Can Semantic Learning Tighten a Mathematical Bound?

A Frozen Prospective Test in Boolean Formula Complexity

J. R. Landers

Abstract

Machine learning can predict exact Boolean formula complexity surprisingly well from semantic properties of a truth table. Prediction, however, is not a bound. We test a more demanding use: predict how much slack remains in a certified constructive upper bound, subtract a held-out safety correction, and tighten the bound only when a useful margin remains. The model uses 23 truth-table and construction-aware features, is trained on the complete four-input universe and known five-input cases through exact cost 12, and is calibrated on a frozen cost-13 sample. Before computing cost 14, we freeze the model, three risk-stratified corrections, sampling rule, and success criteria. Exhaustive enumeration then discovers 89,637,411 functions of exact cost 14. On 120 prospectively selected cases, the method lowers the constructive upper endpoint by 0.879 gates on average and improves 82.5% of functions. Yet it covers the exact answer only 95.0%, below the prespecified 99% target; all six misses occur among the 80 functions selected for the largest predicted improvements, and the largest miss is 0.401 gate. The experiment therefore finds real semantic signal but rejects the proposed calibration rule. The lesson is constructive: learning can identify where a bound is loose, while safe refinement requires calibration that explicitly matches both complexity growth and model-based selection.

1 From prediction to a sharper question

For a Boolean function $f : \{0,1\}^n \to \{0,1\}$, let $K(f)$ be the minimum number of gates in a tree formula over $\mathrm{AON} = \{\mathrm{AND}, \mathrm{OR}, \mathrm{NOT}\}$. Exact synthesis is easy to state and hard to scale: enumerate every function reachable with zero gates, then one gate, and so on, retaining a function only when it first appears. The first layer containing f gives $K(f)$. The number of reachable functions grows quickly enough that even five inputs become a serious finite computation; standard background on formula measures and bounds is given in [1].

Earlier experiments changed what could be asked of these exact data. A 66-coordinate *semantic construction profile*—influences, sensitivity, certificates, decision trees, algebraic normal form, Fourier structure, prime cubes, restrictions, and hypercube topology—predicts much of the variation in exact four-input formula size. A later, compact model placed exact size inside a classical lower–upper envelope and used held-out residuals to form much narrower empirical intervals. These results established that semantic structure carries cost information. They did not establish a new upper bound: a prediction may be accurate on average and still fall below the truth.

The present experiment asks the next, sharper question. Suppose a separate mathematical construction already proves

$$K(f) + 1 \le U(f). \tag{1}$$

Can a learned model identify, with a calibrated safety margin, how much of $U(f)$ is unnecessary? This shifts the target from complexity itself to the *residual slack*

$$s(f) = U(f) - \big(K(f) + 1\big) \ge 0. \tag{2}$$

The construction U is the best of the existing analytic endpoint, exact four-variable restriction constructions, and disjoint-variable factorizations. It is valid independently of the model. Learning enters only after this certified baseline has been computed.

This distinction creates a useful scientific fork. If semantic features do not predict s, then their earlier success concerns localization but not bound improvement. If they do predict s but calibration fails, then the signal is useful for discovery but is not yet a certificate. Only prediction *and* transportable one-sided calibration would support a new empirical bound rule.

2 Selective residual-slack refinement

Let $\phi(f) \in \mathbb{R}^{23}$ be the frozen feature vector. It contains the Khrapchenko boundary product, decision-tree leaf count, four certificate statistics, degree-wise ANF support, degree-wise counts of negative Fourier coefficients, AND/OR factorability, canalization, analytic-envelope width, and the gains supplied by restriction and factor constructions. These are semantic in the relevant sense: they are computed from the function, not from a minimum formula supplied as input. The exact gate count appears only as a training or calibration label.

A histogram gradient-boosted regressor M [2] predicts

$$\widehat{s}(f) = \max\{0, M(\phi(f))\}.$$

The model is trained on all four-input functions and the available exact five-input observations through cost 12. Cost 13 is held out for calibration. Predictions are divided into three observable strata,

$$g(f) = \begin{cases} \text{low}, & \widehat{s} \le 2, \\ \text{medium}, & 2 < \widehat{s} \le 3, \\ \text{high}, & \widehat{s} > 3. \end{cases}$$

Within stratum g, the dangerous residual is $d(f) = \widehat{s}(f) - s(f)$: positive d means the model proposes removing more slack than actually exists. The correction is

$$q_g = \max_{f \in \mathcal{C}_g} (d(f))_+ + \delta, \tag{3}$$

where $\mathcal{C}_g$ is the cost-13 calibration subset and $\delta = 0.260819$ is the largest violation observed in the preceding cost-12-to-cost-13 transition. The learned reduction and selective endpoint are

$$r(f) = \left(\widehat{s}(f) - q_{g(f)}\right)_+, \tag{4}$$

$$U_{\text{sel}}(f) = U(f) - r(f). \tag{5}$$

with reductions below 0.25 set to zero. This abstention rule avoids making numerically tiny claims.

Proposition 1 (What is automatic, and what is not)**.** *For every function, $U_{\text{sel}}(f) \le U(f)$, so the method never loosens the construction endpoint. It is a valid upper bound exactly when $r(f) \le s(f)$. Equation (3) guarantees this on the calibration data, but does not by itself guarantee validity on a later complexity layer.*

Proof. Since $r \ge 0$, the first claim follows from (5). Using (2), $K + 1 \le U - r$ is equivalent to $r \le U - (K + 1) = s$. The maximum in (3) makes $\widehat{s} - q_g \le s$ within each calibration stratum; the final statement follows because cost-14 residuals were not used in q_g. □

The proposition is deliberately modest. The original U remains a proved bound. The smaller U_{sel} is a prospectively testable empirical refinement. Calling it certified before a dimension- or layer-general error argument would confuse the role of the construction with the role of the model.

3 Frozen prospective test

The complete protocol was serialized before cost-14 labels were generated: feature order, model hash, random seed, stratum thresholds, corrections, abstention threshold, cohort sizes, and success criteria. Exact enumeration found 89,637,411 functions first reached at cost 14 and 173,980,568 functions cumulatively through cost 14. Enumeration took 14,258.7 seconds on a single-threaded cloud run.

From the exact cost-14 layer, a seeded random candidate pool of 1,000 functions was drawn. The *targeted* cohort contains the 80 candidates with the largest positive proposed reductions. A disjoint random-reference cohort contains 40 functions. Selection used predictions and semantic features, never cost-14 errors. Success required at least 99% overall coverage, maximum violation at most one gate, positive mean reduction beyond the construction, and improvement on at least 10% of functions.

Table 1: Frozen cost-14 results. A miss means $K(f) + 1 > U_{\mathrm{sel}}(f)$. Reductions are measured beyond the certified construction endpoint.

Cohort	N	Coverage	Misses	Mean reduction	Worst miss
Targeted	80	92.5%	6	1.1744	0.4007
Random reference	40	100.0%	0	0.2895	0
Overall	120	95.0%	6	0.8795	0.4007

Table 1 gives a clean but mixed result. The model finds substantial opportunity: it improves 99 of 120 functions (82.5%) and removes 0.8795 gates from the construction endpoint on average. Its ranking is also meaningful. Every targeted function is improved, and the targeted mean reduction is four times the random-reference reduction. Semantic structure therefore predicts where the constructive bound is loose.

The calibration claim fails. Six targeted endpoints fall below the exact answer, giving 95.0% rather than the required 99% coverage. All random references are covered. The violations are small—the maximum is 0.4007 gate—and remain below the prespecified one-gate severity limit, but their count is decisive. The frozen protocol therefore fails as a whole. A rule cannot trade away its coverage requirement merely because the misses are numerically modest.

This result extends an instructive progression. The preceding frozen cost-13 experiment covered 119 of 120 functions (99.17%), improved the construction by 0.6139 gate on average, and had one 0.2608-gate miss. Treating that miss as cross-layer drift and adding it to every cost-13 stratum produced perfect calibration-set coverage, but did not transport to the selected cost-14 tail. Mean improvement grew from 0.6139 to 0.8795 exactly as coverage fell. The model became more useful and less safe at the same time.

4 What the failure teaches

The simplest reading is not that semantic learning failed. If the predictor were noise, the targeted cohort would not show much larger reductions than a random cohort. Rather, the calibration set and test query ask different questions. Equation (3) controls the largest observed error inside broad prediction strata. The prospective test then deliberately chooses the most aggressive predictions from 1,000 new functions at a harder exact-cost layer. That top-ranked conditional tail need not resemble the average member of the same low/medium/high stratum. This is selection-induced distribution shift, combined with layer-to-layer drift.

The six misses also sharpen the next formal target. A useful refinement theorem should control

$$\Pr(r(f) > s(f) \mid f \text{ is selected by the policy at layer } k)\,,$$

not merely an unconditioned or coarsely stratified residual. Possible routes include calibration on the complete selection pipeline, finer strata indexed by predicted slack and construction type, or a proved

upper model for how the dangerous residual changes with exact-cost layer. General split-calibration methods provide a starting point [3, 4], but their sampling assumptions must match this selected, cross-layer query. A post-hoc extra correction of 0.4007 would cover this particular sample, but it would be a diagnostic, not prospective evidence; cost 14 has already been observed.

There is also a computational lesson. The exact enumerator recomputes large Python sets serially and reports no within-layer progress. The four-hour run was valid but operationally opaque. Checkpointed layers, compact bitsets or sorted integer arrays, and parallelized gate products would make the next test both cheaper and easier to audit. Better engineering will not repair calibration, but it will allow more informative prospective tests.

5 Conclusion

This experiment crosses an important boundary. Earlier work showed that semantic features predict Boolean formula complexity and can narrow analytic envelopes. Here the same idea is asked to remove slack from a certified construction on a harder, untouched exact layer. It succeeds at discovery: the model reliably locates functions with larger removable slack and tightens most sampled endpoints. It fails at safety: a correction that looked conservative through cost 13 covers only 95% after targeted cost-14 selection.

The negative result is the useful result. It separates two claims that are easy to blur. Semantic learning can guide the search for better bounds; current calibration does not yet make those improvements uniformly reliable. The next advance should preserve the learned ranking while calibrating the actual selective query above. That is a narrower and more mathematical goal than simply fitting a more accurate predictor.

Reproducibility. The exact predictions, cohort labels, frozen protocol and model hashes, analysis JSON, cloud log, and generated report are stored with the companion experiment. An independent test suite recomputes the reported metrics and checks that the learned endpoint never exceeds the construction endpoint.

References

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

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

[3] V. Vovk, A. Gammerman, and G. Shafer, *Algorithmic Learning in a Random World*. Springer, 2005.

[4] J. Lei, M. G'Sell, A. Rinaldo, R. Tibshirani, and L. Wasserman, "Distribution-free predictive inference for regression," *J. Amer. Stat. Assoc.*, 113(523):1094–1111, 2018.