CERTIFIED CONTRAST EXPLANATIONS FOR NEURO-SYMBOLIC TREATMENT-EFFECT MODELS

Anonymous authors
Paper under double-blind review

ABSTRACT

We study treatment-effect models through four desiderata: predictive adequacy, intrinsic interpretability, effect-specific informativeness, and verifiable fidelity. Our models combine sparse modified Łukasiewicz logic circuits with compact symbolic refinements discovered from held-out residuals. Logic-S shares one circuit across treatment arms. Hard literal selection controls circuit complexity, while the refinement adds expressive capacity through a few permitted nonlinear terms. We introduce a contrast-aware pruning certificate that identifies shared outcome structure whose deletion preserves the fitted treatment effect throughout a declared input domain. The guarantee accounts for nonlinear clamps, outcome scaling, crossfit aggregation, and accepted symbolic corrections. We evaluate this architecture on 1,000 IHDP simulations and 135 ACIC variants; paired IHDP ablations show that symbolic refinement improves the centered effect surface. In an audit of 30 IHDP simulations, Logic-S ensembles permit a median 77.4% reduction in logically simplified base-circuit nodes at zero certified contrast error; a matched outcome-preserving control permits no further deletions. A historical stroke-trial case study illustrates how to compare estimators against randomized outcomes, inspect the learned benefit structure, and verify that the simplified explanation preserves the fitted contrast. Exact specialization of a factorial Logic-S fit separates aspirin, heparin, and interaction contrasts. It estimates a 1.55-point aspirin risk reduction against a randomized 1.99-point estimate, a near-zero net heparin effect, and aspirin–heparin interactions whose five-seed consensus is bounded throughout the input box within 0.24 percentage points of zero. A separate compact aspirin fit exposes output saturation as the source of its small displayed variation.

1 INTRODUCTION

Since the late 2010s, much work on heterogeneous treatment effects (HTE) has focused on black-box estimators. Their appeal is the ability to capture complex nonlinearities and scale in ways that may evade traditional or more interpretable methods.

HTE estimators are used to investigate which covariate-defined groups respond differently to treatment, and which interactions govern those differences.

Since 2016, the literature has largely pursued marginal gains in predictive performance on semi-synthetic benchmarks. Many approaches use doubly robust estimation, which can remain consistent when either the treatment-assignment model or the outcome model is correctly specified, under the required identification assumptions (Bang & Robins, 2005). Such guarantees still leave considerable freedom to put most available variables into a black box. This dependence on nonlinearities, without meaningful ways to explain them, poses risks. Outside careful specification, spurious interactions may govern the heterogeneity practitioners wish to model. Benchmark design and evaluation metrics can also favor particular inductive biases, leaving real-world uses of conditional average treatment effect (CATE) estimates untested (Curth et al., 2021; Crabbé et al., 2022). Post-hoc explainers such as SHAP and LIME (Lundberg & Lee, 2017; Ribeiro et al., 2016) offer little insight into learned causal structure beyond variable importance. Black-box models may capture causal structure well, but without methods to inspect that structure, practitioners have little reason to trust them beyond their expert initialization. Benchmarking is useful, but it should be one star in a constellation of

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

criteria for model selection. We therefore ask what a treatment-effect model must provide to be useful as an object of causal interpretation.

We propose four desiderata. **Predictive adequacy** requires treatment-effect estimates competitive with strong statistical and neural baselines, with sufficient stability and expressivity to model the underlying causal mechanisms. **Intrinsic interpretability** requires an explicit symbolic representation of the fitted CATE that can be inspected directly. **Effect-specific informativeness** requires exposing the covariates and interactions through which the model expresses effect heterogeneity, distinguishing them from shared outcome structure. **Verifiable fidelity** requires a checkable guarantee that simplifying the explanation preserves the fitted treatment contrast over a declared input domain.

A correctly specified linear treatment-effect model can satisfy these properties in a simple regime. Our aim is to retain them within a richer function class.

Our architecture builds on the symbolic computation of Neural Reasoning Networks (NRNs; Carrow et al., 2025). R-NRN omits inactive branches to produce concise sample-level explanations. Our target is the complete treatment contrast over a declared domain, where structural branches cannot be ignored because they are inactive for one input. We implement a small logical DNF block with hard top-K literal selection and straight-through gradients, encoding the inductive bias that a few short, weighted rules can explain much of the relevant variation. Logic-S uses one treatment-conditional circuit. A restricted symbolic search on held-out residuals adds a few explicit terms. The complete deployed contrast remains symbolic, including crossfit aggregation and accepted symbolic corrections. On our benchmark configurations, the deployed Logic-S ensemble uses fewer than 3% of DragonNet's fitted parameters while remaining competitive in predictive error.

Our central technical contribution is a certificate for contrast-preserving simplification. It identifies shared outcome structure whose deletion has equal effects on both arms throughout the declared domain. The proof accounts for nonlinear clamp boundaries, arm-specific outcome scales, and the composition of cross-fitted models. Exact rational arithmetic preserves the certified symbolic object through simplification and serialization.

Shared IHDP and ACIC benchmarks evaluate prediction; paired ablations isolate symbolic refinement's contribution to average and heterogeneous effects. A matched certification study and worked example measure and illustrate zero-error contrast simplification. In a historical randomized stroke trial, a factorial fit can be specialized into aspirin, heparin, and aspirin–heparin interaction contrasts. It recovers a small aspirin benefit, no meaningful net heparin effect on six-month death or dependency, and a uniformly small interaction. A separate compact aspirin fit has a nearly constant contrast across the observed cohort; inspection attributes its small variation to output saturation. The case study connects fidelity to a practical use by making both positive and null learned effects available for substantive scrutiny.

2 Related Work

Logical Neural Networks, Logic Tensor Networks, and NRNs connect neural computation with explicit logical semantics (Riegel et al., 2020; Badreddine et al., 2022; Carrow et al., 2025). We adopt this architectural approach for outcome models. R-NRN uses sparse connectivity, bandit-guided pruning, and regrowth to provide simplified sample-level explanations. Its global logic can remain dense. The reported API examples make global displays manageable by restricting them to the top 10% of logic by importance (Carrow et al., 2025). A faithful global representation is of limited practical use when it cannot be inspected. We enforce parsimony by construction and expand expressivity with a symbolically searched correction. Our pruning certificate preserves the complete numerical contrast throughout a declared input domain, including paths that are inactive for a given input. The reported NRN experiments focus on classification; here related logical units model continuous outcomes. NESTER uses a domain-specific language to compose learned neural modules for effect estimation (Reddy & Balasubramanian, 2024). The program exposes module composition, while the modules' internal outcome mappings can remain opaque neural networks.

Bayesian Causal Forests separately regularize prognosis and effect modification, allowing shrinkage toward homogeneous effects (Hahn et al., 2020); neural estimators also encode shared response structure (Curth & van der Schaar, 2021). Neural Additive Models (NAMs) expose feature-response

functions (Agarwal et al., 2021). Causal variants distinguish prognostic and moderating functions (Caron et al., 2022) or combine additive outcome models with covariate balancing (Chen et al., 2025). Causal Rule Forest learns Boolean representations for effect estimation (Hsu et al., 2024); rule ensembles support interpretable subgroup discovery (Bargagli-Stoffi et al., 2024) and separate prognostic and treatment rules (Hiraishi et al., 2024). causalKANs uses pruning and symbolic approximation of learned splines (Almodóvar et al., 2025).

DISCRET synthesizes patient-specific subgroup rules used directly for effect estimation, evaluating explanation-prediction consistency (Wu et al., 2024). Our guarantee concerns numerical fidelity. Deleting fitted outcome computation preserves the complete contrast over a declared domain.

Symbolic regression directly supplies explicit expressions (Schmidt & Lipson, 2009; Cranmer, 2023). Our refinement step uses restricted symbolic search to select terms from a fixed dictionary over a screened feature set. This limits search and expression complexity. We compare standalone restricted search on the full IHDP and independent synthetic cohorts, with evolutionary search in a separate timing pilot (Appendix C.7). Our restricted symbolic search performs comparably to gplearn's symbolic regression on our selected benchmarks, and is often faster.

Neural verification methods such as CROWN and DeepPoly propagate bounds over input regions (Zhang et al., 2018; Singh et al., 2019). ReluDiff exploits relationships between networks to bound their output differences (Paulsen et al., 2020), and lumpability-based compression preserves network behavior exactly (Ressi et al., 2024). Our certificate specializes verification to branch deletion in a treatment contrast, permitting outcome levels to change when their raw-unit contributions cancel across arms, accounting for clamps, scales, and crossfit aggregation.

Our predictive evaluation uses causal forests, TARNet, and DragonNet (Wager & Athey, 2018; Shalit et al., 2017; Shi et al., 2019).

3 Sparse Logic Treatment-Effect Models

We formulate treatment effects in an acyclic structural causal model (Pearl, 2009). Let $X \in [0,1]^d$ denote d preprocessed pretreatment covariates, $T \in \{0,1\}$ treatment, and $Y \in \mathbb{R}$ the observed outcome. The intervention $\mathrm{do}(T=t)$ replaces the treatment assignment mechanism with the constant t, leaving the remaining structural assignments unchanged. Define

$$\mu_t(x) = \mathbb{E}[Y \mid \mathrm{do}(T=t), X=x], \qquad \tau(x) = \mu_1(x) - \mu_0(x). \tag{1}$$

Here $\tau : [0,1]^d \to \mathbb{R}$ is the CATE. If X blocks all back-door paths from T to Y and $0 < \Pr(T = 1 \mid X = x) < 1$ on the population support, then $\mu_t(x) = \mathbb{E}[Y \mid X = x, T = t]$. We fit these identified outcome regressions by factual squared error and retain an explicit expression for their contrast. Identification rests on the assumed causal structure; the learned circuit is a computational DAG, and its simplification certificate concerns the fitted contrast.

Logic backbone. The backbone is a weighted disjunctive-normal-form (DNF) block of C conjunctions feeding one disjunction. Numerical inputs are imputed and min–max scaled on training data, with later values clipped to $[0,1]$; categorical inputs are one-hot encoded. A block takes $u \in [0,1]^{d_u}$. Logic-S uses $u = (x,t)$, $d_u = d+1$, specializing the same fitted circuit at each treatment value.

Clause $c \in \{1,\dots,C\}$ learns logits $\alpha_c \in \mathbb{R}^{d_u}$ and weights $w_c \in \mathbb{R}^{d_u}$. Its mask $m_c \in \{0,1\}^{d_u}$ selects exactly $K \le d_u$ coordinates globally, reserving required slots first. For Logic-S, treatment is required in the first two clauses with opposite polarities and weight magnitude at least 0.05. Other memberships and signs are learned.

Write $\ell_{cj}(u) = u_j$ if $w_{cj} \ge 0$ and $1 - u_j$ otherwise, and $[v]_0^1 = \min(1, \max(0, v))$. The forward computation is

$$h_c(u) = \left[1 - \sum_{j=1}^{d_u} m_{cj} |w_{cj}| \{1 - \ell_{cj}(u)\}\right]_0^1, \qquad f_\theta(u) = \left[\sum_{c=1}^{C} \gamma_c h_c(u)\right]_0^1, \tag{2}$$

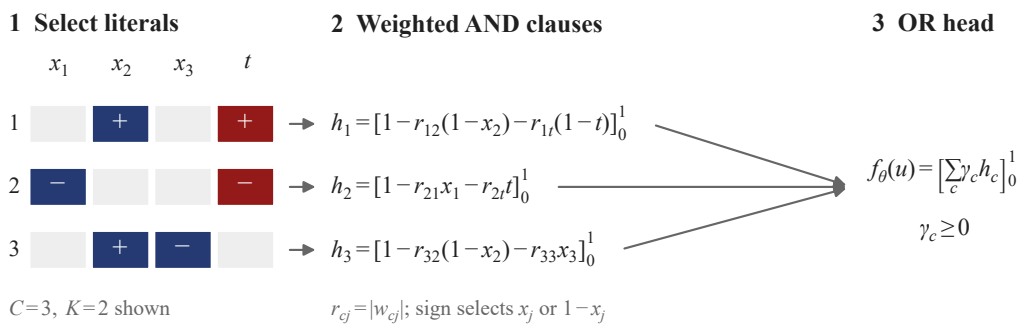


Figure 1: Logic-S DNF block with $d = 3$, $d_u = 4$, $C = 3$, and $K = 2$. Hard masks select literal membership; $r_{cj} = |w_{cj}|$ denotes strength, with $j = t$ labeling the treatment coordinate. Its polarity is opposite in the first two clauses. The nonnegative OR head combines clause outputs.

Here $v \in \mathbb{R}^C$ is the raw head vector, $\gamma_c = \max(0, v_c)$, and $\theta = (\alpha, w, v)$ collects all parameters; h_c, f_θ are scalar maps to $[0, 1]$. The AND offset is one and the OR offset is zero; there is no additional learned bias or hidden feature extractor.

Learning discrete structure. Masks are hard during training and inference. We use straight-through gradients (Bengio et al., 2013), AdamW (Loshchilov & Hutter, 2019), and validation factual MSE to select checkpoints and restarts. Appendix A gives the gradient surrogates, loss, regularization, and training budgets.

Residual-guided symbolic refinement. On discovery rows (x_i, t_i, y_i), let $r_i = y_i - \hat{\mu}_{t_i}(x_i)$ be the scalar residual of the fitted base arm $\hat{\mu}_t : [0, 1]^d \to \mathbb{R}$. Rank covariates by $\max_t |\operatorname{corr}(x_j, r \mid T = t)|$ and retain a fixed number of features. This residual screen is separate from the DNF's learned masks. The dictionary contains individual coordinates, squares, cubes, $\sin(\pi x_j)$, $\cos(\pi x_j)$, pairwise sums, differences, products, and $\exp(a x_j x_k)$ for $a \in \{1, 1.5, 2\}$. Greedy selection adds terms by treatment-stratified cross-validation of factual residual error, with a complexity penalty. For m selected terms, $z : [0, 1]^d \to \mathbb{R}^m$, scalar intercepts b_0, b_t, and coefficient vectors $b_z, b_{tz} \in \mathbb{R}^m$, refit

$$c(x, t) = b_0 + z(x)^\top b_z + t\{b_t + z(x)^\top b_{tz}\}. \tag{3}$$

The small and large profiles retain at most four and six terms, respectively. Discovery-improvement and held-out acceptance-loss thresholds determine deployment. Accepted additive models return $\hat{\tau}(x) = \hat{\tau}_b(x) + b_t + z(x)^\top b_{tz}$; rejection returns the base, whose contrast is $\hat{\tau}_b = \hat{\mu}_1 - \hat{\mu}_0$. ATE is the mean of this contrast over the evaluation population. Both stages use factual outcomes without propensity or doubly robust targets. Correction coefficients use ridge penalty 0.01 on fold-standardized terms, with unpenalized intercepts, throughout search and fitting (Appendix C.2). Appendix A specifies thresholds, data roles, and the arm-predicate variant, which inserts terms into a refitted circuit.

Crossfit deployment. The CF3 variant fits three treatment-conditional blocks on treatment-stratified development folds. Held-out-fold predictions supply residuals for symbolic search. Outer-fold comparisons supply acceptance evidence, followed by one final correction fit on all development residuals. Held-out outcomes also select base checkpoints, so intermediate estimates are not strictly independent. Test outcomes remain untouched. Deployment averages base outcomes in original units and adds the correction:

$$\hat{\tau}(x) = \frac{1}{F} \sum_{r=1}^{F} \{a_{r1} - a_{r0} + s_{r1} f_{r1}(x) - s_{r0} f_{r0}(x)\} + \Delta(x), \tag{4}$$

Here r indexes $F = 3$ folds; $f_{rt} : [0, 1]^d \to [0, 1]$ is the arm-specialized circuit, $a_{rt} \in \mathbb{R}$ its outcome offset, and $s_{rt} > 0$ its outcome scale. Within each Logic-S fold, $a_{r0} = a_{r1}$ and $s_{r0} = s_{r1}$. The scalar correction $\Delta(x) = b_t + z(x)^\top b_{tz}$ is zero upon rejection. Algorithms 1–3 give executable pseudocode and source pointers in Appendix A.1.

4 CERTIFIED CONTRAST SIMPLIFICATION

We compile each outcome block into a directed acyclic graph (DAG) of variables, constants, negations, scalar dilations, and clamped sums. Treatment specialization yields an arm pair for each fold. The certificate covers $D = [0,1]^d$, including the continuous relaxation of categorical coordinates. Shared outcome structure can be removed when its deletion changes both arms equally in outcome units. Sharing alone is insufficient because downstream clamps can respond differently in the two arms. We consider only deletions that preserve the contrast exactly on D.

Proposition 1 (Soundness of contrast-preserving deletion)**.** *Let $f_0, f_1 : D \to [0,1]$ be arm circuits, with offsets $a_0, a_1 \in \mathbb{R}$, scales $s_0, s_1 > 0$, and contrast $\tau_b = a_1 - a_0 + s_1 f_1 - s_0 f_0$. For a candidate branch $v : D \to [0,1]$, suppose the interval-and-response procedure in Appendix B returns resolved coefficients $c_0, c_1 \in \mathbb{Q}$. An arm without the candidate has coefficient zero. If $s_1 c_1 = s_0 c_0$, replacing every occurrence of the candidate by zero gives a reduced contrast $\tilde{\tau}_b : D \to \mathbb{R}$ satisfying*

$$\tilde{\tau}_b(x) = \tau_b(x) \qquad \textit{for every } x \in D.$$

Appendix B proves the response rules by structural induction. Full deletion changes arm t by $-c_t v(x)$. The scaled changes then cancel in the contrast. The procedure encloses the full deletion segment and propagates intervals and response coefficients through both arms. Unresolved candidates are retained, and bounds are recomputed after each accepted deletion. The test is sufficient and conservative; it need not find a smallest equivalent circuit. Separate outcome bounds instead give $\|v\|_\infty(|s_1 c_1| + |s_0 c_0|)$, which can be positive even when the contrast change is zero, where $\|v\|_\infty = \sup_{x \in D} |v(x)|$.

Ensembles and correction. Each accepted deletion preserves the current fold contrast, so any sequence of accepted deletions preserves the original. With the correction, offsets, and averaging weights unchanged, averaging the preserved fold contrasts gives $\tilde{\tau}(x) = \hat{\tau}(x)$ for every $x \in D$. Here $\hat{\tau}, \tilde{\tau} : D \to \mathbb{R}$ are the original and reduced deployed contrasts. Their ATEs agree under any distribution supported on D, although individual outcome levels may change. Appendix B gives the derivation.

Post-fit organization. With the correction split in Algorithm 5, the deployed arms admit $F_t = P + C_t$, where $F_t, P, C_t : D \to \mathbb{R}$, $t \in \{0,1\}$, and $C_1 - C_0 = \hat{\tau}$. The shared contrast-null component P records the removed outcome structure alongside the retained contrast representation. This organization requires no retraining and is not claimed to be unique or minimal.

The certificate applies to the exported DAGs in real arithmetic, interpreting binary floating-point coefficients as exact rationals. Floating-point evaluation and agreement with the neural implementation are checked separately. Appendix B gives the proofs and numerical checks; Algorithm 4 specifies the pruning procedure.

5 EVALUATION

Data and protocol. We evaluate 1,000 IHDP response-surface-B simulations (Hill, 2011) generated using the NPCI R package (Dorie, 2016), and 135 ACIC-2018 variants with fewer than 10,000 observations (Shimoni et al., 2018). The heterogeneous ACIC comparison uses the 61 variants with nonconstant ground-truth training CATE; we also report all 135 variants. For IHDP, ACIC, and the independent synthetic setups, a treatment-stratified split assigns 10% to a common untouched test set; methods differ only in how they use the remaining development rows (Appendix C). IHDP and ACIC use fitting seeds 42–44 on fixed splits; synthetic comparisons and certification use seed 42. We normalize PEHE and absolute ATE error by training-outcome SD. Primary means average seeds within each simulation, then simulations equally; case-bootstrap intervals retain all three seeds. Outcome surfaces vary significantly in scale, which is why we normalize for reporting purposes, however, the models themselves predict on the original scale. Appendix C gives the configurations, data budgets, full tables, and uncertainty estimates. Certification uses the first 30 generated IHDP simulations; the certificate is conditional on the fitted model and does not depend on how its coefficients were fitted.

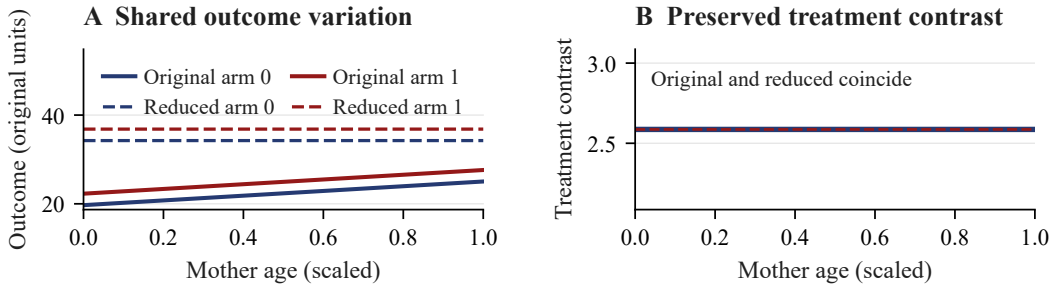


Figure 2: Third fold of the illustrative Logic-S (CF3) ensemble. The original outcome curves (solid) and reduced arm representations (dashed) differ by a shared component, preserving their contrast. Other inputs are fixed at training medians; mother's age gives the largest outcome range among the continuous inputs at this anchor. The certificate covers the full input box. Reduced arms explain the contrast; the fitted outcome predictor is unchanged. The full ensemble also retains the final correction.

5.1 Predictive adequacy

We use these benchmarks to assess predictive adequacy alongside established estimators. Tables 9 and 10 report mean normalized errors across all three seeds and all cases, without trimming. On IHDP, Logic-S + correction (CF3) has mean PEHE 0.346, compared with 0.250 for DragonNet + TARReg and 0.290 for treatment-structured SR. On all 135 ACIC variants, single-fit Logic-S + correction has mean PEHE 0.271, compared with 0.340 for TARNet; S-CF3 has mean 0.205. On heterogeneous ACIC, the respective S and S-CF3 means are 0.334 and 0.216. S-CF3 has mean absolute ATE error 0.052 on heterogeneous ACIC and 0.059 on all ACIC variants, compared with TARNet's 0.052 and 0.094.

Under fixed CPU fitting conditions across ten seeds, IHDP S-CF3 has prediction-ranking correlation 0.918 but exact correction-term overlap 0.184. Thus prediction stability need not imply a stable symbolic expression (Appendix C.3).

5.2 Intrinsic interpretability

The representation includes the fitted rules, all coefficients, and the complete deployment composition. For an illustrative simulation from the certification cohort, selected for compact exposition, the S ensemble reduces from 257 to 42 distinct base nodes, while retaining its six-term correction. Its full effect is

$$\hat{\tau}(x) = \frac{1}{3}\sum_{r=1}^{3} s_r d_r(x) + \Delta(x), \tag{5}$$

Here $x \in [0,1]^{25}$, $s_r > 0$ is the fold's outcome scale, $d_r : [0,1]^{25} \to [-1,1]$ its unscaled arm difference, and Δ the correction contrast. The first two d_r use three literals each; the third is constant. Appendix E.4 gives their expressions and the six correction terms with coefficients. Formula size counts distinct DAG nodes, with correction terms reported separately; it is not a direct measurement of human comprehension.

5.3 Effect-specific informativeness

Figure 2 shows why inspecting the contrast is useful. The third fold's outcome functions admit the decomposition

$$\hat{\mu}_t^{(3)}(x) = p_3(x) + a_3 + s_3 c_{3t}, \qquad \hat{\tau}^{(3)}(x) = s_3(c_{31} - c_{30}). \tag{6}$$

Here $p_3 : [0,1]^{25} \to \mathbb{R}$ is the removed shared outcome component, $a_3 \in \mathbb{R}$ the outcome offset, and $c_{3t} \in [0,1]$ the retained arm constant. Mother's age changes both outcome predictions through shared structure $p_3(x)$ and has no effect on this fold's contrast. It still appears in the second fold

Table 1: Logic-S model size through the certification pipeline on 30 IHDP ensembles: medians [95% bootstrap intervals]. Counts aggregate all crossfit base circuits. Compiled precedes exact rewriting; Logical follows outcome-preserving rewriting and arm specialization; Contrast adds zero-error cross-arm pruning. Corrections remain outside the base graph.

Model	Base params	Compiled nodes	Logical nodes	Contrast nodes
Logic-S + correction (CF3)	2,544	671.5 [669.5, 674.5]	409.0 [367.0, 450.0]	96.5 [45.0, 185.0]

and in the correction. The explicit representation makes these component-specific roles available for inspection.

We quantify the correction's contribution across the 1,000 IHDP cases at seed 42. On n test rows, let $\hat{\tau}_i = \hat{\tau}(x_i)$, $\tau_i = \tau(x_i)$, $e_i = \hat{\tau}_i - \tau_i$, and $\bar{e} = n^{-1} \sum_{i=1}^{n} e_i$. The identity

$$\mathrm{PEHE}^2 = \frac{1}{n} \sum_{i=1}^{n} e_i^2 = \bar{e}^2 + \frac{1}{n} \sum_{i=1}^{n} (e_i - \bar{e})^2$$

separates average-effect error from error in the centered effect surface. Adding the accepted correction reduces centered RMSE by a paired median 0.527 (95% CI $[0.510, 0.545]$) for Logic-S CF3, with PEHE improving in 885/1,000 cases. These comparisons isolate refinement of the same fitted base; the centered-RMSE reduction shows that the correction improves heterogeneity estimation.

IHDP also provides known effect-modifier sets. Among the 892 nonempty accepted S-CF3 corrections, median precision of named covariates is 0.857 (95% CI $[0.833, 0.875]$). Paired excess precision above each simulation's random-selection chance rate is 0.420. Recall across all 1,000 cases, counting empty corrections as zero, is 0.700. This evaluates covariates named by nonzero correction terms; the complete effect representation includes the base as well.

On the matched 30-case S certification cohort, the complete contrast's median named set shrinks from 25 to 15 covariates, including all folds and the correction. Removing a name establishes that the learned contrast is independent of that input on the declared box. Retained names may still cancel functionally. The median logical, contrast, and identified-null counts are 409, 96.5, and 284 nodes. The baseline component names all 25 covariates, while the contrast representation names 15 and leaves a median 10 baseline-only; variables may occur in both components. Appendix D gives definitions, intervals, and support audits.

On heterogeneous ACIC, the correction is accepted in 25/61 S-CF3 cases. PEHE improves in 14 and worsens in 11; the remainder retain the base. Factual validation can therefore select refinements that do not improve CATE, a qualification we retain when interpreting effect estimates.

5.4 Verifiable fidelity

Table 1 follows the Logic-S ensemble through compilation, logical rewriting, and zero-error contrast pruning, retaining accepted corrections. TARNet and DragonNet provide single-fit parameter comparisons under the same input dimension. Including the maximum 14 coefficients of a six-term correction, the deployed S ensemble contains at most 2,558 fitted scalars, approximately 57 times fewer than DragonNet's 146,404. Logic counts aggregate all crossfit folds. Ordinary logical rewrites remove a median 39.0% of compiled S nodes. The subsequent contrast stage removes a further 77.4% of the S representation, with reductions in 26/30 cases. The full staged reduction is 85.6%. Every returned contrast has an exact zero bound over the declared box.

An outcome-preserving branch-deletion control, applied after logical simplification with the same candidates, interval tests, and size criterion, permits no additional deletions. The S contrast reductions therefore depend on equal nonzero responses across arms (Appendix B). Numerical evaluation and neural-export checks are reported separately in Appendix B.

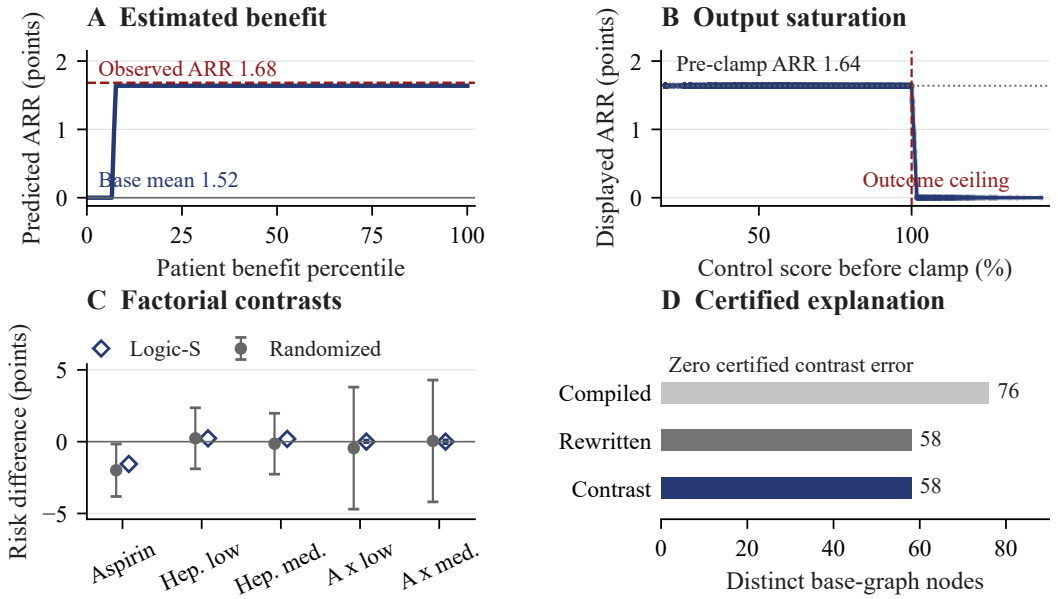


Figure 3: Stroke-trial case study. A–B trace constant pre-clamp aspirin benefit into variation at the outcome ceiling. C compares randomized factorial risk differences (95% CIs) with Logic-S; blue interaction bars are whole-box solver bounds. D shows exact rewriting and certification. A–B and D use the restricted aspirin cohort; C uses the broader factorial cohort.

5.5 Factorial aspirin and heparin case study

The International Stroke Trial (IST) randomized 19,435 patients to aspirin, subcutaneous heparin, both, or neither (International Stroke Trial Collaborative Group, 1997). This factorial design allows two treatment effects and their interaction to be checked against randomized outcomes. For aspirin assignment $a \in \{0, 1\}$ and randomized heparin $h \in \{N, L, M\}$ (none, low dose, or medium dose), we specialize a fitted outcome circuit into six programs $F_{ah}(x) = F(x, \mathrm{do}(A = a), \mathrm{do}(H = h))$. They give the aspirin contrast $F_{1h} - F_{0h}$ and heparin contrasts $F_{ah} - F_{aN}$. Their interaction is $I_h = (F_{1h} - F_{0h}) - (F_{1N} - F_{0N})$. Negative risk differences lower the probability of death or dependency.

Randomized and fitted effects. The factorial audit retains 12,291 post-pilot patients with CT before randomization and observed six-month outcomes. A three-literal Logic-S model has enough capacity to place aspirin, randomized heparin, and a clinical modifier in one clause. We fit three-fold models for five procedural seeds and evaluate their consensus on one held-out set. The randomized marginal risk differences are -1.99 points for aspirin (95% CI $[-3.82, -0.17]$), $+0.23$ for low-dose heparin ($[-1.89, 2.36]$), and -0.14 for medium-dose heparin ($[-2.27, 1.98]$). Logic-S estimates -1.55, $+0.23$, and $+0.19$ points, respectively. Thus it recovers the small aspirin benefit and the trial's lack of net heparin benefit on this outcome.

The randomized aspirin–heparin interactions are -0.45 points for low dose (95% CI $[-4.70, 3.80]$) and $+0.05$ for medium dose ($[-4.19, 4.29]$). Logic-S estimates $+0.0005$ and -0.0007 points. Mixed-integer encodings find an identically zero interaction throughout the input box in 10/15 low-dose and 9/15 medium-dose folds. Foldwise solver ranges bound the consensus by $[-0.157, 0.205]$ and $[-0.239, 0.222]$ points, respectively, subject to numerical solver tolerance. This fitted magnitude certificate agrees with the IST finding of no interaction in the main outcomes (International Stroke Trial Collaborative Group, 1997); Appendix E gives the full audit.

Inspecting aspirin constancy. We separately retain 5,978 patients with no recent or randomized heparin and fit one compact Logic-S base with eight two-literal clauses and 744 parameters. Its pre-clamp aspirin benefit is 1.639 points for every observed profile; after clamping, the mean is

1.52 points (SD 0.41; range 0–1.64), near the randomized 1.68 points. Of the cohort, 92.4% retain the full value; the remainder reach the upper outcome clamp. The circuit thereby distinguishes a shared benefit from bounded-output variation, consistent with a 39,166-patient analysis that found no variation across strata of predicted thrombotic and hemorrhagic risk (Thompson et al., 2015).

The residual search's five-term correction misses its 2% factual-loss gate (1.876%). Forcing it preserves the mean but raises benefit SD from 0.41 to 6.67 points, with its largest interaction supported by 19 patients. Exact rewriting reduces the retained base from 76 to 58 nodes; saturation prevents further zero-error deletion. Symbolic and neural evaluations agree within 1.22×10^{-7}. Appendix E retains the stress test and external comparators.

6 Discussion

The main contribution is a checkable connection between a fitted treatment contrast and its reduced symbolic representation. This separates two questions that an interpretable estimator must answer. Statistical evaluation asks whether the fitted contrast estimates the treatment effect well; certification asks whether the explanation preserves what the fitted model computes. The zero-error guarantee resolves the second question on the declared domain. Identification assumptions, estimation error, and validation of treatment-effect heterogeneity remain necessary for the first.

The IHDP results show that the correction contributes substantially to predictive accuracy. The uncorrected bases have high median as well as mean PEHE, so their limitations extend beyond a few extreme runs. Short logical rules provide an inspectable starting representation, but the permitted nonlinear terms supply much of the accuracy in this setting. Because the certificate retains the accepted correction, fidelity applies to the complete estimator. Interpretability must likewise be assessed on the complete expression, including its terms, coefficients, and fold composition. A small base graph alone does not establish that the deployed model is easy to read.

A priority for future work is to align learning and model selection more directly with the treatment contrast. Our bases and correction gates optimize factual prediction, which can favor accurate shared outcome structure without improving CATE. Established approaches use causal regularization, propensity and outcome nuisance estimates, or doubly robust objectives to improve treatment-effect estimation (Shalit et al., 2017; Shi et al., 2019; Nie & Wager, 2021; Bang & Robins, 2005). The present architecture does not yet incorporate these causal learning components. Its results therefore leave open how much accuracy could improve by combining them with an explicit, certifiable contrast representation. Effect-targeted training objectives and validation criteria using cross-fitted nuisance estimates are concrete next steps. They should be evaluated jointly with correction acceptance and tail behavior, using fixed data budgets and untouched test sets. Improving the base's effect representation may also reduce the burden placed on the symbolic correction.

The seed study shows that prediction stability does not imply stability of selected expressions (Appendix C.3). Future analyses should distinguish interchangeable symbolic forms from materially different effect surfaces, and measure agreement in the complete certified contrast. The library ablation provides a concrete starting point. Restricting operator families can increase term agreement, but its predictive effects vary by setting (Appendix C.4). Libraries organized around low-order interactions, together with selection criteria that account for expression complexity and seed sensitivity, could make this tradeoff more explicit. Such choices require validation across response surfaces beyond the fixed covariate populations studied here.

The IST analysis illustrates how explicit contrasts expose small learned effects, shared outcome structure, and variation induced by output clamps. Its adapted analyses require independent clinical evaluation. More generally, practitioner studies are needed to assess whether the representations help users inspect and challenge fitted effects. On the certification side, a bounded contrast-error budget could permit further simplification when exact cancellation is unavailable. The resulting certificate would need to account for accumulated error through nonlinear circuits, outcome scaling, and ensemble aggregation while remaining separate from statistical uncertainty.

AI USE STATEMENT

The research questions, overall conceptual framing, and project direction originated with the authors. Generative AI contributed to experimental design, including proposing the Nie–Wager benchmark comparison and implementing its evaluation, and contributed substantially to developing and formalizing the contrast-preservation certificate. AI tools also assisted with implementation, debugging, analysis, manuscript editing, figures, and reproducibility materials. The authors are responsible for the final claims, proofs, experimental results, and interpretation.

REPRODUCIBILITY STATEMENT

Appendix A specifies the model and training procedure. Appendix B describes the rational certificate and tests. Appendices C and D document evaluation and paired analyses. The accompanying `logic-cate` package provides fitting, correction, and exact certification through public APIs. Simple drivers include 15 complete IHDP/ACIC example datasets, including IHDP 001–012.

REFERENCES

Rishabh Agarwal, Levi Melnick, Nicholas Frosst, Xuezhou Zhang, Ben Lengerich, Rich Caruana, and Geoffrey E. Hinton. Neural additive models: Interpretable machine learning with neural nets. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 34, 2021. URL https://papers.neurips.cc/paper/2021/hash/251bd0442dfcc53b5a761e050f8022b8-Abstract.html.

Alejandro Almodóvar, Patricia A. Apellániz, Santiago Zazo, and Juan Parras. causalKANs: Interpretable treatment effect estimation with Kolmogorov–Arnold networks. *arXiv preprint arXiv:2509.22467*, 2025.

Samy Badreddine, Artur d'Avila Garcez, Luciano Serafini, and Michael Spranger. Logic tensor networks. *Artificial Intelligence*, 303:103649, 2022. doi: 10.1016/j.artint.2021.103649.

Heejung Bang and James M. Robins. Doubly robust estimation in missing data and causal inference models. *Biometrics*, 61(4):962–973, 2005. doi: 10.1111/j.1541-0420.2005.00377.x.

Falco J. Bargagli-Stoffi, Riccardo Cadei, Kwonsang Lee, and Francesca Dominici. Causal rule ensemble: Interpretable discovery and inference of heterogeneous treatment effects, 2024. URL https://arxiv.org/abs/2009.09036v6. Version 6; first posted in 2020.

Yoshua Bengio, Nicholas Léonard, and Aaron Courville. Estimating or propagating gradients through stochastic neurons for conditional computation. *arXiv preprint arXiv:1308.3432*, 2013.

Alberto Caron, Gianluca Baio, and Ioanna Manolopoulou. Interpretable deep causal learning for moderation effects. In *ICML Workshop on Interpretable Machine Learning in Healthcare*, 2022. URL https://arxiv.org/abs/2206.10261.

Stephen Carrow, Kyle Erwin, Olga Vilenskaia, Parikshit Ram, Tim Klinger, Naweed Khan, Ndivhuwo Makondo, and Alexander G. Gray. Neural reasoning networks: Efficient interpretable neural networks with automatic textual explanations. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pp. 15669–15677, 2025. doi: 10.1609/aaai.v39i15.33720.

Kan Chen, Qishuo Yin, and Qi Long. Covariate-balancing-aware interpretable deep learning models for treatment effect estimation. *Statistics in Biosciences*, 17(1):132–150, 2025. doi: 10.1007/s12561-023-09394-6.

Jonathan Crabbé, Alicia Curth, Ioana Bica, and Mihaela van der Schaar. Benchmarking heterogeneous treatment effect models through the lens of interpretability. In *Advances in Neural Information Processing Systems (NeurIPS) Datasets and Benchmarks Track*, 2022. arXiv:2206.08363.

Miles Cranmer. Interpretable machine learning for science with PySR and SymbolicRegression.jl. *arXiv preprint arXiv:2305.01582*, 2023.

Alicia Curth and Mihaela van der Schaar. On inductive biases for heterogeneous treatment effect estimation. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 34, 2021.

Alicia Curth, David Svensson, Jim Weatherall, and Mihaela van der Schaar. Really doing great at estimating CATE? A critical look at ML benchmarking practices in treatment effect estimation. In *Advances in Neural Information Processing Systems (NeurIPS) Datasets and Benchmarks Track*, 2021.

Vincent Dorie. NPCI: Non-parametrics for causal inference. R package, GitHub repository, 2016. URL `https://github.com/vdorie/npci`.

P. Richard Hahn, Jared S. Murray, and Carlos M. Carvalho. Bayesian regression tree models for causal inference: Regularization, confounding, and heterogeneous effects. *Bayesian Analysis*, 15 (3):965–1056, 2020. doi: 10.1214/19-BA1195.

Jennifer L. Hill. Bayesian nonparametric modeling for causal inference. *Journal of Computational and Graphical Statistics*, 20(1):217–240, 2011.

Mayu Hiraishi, Ke Wan, Kensuke Tanioka, Hiroshi Yadohisa, and Toshio Shimokawa. Causal rule ensemble method for estimating heterogeneous treatment effect with consideration of prognostic effects. *Statistical Methods in Medical Research*, 33(6):1021–1042, 2024. doi: 10.1177/09622802241247728.

Chan Hsu, Jun-Ting Wu, and Yihuang Kang. Causal rule forest: Toward interpretable and precise treatment effect estimation. In *IEEE International Conference on Information Reuse and Integration for Data Science (IRI)*, 2024. arXiv:2408.15055.

International Stroke Trial Collaborative Group. The International Stroke Trial (IST): A randomised trial of aspirin, subcutaneous heparin, both, or neither among 19,435 patients with acute ischaemic stroke. *The Lancet*, 349(9065):1569–1581, 1997. doi: 10.1016/S0140-6736(97)04011-7.

Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. In *International Conference on Learning Representations (ICLR)*, 2019.

Scott M. Lundberg and Su-In Lee. A unified approach to interpreting model predictions. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 30, 2017. URL `https://proceedings.neurips.cc/paper_files/paper/2017/hash/8a20a8621978632d76c43dfd28b67767-Abstract.html`.

Xinkun Nie and Stefan Wager. Quasi-oracle estimation of heterogeneous treatment effects. *Biometrika*, 108(2):299–319, 2021.

Brandon Paulsen, Jingbo Wang, and Chao Wang. ReluDiff: Differential verification of deep neural networks. In *Proceedings of the 42nd International Conference on Software Engineering (ICSE)*, pp. 714–726, 2020. doi: 10.1145/3377811.3380337. URL `https://arxiv.org/abs/2001.03662`.

Judea Pearl. Causal inference in statistics: An overview. *Statistics Surveys*, 3:96–146, 2009. doi: 10.1214/09-SS057.

Abbavaram Gowtham Reddy and Vineeth N Balasubramanian. NESTER: An adaptive neurosymbolic method for causal effect estimation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pp. 14793–14801, 2024. doi: 10.1609/aaai.v38i13.29398.

Dalila Ressi, Riccardo Romanello, Sabina Rossi, and Carla Piazza. Compressing neural networks via formal methods. *Neural Networks*, 178:106411, 2024. doi: 10.1016/j.neunet.2024.106411.

Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. "Why should I trust you?": Explaining the predictions of any classifier. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pp. 1135–1144, 2016. doi: 10.1145/2939672.2939778.

Ryan Riegel, Alexander Gray, Francois Luus, Naweed Khan, Ndivhuwo Makondo, Ismail Yunus Akhalwaya, Haifeng Qian, Ronald Fagin, Francisco Barahona, Udit Sharma, Shajith Ikbal, Hima Karanam, Sumit Neelam, Ankita Likhyani, and Santosh Srivastava. Logical neural networks. *arXiv preprint arXiv:2006.13155*, 2020.

Peter Sandercock, Maciej Niewada, and Anna Członkowska. International stroke trial database (version 2). Dataset, University of Edinburgh, Department of Clinical Neurosciences, 2011.

Michael Schmidt and Hod Lipson. Distilling free-form natural laws from experimental data. *Science*, 324(5923):81–85, 2009.

Uri Shalit, Fredrik D. Johansson, and David Sontag. Estimating individual treatment effect: generalization bounds and algorithms. In *International Conference on Machine Learning (ICML)*, volume 70 of *Proceedings of Machine Learning Research*, pp. 3076–3085, 2017. URL https://proceedings.mlr.press/v70/shalit17a.html.

Claudia Shi, David M. Blei, and Victor Veitch. Adapting neural networks for the estimation of treatment effects. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 32, 2019.

Yishai Shimoni, Chen Yanover, Ehud Karavani, and Yaara Goldschmidt. Benchmarking framework for performance-evaluation of causal inference analysis, 2018. URL https://arxiv.org/abs/1802.05046.

Gagandeep Singh, Timon Gehr, Markus Püschel, and Martin Vechev. An abstract domain for certifying neural networks. *Proceedings of the ACM on Programming Languages*, 3(POPL):41:1–41:30, 2019. doi: 10.1145/3290354.

Trevor Stephens. gplearn: Genetic programming in Python, with a scikit-learn inspired API, 2026. URL https://pypi.org/project/gplearn/0.4.3/. Version 0.4.3.

Douglas D. Thompson, Gordon D. Murray, Livia Candelise, Zhengming Chen, Peter A. G. Sandercock, and William N. Whiteley. Targeting aspirin in acute disabling ischemic stroke: An individual patient data meta-analysis of three large randomized trials. *International Journal of Stroke*, 10(7):1024–1030, 2015. doi: 10.1111/ijs.12487.

Stefan Wager and Susan Athey. Estimation and inference of heterogeneous treatment effects using random forests. *Journal of the American Statistical Association*, 113(523):1228–1242, 2018.

Yinjun Wu, Mayank Keoliya, Kan Chen, Neelay Velingker, Ziyang Li, Emily J. Getzen, Qi Long, Mayur Naik, Ravi B. Parikh, and Eric Wong. DISCRET: Synthesizing faithful explanations for treatment effect estimation. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pp. 53597–53618, 2024. URL https://proceedings.mlr.press/v235/wu24n.html.

Jiahui Xu, Tanvi Shinkre, and Jennie Brand. htetree: Causal inference with tree-based machine learning algorithms, 2026. URL https://CRAN.R-project.org/package=htetree. R package version 0.1.23.

Huan Zhang, Tsui-Wei Weng, Pin-Yu Chen, Cho-Jui Hsieh, and Luca Daniel. Efficient neural network robustness certification with general activation functions. In *Advances in Neural Information Processing Systems*, volume 31, 2018. URL https://papers.nips.cc/paper_files/paper/2018/hash/d04863f100d59b3eb688a11f95b0ae60-Abstract.html.

A Model and Training Details

Notation and dimensions

Inputs are preprocessed unless stated otherwise. The domain $D = [0,1]^d$ includes continuous relaxations of one-hot coordinates. Lowercase x_i, t_i, y_i are observations of X, T, Y; hats mark estimates and tildes mark pruned functions, except for scaled targets $\tilde{y}$ and invented inputs $\widetilde{z}_t$. Local algorithm and example symbols are defined with their expressions. Correlations are empirical Pearson correlations; pairwise dictionary indices are distinct, and trigonometric π is the circle constant. Specialized S exports may retain an unused treatment input.

Table 2: Shared notation. Outcomes and contrasts are scalar.

Symbol	Domain or dimension	Meaning
d, d_u, C, K	Positive integers; $K \le d_u$	Covariates, block inputs, clauses, literals per clause.
$x, X; u$	D; $[0,1]^{d_u}$	Covariates; treatment-conditional inputs, with $d_u = d + 1$.
$t, T; Y$	$\{0,1\}$; $\mathbb{R}$	Treatment; observed outcome.
$\mu_t, \hat{\mu}_t$	$D \to \mathbb{R}$	Interventional mean; fitted base arm.
$\tau, \hat{\tau}_b, \hat{\tau}$	$D \to \mathbb{R}$	True, fitted base, and deployed treatment contrasts.
$\alpha, w; v, \gamma$	$\mathbb{R}^{C \times d_u}$; $\mathbb{R}^C$	Selection logits, signed weights; raw and nonnegative head weights.
θ	$\mathbb{R}^{C(2d_u+1)}$	Flattened block parameters; constraints restrict this ambient space.
m, g, m^{ST}	$C \times d_u$ arrays	Binary masks, soft gates in $[0,1]$, straight-through masks.
$\kappa_c; \ell_{cj}, h_c$	$\mathbb{R}$; $[0,1]^{d_u} \to [0,1]$	Selection threshold; signed literal and clause activation.
f_θ	$[0,1]^{d_u} \to [0,1]$	Block output before inverse scaling.
G_j, π_j	Nonnegative scalars	Soft-gate usage and normalized usage, $j = 1, \ldots, d_u$.
$B; \tilde{y}_i; r_i; s_j$	Row-index set; $\mathbb{R}$; $\mathbb{R}$; $[0,1]$	Minibatch; scaled target; factual residual; screening score.
$m, M; z(x)$	Integers $0 \le m \le M$; $\mathbb{R}^m$	Selected terms, cap; term values (unindexed m counts terms).
$b_0, b_t; b_z, b_{tz}$	$\mathbb{R}$; $\mathbb{R}^m$	Shared/treatment intercepts and shared/interaction coefficients.
$c; \Delta$	$D \times \{0,1\} \to \mathbb{R}$; $D \to \mathbb{R}$	Outcome correction; its treatment contrast, zero on rejection.
$F; r; a_{rt}, s_{rt}$	$\mathbb{N}$; $1, \ldots, F$; $\mathbb{R}, \mathbb{R}_{>0}$	Fold count and index; outcome offset and scale.
f_{rt}	$D \to [0,1]$	Arm-specialized fold circuit.
$v; \lambda; c_0, c_1$	$D \to [0,1]$; $[0,1]$; $\mathbb{Q}$	Candidate branch, deletion parameter, resolved arm sensitivities (local notation).
$\tau_b, \tilde{\tau}_b, e_r$	$D \to \mathbb{R}$	Original/reduced fold contrasts; their difference.
$\lVert f \rVert_\infty$	$[0, \infty]$	$\sup_{x \in D} \lvert f(x) \rvert$ for scalar f.
F_t, R_t, C_t, P_t, P, B	$D \to \mathbb{R}$	Original, reduced-base, contrast arms; arm remainder, shared null part, baseline.
$\alpha_{\mathrm{ATE}}; h$	$\mathbb{R}$; $D \to \mathbb{R}$	Mean fitted contrast; centered heterogeneity.
$n; i; \hat{\tau}_i, \tau_i, e_i$	$\mathbb{N}$; $1, \ldots, n$; $\mathbb{R}$	Row count/index; fitted/true effect and error. Bars denote row means.
PEHE, R_{het}	$\mathbb{R}_{\ge 0}$	Effect RMSE and centered effect RMSE, in outcome units before normalization.

Function class. The dictionary terms are analytic in continuous inputs. Logic additionally represents hinges: for $x \in [0,1]$ and $t \in \{0,1\}$, $f(x,t) = \frac{1}{2}[2x + t - 2]_0^1$ gives $f(x,1) - f(x,0) = [x - \frac{1}{2}]_+$. A finite sum of analytic dictionary terms cannot equal this hinge on an interval containing $1/2$, although it can approximate it. This capacity distinction implies no estimation advantage.

Table 3: The four Logic-S configurations in the IHDP/ACIC panel. Each uses a treatment-conditional circuit. CF3 averages three base fits and adds one final accepted correction. Base-only CF3 ablations remove that correction from the same fitted ensemble.

Paper name	Base	Symbolic deployment	Gate
Logic-S	S	None	None
Logic-S + correction	S	Additive	Held-out
Logic-S + correction (CF3)	S ensemble	One correction	Crossfit
Logic-S + arm predicates	S refit	Arm-specific inputs	Held-out

Table 4: Fixed Logic model profiles. Symbolic cross-validation uses five folds, clamped to available arm counts. All benchmark CF3 models and the matched certification audit use the large profile.

Setting	Single-fit	CF3
Clauses / literals per clause	8 / 2	16 / 4
Maximum epochs / restarts	100 / 2	150 / 3
Maximum symbolic terms	4	6
IHDP / ACIC screened features	24 / 48	25 / 25
Discovery improvement	0.05	0.02
Marginal term improvement	0.005	0.002
Acceptance improvement	0.02	0.01

Predicate vocabulary and scaling. Initial literals use the preprocessed input coordinates directly, with negation supplied by the sign of a clause weight. The backbone does not first learn a separate bank of threshold predicates. Numerical missing values use training medians; categorical values are imputed and one-hot encoded, with unseen categories ignored. Numerical min–max scaling is fit on the original training partition and clips later inputs to $[0, 1]$. This same fitted transform is reused in the crossfit variants. Training outcomes are converted to float32 before obtaining their minimum $y_{\min}$ and maximum $y_{\max}$; the outcome offset is $a = y_{\min}$, the scale is $s = \max(y_{\max} - y_{\min}, 10^{-6})$, and $\tilde{y}_i = (y_i - a)/s$. Validation targets use the same affine transformation without clipping. Predictions are transformed back to original units before subtracting arms or fitting a symbolic correction.

Selection, polarity, and logical compilation. We note the capability to initialize the S-learner with logic constraints. Requiring bipolar treatment clauses ensures treatment appears in designated clauses, though zero head weights or saturation can suppress its influence on the output. We note expert initialization with logic as an area of future investigation, particularly integration with causal discovery routines.

Every clause owns a length-d_u selection-logit vector and a separate length-d_u signed-weight vector. Its top-K mask depends only on learned logits; it is independent of x and of the absolute literal weights. Required memberships reserve slots before the remaining indices are ranked. The first two S clauses require treatment with opposite signs; the required signed weights are projected to their prescribed sign and magnitude at least 0.05 after every optimizer step. Unselected coordinates have zero forward contribution; the mask multiplies $|w|$ directly, so unselected coordinates can receive a recruitment gradient through Eq. (7). Polarity is the hard sign of the weight, with zero treated as positive. Inference uses the learned masks and ordinary clamps. Clause dropout is zero in the retained configurations.

For truth values $a, b \in [0, 1]$ and dilation coefficient $\rho \in [0, 1]$, the logical operators are $\neg a = 1 - a$, $\nabla_\rho a = \rho a$, $a \oplus b = \min(1, a + b)$, and $a \odot b = \max(0, a + b - 1)$. For a selected literal $\ell : [0, 1]^{d_u} \to [0, 1]$ with strength $r \in [0, 1]$, its weighted AND contribution compiles as $\neg\nabla_r\neg\ell$, and clauses combine using Łukasiewicz conjunction. The head compiles using disjunction of weighted clauses. If the strength $r > 1$, write $r = q + \rho$ with integer part $q = \lfloor r \rfloor \in \mathbb{N}$ and fractional part $\rho \in [0, 1)$; repeated truncated sums represent the integer part and a dilation represents the fractional part. The circuit therefore uses only variables, constants, negation, dilation, and clamped sums. Lattice min/max are supported by the general formula library but are not extra learned layers in this DNF. Weights exceeding one compile through repeated truncated sums and a fractional dilation.

Shared subexpressions remain shared in the compiled DAG. Forward and compiled predictions are checked on validation inputs; equality here is the mathematical semantics, subject to floating-point evaluation error.

Learning discrete structure. Masks are hard during training and inference. A sigmoid supplies straight-through gradients for free slots (Bengio et al., 2013):

$$g_{cj} = \sigma((\alpha_{cj} - \mathrm{sg}(\kappa_c))/0.5)\,, \qquad m_{cj}^{\mathrm{ST}} = m_{cj} + g_{cj} - \mathrm{sg}(g_{cj}), \tag{7}$$

where $\sigma(a) = (1 + e^{-a})^{-1}$, sg stops gradients, and $\kappa_c \in \mathbb{R}$ is the last selected free-slot logit. Required slots remain fixed. The AND/OR clamps use identity surrogate gradients in saturated regions. For minibatch indices B and scaled targets $\tilde{y}_i \in \mathbb{R}$, the loss, alongside AdamW (Loshchilov & Hutter, 2019) weight decay 10^{-4}, is

$$\frac{1}{|B|}\sum_{i\in B}(f_\theta(u_i) - \tilde{y}_i)^2 + 10^{-4}\left(\sum_{c,j}|w_{cj}| + \sum_c \gamma_c\right) + 0.01\sum_j \pi_j \log(\pi_j + 10^{-12}), \tag{8}$$

where $\pi \in \mathbb{R}^{d_u}_{\geq 0}$ is normalized soft-gate usage across clauses (Appendix A). The final term encourages coverage of different features. We use AdamW and select the checkpoint and restart by validation factual MSE. The small profile uses $(C, K) = (8, 2)$, 100 epochs, and two restarts; the large profile uses $(16, 4)$, 150 epochs, and three restarts per block.

Initialization and optimization. Selection logits are initialized independently as $\mathcal{N}(0, 0.3^2)$, signed clause weights as $\mathcal{U}(-1/K, 1/K)$, and raw head weights as $\mathcal{U}(0.5/C, 1.5/C)$. The retained base fits use no selection prior. Restarts use seed $s_{\mathrm{seed}} + 1000r$, where s_{seed} is the configured integer seed and $r \in \{0, \ldots, R-1\}$ indexes R restarts. Here $\mathcal{N}$ and $\mathcal{U}$ denote normal and uniform distributions. The initialized mask changes during training. Selection logits, signed weights, and head weights are jointly optimized. For the load-balance term, let $G_j = \sum_{c=1}^{C} g_{cj}$ and $\pi_j = G_j/(\sum_{l=1}^{d_u} G_l + 10^{-12})$, for $j = 1, \ldots, d_u$, using the constrained soft gates (required slots equal one). The penalty is negative entropy, so minimizing it encourages dispersed usage, without requiring different clauses to select disjoint features. L1 acts on all signed clause weights, including currently unselected coordinates, and on nonnegative head weights.

Table 5: Shared DNF training settings in the retained factual configurations. Architecture size and epoch/restart budgets are in Table 4.

Setting	Value
Optimizer / learning rate	AdamW / 0.003
Batch size / weight decay	256 / 10^{-4}
Adam moments / numerical epsilon	$(0.9, 0.999)$ / 10^{-8}
L1 / load-balance coefficient	10^{-4} / 0.01
Selection temperature	0.5, fixed
Validation interval / patience	10 / 40 epochs
Required-literal minimum magnitude	0.05
Clause dropout / selection prior	0 / none

Minibatches are reshuffled each epoch. Validation MSE is evaluated every 10 epochs. Every strictly lower validation MSE replaces the best checkpoint; an improvement exceeding 10^{-5} in scaled MSE resets patience. A restart stops after 40 epochs without such improvement, or its epoch cap. All configured restarts are attempted; the best validation checkpoint across them is deployed. AdamW weight decay also acts on selection logits. The training forward pass uses hard masks and hard clamp values, with the two surrogate gradients described in the main text.

Symbolic search and acceptance. For covariate index $j \in \{1, \ldots, d\}$, screening computes the scalar score $s_j \in [0, 1]$ on discovery rows,

$$s_j = \max_{t\in\{0,1\}} |\operatorname{corr}(x_j, r \mid T = t)|.$$

Constant features receive zero score. The feature limit is distinct from the DNF width K. The single-S runs typically retain 24 IHDP or 48 ACIC coordinates, while the large crossfit runs retain at most 25. Each run stores its limit, selected indices, and scores. Feature order is retained when constructing pairwise differences. The search never reads the DNF's top-K masks as a shortlist.

Candidates with nonfinite values or empirical standard deviation below 10^{-10} are discarded. At each greedy step, every remaining dictionary term is tentatively added and the coefficients in Eq. (3) are refit jointly using the fold-standardized penalized fit in Appendix C.2 on each symbolic CV training fold. For m selected terms and n_I fitting rows, the design matrix is in $\mathbb{R}^{n_I \times (2m+2)}$, with row $[1, z(x_i)^\top, t_i, t_i z(x_i)^\top]$. Validation MSE is averaged equally across treatment-stratified folds, normally five (clamped to arm counts if necessary). Candidate ranking adds 10^{-5} times total expression complexity: variable 1, square 2, cube/sine/cosine/pairwise algebra 3, and exponential-product 5. The baseline fits the intercept-only arm correction $b_0 + tb_t$. After the first selected term, expansion stops if marginal relative CV improvement is below the profile threshold; the final selected set must also meet the total discovery-improvement threshold. Final coefficients are fit using all discovery rows. A first term can be selected internally even when the discovery gate subsequently rejects deployment. The optional effect-specific discovery gate is unset in the reported benchmarks; deployment uses factual-loss improvement.

Let L_b, L_c be acceptance-set factual MSE before and after the additive correction, and $Q \geq 0$ its scalar total expression complexity. With minimum relative gain $\eta_{\text{acc}} \geq 0$, deployment requires the discovery gate, $(L_b - L_c)/\max(L_b, 10^{-12}) \geq \eta_{\text{acc}}$, and $L_b - L_c - 10^{-5}Q > 0$. The additive variant has no predicate-use test. Acceptance retains the correction verbatim; rejection returns the base. The profile-specific thresholds are in Table 4.

Arm-predicate refitting. Selected expressions are scaled to $[0, 1]$ using the fitting rows. For each term, the refit receives one control-gated and one treatment-gated coordinate. The control-gated copy is complemented according to the fitted control coefficient sign; the treatment-gated copy uses the fitted treatment-interaction coefficient sign. Thus these have the form $(1-t)q_{j0}(x)$ and $tq_{j1}(x)$, where $q_{jt} : [0,1]^d \to [0,1]$, $j = 1, \ldots, m$, is either the scaled jth term or its complement. The resulting vector $\widetilde{z}_t(x) = ((1-t)q_{10}(x), tq_{11}(x), \ldots, (1-t)q_{m0}(x), tq_{m1}(x)) \in [0,1]^{2m}$ augments (x, t) to $d + 1 + 2m$ inputs; $\theta' \in \mathbb{R}^{C(2(d+1+2m)+1)}$ denotes the new circuit parameters for the same C clauses. Their literal memberships are required in designated clauses and their DNF polarities are positive. The two required treatment literals remain in place. Thus each required discovered input consumes a top-K slot. A new outcome circuit is fit from initialization; acceptance compares that candidate with the base and also checks operative predicate use and treatment dependence. This is a different estimator from adding the original regression correction.

Crossfit data roles. Treatment-stratified F-fold splitting is applied to the pooled development rows. Each S block fits on $F - 1$ folds and uses the held-out fold for checkpoint selection and arm-outcome prediction. This held-out prediction therefore excludes that row from gradient fitting, but not from checkpoint selection. A second pass searches corrections on the outer fitting indices using the assembled held-out-fold residuals, and scores each correction on the outer held-out indices. Averaging row-level squared errors supplies L_b, L_c for the final gate. One final dictionary search and coefficient fit use all development residuals; its discovery gate and complexity, together with the outer comparison, determine acceptance. This procedure does not claim a fresh independent nuisance fit inside every symbolic fold. Test outcomes are never used. Deployment averages base outcomes after each block's own inverse scaling, then adds the one accepted final correction.

A.1 Algorithms and source map

Training settings are in Table 5; local notation and implementation entry points follow the algorithms.

Local notation. Data sets contain rows in $D \times \{0,1\} \times \mathbb{R}$; $X_{\text{test}} \in [0,1]^{n_{\text{test}} \times d}$. Restart and epoch indices are r, e, with integer budgets R, E. Losses $L, L_r, L^* \in [0, \infty]$ are current, restart-best, and overall-best validation MSE; θ^* has the parameter dimension in Table 2 and G is its DAG. In symbolic search, I indexes n_I rows, $p \leq d$ is the feature cap, and S the ordered term set. Candidates $q : D \to \mathbb{R}$ augment $z_S : D \to \mathbb{R}^{|S|}$; q^* minimizes penalized CV loss. Each CV design has n_{train}

Algorithm 1 Fit one sparse factual Logic block

Require: Fitting/checkpoint rows $D_{\text{fit}}, D_{\text{val}}$; clauses C, width K, restarts R, epoch cap E
Ensure: Selected parameters θ^* and compiled DAG G

$L^* \leftarrow \infty$
for $r = 0, \ldots, R-1$ **do**
  Initialize parameters, required slots, $L_r \leftarrow \infty$, and local patience
  for $e = 1, \ldots, E$ **do**
    for minibatch B from a fresh permutation of D_{fit} **do**
      Form hard top-K masks respecting slot constraints; evaluate Eq. (2)
      Differentiate Eq. (8) with the stated mask/clamp surrogate gradients
      Apply AdamW; project required signs and minimum magnitudes
    end for
    At checkpoint epochs, evaluate L on D_{val}; if $L < L^*$, save $(L^*, \theta^*) \leftarrow (L, \theta)$
    Reset patience and update L_r when $L < L_r - 10^{-5}$; otherwise advance patience; stop at its limit
  end for
end for
Compile θ^* to G, check validation agreement, and return (θ^*, G)

Algorithm 2 Greedy factual symbolic-correction search

Require: Rows I with $(x_i, t_i, y_i, \hat{\mu}_0(x_i), \hat{\mu}_1(x_i))$; feature cap p, term cap M, discovery thresholds
Ensure: Correction c_I and discovery decision g_I

Compute residuals $r_i = y_i - \hat{\mu}_{t_i}(x_i)$; screen top p features by within-arm correlation
Build the dictionary; discard nonfinite/constant terms; construct treatment-stratified CV folds
$S \leftarrow \varnothing$; score the arm-intercept correction $b_0 + tb_t$
for $m = 1, \ldots, M$ **do**
  For each remaining q, fit $[1, z_S, q, t, tz_S, tq]$ using the penalized fit in Appendix C.2
  Choose q^* minimizing CV factual residual MSE plus complexity penalty; break ties by name
  if $m > 1$ and marginal relative improvement is below threshold **then**
    break
  end if
  $S \leftarrow S \cup \{q^*\}$
end for
Refit Eq. (3) on I; set g_I from discovery improvement (and the optional effect-specific gate, unset here); return (c_I, g_I)

Algorithm 3 Fit and deploy a cross-fitted Logic ensemble

Require: Development data $\mathcal{D}$, test inputs X_{test}, folds F
Ensure: Arm predictions and their treatment contrast

Split $\mathcal{D}$ into F treatment-stratified folds $\mathcal{D}_r$
for $r = 1, \ldots, F$ **do**
  Fit one treatment-conditional block on $\mathcal{D} \setminus \mathcal{D}_r$, using $\mathcal{D}_r$ for checkpoint selection; predict both arms on $\mathcal{D}_r$ and X_{test}
end for
for $r = 1, \ldots, F$ **do**
  Run Algorithm 2 on residuals for $\mathcal{D} \setminus \mathcal{D}_r$; evaluate its factual correction on $\mathcal{D}_r$
end for
Aggregate outer squared errors into L_b, L_c; search all development residuals for final (c, g)
Accept iff $g = 1$, $(L_b - L_c)/\max(L_b, 10^{-12}) \geq \eta_{\text{acc}}$, and $L_b - L_c - \lambda_{\text{comp}} Q(c) > 0$
Average base arms after fold-specific inverse scaling; add $c(x, t)$ if accepted, zero otherwise
Return both arms and their difference

Algorithm 4 Exact contrast pruning for one fold

Require: Arm DAGs (G_0, G_1), scales $(s_0, s_1) > 0$, domain $D = [0,1]^d$
Ensure: Reduced DAGs $(\widetilde{G}_0, \widetilde{G}_1)$ preserving the scaled contrast on D
Convert binary coefficients to exact rationals; normalize both DAGs
for candidate dilation branch q in deterministic order **do**
    Compute deletion responses c_t on the current DAGs; use zero if $q \notin G_t$
    if either response is unresolved or $s_0 c_0 \neq s_1 c_1$ **then**
        continue
    end if
    Replace q by zero and normalize, obtaining $(\bar{G}_0, \bar{G}_1)$
    Adopt the proposed DAGs only if their joint distinct-node count is smaller
end for
Return (G_0, G_1) and the exact bound zero

Algorithm 5 Organize a certified fit

Require: Original arms (F_0, F_1), reduced base arms (R_0, R_1), correction $c(x,t) = p(x) + t\Delta(x)$
Ensure: Baseline B, contrast $\hat{\tau}$, null component P, mean α_{ATE}, centered effect h
$C_0 \leftarrow R_0$; $C_1 \leftarrow R_1 + \Delta$; $P_t \leftarrow F_t - C_t$ (the certificate gives $P_0 = P_1$)
$P \leftarrow P_0$; $B \leftarrow F_0$; $\hat{\tau} \leftarrow C_1 - C_0$
$\alpha_{\mathrm{ATE}} \leftarrow n^{-1} \sum_i \hat{\tau}(x_i)$; $h \leftarrow \hat{\tau} - \alpha_{\mathrm{ATE}}$
Check $F_t = P + C_t = B + t\hat{\tau}$ and $n^{-1} \sum_i h(x_i) = 0$
Return components, supports, and source hashes

rows and $2|S| + 4$ columns. Returned corrections $c_I : D \times \{0,1\} \to \mathbb{R}$ have decisions $g_I \in \{0,1\}$. Deployment uses nonnegative losses L_b, L_c, binary discovery decision g, complexity $Q(c) \geq 0$, penalty $\lambda_{\mathrm{comp}} = 10^{-5}$, and profile threshold $\eta_{\mathrm{acc}} \geq 0$. In pruning, $G_t, \bar{G}_t, \widetilde{G}_t$ denote current, proposed, and returned arm DAGs for maps $D \to [0,1]$; the candidate $q : D \to [0,1]$ has responses $c_t \in \mathbb{Q}$ when resolved. Organization components are scalar maps on D; $p(x) = b_0 + z(x)^\top b_z$, and R_t includes offsets, scales, and averaging. Rejection sets $p = \Delta = 0$; $x_1, \ldots, x_n \in D$ are the evaluation rows.

B Exact Certification and Worked Example

Object being certified. The certificate starts from the treatment-specialized outcome DAGs. Each finite binary coefficient is converted to its exact rational value. The certified contrast includes every fold's arm scales and offsets and the accepted symbolic correction in original outcome units. It is the real-arithmetic interpretation of these exported programs. The neural implementation evaluates floating-point operations, whose ordering can give small discrepancies from that interpretation. We retain both the analytic guarantee and measured implementation discrepancies.

Exact simplification. The rational backend implements identities for negation, scalar dilation, and truncated sums. Constants and combined scalar weights remain rational. For a truth-valued function $\varphi : D \to [0,1]$ and scalar dilation coefficients $r, s \in [0,1]$, the pointwise identity $(r\varphi) \oplus (s\varphi) = (r+s)\varphi$ holds when $r + s \leq 1$, where the comparison and the addition are exact. Nested truncated sums of the same type flatten associatively. Complement pairs and neutral elements are eliminated by their logical identities. Equal subgraphs are interned by their complete structure, input indices, and rational coefficients. There is no decimal rounding of structural fingerprints. Serialization stores numerator/denominator values as rational strings, so reading an explanation back preserves its coefficients.

Deletion-response propagation. Fix a candidate node q whose original scalar value is $v(x)$ for $x \in D = [0,1]^d$. Replace every occurrence of q by $(1-\lambda)v(x)$, where $\lambda \in [0,1]$ is the deletion parameter. For any node n, write $g_{n,\lambda} : D \to [0,1]$ for its resulting value. The bottom-up pass records an interval $I_n = [L_n, U_n] \subseteq [0,1]$ and a response coefficient $d_n \in \mathbb{Q} \cup \{\perp\}$, where $\perp$ denotes an unresolved response. A resolved d_n is constant across both covariates and deletion

parameters. The invariants are

$$g_{n,\lambda}(x) \in I_n, \qquad d_n \neq \bot \implies g_{n,\lambda}(x) = g_{n,0}(x) - \lambda d_n v(x) \quad (x \in D,\ \lambda \in [0,1]). \tag{9}$$

All interval endpoints and resolved coefficients use exact rational arithmetic. The propagation rules are as follows.

1. At the candidate q, enclose its original value interval together with zero and set $d_q = 1$. This overrides its usual operator rule.
2. Other input variables have interval $[0, 1]$ and response zero; a constant $b \in [0, 1]$ has interval $[b, b]$ and response zero.
3. Negation sends $[L, U]$ to $[1-U, 1-L]$ and a resolved coefficient d to $-d$. Scalar dilation by $r \in [0, 1]$ sends them to $[rL, rU]$ and rd. An unresolved child response remains unresolved.
4. A clamped-sum node with $k \geq 1$ children $n_1, \ldots, n_k$ has the form $\psi(\sum_{i=1}^{k} g_{n_i,\lambda})$, where $\psi(a) = \min(1, a)$ for a disjunction or $\psi(a) = \max(0, a - (k-1))$ for a conjunction. Set $A = \sum_i L_{n_i}$, $B = \sum_i U_{n_i}$, and $I_n = [\psi(A), \psi(B)]$. If every child response is zero, return $d_n = 0$. Otherwise, if $[A, B]$ lies in the affine regime of ψ and every child response is resolved, return $d_n = \sum_i d_{n_i}$. If $[A, B]$ lies in its constant regime, return $d_n = 0$, even if a child response is unresolved. Return $\bot$ in the remaining cases. The affine/constant tests are $B \leq 1/A \geq 1$ for disjunction and $A \geq k-1/B \leq k-1$ for conjunction, with the affine test considered first when its child responses are resolved.

Lemma 1 (Soundness of deletion-response propagation)**.** *For an arm circuit over variables, constants, negations, scalar dilations, and clamped sums, the interval-and-response rules above return a scalar c_t only if*

$$f_{t,\lambda}(x) = f_t(x) - \lambda c_t v(x) \quad (t \in \{0,1\},\ x \in D,\ \lambda \in [0,1]),$$

where $f_{t,\lambda}$ replaces every occurrence of the candidate branch $v(x)$ by $(1-\lambda)v(x)$. Thus $\lambda = 0$ is the original circuit and $\lambda = 1$ is full deletion.

Proof of Lemma 1. Induct in topological order on the DAG. At q, the subgraph computing the original $v(x)$ is unchanged, so its original interval is valid; adjoining zero encloses $(1-\lambda)v(x)$ and the response identity holds with $d_q = 1$. Other leaves satisfy both invariants directly. Negation and dilation preserve interval containment and transform the response identity by subtraction from one or multiplication by the fixed scalar, respectively.

At a clamped sum, the inductive interval bounds enclose the child sum in $[A, B]$. Monotonicity of ψ therefore gives the stated output interval. If all child responses are zero, their values are unchanged for each fixed x, so the output is unchanged regardless of clamp regime. In the affine regime, substituting the child response identities gives $d_n = \sum_i d_{n_i}$. In the constant regime, the output is fixed over the entire enclosure, so its response is zero regardless of unresolved child responses. If both regimes meet at a singleton boundary and the affine rule applies, the same identity still holds on that singleton. When the rule returns $\bot$, only interval containment is asserted. Repeated children and reconvergent paths are accounted for by the child sum, with each incoming occurrence counted. This establishes Eq. (9) at every node. At the arm root, set $c_t = d_n$; if the candidate is absent, the response is zero. □

Proof of Proposition 1. At full deletion, Lemma 1 gives $f_{t,1}(x) - f_t(x) = -c_t v(x)$ for every $x \in D$. The offsets are unchanged, so

$$\tilde{\tau}_b(x) - \tau_b(x) = -(s_1 c_1 - s_0 c_0) v(x) = 0,$$

using the certified equality $s_1 c_1 = s_0 c_0$. □

The restricted operator set contains no multiplication of two changing branch-dependent values. Within a resolved affine regime, its response is therefore exactly affine in the substituted branch value for each fixed x. This assertion does not require the circuit to be affine in x. Nonlinear symbolic corrections are retained unchanged and are outside this propagation pass. Interval overapproximation can prevent a successful check but cannot invalidate an accepted one. Structural

interning likewise recognizes sufficient instances of functional identity without requiring a complete equivalence test.

The implementation deletes a branch only when both raw-unit output sensitivities are resolved and equal. After deletion, exact simplification is repeated and subsequent candidates are checked against the changed graphs. Candidates are visited in a deterministic order. The certificate is sufficient and need not find a smallest equivalent representation. Each zero-error deletion preserves the current contrast, so induction over accepted deletions preserves the original contrast. Greedy order can affect the returned size even though every accepted deletion preserves the contrast exactly.

Composition. Let $e_r : D \to \mathbb{R}$ denote the original-minus-reduced raw contrast in fold $r \in \{1, \dots, F\}$, where F is the number of folds. The accepted deletions give $e_r(x) = 0$ for every r and $x \in D$. Because the correction, arm offsets, and averaging weights are retained, the original and reduced deployed contrasts satisfy

$$\hat{\tau}(x) - \tilde{\tau}(x) = \frac{1}{F} \sum_{r=1}^{F} e_r(x) = 0 \qquad (x \in D).$$

For any probability distribution supported on the box, integrating the same equality preserves the model's ATE. This statement is independent of whether the original effect estimate is accurate.

Matched audit. The fixed cohort consists of IHDP 001–030, seed 42, for the large-profile Logic-S CF3 architecture. The rational audit reconstructs the certificate from each fitted program. Each fit retains all three folds and its single final correction. Distinct base-node counts exclude fixed affine and averaging wrappers; correction terms are reported separately and retained unchanged. Table 6 deduplicates identical nodes across all folds and arms. Table 1 instead sums the per-fold counts, giving different size summaries for the same pruning results.

Outcome-preserving control. We rerun the deletion procedure from each original normalized arm pair, using the same candidate ordering, interval propagation, and requirement that a deletion reduce DAG size. The control requires $c_0 = c_1 = 0$, preserving each outcome separately. Contrast pruning instead requires $s_0 c_0 = s_1 c_1$ and can accept equal nonzero responses. Both procedures recompute sensitivities after accepted deletions. This comparison isolates the role of cross-arm cancellation within the stated sufficient rule; it does not compare globally minimal representations.

Table 6: Outcome-preserving control and contrast-preserving pruning on the same 30 fitted ensembles. Both rules use exact interval tests, identical candidate ordering, and the same DAG-size criterion. Medians and 95% bootstrap intervals.

Model	Rule	Nodes retained	Removed (%)
Logic-S + correction (CF3)	Preserve outcomes	347.0 [309.5, 378.5]	0.0 [0.0, 0.0]
Logic-S + correction (CF3)	Preserve contrast	92.0 [44.0, 174.5]	74.2 [51.4, 83.8]

Table 7: Floating-point diagnostics, separate from the zero real-arithmetic pruning bound. Export gaps are normalized by training-outcome SD. The pruning residual is in original outcome units.

Model	Maximum pruning residual	Maximum export gap
Logic-S + correction (CF3)	2.37×10^{-14}	7.23×10^{-7}

Verification. The rational proof supplies the uniform guarantee. Test rows and 20,000 seeded box points check numerical implementation. Regression tests cover unequal scales, clamp crossings, saturation, near-equal coefficients, constant folding, serialization, and random rational circuits. Tamper tests reject altered deployment coefficients, names, sensitivities, and DAGs.

C Benchmark Scope and Full Comparisons

Nine configurations, including CRF + CT, share 1,000 IHDP and 135 ACIC evaluation keys at fitting seeds 42–44. Treatment-structured SR adds an tenth configuration on all IHDP cases and the 61 heterogeneous ACIC variants. No case is removed based on prediction quality. Audits verify prediction hashes, test identifiers, ground truth, and normalization scales.

The evaluation boundary is uniform across methods. For every IHDP, ACIC, and Nie–Wager case, the deterministic seed-42 splitter stratifies by treatment and allocates 50%, 15%, 15%, 10%, and 10% to training, validation, discovery, acceptance, and test. Integer counts use largest-remainder rounding within each treatment arm. Every method is scored on the exact same untouched test indices. The estimators use the remaining rows according to their fitting procedure:

Table 8: Development-row allocation as a percentage of all rows. All methods use training-fitted preprocessing and common untouched test rows.

Estimator	Development use	Test
Single Logic family	50% base fit; 15% checkpoint; corrected variants use 15% discovery and 10% acceptance	10%
Logic CF3	90% pooled into three treatment-stratified folds; correction search uses development-wide OOF residuals	10%
Neural baselines	75% gradient fit; 15% checkpoint selection	10%
Causal forest	90% fit; five-fold OOF nuisance residuals within development	10%
CRF + CT (reimplementation)	90% forest-rule construction; internal honest splits and final-tree cross-validation	10%
Restricted SR	90% pooled search around arm means; five-fold internal term selection	10%

Normalization uses each case's training-outcome SD; these scales are identical across paired models.

TARNet and the two DragonNet variants fit on the pooled training, discovery, and acceptance partitions (approximately 75%) and use the original validation partition (15%) for checkpoint selection. Their three shared ELU layers have width 200; each outcome head has two 100-wide ELU layers and a scalar output. Training uses 30 Adam epochs at 10^{-3} followed by 50 momentum-SGD epochs at 10^{-5}, with early-stopping patience 2 and 20, batch size 64, and an L_2 penalty of 0.01. Targeted regularization has weight one. These settings are fixed across datasets. The causal forest uses all 90% development rows, 400 honest trees, minimum leaf size 10, minimum split size 20, and subsampling fraction 0.45. Five-fold ridge/logistic nuisance fits supply out-of-fold marginal outcome and propensity residuals; propensities are clipped to $[0.025, 0.975]$. The ridge penalty and inverse logistic penalty are one. Preprocessing for every method is fit on the original training partition; these baselines use standardized coordinates and Logic uses unit-box coordinates. All share the untouched test set.

C.1 Primary seed-averaged comparisons

Tables 9 and 10 average the three fitting seeds within each case, then cases equally. Intervals use 5,000 case-bootstrap resamples retaining all three seeds. The Logic-S and CF3 base rows use the same fitted circuits as their corrected counterparts. All 6,810 corrected-model runs are included.

Medians summarize the typical response surface after averaging its three fitting seeds; means also reflect the magnitude of tail errors.

The fixed ACIC asset selector contains 300 available variants. Its pre-outcome size filter retains the 135 variants with fewer than 10,000 observations. Of these, 46 have 1,000 rows, 44 have 2,500, and 45 have 5,000. The ACIC heterogeneity partition uses the noiseless training CATE and describes the target of evaluation. The 74 remaining variants have training CATE SD at most 10^{-6}. The all-case analysis is retained because performance on constant effects is also relevant. This partitioning is never used by the fitting or acceptance procedures. IHDP and ACIC reuse covariate populations across response surfaces; bootstrap intervals therefore describe variation across the sampled response surfaces, conditional on those covariates and fitting seed.

Table 9: Primary comparison: mean normalized PEHE, averaging fitting seeds 42–44 within each case before summarizing across cases. Brackets give 95% case-bootstrap intervals (5,000 resamples, all three seeds retained). Logic-S bases and their corrected counterparts share the same fitted circuits. SR covers only heterogeneous ACIC.

Model	IHDP	ACIC heterogeneous	ACIC all
Causal Forest	0.564 [0.549, 0.579]	120.840 [0.146, 362.170]	54.874 [0.263, 163.939]
CRF + CT (reimplementation)	0.714 [0.696, 0.731]	14.906 [0.380, 43.892]	7.207 [0.595, 20.320]
TARNet	0.284 [0.278, 0.290]	0.316 [0.238, 0.434]	0.340 [0.285, 0.408]
DragonNet	0.262 [0.257, 0.267]	0.343 [0.255, 0.468]	0.421 [0.360, 0.492]
DragonNet + TARReg	0.250 [0.245, 0.256]	0.343 [0.258, 0.467]	0.400 [0.341, 0.469]
Logic-S (base)	0.899 [0.880, 0.917]	0.290 [0.189, 0.427]	0.312 [0.241, 0.398]
Logic-S + correction	0.509 [0.500, 0.518]	0.334 [0.240, 0.463]	0.271 [0.215, 0.341]
Logic-S + arm predicates	0.734 [0.718, 0.750]	0.304 [0.211, 0.439]	0.375 [0.304, 0.459]
Logic-S (CF3 base)	0.888 [0.868, 0.907]	0.253 [0.148, 0.395]	0.234 [0.170, 0.311]
Logic-S + correction (CF3)	0.346 [0.339, 0.353]	0.216 [0.130, 0.346]	0.205 [0.157, 0.270]
Treatment-structured SR	0.290 [0.283, 0.298]	11.518 [0.208, 34.078]	n/a

Table 10: Primary comparison: mean normalized absolute ATE error, averaging fitting seeds 42–44 within each case before summarizing across cases. Brackets give 95% case-bootstrap intervals (5,000 resamples, all three seeds retained). Logic-S bases and their corrected counterparts share the same fitted circuits. SR covers only heterogeneous ACIC.

Model	IHDP	ACIC heterogeneous	ACIC all
Causal Forest	0.086 [0.082, 0.090]	81.761 [0.034, 245.206]	37.007 [0.059, 110.872]
CRF + CT (reimplementation)	0.101 [0.097, 0.105]	9.219 [0.093, 27.441]	4.485 [0.306, 12.750]
TARNet	0.048 [0.045, 0.050]	0.052 [0.038, 0.068]	0.094 [0.073, 0.117]
DragonNet	0.049 [0.047, 0.051]	0.066 [0.046, 0.092]	0.145 [0.113, 0.179]
DragonNet + TARReg	0.050 [0.048, 0.052]	0.065 [0.045, 0.089]	0.130 [0.103, 0.161]
Logic-S (base)	0.346 [0.336, 0.357]	0.105 [0.085, 0.127]	0.223 [0.168, 0.287]
Logic-S + correction	0.107 [0.102, 0.112]	0.089 [0.069, 0.111]	0.088 [0.070, 0.111]
Logic-S + arm predicates	0.242 [0.234, 0.250]	0.089 [0.071, 0.110]	0.221 [0.165, 0.284]
Logic-S (CF3 base)	0.289 [0.278, 0.300]	0.067 [0.052, 0.085]	0.146 [0.105, 0.193]
Logic-S + correction (CF3)	0.060 [0.058, 0.063]	0.052 [0.038, 0.066]	0.059 [0.049, 0.070]
Treatment-structured SR	0.040 [0.038, 0.042]	8.805 [0.029, 26.348]	n/a

Table 11: Median normalized PEHE, averaging fitting seeds 42–44 within each case before summarizing across cases. Brackets give 95% case-bootstrap intervals (5,000 resamples, all three seeds retained). Logic-S bases and their corrected counterparts share the same fitted circuits. SR covers only heterogeneous ACIC.

Model	IHDP	ACIC heterogeneous	ACIC all
Causal Forest	0.557 [0.543, 0.580]	0.095 [0.079, 0.131]	0.086 [0.066, 0.117]
CRF + CT (reimplementation)	0.736 [0.716, 0.751]	0.363 [0.317, 0.421]	0.421 [0.363, 0.474]
TARNet	0.273 [0.267, 0.279]	0.218 [0.184, 0.288]	0.214 [0.184, 0.281]
DragonNet	0.252 [0.247, 0.258]	0.233 [0.176, 0.309]	0.304 [0.255, 0.358]
DragonNet + TARReg	0.240 [0.233, 0.246]	0.240 [0.186, 0.319]	0.294 [0.229, 0.331]
Logic-S (base)	0.958 [0.941, 0.974]	0.150 [0.094, 0.201]	0.134 [0.105, 0.170]
Logic-S + correction	0.488 [0.479, 0.499]	0.241 [0.164, 0.299]	0.164 [0.138, 0.205]
Logic-S + arm predicates	0.740 [0.720, 0.758]	0.184 [0.125, 0.274]	0.255 [0.204, 0.299]
Logic-S (CF3 base)	0.960 [0.945, 0.973]	0.092 [0.061, 0.135]	0.077 [0.059, 0.106]
Logic-S + correction (CF3)	0.332 [0.324, 0.340]	0.091 [0.070, 0.153]	0.117 [0.089, 0.144]
Treatment-structured SR	0.263 [0.256, 0.271]	0.172 [0.134, 0.228]	n/a

Table 12: Median normalized absolute ATE error, averaging fitting seeds 42–44 within each case before summarizing across cases. Brackets give 95% case-bootstrap intervals (5,000 resamples, all three seeds retained). Logic-S bases and their corrected counterparts share the same fitted circuits. SR covers only heterogeneous ACIC.

Model	IHDP	ACIC heterogeneous	ACIC all
Causal Forest	0.071 [0.066, 0.077]	0.032 [0.020, 0.039]	0.029 [0.019, 0.037]
CRF + CT (reimplementation)	0.086 [0.083, 0.089]	0.072 [0.059, 0.096]	0.126 [0.096, 0.176]
TARNet	0.038 [0.036, 0.040]	0.032 [0.024, 0.044]	0.039 [0.031, 0.051]
DragonNet	0.039 [0.037, 0.041]	0.035 [0.029, 0.047]	0.057 [0.046, 0.083]
DragonNet + TARReg	0.041 [0.039, 0.043]	0.037 [0.027, 0.050]	0.059 [0.045, 0.076]
Logic-S (base)	0.322 [0.310, 0.336]	0.079 [0.056, 0.103]	0.093 [0.073, 0.119]
Logic-S + correction	0.084 [0.079, 0.088]	0.061 [0.053, 0.071]	0.053 [0.045, 0.062]
Logic-S + arm predicates	0.222 [0.212, 0.231]	0.057 [0.050, 0.081]	0.086 [0.069, 0.123]
Logic-S (CF3 base)	0.277 [0.265, 0.289]	0.042 [0.027, 0.061]	0.051 [0.037, 0.067]
Logic-S + correction (CF3)	0.048 [0.045, 0.051]	0.030 [0.023, 0.046]	0.034 [0.028, 0.043]
Treatment-structured SR	0.032 [0.030, 0.035]	0.020 [0.016, 0.030]	n/a

C.2 Correction coefficient specification

For each fitting fold, selected dictionary columns are centered and scaled using that fold's pooled training rows, with scale floored at 10^{-8}. Let $\widetilde{z}_i$ denote these standardized terms, r_i the factual base residual, and $I_a = \{i \in I : t_i = a\}$. Each arm solves

$$(\widehat{\alpha}_a, \widehat{\beta}_a) = \arg\min_{\alpha_a, \beta_a} \frac{1}{|I_a|} \sum_{i \in I_a} (r_i - \alpha_a - \widetilde{z}_i^\top \beta_a)^2 + 0.01 \|\beta_a\|_2^2, \qquad a \in \{0, 1\}.$$

Intercepts are unpenalized. Coefficients are mapped back to the original term scale and then to Eq. (3). The penalty acts on each arm's slope vector, so treatment-difference coefficients are penalized only indirectly. Fold-local scaling and ridge apply throughout candidate scoring, acceptance searches, and the final fit. Candidate and acceptance scores remain factual validation losses with the existing complexity costs.

C.3 Seed stability with fixed fitting conditions

We additionally hold the implementation, CPU device, and data split fixed and vary only fitting seeds (45–54). The sample comprises 10 randomly selected IHDP cases, five heterogeneous and five constant-effect ACIC cases, and a separate stress set of 11 previously identified ACIC failure cases. Two ACIC cases belong to both samples and are fitted once. Across the 29 distinct cases, both S variants complete all ten seeds (580 fits). Table 13 reports their prediction and support stability.

In Table 13, A is the fraction of all case–seed fits that accept a correction. J_T and J_F measure Jaccard overlap of correction terms and their constituent covariates, respectively. Jaccard overlap is intersection size divided by union size; 1 denotes identical sets and 0 disjoint sets. We average over seed pairs within each case, then equally over cases. Both corrections must be accepted and their union nonempty. The count n_J gives the cases with at least one eligible pair. A term includes its operators, covariates, and constants but not its fitted coefficient. Commutative terms are identified independent of operand order, and differences up to sign.

V measures prediction variability. For each test patient we compute the sample SD of the predicted effect across ten seeds, take the root mean square over patients, and divide by training-outcome SD. The table gives the median across cases; lower values indicate more stable predictions. ρ is the Spearman correlation of predicted effects across test patients for two seeds, summarized by the median over seed pairs within each case and then the median across cases. Higher values indicate more consistent patient rankings. Its subscript gives the number of cases with an eligible pair. A pair is excluded if either predicted contrast has normalized SD at most 10^{-5}. Larger J_T and J_F indicate greater support agreement; A and n_J describe acceptance and coverage, not stability.

On IHDP, S-CF3 has ranking correlation 0.918, feature overlap 0.687, and exact term overlap 0.184. Thus stable predictions need not imply stable symbolic terms. These measures concern correction support; the complete pruned circuit and recovery of true effect modifiers are outside their scope.

C.4 Correction-library ablation

We run an exploratory library comparison on all 20 randomly selected cases in the fixed-condition seed study (10 IHDP, five heterogeneous ACIC, and five constant-effect ACIC), using all ten fitting seeds and both S methods. All 400 matched jobs (1,200 correction procedures) share neural bases within each library comparison. Ridge 0.01, feature screening, model-specific term limits, complexity costs, and acceptance thresholds remain fixed. Each library reruns selection, validation, and final fitting.

The full dictionary includes variables, squares, cubes, pairwise products, sums, differences, trigonometric functions, and scaled exponentials of products. The quadratic library retains variables, squares, and pairwise products; the interaction library retains variables and pairwise products. Neither restricted library imposes strong hierarchy. With 25 screened covariates, the search decreases from 1,925 to 350 or 325 candidates. This changes operator families and library size together, so the comparison does not isolate any single family.

Table 14 averages errors over seeds within each case and then equally over cases. Table 15 reports correction-term overlap using the same within-case definition as the seed-stability analysis. It con-

Table 13: Seed stability across ten CPU fitting seeds (45–54). A is correction acceptance rate; J_T and J_F are term and covariate Jaccard overlaps over accepted pairs in n_J eligible cases. V is normalized prediction variability (lower is more stable), and ρ is Spearman ranking correlation (higher is more stable), with eligible-case counts in subscripts. The text defines the within-case and across-case summaries.

Logic model	A	J_T	J_F	n_J	V	ρ
IHDP random (10 cases)						
S + correction	99%	0.119	0.577	10	0.285	0.741_{10}
S + correction (CF3)	96%	0.184	0.687	10	0.203	0.918_{10}
ACIC heterogeneous random (5 cases)						
S + correction	50%	0.152	0.336	3	0.173	0.319_{3}
S + correction (CF3)	40%	0.319	0.572	2	0.019	0.717_{3}
ACIC constant random (5 cases)						
S + correction	98%	0.085	0.248	5	0.114	0.264_{5}
S + correction (CF3)	80%	0.068	0.287	5	0.285	0.298_{5}
ACIC stress (11 cases)						
S + correction	87%	0.128	0.309	11	0.267	0.433_{11}
S + correction (CF3)	85%	0.160	0.353	10	0.255	0.200_{10}

ditions on both corrections being accepted, with eligible-case counts shown. A smaller dictionary increases chance overlap; these unadjusted scores measure syntactic agreement only. The small, fixed response-surface samples support an exploratory comparison of accuracy and representation stability.

For IHDP S-CF3, reducing the library to interactions raises mean PEHE from 0.331 to 0.370 while accepted-term overlap rises from 0.184 to 0.365. On constant-effect ACIC, the same restriction lowers mean PEHE from 0.321 to 0.276. On heterogeneous ACIC, S-CF3 improves slightly (0.154 to 0.145), whereas single-fit S worsens (0.246 to 0.291). No library is uniformly best; the differences describe an accuracy–representation tradeoff.

Table 14: Mean normalized PEHE (lower is better) in the expanded library ablation: ten fitting seeds, 10 IHDP and five cases per ACIC stratum. Full, quadratic, and interaction libraries share the same neural bases and ridge 0.01.

Cohort	Logic model	Full	Quadratic	Interactions
IHDP	S	0.507	0.546	0.550
IHDP	S-CF3	0.331	0.369	0.370
ACIC heterogeneous	S	0.246	0.289	0.291
ACIC heterogeneous	S-CF3	0.154	0.148	0.145
ACIC constant	S	0.489	0.433	0.435
ACIC constant	S-CF3	0.321	0.272	0.276

Table 15: Accepted correction-term overlap (higher is greater syntactic agreement; subscripts give eligible-case counts) in the expanded library ablation: ten fitting seeds, 10 IHDP and five cases per ACIC stratum. Full, quadratic, and interaction libraries share the same neural bases and ridge 0.01.

Cohort	Logic model	Full	Quadratic	Interactions
IHDP	S	0.119_{10}	0.211_{10}	0.217_{10}
IHDP	S-CF3	0.184_{10}	0.361_{10}	0.365_{10}
ACIC heterogeneous	S	0.152_{3}	0.166_{3}	0.198_{3}
ACIC heterogeneous	S-CF3	0.319_{2}	0.254_{2}	0.237_{2}
ACIC constant	S	0.085_{5}	0.191_{5}	0.171_{5}
ACIC constant	S-CF3	0.068_{5}	0.115_{5}	0.103_{5}

C.5 Causal Rule Forest comparator

Performance is reported in Tables 9 and 10.

We add CRF + CT, the Causal Rule Forest representation followed by a causal-tree estimator described by Hsu et al. (2024). We could not locate public author code and therefore report a documented reimplementation using the `htetree`,(Xu et al., 2026), causal-tree backend (version 0.1.23). It fits 200 honest trees, each using one randomly selected covariate, bucket splitting, and a per-arm minimum leaf size of one, matching the reported forest settings. Each tree supplies a categorical leaf-membership feature. A final causal tree combines these features; leaf identifiers are not treated as ordered numeric inputs.

The source paper does not fully specify representation depth, subsample size, or final-tree fitting settings. We fix one forest layer, sample 10% of development rows without replacement per tree, and split each sample into treatment-stratified growth and estimation halves. The final tree uses all development rows, split into growth and estimation halves, a per-arm minimum leaf size of two, and pruning at minimum 10-fold causal-tree validation error. These are reconstruction choices, not claimed author settings. Honesty is local to each tree. The upstream rule representation uses development outcomes from both final-tree halves. No fitted nuisance regressions or doubly robust targets are supplied. If every rule is constant, the final estimator is the estimation-half arm-mean difference.

C.6 Independent synthetic datasets

We follow the four response-surface setups of Nie & Wager (2021), using 1,500 observations, 12 independent covariates, and unit Gaussian noise. For each setup, dataset seeds 2000–2019 regenerate the covariates, treatments, and noise. All methods use fitting seed 42 and the same five-way split procedure, with 750 training, 225 validation, 225 discovery, 150 acceptance, and 150 common test rows per generated dataset. Inputs are uniform on $[0,1]$ in setup A and standard normal in B–D. In this paragraph, $X = (X_1, \dots, X_{12}) \in \mathbb{R}^{12}$ denotes raw covariates before preprocessing, and x is a realization. The scalar maps $b, \tau : \mathbb{R}^{12} \to \mathbb{R}$ give the arm midpoint and treatment effect, while $e : \mathbb{R}^{12} \to [0,1]$ is treatment propensity on the covariate support. For scalar $v \in \mathbb{R}$, define $[v]_+ = \max(0, v)$, $[v]_a^b = \min(b, \max(a, v))$ for real bounds $a \le b$, and $\mathrm{sp}(v) = \log(1 + e^v)$. The generating model is

$$T \mid X \sim \mathrm{Bernoulli}(e(X)), \qquad Y = b(X) + (T - 1/2)\tau(X) + \epsilon, \qquad \epsilon \sim N(0,1).$$

Here $T \in \{0,1\}$, $Y \in \mathbb{R}$, and the scalar Gaussian noise ϵ is independent of (X, T). Subscripts A–D identify the four setups:

$$\begin{aligned}
&b_A(x) = \sin(\pi x_1 x_2) + 2(x_3 - 1/2)^2 + x_4 + x_5/2, && \tau_A(x) = (x_1 + x_2)/2,\\
&e_A(x) = [\sin(\pi x_1 x_2)]_{0.1}^{0.9},\\
&b_B(x) = \max(0, x_1 + x_2, x_3) + [x_4 + x_5]_+, && \tau_B(x) = x_1 + \mathrm{sp}(x_2),\\
&e_B(x) = 1/2,\\
&b_C(x) = 2\,\mathrm{sp}(x_1 + x_2 + x_3), && \tau_C(x) = 1,\\
&e_C(x) = (1 + e^{x_2 + x_3})^{-1},\\
&b_D(x) = \{[x_1 + x_2 + x_3]_+ + [x_4 + x_5]_+\}/2, && \tau_D(x) = [x_1 + x_2 + x_3]_+ - [x_4 + x_5]_+,\\
&e_D(x) = (1 + e^{-x_3} + e^{-x_4})^{-1}.
\end{aligned}$$

The CF3 variant uses the exact large configuration from the IHDP panel, including symbolic capacity and acceptance thresholds. TARNet, DragonNet, and the causal forest use the same configurations, development roles, and fitting routines as the revised primary panel; neural widths are 200/100.

Intervals bootstrap the 20 independently generated datasets within each setup. Paired comparisons match dataset identifiers. Regenerated inputs, assignments, factual and interventional outcomes, and splits are checked across all 80 cases.

No runtime comparison uses these concurrently executed fits.

C.7 Symbolic-only control

The seed-averaged comparison with SR appears in Table 9.

Table 16: Independent synthetic datasets following Nie–Wager setups A and B. Each cell is median normalized PEHE with a 95% bootstrap interval over 20 generated datasets. CF3 configurations match the IHDP panel.

Model	Setup A	Setup B
Causal Forest	0.207 [0.192, 0.228]	0.278 [0.261, 0.298]
CRF + CT (reimplementation)	0.698 [0.622, 0.723]	0.629 [0.603, 0.674]
TARNet	0.215 [0.185, 0.247]	0.223 [0.197, 0.247]
DragonNet	0.199 [0.188, 0.242]	0.214 [0.199, 0.241]
DragonNet + TARReg	0.185 [0.166, 0.214]	0.223 [0.211, 0.245]
Logic-S + correction (CF3)	0.268 [0.241, 0.278]	0.153 [0.134, 0.190]

Table 17: Independent synthetic datasets following Nie–Wager setups C and D. Each cell is median normalized PEHE with a 95% bootstrap interval over 20 generated datasets. CF3 configurations match the IHDP panel.

Model	Setup C	Setup D
Causal Forest	0.051 [0.038, 0.059]	0.646 [0.613, 0.696]
CRF + CT (reimplementation)	0.834 [0.779, 0.909]	1.002 [0.952, 1.027]
TARNet	0.143 [0.124, 0.157]	0.236 [0.211, 0.256]
DragonNet	0.153 [0.132, 0.161]	0.246 [0.235, 0.250]
DragonNet + TARReg	0.162 [0.127, 0.179]	0.228 [0.212, 0.255]
Logic-S + correction (CF3)	0.147 [0.132, 0.190]	0.365 [0.347, 0.380]

The control replaces the logic base with training-arm means and applies the six-term restricted search with unpenalized coefficients, top-25 within-arm screening, five-fold CV, the fixed dictionary, complexity penalty, and improvement gates. With constant arm means, residual screening is equivalent to centered within-arm outcome screening. Candidates use the design $[1, z, t, tz]$ from Eq. (3); rejection returns the arm means. Search uses all 90% development rows and training-fitted preprocessing, without propensity models or doubly robust targets. The search is therefore a treatment-structured estimator.

The deployed symbolic size has median six terms in each evaluated population. The control matches symbolic capacity; Logic additionally retains its circuits.

Timing pilot. The timing pilot compares the small Logic-S additive profile with two restricted-search controls and finite-budget evolutionary search, all using factual outcomes without propensity or doubly robust targets. It covers IHDP simulations 001–012 and five ACIC variants with exactly 5,000 observations and 177 encoded covariates, selected in identifier order by size before outcomes are inspected. All methods share test rows and training-fitted preprocessing. The restricted ACIC search retains the production top-48 correlation screen, giving 7,008 dictionary entries, and selects at most four terms. All 177 covariates remain available to the neural backbone and the evolutionary methods. The latter use genetic programming as implemented in gplearn (Stephens, 2026), with S and T outcome formulations.

The restricted original-split control substitutes training-arm means for the neural base and uses the same discovery and acceptance data roles. The pooled control searches on all development rows with cross-validation. Evolutionary search evaluates three parsimony settings with a population of 1,000 for 20 generations, selects by validation factual MSE, then refits with population 2,000 for 20 generations. Its broader operator set remains limited by depth, population, and generation budgets.

On ACIC, median wall time is 16.48 seconds for Logic, 29.11 and 55.78 seconds for evolutionary S and T, and 31.67 seconds for pooled restricted search. Restricted search has lower PEHE than Logic in all five cases; the original-split control does so in three. These small, configuration-specific comparisons do not establish a general accuracy advantage or scaling law.

Table 18: Factual-training symbolic-search pilot. Medians of normalized errors and CPU wall time. Pooled standalone methods use more rows for symbolic fitting. These runs use the small Logic-S profile and are separate from the large CF3 benchmark.

Dataset	Method	PEHE	ATE error	Seconds
IHDP (12)	Logic-S + correction	0.437	0.089	2.59
IHDP (12)	Restricted SR (original splits)	0.454	0.082	1.21
IHDP (12)	Restricted SR (pooled)	0.216	0.019	2.04
IHDP (12)	Evolutionary SR-S (tuned GP)	0.534	0.168	20.01
IHDP (12)	Evolutionary SR-T (tuned GP)	0.406	0.084	39.21
ACIC (5)	Logic-S + correction	0.111	0.024	16.48
ACIC (5)	Restricted SR (original splits)	0.133	0.029	9.12
ACIC (5)	Restricted SR (pooled)	0.033	0.026	31.67
ACIC (5)	Evolutionary SR-S (tuned GP)	0.097	0.070	29.11
ACIC (5)	Evolutionary SR-T (tuned GP)	0.524	0.082	55.78

D Refinement and Effect Support

This section audits the seed-42 fitted models.

For each seed-42 test set we load the deployed contrast, the corresponding base-only contrast, and the noiseless treatment effect. Rejected corrections return exactly the base predictions. For the n test rows, define $\bar{\hat{\tau}} = n^{-1}\sum_{i=1}^{n}\hat{\tau}(x_i)$ and $\bar{\tau} = n^{-1}\sum_{i=1}^{n}\tau(x_i)$. With $\hat{\tau}_i = \hat{\tau}(x_i)$ and $\tau_i = \tau(x_i)$, centered effect RMSE is

$$R_{\text{het}} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}\{(\hat{\tau}_i - \bar{\hat{\tau}}) - (\tau_i - \bar{\tau})\}^2}.$$

The identity $\text{PEHE}^2 = (\bar{\hat{\tau}} - \bar{\tau})^2 + R_{\text{het}}^2$ holds for every base and deployed prediction array. The paired change is the deployed error minus its own fitted-base error, so negative values favor refinement. Rejected corrections contribute zero change to the population summaries.

Table 19: Paired median changes from adding the correction to its own fitted base. Negative changes favor refinement. Brackets give 95% bootstrap intervals; rejected corrections contribute zero.

Dataset / model	Δ PEHE	Δ centered RMSE
IHDP / S + correction	-0.447 [-0.460, -0.432]	-0.346 [-0.364, -0.317]
IHDP / S + correction (CF3)	-0.597 [-0.614, -0.580]	-0.527 [-0.545, -0.510]
ACIC-het / S + correction	0.000 [0.000, 0.000]	0.000 [0.000, 0.016]
ACIC-het / S + correction (CF3)	0.000 [0.000, 0.000]	0.000 [0.000, 0.000]

For effect-modifier recovery we reconstruct the known response-surface-B coefficients from IHDP's noiseless arm means. The treated surface is linear in the original covariates and the logarithm of the control surface is affine. Both fits must achieve $R^2 > 1-10^{-9}$ before a truth set is used. A covariate with a nonzero generating coefficient belongs to the effect-modifier set. Here R^2 is the coefficient of determination for the corresponding generating-surface regression. The correction's named set contains every coordinate appearing in an accepted term with a nonzero treatment coefficient. This definition concerns syntactic support; separate terms could in principle cancel functionally.

Precision is computed only for nonempty named sets, whose number is reported. Recall includes empty corrections as zero when the truth set is nonempty. Each simulation's chance precision is the fraction of its 25 covariates that are true modifiers. We bootstrap the per-simulation precision-minus-chance difference using the same nonempty cohort. This gives a support-recovery comparison conditional on the IHDP generator; it does not evaluate the causal meaning of an individual symbolic interaction.

Complete-contrast support. On the 30-case certification cohort, support is the union of covariates across all base arms and folds and nonzero correction-effect terms. A removed name establishes in-

Table 20: Covariates named by nonzero accepted correction terms on 1,000 IHDP cases. Precision is conditional on nonempty named sets; recall includes empty sets as zero. Brackets give 95% bootstrap intervals.

Model	Nonempty	Precision	Recall
Logic-S + correction	897/1,000	0.833 [0.833, 0.833]	0.500 [0.500, 0.500]
Logic-S + correction (CF3)	892/1,000	0.857 [0.833, 0.875]	0.700 [0.700, 0.727]

dependence of the fitted contrast from that input on the box. Causal irrelevance would require further assumptions. Retained names may cancel functionally; these are conservative syntactic supports.

Post-fit organization. Algorithm 5 assigns the shared correction to the contrast-null component and its treatment-dependent part to the contrast. Null-node counts measure base structure removed between logical and contrast stages; correction terms are outside those counts. Baseline support combines original control-arm programs with nonzero shared correction terms. Contrast support combines reduced arms with nonzero treatment-correction terms. The supports may overlap.

Table 21: Certified post-fit organization of the same 30 IHDP ensembles. Logical is the arm-specialized base graph; Contrast is the exact zero-budget treatment-effect representation; Null is the graph structure assigned to the contrast-null complement. Variable counts include the accepted correction and are conservative syntactic supports. Entries are medians with 95% bootstrap intervals.

Model	Logical nodes	Contrast nodes	Null nodes	Baseline vars	Contrast vars
Logic-S CF3	409.0 [367.0, 450.0]	96.5 [45.0, 185.0]	284.0 [239.0, 324.5]	25.0 [25.0, 25.0]	15.0 [13.0, 21.5]

E Stroke-Trial Case Study

E.1 Factorial aspirin–heparin audit

The broad audit retains 12,291 post-pilot IST patients with CT before randomization and observed six-month death-or-dependency status. Randomized aspirin and heparin assignments define six valid intervention cells. Prior heparin use, `RHEP24`, remains a pretreatment covariate; `RXHEP` is the randomized intervention. We fit the compact Logic-S base with three selected literals per clause, three crossfit folds, and procedural seeds 42–46. The third literal permits a clause containing aspirin, randomized heparin, and a clinical modifier. It appears in only two of 15 folds, with different modifiers.

For each fold, exact arm specialization first fixes aspirin and then sets the three heparin indicators to one of their valid one-hot assignments. This produces the six programs F_{ah} used in the main-text factorial contrasts. Computing the interaction by subtracting aspirin contrasts or by subtracting heparin contrasts agrees exactly in floating-point evaluation.

Table 22: Factorial six-month death/dependency contrasts in percentage points. Negative values favor the first active treatment. Randomized intervals are unadjusted Wald intervals standardized over the other randomized factor. Logic-S is the five-seed held-out consensus point estimate. Whole-box ranges average the 15 foldwise MILP lower and upper bounds and are conservative.

Contrast	Randomized RD (95% CI)	Logic-S RD	Whole-box interaction range
Aspirin	$-1.99\ [-3.82, -0.17]$	-1.55	n/a
Low-dose heparin	$+0.23\ [-1.89, 2.36]$	$+0.23$	n/a
Medium-dose heparin	$-0.14\ [-2.27, 1.98]$	$+0.19$	n/a
Aspirin × low dose	$-0.45\ [-4.70, 3.80]$	$+0.0005$	$[-0.157, 0.205]$
Aspirin × medium dose	$+0.05\ [-4.19, 4.29]$	-0.0007	$[-0.239, 0.222]$

Across patients, fitted interaction means/SDs are $0.00053/0.0083$ points for low dose and $-0.00072/0.0101$ for medium dose. All 30 mixed-integer interaction optimizations terminate successfully; 10/15 low-dose and 9/15 medium-dose folds have zero interaction over the box, subject to solver tolerance. The ordinary exact deletion routine removes no further nodes; the bounds constrain interaction magnitude only.

E.2 Aspirin-only evaluation and explanatory fit

The public IST database (Sandercock et al., 2011) supplies 5,978 post-pilot patients with CT before randomization, observed aspirin assignment and six-month outcome, no heparin in the preceding 24 hours, and allocation to avoid trial heparin. Observed aspirin ARR is 1.68 points (95% CI $[-0.81, 4.17]$). This retrospective demonstration does not encode contemporary treatment eligibility or contraindications.

Five outer folds produce OOF predictions for all patients, repeated for seeds 42–44. Logic-S fits only the outer training rows; three inner folds generate residuals for symbolic search. Held-out outcomes enter neither their base fit nor correction search. SR and S-RF use the same outer partitions. All fit factual outcomes without nuisance regressions or doubly robust targets.

The primary Logic-S result uses the base prediction from every outer fold. We also retain a diagnostic in which the regularized correction candidate is forced into every fold and centered on its correction-training rows. If $c_0, c_1 : [0,1]^d \to \mathbb{R}$ are the fitted arm corrections on n correction-training rows, $\rho = n^{-1}\sum_{i=1}^{n} t_i$ is their treatment prevalence, and $\bar{h} = n^{-1}\sum_{i=1}^{n}[c_1(x_i) - c_0(x_i)]$ their scalar mean correction contrast, the diagnostic uses

$$c_0^{\mathrm{ctr}}(x) = c_0(x) + \rho\bar{h}, \qquad c_1^{\mathrm{ctr}}(x) = c_1(x) - (1-\rho)\bar{h}.$$

This gives zero mean correction contrast on the training rows while retaining the candidate heterogeneity. The configured 2% gain gate accepted only 5 of 15 outer-fold corrections across the three procedural seeds. We report the base consistently and retain the forced candidate as a stress test. The configuration and this presentation were selected after examining earlier fits on the same cohort. Outer-fold exclusion protects individual fits, but does not remove this analysis-level adaptation; the comparison is exploratory.

Table 23: Full-cohort OOF evaluation in the stroke-trial cohort. Predicted ARR is control minus aspirin risk; positive means benefit. Rank interaction is on the adverse-outcome scale per SD of benefit rank; negative supports the predicted ordering. All risk differences are in percentage points. The last two columns estimate ARR as the control minus aspirin adverse-outcome rate within the lowest and highest thirds of the three-seed mean predicted-benefit score. The column p is the two-sided HC3 Wald interaction p-value. Intervals and tests are descriptive, conditional on the learned scores.

Model	MSE	Pred. ARR	Rank int.	p	ARR Low-score third	ARR High-score third
Logic-S	0.181	2.31	0.42	0.736	2.67	1.29
Logic-S + forced correction	0.180	2.32	-1.39	0.269	1.23	4.57
Restricted SR	0.178	2.59	-1.49	0.231	1.05	5.23
S-RF	0.180	3.03	-1.60	0.216	-0.60	4.28

The three-seed mean Logic-S score has observed ARR 2.67, 1.43, and 1.29 points across increasing thirds and an unresolved continuous interaction ($p = 0.736$); SR's interaction is also unresolved ($p = 0.231$), despite its lower factual MSE. Mean seed Spearman correlations are 0.15, 0.56, and 0.74 for Logic-S, SR, and the forest; top-tertile Jaccard similarities are 0.23, 0.49, and 0.55. Their consensus benefit SDs are 1.10, 3.06, and 6.77 points. Stability alone does not validate benefit rankings. The Logic-S OOF scores are not constant (predicted means 1.10, 2.47, 3.36 across thirds), so their rank instability cannot be attributed to the final explanatory fit's clamps.

Interpretation of ranking checks. HC3 tests and within-arm bootstrap intervals condition on learned scores; they omit fitting and selection uncertainty, and training sets overlap. Repeated seeds

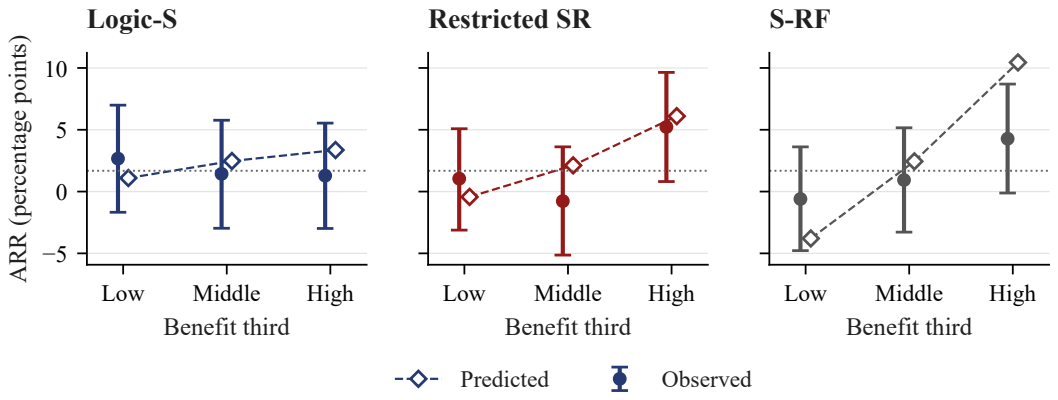


Figure 4: Randomized benefit-ranking check using the three-seed mean of OOF scores. Filled points are observed control-minus-aspirin risk differences; diamonds are mean predicted ARR. Intervals bootstrap randomized arms within each score tertile, holding scores fixed. Dotted horizontal lines show the full-cohort randomized ARR; dashed segments join predicted group means.

assess procedural sensitivity on the same patients. The 39,166-participant IST/CAST/MAST analysis found no evidence of benefit variation across its risk strata (Thompson et al., 2015). It includes IST and neither independently replicates this cohort nor establishes constant CATE for every profile. The forest's -0.60 to 4.28 and SR's 1.05 to 5.23 low-to-high tertile ARRs are noisy observed subgroup effects. They do not describe the spread of predicted individual benefit.

Final explanatory fit and rejected correction. The final eight-clause, two-literal Logic-S base has 744 parameters. Its pre-clamp ARR is 1.638678 points on every observed profile (SD below 10^{-14} points). After clamping, mean ARR is 1.52 (SD 0.41, range 0–1.64). Control scores reach the upper clamp for 7.63% of patients, both scores for 6.59%; 92.4% retain the full benefit. All observed variation arises from saturation. This establishes fitted constancy before clamping over the observed cohort. It does not address box-wide constancy or true effect homogeneity.

This case uses 120 base epochs and three restarts, with learning rate 0.003 and L1 weight 10^{-4}. The dictionary search screens at most 25 coordinates, selects at most six terms using five-fold symbolic CV, and uses complexity penalty 10^{-5} and marginal-gain threshold 0.002. The final base uses 5,380 training and 598 validation rows. A five-term correction fitted to seed-42 OOF base residuals achieved 1.876% cross-validated factual-loss improvement relative to a residual model with an intercept and treatment indicator, and failed the configured 2% search-acceptance gate. This comparison uses search cross-validation; it provides neither an independent deployment test nor a test of effect heterogeneity. For a stress test, we center and force this rejected candidate. Here $x \in [0,1]^{45}$ contains encoded covariates, $z(x) \in \mathbb{R}^5$ contains the five selected terms, and $b_z, b_{tz} \in \mathbb{R}^5$, $b_0, b_t \in \mathbb{R}$ are the coefficients in Table 24, defining $c(x,t) = b_0 + z(x)^\top b_z + t\{b_t + z(x)^\top b_{tz}\}$. Its raw mean contrast is $\bar{h} = -0.0047698$ before centering, so the correction to benefit is $-b_t - z(x)^\top b_{tz} + \bar{h}$. The shared coefficients affect outcome levels but cancel in benefit. These squared-error outcome fits are not probability-calibrated, and the additive correction is not constrained to keep both predictions in $[0,1]$. The forced candidate retains the base mean ARR but raises its benefit SD from 0.41 to 6.67 points and expands its range from $[0, 1.64]$ to $[-13.66, 88.99]$. Its largest interaction, "fully alert" times "arm/hand deficit: cannot assess," contributes 73.1 points and is nonzero for 19 of 5,978 patients. This sparse support makes the fitted interaction sensitive to a small number of outcomes; the gate does not diagnose its validity.

Exact specialization and rewriting reduce the final base from 76 to 58 nodes, with no further exact contrast deletions. The rational bound is zero on the input box; the largest neural/export gap is 1.22×10^{-7}. The guarantee concerns fidelity; the reduced circuit need not be minimal.

Table 24: Gate-rejected stroke correction used only as a stress test, before centering. Age is training-scaled to $[0, 1]$; all other entries are binary indicators of the named state. Coefficients use adverse-outcome units, with b_{tz} multiplying treatment.

Term	Shared coefficient	Treatment coefficient
Intercept	-0.227119	-0.010584
Age + unconscious	0.184001	0.024865
Leg/foot deficit + other deficit	0.001576	0.080923
Visible infarct × no hemianopia	0.066501	-0.003348
Age + hemianopia	0.099901	-0.089798
Fully alert × arm/hand deficit cannot assess	0.256168	-0.731082

E.3 Symbolic comparator term support

Each six-term SR fit uses 7–10 clinical variables, combining binary conjunctions, sums or differences, and indicator-gated exponential age curves. Term count thus understates the number of variables a reader must interpret.

In the first seed's first fold, the conjunction of no arm/hand deficit and hemianopia contributes $+21.39$ percentage points to fitted benefit. It occurs in 71 training and 17 held-out patients. A second fit selects the same conjunction at $+20.92$ points; the other 13 omit it. These are additive contributions to the fitted contrast. Total patient ARRs combine them with the other terms, and the subgroup effects are not independently validated. The SR contrast remains explicit. Shared prognostic coefficients cancel algebraically. This audit concerns the support and interpretation of its fitted heterogeneity.

E.4 Worked contrast and extracted DAG

The worked example in Section 5.2 is IHDP 005, seed-42 S-CF3. For $x \in [0, 1]^{25}$ after preprocessing, the following expressions give its three unscaled fold contrasts d_r, outcome scales s_r, and correction contrast Δ. Here $[v]_+ = \max(0, v)$ and the complete contrast is $\frac{1}{3}\sum_{r=1}^{3} s_r d_r(x) + \Delta(x)$. Coefficients are rounded to six decimals for display; the certificate uses the exact coefficients before rounding.

$$
\begin{aligned}
L_1(x) &\simeq 0.382321(1 - x_{19}) + 0.376237x_{24} + 0.204068x_5,\\
L_2(x) &\simeq 0.427527(1 - x_6) + 0.337107(1 - x_9) + 0.167571x_{11},\\
d_1(x) &\simeq -0.002595 + 0.084887\big\{1 - L_1(x) - [0.765866 - L_1(x)]_+\big\},\\
d_2(x) &\simeq 0.069723\big\{1 - L_2(x) - [0.677141 - L_2(x)]_+\big\},\\
d_3 &\simeq 0.029738,\\
(s_1, s_2, s_3) &\simeq (87.613526, 87.907792, 86.972054).
\end{aligned}
$$

$$
\begin{aligned}
\Delta(x) \simeq\ & 43.724554 - 14.165122\exp(2x_6x_2) - 7.027227(x_4 + x_{13})\\
& - 6.322231\exp(2x_6x_{19}) + 0.883253(x_{14} + x_{23})\\
& - 4.367715\exp(2x_4x_6) + 3.715164(x_6 - x_2).
\end{aligned}
$$

The extracted computation graph. Figure 5 shows the first fold's complete retained base graph.

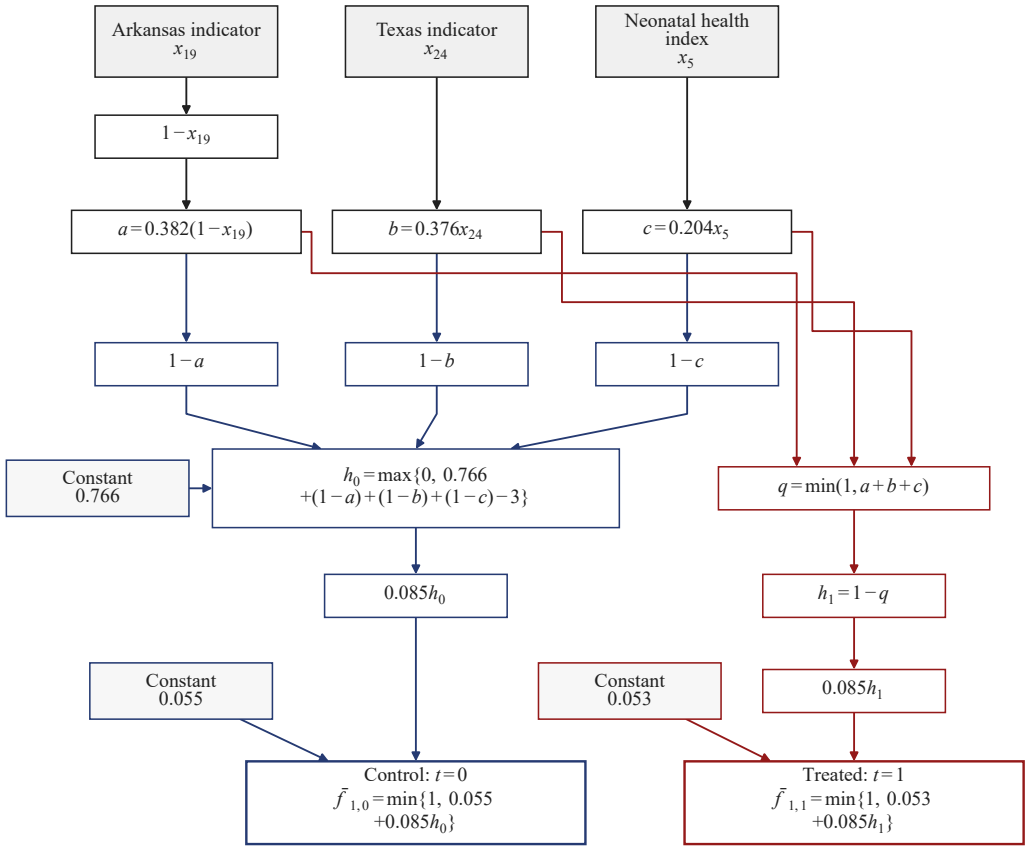


Figure 5: Extracted computation DAG for IHDP 005, seed 42, first fold of Logic-S + correction (CF3). Exact contrast pruning reduces 111 logically normalized base nodes to the 20 shown here. Every retained edge is drawn, and shared subexpressions appear once. Treatment enters the original shared circuit and is fixed at $t = 0$ and $t = 1$ before this graph is extracted. Gray input boxes identify covariates; blue and red paths lead to the control and treated scores. Each node displays its actual expression. Complementation subtracts its incoming value from one, and each clamped sum combines all incoming values. The four-input conjunction subtracts three. The local scalar intermediates $a, b, c, q, h_0, h_1 : D \to [0, 1]$ are defined inside the nodes; their dependence on x is suppressed for readability. Arrows carry node values; crossings do not join paths. The outputs $\bar{f}_{1,0}, \bar{f}_{1,1}$ are the pruned control and treated base scores. The certificate preserves their scaled difference throughout the unit box of 26 coordinates, whose treatment coordinate is unused after specialization. Individual arm scores may change. Node counts exclude outcome rescaling, subtraction, ensemble averaging, and the unchanged correction Δ. Coefficients are rounded to three decimals for display; the guarantee uses the exact rational values.