A Semantic Construction Profile for Exact Boolean Formula Complexity

J. R. Landers

Abstract

Prior exhaustive synthesis showed that twelve elementary truth-table invariants predict 77.7% of exact four-input NAND/NOR formula-size variance and 84.3% for AND/OR/NOT under strict descriptor holdout. We investigate the residual with 54 additional, input-renaming-invariant coordinates drawn from formula and query complexity: decision trees, certificates, sensitivity, algebraic normal form (ANF), signed Fourier structure, prime cubes, restrictions, and hypercube topology. The resulting 66-dimensional *semantic construction profile* forms 3,958 classes among all 65,536 functions, nearly matching the 3,984 input-permutation orbits. A nonlinear regressor evaluated by five-fold holdout of complete profile classes explains 95.7%, 94.7%, and 97.3% of exact minimum-size variance in NAND, NOR, and AND/OR/NOT. Conditional permutation importance is dominated by decision-tree leaf count, followed by certificate size, the Khrapchenko product, and linear ANF support. Thus the missing signal is structured: exact formula cost is most predictable from a function's recursive decomposability, with boundary and algebraic structure providing secondary information. The result identifies ingredients for a theory, but does not turn the learned relation into an exact formula or a new asymptotic bound.

1 From semantic correlation to construction potential

Takeaway. The earlier descriptor measured coarse shape. The expanded profile measures ways a function can be decomposed, certified, or bounded.

For a gate basis $\mathcal{L}$, recall the exact tree-formula complexity

$$K_{\mathcal{L}}(f) = \min\{|e| : \mathrm{eval}_{\mathcal{L}}(e) = f\}. \tag{1}$$

The target is known exactly for every $f : \{0,1\}^4 \to \{0,1\}$ in NAND, NOR, and ordered AND/OR/NOT by minimum-layer enumeration. The original semantic map $\psi(f) \in \mathbb{R}^{12}$ contained signed bias, four sorted influences, five Walsh–Fourier energies by degree, algebraic degree, and input-permutation stabilizer size. It was compact, but it discarded how spectral mass is signed, how ANF support is distributed, and how the truth table decomposes under queries or restrictions.

Definition 1 (Semantic construction profile)**.** Let

$$\Phi(f) = \big(\psi(f), D(f), C(f), S(f), A(f), W(f), P(f), R(f), H(f)\big) \in \mathbb{R}^{66}, \tag{2}$$

where D, C, S, A, W, P, R, H respectively summarize exact decision trees, certificates, local and higher-order sensitivity, ANF support, signed and shape-sensitive Fourier data, prime cubes, restrictions and factorability, and monochromatic hypercube topology.

All coordinates are functions of the truth table; no minimum formula or gate count is used. Sorting or aggregation makes every coordinate invariant under renaming inputs. The term "construction profile" is operational: each family records either a known obstruction to small formulas or a constructive route to one. Decision-tree leaves expand directly into a formula; certificate complexity is linked to decision-tree complexity; the Khrapchenko product is a De Morgan formula lower bound; prime implicants induce DNF/CNF upper bounds; and random restrictions are a classical formula-size method [1, 2, 3, 4].

2 Evaluation without semantic twins

The expanded map partitions the complete universe into 3,958 exact profile classes, compared with 675 original descriptor classes and 3,984 orbits under input permutation. Near-orbit resolution creates a leakage risk: a random function split could place an identical profile in both train and test. We therefore split *profile classes*, not functions,

Profile family	Coordinates	Mathematical connection
Original descriptor	12	influence, Fourier concentration, symmetry
Decision trees/certificates	6	constructive trees; query inequalities
Local/higher sensitivity	9	Khrapchenko and boundary complexity
ANF support	5	$\mathbb{F}_2$ gate composition
Fourier phase/shape	17	information lost by degree-wise energy
Prime cubes	6	canonical DNF/CNF constructions
Restrictions/factorability	8	shrinkage and recursive decompositions
Cube topology	3	fragmentation beyond boundary size

Table 1: The 66-coordinate profile. The rows partition the coordinates; the first row is the complete original descriptor.

into five folds. Every class is represented once in fitting, weighted by its number of member functions. At evaluation, the within-class squared error is restored, so a model receives no credit for variation that Φ cannot distinguish.

For fold j, a histogram gradient-boosting regressor m_{-j} is trained on the other four folds and evaluated by function-weighted

$$R_j^2 = 1 - \frac{\sum_f (K_{\mathcal{L}}(f) - m_{-j}(\Phi(f)))^2}{\sum_f (K_{\mathcal{L}}(f) - \overline{K}_{\mathcal{L},j})^2}. \tag{3}$$

This tests generalization to unseen semantic profiles. It differs from the earlier tie-aware k-nearest-neighbor experiment in both descriptor and model, so the comparison is evidence of recovered signal rather than a one-variable ablation.

Importance is computed only on held-out classes. For coordinate q, we permute q among test profiles and report

$$I_q = \mathbb{E}_{j,r}\big[R_j^2 - R_{j,r}^{2,\mathrm{perm}(q)}\big], \tag{4}$$

over five folds and eight permutations per fold. We also permute all coordinates in a conceptual family jointly. These are *conditional* importances: a low value may mean that correlated coordinates substitute for one another, not that the underlying concept is irrelevant.

3 Results

Takeaway. The enriched profile closes most of the finite prediction gap, and the learned dependence has a clear hierarchy.

Language	Original strict R^2	Expanded held-out R^2	Fold s.d.
NAND	.777	.957	.005
NOR	.777	.947	.008
AND/OR/NOT	.843	.973	.004

Table 2: Prediction of exact minimum gate count over all 65,536 functions. Both evaluations exclude exact descriptor twins; the expanded result uses five-fold holdout of complete 66-feature profile classes.

The original descriptor ceiling was .893 for NAND/NOR and .931 for AND/OR/NOT. The expanded profile ceiling is .9996, .9996, and .9999: only 15, 14, and 2 profile classes, respectively, retain any target ambiguity. The nonlinear model remains below those ceilings, showing that a small amount of error is statistical/modeling error rather than missing truth-table signal. Nevertheless, mean held-out R^2 rises by .180 for NAND, .169 for NOR, and .130 for AND/OR/NOT relative to the original strict result.

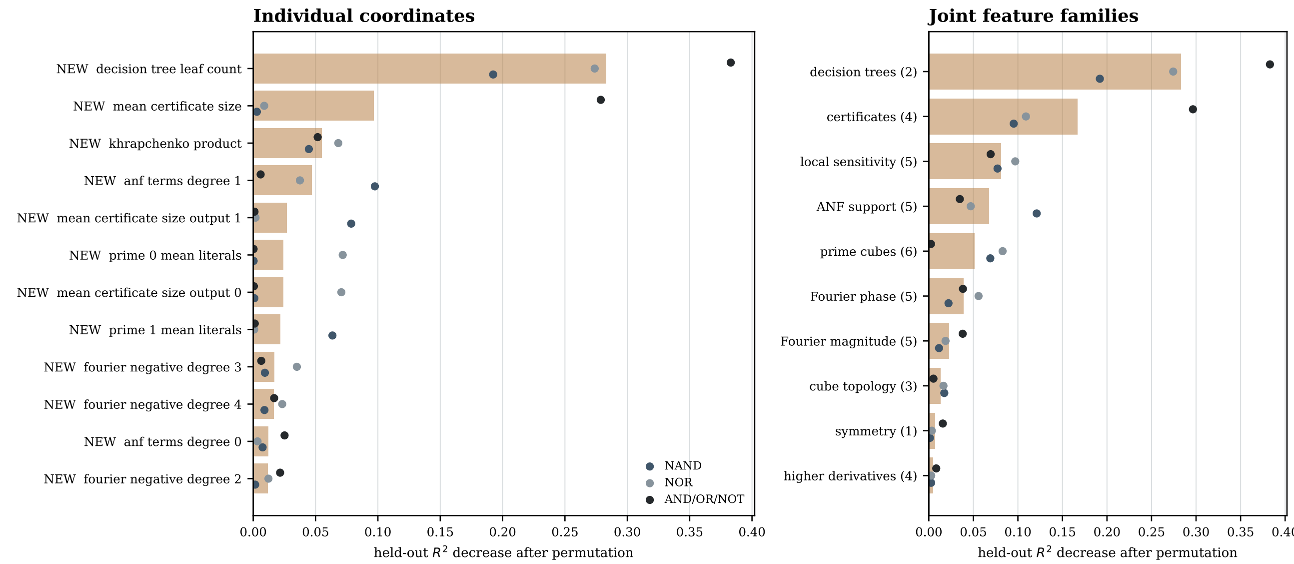

Figure 1: Conditional permutation importance on unseen profile classes. Bars average the three gate languages; points give language-specific values. "New" coordinates were absent from the original semantic descriptor.

Decision-tree leaf count is the dominant individual coordinate, with mean importance .283. Mean certificate size follows at .097, then the Khrapchenko product (.055) and the number of linear ANF monomials (.047). Jointly, decision-tree coordinates score .283, certificates .167, local sensitivity .081, ANF support .068, prime cubes .051, and Fourier phase .039. The original Fourier-magnitude family remains useful (.023), while individual influences become conditionally small (.002 jointly) because certificates, sensitivity, and Khrapchenko coordinates encode overlapping boundary information.

Standalone association tells the same broad story without conditioning on the regressor. Decision-tree leaf count, the Khrapchenko product, and mean certificate size each explain about 74% of target variance under an adjusted nonparametric correlation ratio. No single coordinate suffices, however: their conditional roles separate once competing semantic summaries are held fixed.

4 What the model appears to measure

The hierarchy suggests the following finite structural hypothesis:

$$\text{formula cost} \;\approx\; \text{recursive decomposition cost} + \text{boundary obstruction} + \text{algebraic/phase correction}. \tag{5}$$

The first term is witnessed by decision-tree leaves and certificates. These features ask how quickly restrictions force the output and therefore how many semantic cases a tree formula must separate. The second is witnessed by the Khrapchenko product and local sensitivity, which measure the edge boundary between zero and one inputs and have direct lower-bound precedent. The third distinguishes functions with similar boundary magnitude but different ANF support or Fourier signs.

The correction is not merely a passive correlation. XORing linear parity into a function leaves all higher-degree ANF coefficients fixed. Among 17,024 such pairs that also preserve the original descriptor exactly, NAND cost changes by 1.271 gates on average and by as many as eight; matched non-parity pairs average 1.167. The analogous NOR values are 1.268 and 1.167. For AND/OR/NOT, however, parity-related pairs are *closer* than controls (.664 versus .813). Phase sensitivity is therefore real but basis-relative; "parity burden" is not a universal scalar explanation.

Equation (5) is a research program rather than a closed form. A theorem would need explicit upper and lower inequalities connecting these quantities to $K_{\mathcal{L}}$, while controlling their redundancy and basis dependence. The present result supplies empirical targets for such inequalities: begin with recursive restriction trees, add boundary lower bounds, then characterize when algebraic phase makes those bounds loose.

5 Toward an analytic statement

The importance ordering points to a more precise theorem target. Let $U(f)$ be the minimum number of leaves in a deterministic decision tree for f, and let $B(f)$ be its Khrapchenko product (the product of the average numbers of bichromatic hypercube edges incident to zero and one inputs). A decision tree can be compiled recursively by Shannon expansion into a formula, giving a basis-dependent upper bound linear in $U(f)$. Conversely, $B(f)$ lower-bounds De Morgan formula leaf size and therefore, up to constant-factor translations, the formula measures considered here [1]. The established machinery thus has the schematic form

$$c_{\mathcal{L}}B(f) \;\le\; K_{\mathcal{L}}(f)+1 \;\le\; C_{\mathcal{L}}U(f), \tag{6}$$

for suitable conventions and basis-dependent constants.

Our leading features sit at the two ends of (6): U is first and B third in individual conditional importance. Mean certificate size, ranked second, describes how rapidly the upper construction terminates on typical inputs. This suggests that the regressor is learning where $K_{\mathcal{L}}(f)$ lies inside a classical lower–upper interval rather than discovering an unrelated numerical coincidence. ANF support and Fourier phase then predict when boundary-based lower information or decision-tree upper information is loose.

Because the four-input universe is complete, this claim can be sharpened without fitting another black-box predictor. One can first compute the best valid constants in (6) for each basis, then regress only the normalized position inside the resulting interval. Symbolic search can propose a small certificate/phase correction, and exhaustive verification can either prove the candidate for all four-input functions or return exact counterexamples. The useful output would be an inequality or approximation guarantee with a short semantic vector, not a 66-coordinate identity test. Five-input tests would then distinguish a genuine scaling law from a finite-universe fit.

6 Limitations and next tests

The universe is exhaustive but has only four inputs. The profile is almost an orbit identifier, so its near-unit ceiling does not itself constitute a compact theory. Gradient boosting and permutation importance describe predictive dependence, not causation; importances can move among correlated features. Finally, exact decision-tree quantities become expensive as n grows.

The decisive next experiment is therefore not to add more four-input coordinates. It is to freeze a compact profile—decision-tree leaves, certificate means, a Khrapchenko term, low-degree ANF support, and signed Fourier counts—and test selected five-input families under the same profile-class holdout. In parallel, inequality search can test whether a small combination sandwiches exact complexity within a basis-dependent additive or multiplicative error. Success would turn the observed semantic construction potential into a mathematical theory; failure would localize the remaining obstruction.

Reproducibility. All targets, 66 features, profile classes, folds, and permutation trials are checked in with the companion experiment. Five folds and eight permutations per fold use fixed seeds; no cloud service or model API is required.

References

[1] S. Jukna, *Boolean Function Complexity: Advances and Frontiers*. Springer, 2012.

[2] H. Buhrman and R. de Wolf, "Complexity measures and decision tree complexity: a survey," *Theor. Comput. Sci.*, 288(1):21–43, 2002.

[3] R. Impagliazzo, "The effect of random restrictions on formula size," *Random Struct. Algorithms*, 4(2):121–134, 1993.

[4] A. Chandra and G. Markowsky, "On the number of prime implicants," *Discrete Mathematics*, 24(1):7–11, 1978.

[5] R. O'Donnell, *Analysis of Boolean Functions*. Cambridge University Press, 2014; revised arXiv edition, 2021.

[6] J. H. Friedman, "Greedy function approximation: a gradient boosting machine," *Annals of Statistics*, 29(5):1189–1232, 2001.