Assumption 3.2) or localization along the data-dependent slice relevant to empirical risk minimization (Koltchinskii & Panchenko, 2002; Lei et al., 2016; Kanade et al., 2024).

4.2. A compositional Rademacher lemma

We first state a clean inequality that captures the multiplicative depth effect. We present a proof sketch highlighting the main steps; full details are in Section A.

Proposition 4.3 (Lipschitz calculus for common operators). *Let $f, g : \mathcal{X} \to \mathbb{R}$ be Lipschitz with constants $\text{Lip}(f), \text{Lip}(g)$ w.r.t. $\|\cdot\|_2$ on $\mathcal{X}$. Assume moreover that $|f(x)| \le B_f$ and $|g(x)| \le B_g$ for all $x \in \mathcal{X}$. Then on $\mathcal{X}$:*

$$\text{Lip}(f+g) \le \text{Lip}(f) + \text{Lip}(g), \tag{6}$$

$$\text{Lip}(f \cdot g) \le B_g \,\text{Lip}(f) + B_f \,\text{Lip}(g), \tag{7}$$

$$\text{Lip}(\sin \circ f) \le \text{Lip}(f), \qquad \text{Lip}(\cos \circ f) \le \text{Lip}(f), \tag{8}$$

$$\text{Lip}(\exp \circ f) \le \exp(\sup_{x \in \mathcal{X}} f(x)) \,\text{Lip}(f). \tag{9}$$

Proof. Each inequality follows from the mean value theorem and standard product/chain rules, taking suprema over $x \in \mathcal{X}$. □

Lemma 4.4 (Vector contraction (Maurer)). *Let $\mathcal{G} \subset (\mathcal{X} \to \mathbb{R}^m)$ and let $\Phi : \mathbb{R}^m \to \mathbb{R}$ be L-Lipschitz w.r.t. the Euclidean norm on $\mathbb{R}^m$ with $\Phi(0) = 0$. Define $\Phi \circ \mathcal{G} = \{\Phi \circ g : g \in \mathcal{G}\}$. Then for any sample S,*

$$\mathfrak{R}_S(\Phi \circ \mathcal{G}) \le \sqrt{2}\, L\, \mathfrak{R}^{(m)}_S(\mathcal{G}). \tag{10}$$

Proof sketch. This is Maurer's vector-contraction inequality (Maurer, 2016b), stated for the coordinatewise complexity (5). If $\Phi(0) \ne 0$, subtract the constant shift (Rademacher averages are unchanged by adding fixed offsets to $f(x_i)$). □

Lemma 4.5 (Product classes do not multiply complexity). *Let $\mathcal{F}_1, \mathcal{F}_2 \subset (\mathcal{X} \to \mathbb{R})$ and define the paired class $\mathcal{F}_1 \times \mathcal{F}_2 = \{x \mapsto (f_1(x), f_2(x)) : f_1 \in \mathcal{F}_1, f_2 \in \mathcal{F}_2\}$. With the convention $\mathfrak{R}_S(\mathcal{F}_1 \times \mathcal{F}_2) = \mathfrak{R}^{(2)}_S(\mathcal{F}_1 \times \mathcal{F}_2)$ from (5), for any sample S,*

$$\mathfrak{R}_S(\mathcal{F}_1 \times \mathcal{F}_2) \le \mathfrak{R}_S(\mathcal{F}_1) + \mathfrak{R}_S(\mathcal{F}_2). \tag{11}$$

Proof. Expand $\mathfrak{R}^{(2)}_S$ using independent signs $\sigma^{(1)}_i, \sigma^{(2)}_i$ and use $\sup_{f_1, f_2}(A(f_1) + B(f_2)) = \sup_{f_1} A(f_1) + \sup_{f_2} B(f_2)$. □

Lemma 4.6 (Union over a finite operator vocabulary). *Let $K := |\mathcal{H}_{\text{base}}|$ and write $\mathcal{H}_{\text{base}} = \{h_1, \ldots, h_K\}$. For each k, define $\mathcal{C}_k = \{h_k \circ (g_1, \ldots, g_b) : g_j \in \mathcal{H}^{d-1}_{\text{comp}}\}$. Then $\mathcal{H}^d_{\text{comp}} = \bigcup_{k=1}^K \mathcal{C}_k$ and, for every sample S,*

$$\mathfrak{R}_S(\mathcal{H}^d_{\text{comp}}) \le \sum_{k=1}^K \mathfrak{R}_S(\mathcal{C}_k) \le K \cdot \max_{k \in [K]} \mathfrak{R}_S(\mathcal{C}_k). \tag{12}$$

At each recursive composition level the same finite union appears; hence a factor K (not the number of distinct tree shapes) enters per depth level.

Proof. For fixed (σ_i, x_i) and any $f \in \bigcup_k \mathcal{C}_k$, the sample average is bounded by $\sum_{k=1}^K \sup_{g_1, \ldots, g_b \in \mathcal{H}^{d-1}_{\text{comp}}} \frac{1}{n} \sum_i \sigma_i h_k(g_1(x_i), \ldots, g_b(x_i))$; take expectations. □

Theorem 4.7 (Constant-explicit depth bound for binary trees). *Assume each internal operator is L-Lipschitz w.r.t. $\|\cdot\|_2$ on the relevant range (Assumption 4.1) and satisfies $\Phi(0) = 0$ after centering constants. Then for any sample S and depth $d \ge 1$,*

$$\mathfrak{R}_S(\mathcal{H}^d_{\text{comp}}) \le (K \cdot 2\sqrt{2}\, L)^{d-1}\, \mathfrak{R}_S(\mathcal{H}^1_{\text{comp}}), \tag{13}$$

where $K := |\mathcal{H}_{\text{base}}|$ is the (fixed) operator vocabulary size.

Proof sketch. By Lemma 4.6, $\mathfrak{R}_S(\mathcal{H}^d_{\text{comp}}) \le K \max_h \mathfrak{R}_S(\mathcal{C}_h)$. For fixed h, apply Lemma 4.4 with m=2 (cost $\sqrt{2}\, L_h$) and Lemma 4.5 on the paired child outputs (cost 2). Iterate over depth. □

Corollary 4.8 (Branching factor b). *If each internal node has arity at most b (i.e., it combines at most b children) and is L-Lipschitz w.r.t. the Euclidean norm on $\mathbb{R}^b$, then*

$$\mathfrak{R}_S(\mathcal{H}^d_{\text{comp}}) \le (K \cdot b\sqrt{2}\, L)^{d-1}\, \mathfrak{R}_S(\mathcal{H}^1_{\text{comp}}), \tag{14}$$

where $K := |\mathcal{H}_{\text{base}}|$.

Proof sketch. Apply Lemma 4.6 at each depth level, then Lemma 4.4 with m=b and the b-fold analogue of Lemma 4.5, $\mathfrak{R}_S(\times_{j=1}^b \mathcal{F}_j) \le \sum_{j=1}^b \mathfrak{R}_S(\mathcal{F}_j) \le b \max_j \mathfrak{R}_S(\mathcal{F}_j)$. □

4.3. PAC-style generalization bound

We translate Equation (14) into a risk bound. Let $\mathcal{F} = \mathcal{H}^d_{\text{comp}}$ and assume $f(x)$ and y are bounded in $[-B, B]$, which holds in our synthetic experiments by construction. For squared loss $\ell(\hat{y}, y) = (\hat{y} - y)^2$, the map $\hat{y} \mapsto \ell(\hat{y}, y)$ is $4B$-Lipschitz in $\hat{y}$ whenever $|\hat{y}|, |y| \le B$, and the loss is bounded by $4B^2$.

Theorem 4.9 (Generalization bound for scientific discovery). *Assume Theorems 3.2 and 4.1 and that outputs are bounded by $|f(x)| \le B$ and $|y| \le B$ almost surely for every $f \in \mathcal{H}^d_{\text{comp}}$. Then for any $\delta \in (0, 1)$, with probability at least $1 - \delta$ over $S \sim \mathcal{D}^n$, every $f \in \mathcal{H}^d_{\text{comp}}$ satisfies*

$$R(f) \le \widehat{R}_S(f) + \mathcal{O}\Big(\frac{B\,L^d}{\sqrt{n}}\Big) + B^2\sqrt{\frac{\log(1/\delta)}{2n}}. \tag{15}$$

In particular, holding L and d fixed, the excess risk decays as $1/\sqrt{n}$; holding n fixed, composition amplifies stability through $(Kb\sqrt{2}\,L)^{d-1}$ (Equation (14)) rather than through the cardinality of symbolic structures.

Exponent and arity bookkeeping. Uniform deviation bounds track terms of order $\text{Lip}(\ell)\,(Kb\sqrt{2}\,L)^{d-1}\mathfrak{R}_n(\mathcal{H}^1_{\text{comp}})$ with $K{=}|\mathcal{H}_{\text{base}}|$ and maximum arity b from Corollary 4.8. The displayed $\mathcal{O}(BL^d/\sqrt{n})$ adopts the usual big-O shorthand in which fixed vocabulary/arity factors (K^{d-1}, b^{d-1}), $\text{Lip}(\ell)$, the Maurer constant $\sqrt{2}$ (Maurer, 2016b), and $\mathfrak{R}_n(\mathcal{H}^1_{\text{comp}}) = \mathcal{O}(n^{-1/2})$ under Assumption 3.2 are absorbed.

Proof sketch. Standard Rademacher-based uniform convergence bounds for Lipschitz losses give $\sup_{f\in\mathcal{F}}(R(f) - \widehat{R}_S(f)) \lesssim \text{Lip}(\ell)\mathfrak{R}_n(\mathcal{F}) + B^2\sqrt{\log(1/\delta)/(2n)}$. Apply Corollary 4.8 to obtain $\mathfrak{R}_n(\mathcal{H}^d_{\text{comp}}) \le (Kb\sqrt{2}\,L)^{d-1}\mathfrak{R}_n(\mathcal{H}^1_{\text{comp}})$, then invoke Assumption 3.2 for $\mathfrak{R}_n(\mathcal{H}^1_{\text{comp}}) = \mathcal{O}(n^{-1/2})$ and absorb constants into Equation (15). □

Implications for symbolic regression. Theorem 4.9 suggests a *learnability certificate* for SR-style compositional models: if the operator vocabulary is stable on the relevant range (moderate L) and depth is modest, then small-n datasets can suffice for low excess risk. This does not remove the computational challenge of searching over structures, but it clarifies that statistical generalization need not be the bottleneck.

4.4. From complexity to sample complexity

We next translate the complexity statement into a *sample complexity* statement of the PAC form. For a bounded loss ℓ and function class $\mathcal{F}$, standard symmetrization and concentration yield bounds in terms of $\mathfrak{R}_S(\mathcal{F})$ (Bartlett & Mendelson, 2003; Mohri et al., 2018). We highlight two corollaries that make the dependence on d, L, and n explicit.

Corollary 4.10 (Excess risk scaling). *Under the conditions of Theorem 4.9, for any ERM $\widehat{f} \in \arg\min_{f\in\mathcal{H}^d_{\text{comp}}} \widehat{R}_S(f)$,*

$$\begin{aligned} R(\widehat{f}) - \inf_{f\in\mathcal{H}^d_{\text{comp}}} R(f) \\ = \mathcal{O}\Big(\frac{B\,L^d}{\sqrt{n}}\Big) + B^2\sqrt{\frac{\log(1/\delta)}{2n}}, \end{aligned} \tag{16}$$

with probability at least $1 - \delta$.

Corollary 4.11 (PAC sample complexity for ϵ-excess risk). *Fix d and L. To guarantee $R(\widehat{f}) - \inf_{f\in\mathcal{H}^d_{\text{comp}}} R(f) \le \epsilon$ with probability at least $1 - \delta$, it suffices that*

$$n \gtrsim \frac{B^2L^{2d}}{\epsilon^2} + \frac{B^4}{\epsilon^2}\log\frac{1}{\delta}. \tag{17}$$

Interpretation. The key point is that depth affects sample complexity through multiplicative stability growth—equivalently $(Kb\sqrt{2}\,L)^{2(d-1)}$ when arity and vocabulary are tracked explicitly (Equation (14))—rather than through the (typically much larger) number of distinct symbolic trees of depth d. The shorthand L^{2d} in Corollary 4.11 treats bounded arity as $\mathcal{O}(1)$.

Cardinality versus complexity. Let $\mathcal{T}_d$ denote the finite collection of distinct *shapes* of binary trees of depth d. Uniform convergence requires controlling complexity over *the union* of realizable functions associated with each shape *after* accounting for continuous parameterizations at leaves. Naively counting $|\mathcal{T}_d|$ suggests exponential growth in discrete syntax; however, at each composition level Lemma 4.6 replaces the supremum over the finite vocabulary $\mathcal{H}_{\text{base}}$ by a sum over at most $K{=}|\mathcal{H}_{\text{base}}|$ fixed-operator subclasses, and our bounds isolate the contribution that arises purely from repeated composition along paths. Thus statistical complexity need not track worst-case enumeration when stability constants remain moderate—precisely the regime motivated by physics modeling where operators act locally smoothly on bounded domains.

4.5. A PAC-Bayes view: learning with structure priors

Rademacher bounds are uniform over $\mathcal{H}^d_{\text{comp}}$ and do not exploit *a priori* preferences for simpler trees. An alternative is PAC-Bayes analysis, where one posits a prior π over trees (e.g., favoring smaller depth/size via MDL) and controls the risk of a posterior ρ learned from data (McAllester, 1999; Seeger, 2002; Catoni, 2007; Alquier, 2024). For bounded

losses, a representative bound has the form

$$\mathbb{E}_{f\sim\rho}[R(f)] \le \mathbb{E}_{f\sim\rho}[\widehat{R}_S(f)] + \sqrt{\frac{\mathrm{KL}(\rho\|\pi) + \log\frac{2\sqrt{n}}{\delta}}{2(n-1)}}. \tag{18}$$

Combining such KL penalties with the contraction certificates above yields guarantees trading description length (via π) against multiplicative stability accumulation $(Kb\sqrt{2}\,L)^{d-1}$ on top of $\mathfrak{R}_n(\mathcal{H}^1_{\mathrm{comp}})$ (McAllester, 1999; Seeger, 2002).

Worked PAC-Bayes + stability sketch. Suppose π is a fixed MDL-style prior over tree shapes T (e.g., $\pi(T) \propto \exp(-\mathrm{code}(T))$) together with a Gaussian factor over leaf weights truncated to satisfy Assumption 3.2. Let ρ_T be any data-dependent posterior supported on trees of shape T. Combining Equation (18) with a union bound over shapes weighted by $\pi(T)$ yields an expectation bound under ρ_T whose slack adds $\sqrt{\mathrm{KL}(\rho_T\|\pi)/n}$ to the empirical risk; coupling this with the deterministic stability multiplier $(Kb\sqrt{2}\,L)^{d-1}$ from Equation (14) gives a qualitative trade-off between Occam penalties on syntax and Lipschitz-depth amplification along the chosen architecture.

Norm-based nets versus compositional primitives. Deep-network bounds often scale with parameter norms or Jacobians while treating activation composition implicitly (Bartlett et al., 2017; Neyshabur et al., 2017). Here each activation *is* an interpretable primitive with an explicit calculus (Theorem 4.3), which makes it natural to express guarantees directly through operator-level Lipschitz constants accumulated along tree edges rather than through aggregated weight matrices.

5. Experiments

5.1. Setup

We validate the qualitative prediction of Equation (15): the generalization gap should scale with $(L^d)/\sqrt{n}$. We implement: (i) a synthetic dataset generator for controlled-depth "physics" targets, (ii) a differentiable operator-tree model (no MLPs), (iii) a training loop and evaluation on a large test set, and (iv) logging and plotting of the gap vs. the theoretical term.

Isolation of statistical scaling. We deliberately train under *realizability*: each ground-truth $f^\star$ lies in the hypothesis family at the stated depth with matching operator vocabulary. This design separates statistical scaling from misspecification errors or discrete structure mismatch that dominate realistic SR pipelines (La Cava et al., 2021). Extensions to agnostic settings replace $\inf_{f\in\mathcal{H}^d_{\mathrm{comp}}} R(f)$ by an approximation error term while preserving the same variance-type complexity multiplier.

Data. Inputs are $x \sim \mathrm{Unif}([-1,1])$ with scalar outputs $y = f^\star(x) + \epsilon$. We consider depth $d \in \{1,2,3,4\}$ with representative targets:

$$d=1: f^\star(x) = wx + b, \tag{19}$$

$$d=2: f^\star(x) = \sin(wx + b_0) + b_1, \tag{20}$$

$$d=3: f^\star(x) = \exp(-(ax+b_0))\cdot\sin(wx+b_1), \tag{21}$$

$$d=4: f^\star(x) = \exp(-(ax+b_0))\cdot\sin(wx+b_1) + (cx+d). \tag{22}$$

We sweep $n \in \{50, 100, 500, 1000, 5000\}$ and evaluate on a held-out test set of 10^5 samples.

Model. The hypothesis is a *computation tree* whose nodes are differentiable operators from $\{+, \times, \sin, \cos, \exp, \mathrm{affine}\}$. Each affine node has learnable parameters (a, b), and constants are represented as affine nodes with slope fixed to zero. The implementation estimates a Lipschitz proxy $\widehat{L}$ either by (i) sampling gradient norms $\max_i \|\nabla_x f(x_i)\|_2$ or (ii) composing local operator derivative bounds along the tree.

Training protocol. For each depth d and training size n, we minimize mean squared error with Adam (learning rate 10^{-2}, batch size 256, weight decay 0, up to 5000 epochs). Optimization stops early when training MSE fails to improve by more than 10^{-6} for 200 consecutive epochs. Noise levels match the released configuration (σ=0.01 for $d \in \{1,2\}$ and σ=0.02 for $d \in \{3,4\}$). The generalization gap is $\mathrm{gap} = \mathrm{MSE}_{\mathrm{test}} - \mathrm{MSE}_{\mathrm{train}}$ on the same draws. We estimate $\widehat{L}$ by sampling gradient norms $\max_i \|\nabla_x f_\theta(x_i)\|_2$ on a batch of size 8192 drawn from the same input distribution.

5.2. Results

Across depths and sample sizes, the empirical generalization gap (test MSE minus train MSE) increases with depth and decreases with n. More importantly, when plotted against the complexity proxy $(\widehat{L}^d)/\sqrt{n}$, the gap aligns along an approximately linear upper envelope, consistent with Equation (15). In typical runs, depth-4 tasks show larger $\widehat{L}$ due to $\exp(\cdot)$ interactions, and correspondingly larger gaps at fixed n.

Table 1 spans both columns so numeric columns remain readable.

Reading Table 1 and Figure 2. Table 1 summarizes robust trends by reporting medians over seeds at the smallest and largest training sizes ($n \in \{50, 5000\}$). For shallow targets ($d \le 2$), $\widehat{L}$ is essentially stable in n, so halving the

Table 1. Compact sweep summary (median over seeds).

Depth	n	median $\lvert\text{gap}\rvert$	median $\hat{L}$	median $(\hat{L}^d)/\sqrt{n}$
1	50	1.63e-05	1.30e+00	1.83e-01
1	5000	1.25e-06	1.30e+00	1.83e-02
2	50	1.97e-05	1.77e+00	4.42e-01
2	5000	1.45e-06	1.77e+00	4.43e-02
3	50	8.30e-05	7.40e+00	5.73e+01
3	5000	5.99e-06	7.54e+00	6.07e+00
4	50	1.35e-04	1.01e+01	1.50e+03
4	5000	5.53e-06	1.01e+01	1.46e+02

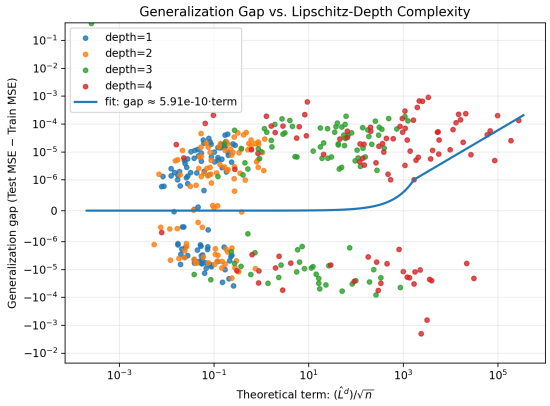


Figure 1. Empirical generalization gap versus the predicted complexity term $(\widehat{L}^d)/\sqrt{n}$ across 400 runs.

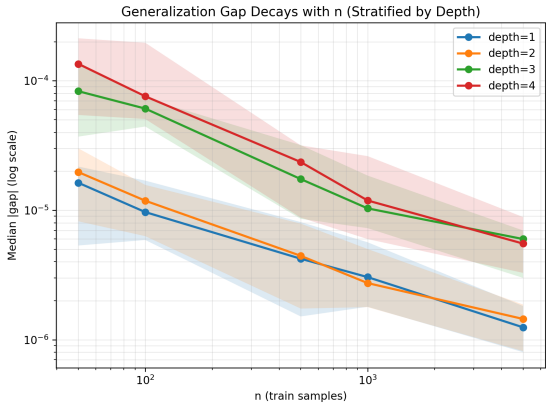


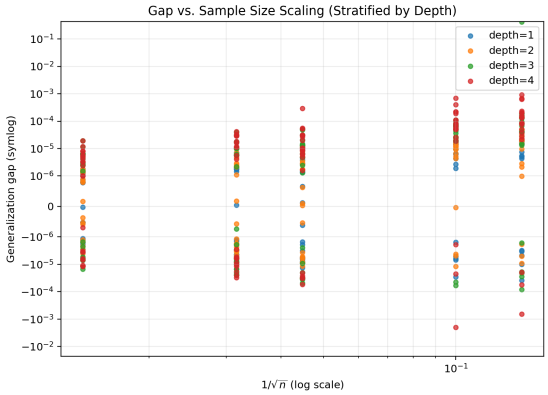
Figure 2. Median $\lvert\text{gap}\rvert$ versus n (IQR bands), stratified by depth.

horizontal uncertainty factor $1/\sqrt{n}$ by increasing n from 50 to 5000 yields comparable reductions in both median $\lvert\text{gap}\rvert$ and median $(\widehat{L}^d)/\sqrt{n}$. At $d = 3$ and especially $d = 4$, the same table shows that $\widehat{L}$ remains order-one to ten while $(\widehat{L}^d)/\sqrt{n}$ spans orders of magnitude across depths because of exponentiation in $\widehat{L}^d$. Nevertheless, increasing n still drives median gaps downward at each depth, matching the $1/\sqrt{n}$ facet of Equation (15). Figure 2 reinforces this story across *all* swept values of n: curves slope downward with sample size, with tighter dispersion (narrower IQR bands) where optimization finds stable minimizers.

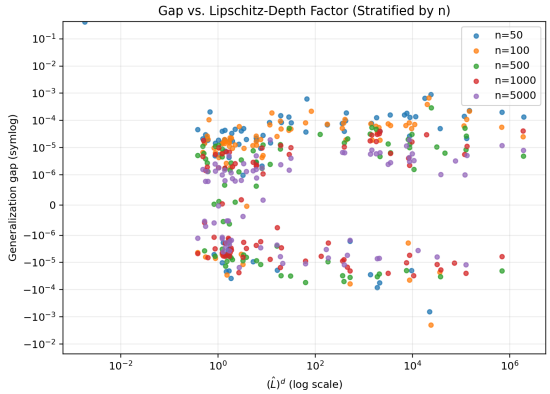
Additional diagnostics (no extra training). To further validate the factorization suggested by Equation (15), we post-process the same runs to produce two complementary views: (i) $\lvert\text{gap}\rvert$ versus $1/\sqrt{n}$ stratified by depth, and (ii) $\lvert\text{gap}\rvert$ versus $(\widehat{L})^d$ stratified by n. We also report a simple goodness-of-fit statistic by regressing the gap onto the single complexity predictor $(\widehat{L}^d)/\sqrt{n}$ with a line through the origin. These diagnostics are computationally negligible and help identify deviations caused by finite optimization error or regime changes in the local Lipschitz proxy. Figure 3 displays both views.

Figure 4 summarizes the full (d, n) grid using median $\log_{10} \lvert\text{gap}\rvert$ (heatmap) and the marginal distribution by depth (boxplots).

Quantitative fit on the full sweep. Using 400 trained models (depths $d \in \{1, 2, 3, 4\}$, sample sizes $n \in \{50, 100, 500, 1000, 5000\}$, 20 random seeds each), we observe a positive association between the predicted term and the observed gap. On a log-log scale, the Pearson correlation between $\log((\widehat{L}^d)/\sqrt{n})$ and $\log \lvert\text{gap}\rvert$ is 0.48. Fitting a power-law model $\log \lvert\text{gap}\rvert \approx \alpha \log((\widehat{L}^d)/\sqrt{n}) + c$ yields $\alpha \approx 0.20$ with $R^2 \approx 0.23$. These summary statistics indicate that the compositional complexity term captures a meaningful fraction of the variance in generalization gaps across tasks and regimes, while leaving room for additional factors such as optimization error, noise level, and local-range effects for $\exp(\cdot)$. The fitted exponent $\alpha \ll 1$ is consistent with substantial nuisance variation—including finite-batch noise in $\widehat{L}$, imperfect convergence, and mismatch between local derivative norms and worst-case Lipschitz constants appearing in uniform bounds—rather than with falsification of the qualitative roles of n and compositional amplification. Per-setting means and standard deviations over seeds appear in Table 2.

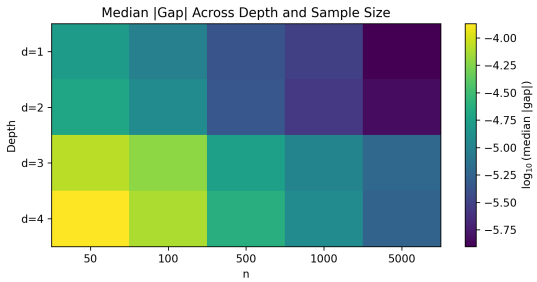


(a) $|\text{gap}|$ vs. $1/\sqrt{n}$, stratified by depth.

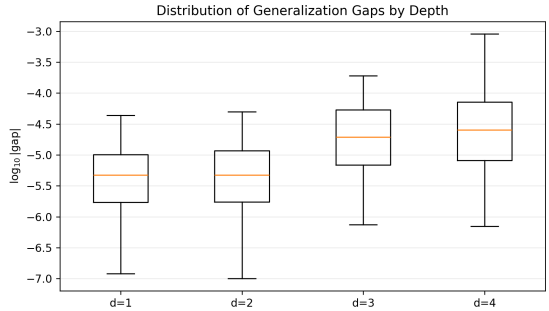


(b) $|\text{gap}|$ vs. $(\widehat{L})^d$, stratified by training size n.

Figure 3. Factorization diagnostics computed from the same logged runs (no retraining).



(a) Heatmap of median $\log_{10}|\text{gap}|$ over depth and n.



(b) Distribution of $\log_{10}|\text{gap}|$ aggregated across (n, seed) within each depth.

Figure 4. Descriptive summaries of the sweep beyond scatter plots (Figure 1).

Reading Figure 3 and Figure 4. Figure 3 factorizes the predicted scaling into $1/\sqrt{n}$ (by depth) and $(\widehat{L})^d$ (by n); both views align qualitatively with Equation (15). Figure 4 summarizes median $\log_{10}|\text{gap}|$ over the (d, n) grid and marginal depth distributions: larger n uniformly reduces gaps, while deeper compositions retain larger medians and heavier tails, consistent with harder optimization under $\exp(\cdot)$.

Limitations. Our empirical study is deliberately controlled: inputs are one-dimensional and uniformly bounded, targets lie in a realizability-aligned architecture family (correct depth and operator set), and distribution shift is absent. The estimator $\widehat{L}$ measures stability of the *trained* map on in-distribution inputs rather than certifying a worst-case global Lipschitz constant over all structured hypotheses at depth d; for compositions involving $\exp(\cdot)$, local slopes can grow rapidly with learned ranges, so $\widehat{L}$ need not reflect uniform worst-case operator stability without additional envelope constraints or interval arithmetic. Finally, uniform bounds such as Equation (15) can be loose compared with hypothesis-dependent bounds when stronger structural priors are available (McAllester, 1999; Maurer & Pontil, 2011).

6. Conclusion

We provided a PAC learning perspective on scientific discovery with compositional function trees. Despite the combinatorial explosion of symbolic structures, the statistical generalization of depth-d compositional hypotheses is controlled by stability: Lipschitz constants accumulate multiplicatively along paths with an explicit $(Kb\sqrt{2}\,L)^{d-1}$ prefactor on $\mathfrak{R}_n(\mathcal{H}^1_{\text{comp}})$ (Equation (14)), yielding excess-risk scalings of order $\mathcal{O}(L^d/\sqrt{n})$ when K, b are treated as $O(1)$. Our experiments with differentiable operator trees confirm that the generalization gap tracks the predicted complexity term. These results support the view that, for stable operator vocabularies and modest depth, scientific discovery from small datasets is statistically feasible; the remaining challenge is computational search and inductive bias design.

Natural extensions include multi-dimensional covariates $x \in \mathbb{R}^p$ with coordinate-wise or norm-based Lipschitz book-

keeping, agnostic risk decompositions when the truth lies outside $\mathcal{H}^d_{\text{comp}}$, and combining PAC-Bayes priors over discrete syntax with the contraction estimates presented here. Closing that loop would align statistical certificates more tightly with end-to-end SR systems that alternate discrete proposals with continuous refitting.

Accessibility

Technical content is fully stated in the main text (Sections 3 to 5); figures and appendix tables are illustrative. Plots use distinct colors and markers for depth-stratified series; numeric values also appear in the referenced equations and tables.

Software and Data

Experiments are implemented in Python using PyTorch; dependency versions are listed alongside the code. Synthetic data are generated in closed form as described in Section 5; we use no confidential or personally identifiable data. Reproducibility materials are available at https://github.com/suayptalha/scientific-pac.

Impact Statement

This work gives learning-theoretic foundations for compositional symbolic models: stability and depth shape statistical complexity in ways that complement purely combinatorial views of expression search. The emphasis on contraction along fixed trees clarifies that bounded-depth estimation with stable primitives need not inherit worst-case syntax-counting intuitions often cited against symbolic regression. The perspective is complementary to algorithm design: it does not prescribe a particular symbolic search procedure, but clarifies the statistical scaling of nonlinear coefficient estimation once a vocabulary and depth budget are fixed. Epistemic risks include trusting compact formulas under distribution shift, misspecification, or confounding; our guarantees do not address causal correctness or out-of-distribution validity. Faster automated hypothesis discovery may also compress publication and decision timelines unless paired with scrutiny suited to the stakes. Openly stating assumed covariate ranges and validating extrapolation beyond training coverage—especially when nonlinear primitives amplify slopes—offers a lightweight mitigation aligned with the paper's local stability viewpoint. Responsible deployment should pair such tools with domain theory, uncertainty quantification, replication, and appropriate governance in high-stakes settings.

References

Alquier, P. User-friendly introduction to PAC-Bayes bounds, 2024. URL https://arxiv.org/abs/2110.11216. Foundations and Trends in Machine Learning.

Bartlett, P. L. and Mendelson, S. Rademacher and gaussian complexities: Risk bounds and structural results. *J. Mach. Learn. Res.*, 3:463–482, 2003. URL https://api.semanticscholar.org/CorpusID:463216.

Bartlett, P. L., Foster, D. J., and Telgarsky, M. J. Spectrally-normalized margin bounds for neural networks, 2017. URL https://arxiv.org/abs/1706.08498.

Brunton, S. L., Proctor, J. L., and Kutz, J. N. Discovering governing equations from data by sparse identification of nonlinear dynamical systems. *Proceedings of the National Academy of Sciences*, 113(15):3932–3937, 2016. doi: 10.1073/pnas.1517384113.

Catoni, O. *PAC-Bayesian Supervised Classification: The Thermodynamics of Statistical Learning*, volume 56 of *Lecture Notes—Monograph Series*. Institute of Mathematical Statistics, Beachwood, Ohio, 2007. ISBN 978-0-940600-68-7. doi: 10.1214/074921707000000391. URL https://doi.org/10.1214/074921707000000391.

Cranmer, M. Interpretable machine learning for science with PySR and SymbolicRegression.jl, 2023. URL https://arxiv.org/abs/2305.01582.

Cranmer, M., Sanchez-Gonzalez, A., Battaglia, P., Xu, R., Cranmer, K., Spergel, D., and Ho, S. Discovering symbolic models from deep learning with inductive biases, 2020. URL https://arxiv.org/abs/2006.11287.

Dudley, R. M. The sizes of compact subsets of hilbert space and continuity of gaussian processes. *Journal of Functional Analysis*, 1:125–165, 1967. URL https://api.semanticscholar.org/CorpusID:122249056.

Kanade, V., Rebeschini, P., and Vaskevicius, T. Exponential tail local Rademacher complexity risk bounds without the Bernstein condition. *Journal of Machine Learning Research*, 25(388):1–43, 2024. URL http://jmlr.org/papers/v25/23-0063.html.

Koltchinskii, V. and Panchenko, D. Empirical margin distributions and bounding the generalization error of combined classifiers. *Annals of Statistics*, 30:1–50, 2002. URL https://api.semanticscholar.org/CorpusID:2307733.

Koza, J. R. *Genetic Programming: On the Programming of Computers by Means of Natural Selection*. MIT Press, Cambridge, MA, 1992. ISBN 978-0-262-11170-6.

La Cava, W., Orzechowski, P., Burlacu, B., Olivetti de França, F., Virgolin, M., Jin, Y., Kommenda, M., and Moore, J. H. Contemporary symbolic regression methods and their relative performance, 2021. URL https://arxiv.org/abs/2107.14351. NeurIPS 2021 Datasets and Benchmarks Track; introduces the SRBench suite.

Langford, J. and Seeger, M. Bounds for averaging classifiers. 02 2001.

Ledoux, M. and Talagrand, M. *Probability in Banach Spaces: Isoperimetry and Processes*. Springer-Verlag, Berlin, 1991. ISBN 978-3-540-52013-9. doi: 10.1007/978-3-642-20212-4.

Lei, Y., Ding, L., and Bi, Y. Local Rademacher complexity bounds based on covering numbers. *Neurocomputing*, 218:320–330, 2016. doi: 10.1016/j.neucom.2016.08.074.

Makke, N. and Chawla, S. Interpretable scientific discovery with symbolic regression: a review. *Artificial Intelligence Review*, 57(1), 2024. doi: 10.1007/s10462-023-10622-0. URL https://doi.org/10.1007/s10462-023-10622-0. Article number 2.

Maurer, A. A chain rule for the expected suprema of Gaussian processes. *Theoretical Computer Science*, 650:109–122, 2016a. doi: 10.1016/j.tcs.2016.07.034.

Maurer, A. A vector-contraction inequality for Rademacher complexities, 2016b. URL https://arxiv.org/abs/1605.00251.

Maurer, A. and Pontil, M. Structured sparsity and generalization, 2011. URL https://arxiv.org/abs/1108.3476.

McAllester, D. A. Some PAC-Bayesian theorems. *Machine Learning*, 37(3):355–363, 1999. doi: 10.1023/A:1007618624809. URL https://doi.org/10.1023/A:1007618624809.

Mohri, M., Rostamizadeh, A., and Talwalkar, A. *Foundations of Machine Learning*. MIT Press, Cambridge, MA, 2nd edition, 2018. ISBN 9780262039406.

Neyshabur, B., Bhojanapalli, S., McAllester, D., and Srebro, N. Exploring generalization in deep learning, 2017. URL https://arxiv.org/abs/1706.08947.

Petersen, B. K., Landajuela, M., Mundhenk, T. N., Santiago, C. P., Kim, S. K., and Kim, J. T. Deep symbolic regression: Recovering mathematical expressions from data via risk-seeking policy gradients, 2021. URL https://arxiv.org/abs/1912.04871. Published at ICLR 2021.

Rissanen, J. Modeling by shortest data description. *Automatica*, 14(5):465–471, 1978. doi: 10.1016/0005-1098(78)90005-5.

Schmidt, M. D. and Lipson, H. Distilling free-form natural laws from experimental data. *Science*, 324(5923):81–85, 2009a. doi: 10.1126/science.1165893.

Schmidt, M. D. and Lipson, H. Symbolic regression of implicit equations. In *Genetic Programming Theory and Practice VII*, Genetic and Evolutionary Computation, pp. 73–85. Springer, New York, 2009b. doi: 10.1007/978-1-4419-1626-6_5. URL https://doi.org/10.1007/978-1-4419-1626-6_5.

Seeger, M. Pac-bayesian generalization error bounds for gaussian process classification. *Journal of Machine Learning Research*, 3, 08 2002. doi: 10.1162/153244303765208386.

Shalev-Shwartz, S. and Ben-David, S. *Understanding Machine Learning: From Theory to Algorithms*. Cambridge University Press, 2014. ISBN 978-1-107-05713-5. doi: 10.1017/CBO9781107298019.

Talagrand, M. *Upper and Lower Bounds for Stochastic Processes: Modern Methods and Classical Problems*. Springer, Heidelberg, 2 edition, 2014. ISBN 978-3-642-54074-5. doi: 10.1007/978-3-642-54075-2.

Truong, L. V. On Rademacher complexity-based generalization bounds for deep learning, 2022. URL https://arxiv.org/abs/2208.04284.

Udrescu, S.-M. and Tegmark, M. AI Feynman: A physics-inspired method for symbolic regression, 2020. URL https://arxiv.org/abs/1905.11481.

Valiant, L. G. A theory of the learnable. *Commun. ACM*, 27:1134–1142, 1984. URL https://api.semanticscholar.org/CorpusID:59712.

van der Vaart, A. W. and Wellner, J. A. *Weak Convergence and Empirical Processes: With Applications to Statistics*. Springer Series in Statistics. Springer, 1996. ISBN 978-0-387-94640-5. doi: 10.1007/978-1-4757-2545-2.

Vapnik, V. N. *Statistical Learning Theory*. Wiley, New York, NY, USA, September 1998.

A. Additional Proof Details

We expand the argument underlying Lemmas 4.4–4.6 and Corollary 4.8. The key tool is Maurer's vector contraction (Maurer, 2016b): for an L-Lipschitz map $\Phi : \mathbb{R}^m \to \mathbb{R}$ with $\Phi(0) = 0$ and a class $\mathcal{G} \subset (\mathcal{X} \to \mathbb{R}^m)$ measured by the coordinatewise complexity (5), $\mathfrak{R}_S(\Phi \circ \mathcal{G}) \leq \sqrt{2}\, L\, \mathfrak{R}_S^{(m)}(\mathcal{G})$. In our tree setting, each internal node computes $h(z_1, \ldots, z_b)$ where $h \in \mathcal{H}_{\text{base}}$ is L_h-Lipschitz and $(z_1, \ldots, z_b)$ are subtree outputs. Define the vector-valued class $\mathcal{G}_{d-1} = \{x \mapsto (g_1(x), \ldots, g_b(x)) : \; g_j \in \mathcal{H}_{\text{comp}}^{d-1}\}$. For fixed $h_k \in \mathcal{H}_{\text{base}}$, set $\mathcal{C}_k = \{h_k \circ g : \; g \in \mathcal{G}_{d-1}\}$. Lemma 4.6 gives $\mathfrak{R}_S(\mathcal{H}_{\text{comp}}^d) \leq K \max_k \mathfrak{R}_S(\mathcal{C}_k)$. For each k, Lemma 4.4 yields $\mathfrak{R}_S(\mathcal{C}_k) \leq \sqrt{2}\, L_k\, \mathfrak{R}_S^{(b)}(\mathcal{G}_{d-1})$. Lemma 4.5 (and its b-fold analogue) gives $\mathfrak{R}_S^{(b)}(\mathcal{G}_{d-1}) \leq b\, \mathfrak{R}_S(\mathcal{H}_{\text{comp}}^{d-1})$ for binary b=2. Combining, $\mathfrak{R}_S(\mathcal{H}_{\text{comp}}^d) \leq Kb\sqrt{2}\, L\, \mathfrak{R}_S(\mathcal{H}_{\text{comp}}^{d-1})$ and iterating yields Equations (13) and (14).

A.1. Proof of Lemma 4.5

Let $S = (x_i)_{i=1}^n$. For the paired class $\mathcal{F}_1 \times \mathcal{F}_2$, define the empirical Rademacher complexity using independent signs for the two coordinates:

$$\mathfrak{R}_S(\mathcal{F}_1 \times \mathcal{F}_2) = \mathbb{E}_{\sigma^{(1)}, \sigma^{(2)}}\Big[\sup_{f_1 \in \mathcal{F}_1, f_2 \in \mathcal{F}_2} \frac{1}{n} \sum_{i=1}^n \sigma_i^{(1)} f_1(x_i) + \sigma_i^{(2)} f_2(x_i)\Big] \tag{23}$$

$$= \mathbb{E}\Big[\sup_{f_1 \in \mathcal{F}_1} \frac{1}{n} \sum_{i=1}^n \sigma_i^{(1)} f_1(x_i)\Big] + \mathbb{E}\Big[\sup_{f_2 \in \mathcal{F}_2} \frac{1}{n} \sum_{i=1}^n \sigma_i^{(2)} f_2(x_i)\Big] \tag{24}$$

$$= \mathfrak{R}_S(\mathcal{F}_1) + \mathfrak{R}_S(\mathcal{F}_2). \tag{25}$$

The second line uses separability of the supremum over independent function choices.

A.2. Deriving Theorem 4.7 via contraction

We show the inductive step in (13). Let $\mathcal{H}_d := \mathcal{H}_{\text{comp}}^d$. By Lemma 4.6, $\mathfrak{R}_S(\mathcal{H}_d) \leq K \max_{h \in \mathcal{H}_{\text{base}}} \mathfrak{R}_S(\mathcal{C}_h)$ where $\mathcal{C}_h = \{h \circ (g_1, g_2) : g_1, g_2 \in \mathcal{H}_{d-1}\}$ and K=$|\mathcal{H}_{\text{base}}|$. Fix h and write $f(x) = h(g_1(x), g_2(x))$. By Maurer's vector contraction (Lemma 4.4) with m=2,

$$\mathfrak{R}_S(\mathcal{C}_h) \leq \sqrt{2}\, L_h\, \mathfrak{R}_S^{(2)}(\mathcal{H}_{d-1} \times \mathcal{H}_{d-1}) = \sqrt{2}\, L_h\, \mathfrak{R}_S(\mathcal{H}_{d-1} \times \mathcal{H}_{d-1}). \tag{26}$$

Lemma 4.5 gives $\mathfrak{R}_S(\mathcal{H}_{d-1} \times \mathcal{H}_{d-1}) \leq 2\, \mathfrak{R}_S(\mathcal{H}_{d-1})$. Thus $\mathfrak{R}_S(\mathcal{C}_h) \leq 2\sqrt{2}\, L_h\, \mathfrak{R}_S(\mathcal{H}_{d-1})$ and taking $\max_h$ yields $\mathfrak{R}_S(\mathcal{H}_d) \leq K \cdot 2\sqrt{2}\, L\, \mathfrak{R}_S(\mathcal{H}_{d-1})$. Iterating gives (13).

A.3. Sketch: PAC-Bayes inequality (18)

For completeness we recall one route to (18). Let $\ell \in [0, 1]$ be a bounded loss and define empirical and true risks as before. PAC-Bayes bounds control $\mathbb{E}_{f \sim \rho}[R(f)]$ in terms of $\mathbb{E}_{f \sim \rho}[\widehat{R}_S(f)]$ plus a complexity penalty involving $\mathrm{KL}(\rho \| \pi)$; see McAllester (1999); Langford & Seeger (2001); Catoni (2007); Alquier (2024) for multiple variants and constants. In our context, π can be chosen to downweight large trees (e.g., via an MDL code length), encouraging posteriors supported on shallow/stable expressions.

B. Symmetrization, Tail Bounds, and Talagrand-type Contractions

This section collects standard ingredients used implicitly when translating Rademacher complexity into high-probability uniform bounds (Bartlett & Mendelson, 2003; Mohri et al., 2018; Ledoux & Talagrand, 1991).

B.1. Rademacher symmetrization

Let $\mathcal{F} \subset (\mathcal{X} \to \mathbb{R})$ and consider the empirical process supremum

$$Z(S) \;:=\; \sup_{f \in \mathcal{F}} \Big| \frac{1}{n} \sum_{i=1}^n \big(f(x_i) - \mathbb{E}[f(X)]\big) \Big|. \tag{27}$$

A symmetrization argument yields

$$\mathbb{E}[Z(S)] \;\le\; 2\,\mathbb{E}_S\big[\mathfrak{R}_S(\mathcal{F})\big], \tag{28}$$

with $\mathfrak{R}_S(\mathcal{F})$ as in the main text (expectation over independent Rademacher signs). Thus controlling $\mathfrak{R}_n(\mathcal{F}) = \mathbb{E}_S[\mathfrak{R}_S(\mathcal{F})]$ controls the first moment of uniform deviations.

B.2. McDiarmid bounded differences

Assume $Z = Z(x_{1:n})$ satisfies bounded differences: changing one coordinate x_i changes Z by at most c_i. Then for all $\varepsilon > 0$,

$$\mathbb{P}\big(Z - \mathbb{E}[Z] \ge \varepsilon\big) \;\le\; \exp\Big(-\frac{2\varepsilon^2}{\sum_{i=1}^n c_i^2}\Big). \tag{29}$$

Applied with $c_i = \mathcal{O}(1/n)$ for bounded-range losses yields Gaussian concentration at rate $\mathcal{O}(n^{-1/2})$ after centering.

B.3. Scalar contraction

Let $\phi : \mathbb{R} \to \mathbb{R}$ be L-Lipschitz and $\phi(0) = 0$. Then for any finite sample S,

$$\mathfrak{R}_S(\phi \circ \mathcal{F}) \;\le\; L\,\mathfrak{R}_S(\mathcal{F}). \tag{30}$$

This is the one-dimensional specialization of Equation (10) used repeatedly when peeling unary operators off a composition chain.

B.4. Unary composition chains

Let $\mathcal{F}_0 \subset (\mathcal{X} \to \mathbb{R})$ and define inductively $\mathcal{F}_k = \{\phi_k \circ f : \; f \in \mathcal{F}_{k-1}\}$ where each $\phi_k : \mathbb{R} \to \mathbb{R}$ is L_k-Lipschitz on the relevant range. Then for every sample S,

$$\mathfrak{R}_S(\mathcal{F}_k) \;\le\; \Big(\prod_{j=1}^{k} L_j\Big)\,\mathfrak{R}_S(\mathcal{F}_0). \tag{31}$$

Proof. Induct on k, applying scalar contraction at each layer. □

B.5. Dudley entropy integral (sketch)

For many geometrically well-behaved classes, chaining yields

$$\mathfrak{R}_n(\mathcal{F}) \;\lesssim\; \inf_{\alpha \ge 0}\Big(\alpha + \frac{1}{\sqrt{n}}\int_\alpha^\infty \sqrt{\log \mathcal{N}(\varepsilon, \mathcal{F}, \|\cdot\|_n)}\,d\varepsilon\Big), \tag{32}$$

where $\mathcal{N}(\varepsilon, \mathcal{F}, \|\cdot\|_n)$ is the ε-covering number of $\mathcal{F}$ in the empirical $L_2(P_n)$ metric (Dudley, 1967; van der Vaart & Wellner, 1996; Talagrand, 2014). Under Assumption 3.2, leaves lie in Lipschitz balls of radius $\mathcal{O}(W_{\max} X_{\max}\sqrt{p})$ over $\mathcal{X}$, while each remaining primitive has bounded subgraph complexity. Thus $\mathcal{H}^1_{\text{comp}}$ is a finite union of uniformly Lipschitz families with bounded envelopes, giving $\mathfrak{R}_n(\mathcal{H}^1_{\text{comp}}) = \mathcal{O}(n^{-1/2})$ by standard VC/Rademacher chaining under fixed vocabulary size (Bartlett & Mendelson, 2003; Mohri et al., 2018).

B.6. Agnostic decomposition

Let $h^\star \in \arg\min_{h \in \mathcal{H}^d_{\text{comp}}} R(h)$ be a population risk minimizer within the class. For any empirical risk minimizer $\widehat{h}$,

$$R(\widehat{h}) - \inf_{h \in \mathcal{H}} R(h) \;=\; \big(R(\widehat{h}) - \widehat{R}_S(\widehat{h})\big) + \big(\widehat{R}_S(\widehat{h}) - \widehat{R}_S(h^\star)\big) + \big(\widehat{R}_S(h^\star) - R(h^\star)\big). \tag{33}$$

The middle term is ≤ 0 by definition of ERM, while the two bracketed differences are controlled by uniform deviations over $\mathcal{H}^d_{\text{comp}}$, yielding the $\mathcal{O}(L^d/\sqrt{n})$ variance term under bounded loss/Lipschitzness after absorbing arity factors as in Equation (14). If the truth lies outside $\mathcal{H}^d_{\text{comp}}$, add an irreducible approximation error $\inf_{h \in \mathcal{H}^d_{\text{comp}}} R(h) - R^\star$ where $R^\star$ is the Bayes risk under squared loss.

Table 2. Full sweep summary (mean±std over seeds).

d	n	Train MSE	Test MSE	$\lvert\text{gap}\rvert$	$\hat{L}$	$(\hat{L}^d)/\sqrt{n}$
1	50	9.44e-05±1.9e-05	1.04e-04±3.3e-06	1.73e-05±1.3e-05	1.28e+00±4.3e-01	1.81e-01±6.1e-02
1	100	9.50e-05±1.3e-05	1.02e-04±2.0e-06	1.19e-05±8.3e-06	1.28e+00±4.3e-01	1.28e-01±4.3e-02
1	500	1.02e-04±6.4e-06	1.01e-04±9.8e-07	5.16e-06±4.2e-06	1.28e+00±4.3e-01	5.72e-02±1.9e-02
1	1000	1.01e-04±4.1e-06	1.01e-04±1.2e-06	3.75e-06±2.6e-06	1.28e+00±4.3e-01	4.05e-02±1.4e-02
1	5000	1.00e-04±1.7e-06	1.00e-04±7.6e-07	1.55e-06±1.2e-06	1.28e+00±4.3e-01	1.81e-02±6.1e-03
2	50	9.08e-05±1.8e-05	1.08e-04±8.4e-06	2.08e-05±1.6e-05	1.75e+00±7.5e-01	5.08e-01±3.8e-01
2	100	9.42e-05±1.2e-05	1.04e-04±3.8e-06	1.21e-05±6.9e-06	1.75e+00±7.5e-01	3.59e-01±2.7e-01
2	500	1.01e-04±6.5e-06	1.01e-04±1.0e-06	5.04e-06±4.0e-06	1.75e+00±7.5e-01	1.60e-01±1.2e-01
2	1000	1.01e-04±4.4e-06	1.01e-04±9.1e-07	3.47e-06±2.5e-06	1.75e+00±7.5e-01	1.14e-01±8.5e-02
2	5000	1.00e-04±1.9e-06	1.01e-04±7.3e-07	1.63e-06±1.2e-06	1.75e+00±7.5e-01	5.08e-02±3.8e-02
3	50	6.24e-02±2.8e-01	8.31e-02±3.7e-01	2.07e-02±9.2e-02	8.12e+00±5.9e+00	2.09e+02±3.5e+02
3	100	3.70e-04±4.6e-05	4.18e-04±1.3e-05	5.90e-05±2.4e-05	8.82e+00±5.8e+00	1.61e+02±2.4e+02
3	500	4.04e-04±2.7e-05	4.05e-04±3.1e-06	2.17e-05±1.6e-05	8.84e+00±5.8e+00	7.27e+01±1.1e+02
3	1000	4.03e-04±1.7e-05	4.03e-04±2.5e-06	1.38e-05±1.0e-05	8.84e+00±5.8e+00	5.15e+01±7.9e+01
3	5000	4.01e-04±7.8e-06	4.02e-04±2.6e-06	6.65e-06±4.7e-06	8.84e+00±5.8e+00	2.30e+01±3.5e+01
4	50	1.37e-03±3.8e-03	1.52e-03±3.6e-03	2.26e-04±2.6e-04	1.18e+01±9.3e+00	2.26e+04±6.4e+04
4	100	1.40e-03±3.8e-03	1.42e-03±3.4e-03	2.25e-04±4.5e-04	1.18e+01±9.3e+00	1.55e+04±4.3e+04
4	500	6.83e-04±8.4e-04	7.03e-04±8.7e-04	3.67e-05±6.3e-05	1.20e+01±9.3e+00	7.05e+03±2.0e+04
4	1000	5.01e-04±2.8e-04	5.04e-04±2.9e-04	1.63e-05±1.1e-05	1.21e+01±9.3e+00	4.99e+03±1.4e+04
4	5000	4.19e-04±3.9e-05	4.20e-04±3.8e-05	6.44e-06±4.8e-06	1.21e+01±9.3e+00	2.25e+03±6.2e+03

C. Detailed Risk Bound for Bounded Squared Loss

We sketch Theorem 4.9 with constants tracked only at the level of $\mathcal{O}(\cdot)$. Assume $|y| \leq B$ and $|f(x)| \leq B$ uniformly over $\mathcal{H}^d_{\text{comp}}$. For squared loss $\ell(\hat{y}, y) = (\hat{y} - y)^2$, on $[-B, B]^2$ one has $|\partial_{\hat{y}}\ell| \leq 4B$, hence $\ell(\cdot, y)$ is $(4B)$-Lipschitz in $\hat{y}$. Talagrand-type contraction for Lipschitz composition with squared loss implies $\mathfrak{R}_S(\ell \circ \mathcal{H}^d_{\text{comp}}) \leq (4B) \cdot (Kb\sqrt{2}\,L)^{d-1}\mathfrak{R}_S(\mathcal{H}^1_{\text{comp}})$ up to fixed vocabulary/arity factors (Equation (14)). Combining symmetrization, concentration on $\sup_{f\in\mathcal{H}}(R(f) - \widehat{R}_S(f))$, and bounded loss yields Equation (15).

D. Experimental Details and Reproducibility

The training code evaluates every combination of depths $d \in \{1, 2, 3, 4\}$ and sample sizes $n \in \{50, 100, 500, 1000, 5000\}$ with multiple random seeds. Models are trained with Adam on MSE until training loss stops improving (patience-based early stopping). The Lipschitz proxy $\widehat{L}$ is computed on a large batch of test-distribution inputs using either gradient norms or composed operator bounds. The primary empirical summary plots the generalization gap against $(\widehat{L}^d)/\sqrt{n}$.

D.1. Wide-format numeric tables

Table 2 reports mean±std over seeds for train/test error, gap magnitude, $\widehat{L}$, and $(\widehat{L}^d)/\sqrt{n}$. Wide numeric layouts are set in a smaller type size and scaled to the appendix text width so columns remain legible. Table 3 gives the same sweep as a compact median grid of $|\text{gap}|$ over (d, n), consistent with Figure 4.

Figure 3 (factorization panels) and Figure 4 (heatmap and boxplots) appear in the main text.

E. Appendix: Connection to SR Search and MDL

Theorem 4.9 concerns statistical generalization for a fixed hypothesis class. Symbolic regression additionally faces discrete search over structures, often guided by minimum-description-length (MDL) principles and sparsity (Rissanen, 1978; Brunton et al., 2016). A promising direction is to combine our Lipschitz-depth certificate with MDL priors over trees to obtain

Table 3. Median $|\mathrm{gap}|$ (grid over seeds).

	50	100	500	1000	5000
1	1.63e-05	9.67e-06	4.22e-06	3.05e-06	1.25e-06
2	1.97e-05	1.18e-05	4.45e-06	2.74e-06	1.45e-06
3	8.30e-05	6.09e-05	1.74e-05	1.03e-05	5.99e-06
4	1.35e-04	7.58e-05	2.36e-05	1.19e-05	5.53e-06

bounds that trade off depth, stability, and description length.