Research Square

Preprints are preliminary reports that have not undergone peer review.
They should not be considered conclusive, used to inform clinical practice,
or referenced by the media as validated information.

Exact Sequence-Domain Certification of Shifted-Block Tree Ensembles for CRISPR Off-Target Prediction

Illych Alvarez

ialvarez@espol.edu.ec

Universidad Católica de Santiago de Guayaquil

Research Article

Keywords:

Posted Date: November 27th, 2026

DOI: https://doi.org/10.21203/rs.3.rs-11816566/v1

License: This work is licensed under a Creative Commons Attribution 4.0 International License.
Read Full License

Additional Declarations: No competing interests reported.

Exact Sequence-Domain Certification of Shifted-Block Tree Ensembles for CRISPR Off-Target Prediction

Illych Álvarez Álvarez

Escuela Superior Politécnica del Litoral, Facultad de Ciencias Naturales y Matemáticas, Guayaquil, Ecuador

Universidad Católica de Santiago de Guayaquil, Facultad de Economía y Empresa, Guayaquil, Ecuador

Correspondence: ialvarez@espol.edu.ec

Abstract

Background

CRISPR–Cas9 off-target predictors are judged by accuracy, yet their behavior under uncertain nucleotide substitutions is rarely certified. Verification is complicated when mismatch count is a model feature derived from the perturbed sequence. We examine shifted-block boosted-tree ensembles that preserve this functional dependence while restricting positional interactions.

Results

For blocks of length b, the positional interaction graph has pathwidth at most $b-1$. A local factorization and count-augmented dynamic program calculate exact minimum and maximum raw prediction margins and attaining sequences over Cartesian nucleotide cuts, in $O(Lq^b(1+s_\alpha))$ arithmetic updates. Here L is sequence length, q is alphabet size and s_α is the number of count-ambiguous positions. In a reconstructed CHANGE-seq benchmark with 2,873,619 encoded observations, guide-disjoint nested evaluation of 109 eligible guides yielded mean per-guide average precision of 0.4287 (disjoint blocks), 0.4498 (shifted blocks) and 0.4887 (unrestricted boosting). Shifted blocks partly recovered the predictive performance lost under disjoint restrictions. Structural audits of trained ensembles verified the required interaction constraints; the available numerical checks supported finite-domain certification within reported floating-point tolerances.

Conclusions

Shifted-block ensembles enable exact sequence-domain margin optimization without treating a derived mismatch count as freely perturbable. The experiments use substitution-only data and synthetic possibility grades; certification describes the fitted predictor on stated uncertainty sets, not biological cleavage outcomes or calibrated uncertainty coverage.

Keywords: CRISPR off-target prediction; sequence uncertainty; boosted tree ensembles; exact certification; dynamic programming; pathwidth.

1 Background

CRISPR–Cas9 off-target activity depends on the relationship between a guide sequence and a candidate DNA target. Experimental resources such as CHANGE-seq provide large collections for developing and evaluating computational predictors [1]. Recent evaluations have also shown that reported accuracy depends materially on feature representations and the inclusion of explicit mismatch-distance information [2]. Predictive performance, however, is distinct from the question of how a fitted model's score may vary over a specified set of uncertain sequences.

Verification of tree ensembles has been studied using general-purpose search and bounding procedures [3], while feature-partitioned and structurally restricted ensembles provide alternative approaches to tractable verification [4, 5, 6]. Dynamic programming with global discrete statistics likewise has established antecedents [7]. These results motivate an exact treatment of sequence uncertainty that does not regard a mismatch count, which is computed from the sequence itself, as a freely perturbable input feature.

Here we investigate shifted position blocks trained in successive boosting phases, with the derived mismatch-count coordinate isolated from positional splits. We formalize certification over substitution-only nucleotide domains, give an explicit local factorization and a dynamic program that preserves the mismatch identity, and distinguish the resulting mathematical guarantees from numerical implementation error. We then evaluate predictive performance and certification using a reconstructed CHANGE-seq benchmark, while explicitly separating the literal benchmark preparation from a differently deduplicated exploratory reconstruction.

The contributions are an auditable positional pathwidth bound, an exact count-augmented certification recurrence with feasible extremizing sequences, complexity bounds sensitive to count ambiguity and a precise treatment of reusable optimizers across nested cuts. The empirical study documents the performance–certification trade-off for disjoint, shifted and unrestricted ensembles; independent external validation and additional matched computational experiments are outside the present study.

2 Problem formulation

2.1 Sequence-domain constraints and derived predictors

Let $\Sigma = \{\mathrm{A}, \mathrm{C}, \mathrm{G}, \mathrm{T}\}$ be the nucleotide alphabet, $q = |\Sigma| = 4$, and fix a guide $g = (g_1, \ldots, g_L) \in (\Sigma \cup \{\mathrm{N}\})^L$. For a candidate off-target sequence $t \in \Sigma^L$, define the sitewise mismatch contribution

under the benchmark's N convention by

$$\delta_i(a) = \mathbf{1}\{g_i \in \Sigma \text{ and } a \neq g_i\}, \qquad d_g(t) = \sum_{i=1}^{L} \delta_i(t_i). \tag{1}$$

A guide symbol N is resolved to the corresponding candidate base and thus contributes zero to the mismatch count. No substitution of a fixed base in Σ reproduces $\delta_i \equiv 0$, and the guide contributing the majority of positives carries N at position 21. The empirical setting has $L = 23$ aligned positions and admits substitutions only; insertions, deletions and bulges are excluded from the certification domain.

Let $e(a) \in \{0,1\}^q$ denote the one-hot vector of $a \in \Sigma$. For each position define the input block

$$x_i(t_i) = \begin{cases} e(g_i) \vee e(t_i), & g_i \in \Sigma, \\ e(t_i), & g_i = \mathrm{N}, \end{cases} \tag{2}$$

where $\vee$ denotes coordinatewise Boolean union. For a fixed guide, $t_i \mapsto x_i(t_i)$ is injective in both cases. If $g_i \in \Sigma$, matching yields one active coordinate and distinct mismatches yield distinct two-active coordinate patterns; if $g_i = \mathrm{N}$, the block equals the one-hot vector of the candidate base. The full model input $\big(x_1(t_1), \ldots, x_L(t_L), d_g(t)\big)$ has $qL + 1 = 93$ features. Consequently, a tree that splits only on positional features at sites B_j defines a function of t_{B_j} alone, as required below.

The predictor receives the positional encoding of t relative to g and an additional, derived mismatch-count coordinate k. Write $\widetilde{F}(t,k)$ for the raw ensemble margin when this coordinate is supplied independently. The biologically admissible evaluation is the restriction

$$F(t) = \widetilde{F}(t, d_g(t)), \qquad \mathcal{M}_g = \{(t,k) \in \Sigma^L \times \{0,\ldots,L\} : k = d_g(t)\}. \tag{3}$$

Thus, the mismatch count is *not* an additional independently uncertain input. The set $\mathcal{M}_g$ is the consistency domain of the lifted model.

Following foundational fuzzy-set terminology [8], an uncertainty specification supplies sitewise possibility grades $\theta_i : \Sigma \to [0,1]$. Under minimum aggregation, its nonempty α-cut is the Cartesian product

$$A_i(\alpha) = \{a \in \Sigma : \theta_i(a) \geq \alpha\}, \qquad Z_\alpha = \prod_{i=1}^{L} A_i(\alpha), \qquad A_i(\alpha) \neq \varnothing. \tag{4}$$

Possibility grades delimit sets of admissible sequences; they are neither cleavage probabilities nor calibrated coverage probabilities. The required certification outputs are

$$\ell_\alpha = \min_{t \in Z_\alpha} F(t), \quad u_\alpha = \max_{t \in Z_\alpha} F(t), \quad t_\alpha^- \in \arg\min_{t \in Z_\alpha} F(t), \quad t_\alpha^+ \in \arg\max_{t \in Z_\alpha} F(t). \tag{5}$$

Both extremal margins and attaining sequences are required. All formal correctness claims below refer to exact arithmetic applied to the fixed, represented tree parameters; finite-precision implementation issues are considered separately.

2.2 Why independent-coordinate verification changes the domain

Even when the admissible nucleotide sets are retained exactly, relaxing only the functional identity $k = d_g(t)$ changes the feasible domain. Define

$$K_\alpha = \{d_g(t) : t \in Z_\alpha\}, \qquad \widehat{Z}_\alpha = Z_\alpha \times K_\alpha. \tag{6}$$

The consistent pairs $\{(t, d_g(t)) : t \in Z_\alpha\}$ lie inside $\widehat{Z}_\alpha$. The relaxed maximum is therefore at least u_α, and the relaxed minimum is at most ℓ_α. Any independent-feature box containing this product domain can only be looser. This refers to a specific relaxation and does not assume that every generic verifier uses it.

Proposition 1 (The possible box-relaxation gap is unbounded)**.** *There exist bounded-length sequence domains and additive tree-ensemble margins for which the relaxation $Z_\alpha \times K_\alpha$ has an arbitrarily large absolute gap from optimization over $\mathcal{M}_g$.*

Proof. Fix any $L \geq 1$, let $g_1 = \mathrm{A}$ with $A_1(\alpha) = \{\mathrm{A}, \mathrm{C}\}$, and let $A_i(\alpha) = \{g_i\}$ for $i \geq 2$, so every $t \in Z_\alpha$ has $d_g(t) \in \{0, 1\}$ and $K_\alpha = \{0, 1\}$. For any $B > 0$, let a positional tree supported on position 1 have values $T(\mathrm{A}) = B$ and $T(\mathrm{C}) = 0$, and a distance-only tree have $G(0) = 0$ and $G(1) = B$. On the consistent domain, both sequences score B; therefore the true maximum is B. The independent-coordinate relaxation admits $(\mathrm{A}, 1)$, whose margin is $2B$. Its maximum exceeds the true maximum by B, with B arbitrary. Both functions can be implemented by single-split decision trees. □

This proposition identifies a consequence of neglecting a derived feature; it does not assert that every generic verifier uses this relaxation or that all relaxed instances have a nonzero gap. Fair benchmarking must specify exactly which input domain each comparator certifies.

3 Shifted-block tree-ensemble construction

3.1 Auditable restriction on positional interactions

Fix an integer block length $1 \le b \le L$. In phase $r \in \{0, \dots, b-1\}$, partition the positions into consecutive blocks of length at most b with the prescribed phase-dependent offset. An interaction-constrained boosting phase may use positional features from only one such block in any single tree. The mismatch-count input belongs to its own interaction group. Subsequent phases use the accumulated raw margin as the stated base margin; offset conventions must be fixed so that the final score is the sum of the actual learned phase contributions.

Let $\mathcal{P}$ index positional trees and $\mathcal{D}$ index mismatch-count-only trees. After the model has passed a structural audit, its lifted raw margin has the form

$$\widetilde{F}(t,k) = C_0 + \sum_{j \in \mathcal{P}} T_j(t_{B_j}) + \sum_{m \in \mathcal{D}} S_m(k) = \sum_{j \in \mathcal{P}} T_j(t_{B_j}) + G(k), \tag{7}$$

where B_j is the set of sequence positions used by tree j, each B_j is contained in some consecutive interval of length at most b, and $G(k) = C_0 + \sum_{m \in \mathcal{D}} S_m(k)$. A constant tree may be absorbed into G. The offset C_0 includes the audited initial margin and any phase intercepts; it need not be zero merely because a particular training implementation uses a zero-offset convention. A phase-level audit also checks that positions used by each tree belong to one permitted interaction group in that phase. The weaker consecutive-span restriction alone suffices for the mathematical guarantees below, whereas the phase-group audit documents faithful implementation of the proposed training procedure. Denote the number of positional trees by $M = |\mathcal{P}|$. The structural audit must inspect actual tree split variables rather than rely exclusively on training-time interaction settings. A tree mixing k with positional features, or using positions that cannot fit into a b-consecutive interval, invalidates (7) and is rejected.

Define the *positional interaction graph* H on vertices $\{1, \dots, L\}$ by making every tree scope B_j a clique.

Theorem 2 (Auditable positional pathwidth)**.** *Under* (7)*, the interaction graph satisfies*

$$H \subseteq P_L^{b-1}, \qquad \mathrm{tw}(H) \le \mathrm{pw}(H) \le \min\{b-1, L-1\}.$$

For $L \ge b$, the upper bound is sharp over the class of admissible interaction graphs.

Proof. Any two vertices in a tree scope B_j are separated by at most $b-1$ positions, so every

edge of H is an edge of the $(b-1)$st power of the length-L path. The consecutive windows $W_i = \{\max(1, i-b+1), \ldots, i\}$, indexed by $i = 1, \ldots, L$, form a path decomposition: each edge is covered by the window ending at its larger endpoint, and the windows containing any given vertex have consecutive indices. The maximum bag cardinality is $\min\{b, L\}$; therefore the width is at most $\min\{b-1, L-1\}$. Treewidth never exceeds pathwidth. If $L \geq b$, the graph P_L^{b-1} is realizable as a permitted positional interaction graph and contains a clique of size b. Its treewidth and pathwidth both equal $b-1$. □

The theorem is a structural guarantee, not by itself a complete tractability theorem for the consistency-constrained prediction problem.

Figure 1 illustrates the two shifted pair partitions used when $b = 2$.

3.2 Direct local-factor extraction

For $i = 1, \ldots, L$, let $W_i = [\max\{1, i-b+1\}, i]$ and define the endpoint group $\mathcal{P}_i = \{j \in \mathcal{P} : \max B_j = i\}$. Put

$$\phi_i(t_{W_i}) = \sum_{j \in \mathcal{P}_i} T_j(t_{B_j}). \tag{8}$$

Every factor ϕ_i depends on at most b consecutive nucleotides.

Lemma 3 (Exact factorization by rightmost position)**.** *If* (7) *passes the stated audit, then for all* $t \in \Sigma^L$*,*

$$F(t) = G(d_g(t)) + \sum_{i=1}^{L} \phi_i(t_{W_i}). \tag{9}$$

Proof. The disjoint sets $\mathcal{P}_1, \ldots, \mathcal{P}_L$ partition $\mathcal{P}$. For $j \in \mathcal{P}_i$, the permitted interval containing B_j implies $B_j \subseteq W_i$, so the corresponding summand is a function of t_{W_i}. Grouping the positional summands in (7) by $i = \max B_j$ and substituting $k = d_g(t)$ gives (9) without approximation. □

In the direct-tree representation, G already contains the intercept, and no additional anchoring constant is needed. Evaluating each member of $\mathcal{P}_i$ on every assignment of W_i requires at most Mq^b *individual tree evaluations* overall, each with its own split-traversal cost. This is distinct from $O(Lq^b)$ evaluations of the *full ensemble* in a black-box anchored-difference construction. The two methods are computationally and numerically different and must not be described as the same operation.

For the black-box alternative, fix an anchor $a^\circ \in \Sigma^L$ and an independent count input k_0, and

put $t^{(i)} = (t_1, \dots, t_i, a^\circ_{i+1}, \dots, a^\circ_L)$. The telescoping differences

$$h_i(t_{W_i}) = \widetilde{F}(t^{(i)}, k_0) - \widetilde{F}(t^{(i-1)}, k_0) \tag{10}$$

are functions of t_{W_i}: a tree unaffected by site i cancels, while a tree using i has all earlier variables in its scope at most $b-1$ positions to the left and all later variables anchored. Consequently,

$$\widetilde{F}(t,k) = \underbrace{\widetilde{F}(a^\circ, k_0)}_{C^\circ} + \underbrace{[\widetilde{F}(a^\circ, k) - \widetilde{F}(a^\circ, k_0)]}_{G^\circ(k)} + \sum_{i=1}^{L} h_i(t_{W_i}). \tag{11}$$

The tables h_i take $O(Lq^b)$ evaluations of the *full ensemble*, holding k_0 fixed. Although exact algebraically, subtraction of full-model margins can lose precision in floating-point arithmetic. In particular, the probes must not evaluate the consistent margin $F(t) = \widetilde{F}(t, d_g(t))$, because the count input must remain fixed while positional differences are formed.

4 Exact sequence-domain certification

4.1 An augmented-state recurrence

For a fixed nonempty cut Z_α, write $A_i = A_i(\alpha)$. Given a prefix $t_{1:i}$, define

$$u_i = \mathrm{suffix}_{\min\{i,b-1\}}(t_{1:i}), \qquad c_i = \sum_{j=1}^{i} \delta_j(t_j), \tag{12}$$

with $(u_0, c_0) = (\varepsilon, 0)$; when $b = 1$, the suffix is always empty. Write $\tau_i(u,a) = \mathrm{suffix}_{\min\{i,b-1\}}(ua)$, and define $w_i(u,a) = \mathrm{suffix}_{\min\{i,b\}}(ua)$. For $\star \in \{\max, \min\}$, let

$$V_i^\star(u,c) = \operatorname*{opt}_\star \sum_{\substack{t_{1:i} \in \prod_{j=1}^{i} A_j \\ u_i(t)=u,\ c_i(t)=c}}^{i} \phi_j(t_{W_j}), \tag{13}$$

where $\mathrm{opt}_{\max} = \max$ and $\mathrm{opt}_{\min} = \min$; an unreachable state has value $-\infty$ for maxima and $+\infty$ for minima.

Theorem 4 (Exact count-augmented dynamic programming)**.** *For each $\star \in \{\max, \min\}$, initialize $V_0^\star(\varepsilon, 0) = 0$ and all other states as unreachable. For $i = 1, \dots, L$, update every successor (v, c')*

by

$$V_i^\star(v,c') = \operatorname*{opt}_\star \limits_{\substack{(u,c)\ \textit{reachable},\ a\in A_i \\ v=\tau_i(u,a),\ c'=c+\delta_i(a)}} \left[V_{i-1}^\star(u,c) + \phi_i(w_i(u,a))\right]. \tag{14}$$

Then

$$\ell_\alpha = \min_{(u,c)\ \textit{reachable}} \{V_L^{\min}(u,c) + G(c)\}, \qquad u_\alpha = \max_{(u,c)\ \textit{reachable}} \{V_L^{\max}(u,c) + G(c)\}. \tag{15}$$

Saving the maximizing and minimizing predecessor and appended symbol at every reached state permits reconstruction of feasible sequences attaining these extrema.

Proof. We prove the maximizing case by induction on i; the minimizing case reverses the ordering. At $i = 0$, the only feasible prefix is empty, with score zero and state $(\varepsilon, 0)$. Assume the claim holds for prefixes of length $i-1$. Any feasible prefix $t_{1:i}$ ends with a symbol $a \in A_i$ and has a preceding state (u,c). Because ϕ_i depends only on the last at most b symbols, its value on $t_{1:i}$ is $\phi_i(w_i(u,a))$. The successor suffix and count are exactly $\tau_i(u,a)$ and $c + \delta_i(a)$. Suppose, for contradiction, that $t_{1:i}$ is optimal for its successor state while $\sum_{j<i}\phi_j < V_{i-1}^{\max}(u,c)$ for its own predecessor state (u,c), and let $\hat{t}_{1:i-1}$ attain $V_{i-1}^{\max}(u,c)$. Since $a \in A_i$, the extension $\hat{t}_{1:i-1}a$ is feasible; by the two preceding sentences it reaches the same successor state and incurs the same incremental factor $\phi_i(w_i(u,a))$, hence attains a strictly larger total, contradicting optimality of $t_{1:i}$. The prefix score therefore equals $V_{i-1}^{\max}(u,c)$ and the optimum in (14) is attained. Conversely, each listed transition corresponds to a feasible prefix, so the recurrence neither omits better candidates nor admits infeasible ones.

At $i = L$, the count equals $d_g(t)$. By (7) the count component G is a function of the mismatch count alone, so $G(c)$ is constant on the class of feasible sequences sharing a terminal state; combined with Lemma 3, adding it after the positional recurrence yields (15). The retained predecessor of each optimal state is itself an optimal prefix for its state. Backtracking from a terminal optimizer consequently returns a feasible sequence with exactly the stated extremal value. □

4.2 Numerical interpretation of exactness

The theorem establishes exact optimization of the specified mathematical function in exact arithmetic. It does not imply that independently implemented floating-point predictors are

Algorithm 1 Exact certification on one nonempty possibility cut

Require: Potentials ϕ_i; count function G; guide g; cuts A_i; mode $\star \in \{\min, \max\}$

$V \leftarrow \{(\varepsilon, 0) \mapsto 0\}$; allocate backpointer layers
for $i = 1, \ldots, L$ **do**
    $V_{\text{next}} \leftarrow \varnothing$
    for all $((u, c), v) \in V$ **do**
        for all $a \in A_i$ **do**
            $s \leftarrow \tau_i(u, a)$; $h \leftarrow c + \delta_i(a)$
            $w \leftarrow v + \phi_i(w_i(u, a))$
            if (s, h) unseen or w improves the stored value **then**
                $V_{\text{next}}[s, h] \leftarrow w$; save predecessor $((u, c), a)$
            end if
        end for
    end for
    $V \leftarrow V_{\text{next}}$
end for
Choose terminal $(u^\star, c^\star)$ optimizing $V[u, c] + G(c)$
return extremal value and sequence obtained by backtracking

bitwise identical. In particular, subtracting full-ensemble evaluations to obtain small anchored potentials can magnify rounding or cancellation error. Direct grouping of tree outputs by their rightmost sequence position avoids this subtraction and is therefore the preferred implementation when the trained tree structures are available. Any quantitative claim of machine-precision agreement on the *saved real-data models* must be established by rerunning those models with explicitly reported precision, evaluation order, and comparator tolerances.

5 Mathematical guarantees and computational complexity

5.1 Why the additive count matters

A generic factor-graph encoding can represent $k = d_g(t)$ by a single factor with scope $\{t_1, \ldots, t_L, k\}$. The primal graph of that *particular* representation contains the complete graph K_{L+1}, so its primal-graph treewidth is exactly L. This observation explains why the positional bound in Theorem 2 alone does not certify tractable inference under a naively encoded global consistency factor. It is *not* an intrinsic lower bound on the treewidth of every equivalent representation: introducing partial-count variables $c_i = c_{i-1} + \delta_i(t_i)$ decomposes the global condition into local updates. The recurrence of Theorem 4 exploits precisely this additive sufficient statistic and the bounded positional memory. No NP-hardness or general intractability statement is implied by the naive graph representation.

Proposition 5 (A bounded-width extended representation)**.** *Introduce integer partial counts* $c_0 = 0$ *and* $c_i = c_{i-1} + \delta_i(t_i)$ *for* $1 \le i \le L$*, and replace the global count input in the score by*

$G(c_L)$. *Together with the positional factors of Lemma 3, this yields a path decomposition of width at most* $b+1$*, independently of* L.

Proof. For $i = 1, \ldots, L$, take

$$D_i = \{t_j : \max(1, i-b+1) \le j \le i\} \cup \{c_{i-1}, c_i\}.$$

Each positional factor with right endpoint i is contained in D_i, as is the ternary transition enforcing $c_i = c_{i-1} + \delta_i(t_i)$; $G(c_L)$ lies in D_L. Each sequence variable t_j occurs in precisely the consecutive bags $D_j, \ldots, D_{\min(L, j+b-1)}$, while an internal count c_i occurs only in D_i, D_{i+1}, with the evident endpoint changes. Therefore the running-intersection condition holds. Each bag has at most $b+2$ variables, giving width at most $b+1$. The fixed c_0 can be removed. □

The benefit is the structured extended representation, not a claim that an intrinsically intractable problem has been solved without transformation. Nor does Proposition 5 by itself imply the count-sensitive cost of Theorem 7. Two variables of each bag represent counts with domains up to $L+1$ rather than q: a naive bag-configuration enumeration can therefore inspect up to $O(L^3 q^b)$ assignments over the L bags. Theorem 4 instead uses an additive partial count and the reachability result of Lemma 6 to avoid that crude count-state enumeration. Such sufficient-statistic augmentations have precedents; the contribution claimed here is their integration with the auditable shifted-block sequence architecture.

5.2 Complexity sensitive to count ambiguity

Call site i *count-ambiguous* at level α if δ_i is non-constant on $A_i(\alpha)$, that is, if $A_i(\alpha)$ admits both a matching and a mismatching nucleotide. For an unambiguous guide base this is equivalent to $g_i \in A_i(\alpha)$ and $|A_i(\alpha)| \ge 2$. Let

$$s_\alpha = \#\{i : \delta_i \text{ is non-constant on } A_i(\alpha)\}, \tag{16}$$

and define $s_{\alpha,i}$ analogously over positions $1, \ldots, i$.

Lemma 6 (Reachable mismatch counts)**.** *For any* i*, the reachable partial mismatch counts form exactly the integer interval* $\{r_i, r_i+1, \ldots, r_i+s_{\alpha,i}\}$*, where* $r_i = \#\{j \le i : \delta_j \equiv 1 \text{ on } A_j(\alpha)\}$*. In particular, there are exactly* $s_{\alpha,i}+1$ *possible partial mismatch counts.*

Proof. Each site $j \le i$ falls into exactly one of three cases. If $\delta_j \equiv 1$ on $A_j(\alpha)$, every admitted

nucleotide contributes one mismatch; these sites are counted by r_i. If $\delta_j \equiv 0$ on $A_j(\alpha)$, every admitted nucleotide contributes zero; this covers both $A_j(\alpha) = \{g_j\}$ and an ambiguous guide site, at which δ_j vanishes identically. Otherwise δ_j is non-constant on $A_j(\alpha)$ and the contribution may independently be zero or one; these sites are counted by $s_{\alpha,i}$. Since the choices at distinct sites are independent, adding the forced mismatches and selecting any number from zero to $s_{\alpha,i}$ of the count-ambiguous sites realizes precisely the claimed interval. □

Theorem 7 (Count-sensitive certification cost)**.** *After local-potential tabulation, one extremum and its witness on a nonempty cut Z_α can be computed in at most*

$$O\Big(Lq^b(1+s_\alpha)\Big) \tag{17}$$

arithmetic updates. The same asymptotic order applies to computing both extrema and witnesses. A rolling implementation requires $O(q^{b-1}(1+s_\alpha))$ memory for the current numerical value layers, plus $O(Lq^{b-1}(1+s_\alpha))$ predecessor records if all witnesses are reconstructed by direct backtracking. The factor tables require $O(Lq^b)$ entries in a simple dense implementation.

Proof. At layer i, the suffix has at most $q^{\min(i,b-1)}$ possibilities. By Lemma 6, the count has at most $1+s_\alpha$ possibilities, so there are at most $q^{b-1}(1+s_\alpha)$ reachable states. Each state generates at most q nucleotide transitions, each consisting of a bounded number of table lookups and arithmetic operations. Summing over L layers proves (17). Keeping only the present and next value layers proves the rolling-space bound; preserving predecessor records for reconstruction costs an additional layer factor L. Directly tabulating the L local factors has at most Lq^b values. □

For fixed nucleotide alphabet and fixed block length, the resulting algorithm is polynomial in L and the number of admissible count values, rather than exponential in the number of uncertain sites. The dependency q^b is exponential in the prescribed block width and must remain visible when b is allowed to vary. An implementation based on direct tree traversal requires at most Mq^b individual tree evaluations for factor extraction, multiplied by the relevant per-tree traversal cost. Additionally, tabulating $G(0), \ldots, G(L)$ requires $L+1$ evaluations of the complete count-only sum (or equivalent direct enumeration of the count-only trees). For Q separately solved cuts,

shared extraction gives the accounting

$$\begin{aligned} &\text{extraction: } \leq Mq^b \text{ positional-tree evaluations and } L+1 \text{ count-only-sum evaluations,} \\ &\text{DP: } O\left(q^b L \sum_{\alpha \in \mathcal{A}} (1+s_\alpha)\right) \end{aligned} \tag{18}$$

for the set $\mathcal{A}$ of Q levels. Rebuilding factors for each cut instead incurs up to QMq^b individual tree evaluations. Both approaches have the same worst-case order in the dominant DP arithmetic term for fixed parameters; sharing extraction is an *operational saving*, not a general claim of better asymptotic complexity in Q.

For $L = 23$ and $s_\alpha = 4$, the ratio between the coarse bound using $L + 1 = 24$ count slots and the refined count factor $s_\alpha + 1 = 5$ is $24/5 = 4.8$. This is a comparison of theoretical state-count upper bounds, not a measured 4.8-fold wall-clock speedup.

5.3 Nested cuts and exact witness reuse

Proposition 8 (Nested certification intervals)**.** *If $\alpha \leq \alpha'$ and both cuts are nonempty, then*

$$Z_{\alpha'} \subseteq Z_\alpha, \quad \ell_\alpha \leq \ell_{\alpha'}, \quad u_\alpha \geq u_{\alpha'}.$$

Proof. The defining threshold condition implies $A_i(\alpha') \subseteq A_i(\alpha)$ for each site, hence the product-set inclusion. Taking extrema over nested sets yields the stated inequalities. □

Proposition 9 (Exact reuse of a feasible old optimizer)**.** *Let $\alpha \leq \alpha'$ and let t^+ maximize F on Z_α. If $t^+ \in Z_{\alpha'}$, then t^+ also maximizes F on $Z_{\alpha'}$ and $u_{\alpha'} = u_\alpha$. The corresponding statement holds for a minimizing witness t^-.*

Proof. The nested-set inclusion gives $u_{\alpha'} \leq u_\alpha$. Feasibility of t^+ at the higher level gives $u_{\alpha'} \geq F(t^+) = u_\alpha$. The argument for minima reverses the inequalities. □

Earlier fuzzy constraint work also considered successive cut levels [9]. Here, the reuse test follows directly from nested feasible sets and is exact. This permits a level-by-level algorithm with separate feasibility checks for maximizing and minimizing witnesses. Write $\mathcal{R}^+$ and $\mathcal{R}^-$ for the levels at which each prior witness is absent or no longer feasible and the corresponding DP must be rerun. Let $R^+ = |\mathcal{R}^+| \leq Q$ and $R^- = |\mathcal{R}^-| \leq Q$. Given precomputed factor tables, the

complete profile has arithmetic cost

$$O\left(QLq + Lq^b\left[\sum_{\alpha\in\mathcal{R}^+}(1+s_\alpha) + \sum_{\alpha\in\mathcal{R}^-}(1+s_\alpha)\right]\right). \tag{19}$$

The QLq term accounts for constructing the level sets and checking witness feasibility; $R^+ + R^-$ counts individual extremum computations, not distinct optimum values, and the worst case still has $R^+ = R^- = Q$. If the $2Q$ witnesses are explicitly retained, output storage itself requires up to $O(QL)$ nucleotide symbols, separate from working space. The exact savings depend on tie-breaking and the uncertainty pattern. A small synthetic illustration reused an optimizer at 15 of 26 levels; that figure has not yet been reproduced on saved CRISPR models.

5.4 What the guarantees do and do not establish

The results of Sections 3–5 apply to audited models meeting the split restrictions in (7) and to nonempty Cartesian substitution-only sequence cuts. They establish exact extrema of the *represented model margin* and feasible extremizing sequences. They do not establish calibration of synthetic possibility grades, bounds on true Cas9 activity, or external predictive generalization. The decomposition and count augmentation are closely related to established dynamic-programming ideas; the specific contribution tested here is their integration with an auditable shifted-block tree architecture and a derived, internally consistent sequence feature. The novelty assessment must compare this construction directly with earlier constrained-ensemble and global-statistic inference frameworks.

6 Experimental design

6.1 Benchmark reconstruction and reproducibility audit

The empirical basis is the CHANGE-seq compendium of genome-wide Cas9 activity [1], in the processed form distributed with the SysEvalOffTarget pipeline and used in the recent evaluation of feature representations for off-target prediction [2]. Positive observations are off-target sites with more than 100 reads; putative negatives are produced by the publicly released negative-generation procedure. Each candidate is a guide–target pair of length $L = 23$ without insertions or deletions, encoded as in Section 2.

Reconstructing the published preprocessing exposed an index-alignment issue in the negative deduplication loop: a Boolean mask formed from the experimental dataframe is combined with a

duplicated-sequence mask formed from the negative dataframe, so the two masks are not aligned on a common index. Because this affects the composition of the negative set, and consequently downstream prevalence and ranking statistics, we distinguish the constructions explicitly rather than pooling their outcomes. The complete nested guide-disjoint evaluation available here is based on the literal protocol; the count for the faithfully filtered intended-deduplication reconstruction and predictive scores from a separate earlier corrected reconstruction serve different evidential roles:

Literal protocol. The original index-sensitive operations are reproduced exactly. This is the protocol that reaches the published-count checkpoint and is used as the main empirical benchmark, so that the comparison with prior reported figures is made on the same object.

Intended-deduplication protocol. Within-target deduplication is applied to the faithfully filtered input, which is what the surrounding code appears to intend. This yields a non-identical dataset, reported separately.

Counts from these reconstructions must not be interchanged; Section 7.1 reports them side by side. We make no claim about which reading the original authors intended, and do not characterize the discrepancy as an error in any published result. In particular, the exploratory AP row from the earlier corrected reconstruction is not a matched nested-CV analysis of the intended-deduplication protocol.

6.2 Encoded datasets

From the literal checkpoint, nine candidate negatives are excluded before encoding: one with zero mismatch distance, and eight carrying symbols outside Σ. The resulting encoded dataset is the one on which all nested cross-validation reported below is performed. Dataset sizes are given in Table 1. The working analysis reports the SHA-256 digest `ce62ad453b979eaf2a8cfe2c3a18bdf266d789bb23ba0deb72030100b423dce1` for the literal encoded dataset; the canonical encoded-file digest and the deposited archive should be checked for consistency before the associated reproduction claims are made.

6.3 Model configurations

Three configurations are compared, all gradient-boosted tree ensembles [10] with a total budget of 200 trees, learning rate 0.1, maximum depth 10, histogram-based split finding, and class weights computed from the training split only:

free. Unconstrained boosting. Not certifiable by the method of Section 4, and used only as an external performance comparator.

strict2. Disjoint 2-position blocks, a single phase. Certifiable with $b = 2$.

shift2. Two-phase shifted-block boosting with $b = 2$: two successive 100-tree phases, the second fitted with the raw margins of the first supplied as base margin, the final score being the sum of both phase margins under the documented zero-offset convention. Certifiable with $b = 2$.

Every fitted certifiable model is submitted to the structural audit of Section 3 before any certification claim is made. The audit inspects realized split variables rather than training-time interaction settings, and a model failing either audit condition is rejected rather than repaired.

6.4 Nested guide-disjoint evaluation

Partitions are guide-disjoint throughout: no guide contributes rows to both sides of any split. The outer loop is a five-fold guide-disjoint partition, each fold holding out approximately 22 distinct guides. Within each outer training fold, a three-fold guide-disjoint inner loop selects between the two *certifiable* candidates, *strict2* and *shift2*, by larger mean per-guide average precision; *free* is evaluated on the held-out guides but is never a selection candidate, since selecting it would defeat the purpose of the comparison.

Performance is summarized by average precision computed per guide and then averaged over guides, restricted to test guides having at least one positive and one negative candidate. This yields 109 evaluable guides across the five outer folds. Guide-macro averaging is used because row-count dominance is extreme in this benchmark (Section 7.1); it mitigates that dominance without removing heterogeneity in class prevalence across guides, which remains a limitation of the comparison rather than something the protocol corrects.

> **Status of the inferential analysis.** The shifted-block architecture was proposed after exploratory work on this same benchmark. Nested resampling prevents held-out labels from entering within-fold selection, but it does not convert the study into prospective external confirmation. All p-values and intervals below are therefore reported as exploratory.

6.5 Statistical analysis

The primary comparison is the paired per-guide difference between *shift2* and *strict2* over the 109 evaluable guides. We report the mean and median paired difference, the counts of improvements,

deteriorations and ties, a two-sided Wilcoxon signed-rank test, an exact two-sided sign test on the non-tied guides, and a 20,000-replicate guide-level bootstrap interval for the mean difference. Because the five outer folds are the unit at which model selection occurred, we additionally report an exact fold-level sign-flip test, which is deliberately conservative. A sensitivity analysis excludes the dominant guide. The comparison against *free* is reported with the same guide-bootstrap machinery.

6.6 Certification verification protocol

The certification validation protocol combines three complementary checks. The completed historical checks and those requiring re-execution are distinguished in Section 7:

1. **Exhaustive enumeration over the cut.** For cuts small enough to enumerate, every $t \in Z_\alpha$ is scored and the extrema compared against the recurrence of Theorem 4. Enumeration is exhaustive *over* Z_α, not over Σ^L; the number of sequences enumerated is reported for each scenario so the scope of the check is unambiguous.

2. **Structured mixed-integer optimization.** An MILP is formulated on the same discrete domain, with the nucleotide simplex and the functional identity $k = d_g(t)$ both imposed as constraints, so that it certifies the same object rather than a relaxation.

3. **Internal consistency of the profile.** Monotonicity of ℓ_α and u_α across nested levels (Proposition 8) is checked at every level, and witness reuse (Proposition 9) is cross-checked against full recomputation wherever it applies.

Two specifications govern how these checks are run and must be stated with the results, because they determine the achievable agreement.

Potential tabulation route. The historical real-model results are retained with their original numerical agreement levels. The planned verification re-run will tabulate local potentials by direct grouping of tree outputs by rightmost position, as in (8); it has not yet produced replacement real-model results. As noted in Section 4, anchored differences and direct grouping are algebraically equivalent in exact arithmetic but not necessarily numerically equivalent: anchored differences may lose precision by cancelling full-ensemble evaluations. Every real-model comparison will therefore report the extraction route, arithmetic precision, evaluation order and tolerances.

Ambiguous guide sites. The mismatch indicator used by the recurrence must be the one the trained model consumes. For guides carrying the ambiguous symbol N, the implemented convention resolves it by the candidate base, so $\delta_i \equiv 0$ at that site (Section 2). Any certification run on such a guide must use the general δ_i; substituting a fixed base changes d_g and therefore certifies a different function. The guides used in each certification scenario must be identified individually in the reproducibility archive so that this condition is checkable.

6.7 Comparators and the domains they certify

Reported comparisons distinguish verifiers by the domain each one certifies, because a difference in optimal value between two procedures on two different domains is not evidence about either. Anytime bounds for general tree ensembles [3] operate on per-feature boxes and enforce neither the nucleotide simplex nor the identity $k = d_g(t)$; by Proposition 1 the resulting gap is not controlled. Such a verifier is therefore reported as an informative generic baseline, never as an equal-domain comparator. The structured MILP of Section 6.6 is the equal-domain comparator.

A planned complementary experiment will quantify the effect of ignoring the derived-feature consistency constraint by comparing u_α with the relaxed optimum obtained when k ranges independently over the counts realized by Z_α. This isolates the consistency constraint while keeping the model and the positional cut fixed. The real-model outcome is *not reported*.

7 Results

7.1 Reproducibility of the benchmark

Reproducing the original index-sensitive negative preparation yields 2,806,152 candidate negatives at the published-count checkpoint, matching the total reported in the prior evaluation. Applying within-target deduplication to the faithfully filtered input instead yields 2,392,443 negatives before the additional encoding restrictions; an earlier, differently filtered reconstruction contained 2,394,901. Table 1 collects the counts.

Table 1: Dataset reconstruction counts for CHANGE-seq preprocessing protocols.

Protocol	Negatives	Positives	Encoded rows
Literal, published-count checkpoint	2,806,152	67,476	—
Literal, after encoding exclusions	2,806,143	67,476	2,873,619
Intended deduplication, before encoding restrictions	2,392,443	67,476	—
Earlier corrected reconstruction	2,394,901	67,476	—

Notes. Nine candidate negatives are excluded before encoding: one with zero mismatch distance and eight carrying symbols outside Σ. The reconstructions are protocol-dependent and must not be pooled.

Positive counts are concentrated to an unusual degree. The dominant guide (guide ID 70) has sequence `GGTGGTGTGGGCCCAGGAGGNGG` and contributes 40,906 of the 67,476 positives, that is 60.62%. Its empirical positive prevalence in the encoded literal dataset is 35.84% (40,906/114,125), against 0.96% across all other guides combined; under the earlier corrected reconstruction the same figures are 43.26% and 1.12%. Guide-macro averaging is adopted in response to this concentration, and the dominant guide is additionally excluded in a sensitivity analysis.

This guide carries the ambiguous symbol N at position 21. Both the mismatch contribution (1) and the positional encoding (2) explicitly cover that convention, ensuring that the input representation remains defined at the dominant guide.

7.2 Predictive performance under certifiability constraints

Nested selection preferred shifted blocks to disjoint blocks in all five outer folds, without using held-out labels for within-fold selection (Table 2). At outer fold 0 both certifiable models exceeded unconstrained boosting, which shows that the aggregate advantage of the unrestricted model does not hold on every test fold; that fold also contains the dominant guide.

Table 2: Nested guide-disjoint cross-validation on the literal CHANGE-seq dataset.

	Inner selection		Held-out guides		
Outer fold	*strict2*	*shift2*	*free*	*strict2*	*shift2*
0	.4157	.4400	.5025	.5048	.5120
1	.4125	.4468	.4441	.4180	.4306
2	.4146	.4294	.4896	.3975	.4250
3	.4492	.4589	.4875	.4008	.4329
4	.4300	.4454	.5195	.4212	.4479

Notes. All models use 200 trees. Inner columns describe selection scores over three inner folds; outer scores are measured on held-out guides. The unrestricted model is not used for inner selection.

Pooling the 109 evaluable outer-test guides gives mean per-guide average precision of 0.48866 (*free*), 0.42873 (*strict2*) and 0.44985 (*shift2*). The mean paired gain of *shift2* over *strict2* is 0.02112 with median 0.01655, from 74 improvements, 33 deteriorations and 2 ties. The two-sided Wilcoxon signed-rank test gives $p = 8.12 \times 10^{-5}$, the exact two-sided sign test on the 107 non-tied guides gives $p = 9.18 \times 10^{-5}$, and the 20,000-replicate guide bootstrap gives a 95% interval of [0.00953, 0.03341] for the mean difference. The conservative fold-level sign-flip test, with five folds all favouring shifted blocks, gives the smallest attainable value $p = 0.0625$. Excluding the dominant guide gives a mean paired gain of 0.02151 with $p = 6.30 \times 10^{-5}$, so the effect is not an artefact of that guide.

The gap to unconstrained boosting remains −0.03881. The corresponding guide-bootstrap interval is [−0.06492, −0.01395]. Shifted blocks therefore recover 35.2% of the per-guide average

precision lost by imposing disjoint positional restrictions. Table 3 reports the three model classes on the main literal dataset and on an earlier corrected exploratory reconstruction. The ordering of the three model classes is preserved in these two records; nevertheless, the second row is not a matched nested-CV re-evaluation of the faithfully filtered intended-deduplication reconstruction.

Table 3: Guide-macro average precision across two distinct dataset reconstructions.

Protocol	*free*	*strict2*	*shift2*
Earlier corrected reconstruction (2,394,901 negatives)	0.5008	0.4602	0.4722
Literal published-count (2,806,143 encoded negatives)	0.4887	0.4287	0.4498

Notes. The literal row corresponds to nested evaluation; the earlier corrected row is an exploratory reconstruction. These rows are not matched nested cross-validation of the same guide partitions.

The two recorded protocol-specific AP series are visualized in Figure 2; the held-out guide-level paired comparison is shown in Figure 3. The former should not be interpreted as a matched preprocessing experiment.

We do not treat the figure of 0.515 reported in the prior evaluation [2] as a head-to-head comparator, since ensemble size and protocol details differ.

7.3 Structural audit of the trained models

The two saved phases of the *shift2* model contain 100 trees each. Phase 0 comprises 31 count-only and 69 position-only trees; phase 1 comprises 21 count-only and 79 position-only trees. No tree in either phase mixes the count feature with positional features, so audit condition (i) of Section 3 is satisfied. The realized positional interaction graph contains 21 edges, all between positions at separation one, so audit condition (ii) is satisfied with $b = 2$ and the realized graph has treewidth and pathwidth one, consistent with Theorem 2. The audit therefore establishes (7) for these models as a verified property, which is the hypothesis every certification claim below relies on.

7.4 Numerical verification of exact certification

Table 4 reports agreement between the recurrence, exhaustive enumeration over the cut, and the equal-domain MILP, on real-data-trained audited models with synthetic possibility grades. Agreement with the MILP was exact to solver optimality tolerance in every scenario. Agreement with exhaustive enumeration was of the order of 10^{-6}.

Interpretation of the residual. These residuals are a property of the extraction route used, not of Theorem 4, which is an identity in exact arithmetic. The magnitudes in Table 4 are consistent with recovering local potentials by cancelling full-ensemble evaluations in single

Table 4: Numerical certification checks on trained tree ensembles with synthetic uncertainty.

Ambiguous sites	Levels	Max. enumeration error	MILP discrepancy
4	11	1.19×10^{-7}	0
8	22	9.80×10^{-7}	0
12	26	1.29×10^{-6}	0

Notes. Enumeration is exhaustive on the respective Z_α cut, not on all Σ^L. The individual cut sizes were not preserved for each scenario in the supplied record. A separate evaluation on three real guides with cuts of 11–27 levels reported maximum factorization discrepancy 4.3×10^{-6} and zero monotonicity violations.

precision, where the expected absolute error is of order $\varepsilon_{32}\|F\|$. A reference implementation of the recurrence applied to synthetic ensembles with grouped tabulation in double precision agreed with exhaustive enumeration to 7.1×10^{-15} across 40 configurations, with every returned witness feasible and reproducing its reported optimum. The corresponding re-run on the saved CRISPR models, with grouped tabulation and explicitly reported precision and summation order, was not available for the present manuscript; therefore no machine-precision claim is made for those models. Scenarios involving a guide that carries N must be re-run under the general δ_i of Section 2; the guides used in each scenario are listed with the deposited outputs.

7.5 Profile cost

The count-augmented recurrence can reuse one tabulation of local potentials across all levels of a cut profile. Compared with a naive recertification implementation that rebuilds the tables at each level, this avoids repeated model evaluations; an independently implemented verifier could also cache its own tables. In the supplied measurements, a complete single-level certificate required 0.98–2.54 ms and a complete profile of 11–27 levels required 1.64–9.93 ms, both inclusive of tabulation. Consistent with the accounting in Section 5, this is an operational saving in the model-evaluation cost model and not an improvement in asymptotic order: the dominant arithmetic term is unchanged, and what is amortized is tabulation, whose cost per unit is large because each tree evaluation traverses splits.

The equal-domain MILP returned identical optima on the same domains and was slower in the measured cases, though the ratio varied considerably with solver environment. We make no claim of asymptotic advantage over a specialized MILP or over an anytime verifier supplied with equivalent structural information.

Measured savings from witness reuse (Proposition 9) on the saved CRISPR models have not been established by the available real-model outputs. A synthetic illustration reused an optimizer at 15 of 26 levels, with values agreeing exactly with full recomputation at every reused level; this is reported as an illustration and not as a validated figure for the trained models. Likewise, the

measurement of box-relaxation looseness described in Section 6.7 is not reported in the available evaluation.

7.6 Summary of evidential status

Table 5 states, for each claim, what currently supports it. It is included so that no reader has to infer the evidential standing of a number from its surrounding prose.

Table 5: Evidence supporting the mathematical and computational claims.

Claim	Supporting evidence
Positional pathwidth at most $b-1$	Theorem 2; verified by structural audit of the saved models
Exactness of the recurrence	Theorem 4; synthetic enumeration to 7.1×10^{-15}; MILP agreement on real models
Cost $O(Lq^b(1+s_\alpha))$	Theorem 7 with Lemma 6
Shifted $>$ disjoint in predictive performance	Nested guide-disjoint resampling; exploratory, same-benchmark architecture development
Machine-precision agreement on saved models	Not established; historical residuals are of order 10^{-6}
Witness-reuse savings on saved models	Only a synthetic illustration is available
Box-relaxation looseness on saved models	Constructive counterexample proved; no real-model gap measured
Biological coverage of the grades	Not claimed; grades are synthetic throughout

Notes. The table distinguishes formal results, historical evaluations and comparisons for which real-model reruns remain unavailable.

8 Discussion

The principal algorithmic distinction is that the mismatch-count feature is maintained as a deterministic statistic of the sequence rather than relaxed into an independent coordinate. The constrained representation restricts positional tree interactions, but the global count would still couple all positions if written as one high-arity consistency factor. The augmented-state recurrence avoids that direct representation by carrying a partial count together with a short nucleotide suffix. The resulting guarantees apply to the represented ensemble margin and the stated Cartesian substitution-only uncertainty sets; they do not imply that the learned predictor is biologically calibrated.

The empirical comparison illustrates a measurable trade-off. On the literal reconstructed benchmark, the shifted architecture improves guide-macro average precision over disjoint blocks, but remains below unconstrained boosting. The data also have substantial guide-to-guide prevalence heterogeneity, particularly for the dominant guide, and architecture development was informed by exploratory work on the same benchmark. Accordingly, the significance

analyses are exploratory rather than prospective validation. Additional comparisons should retain matched preprocessing, model budgets and guide-disjoint splits before drawing conclusions about generalization.

Certification has separate limitations. The possibility grades in the current examples are synthetic, the admissible domain contains substitutions but no insertions or bulges, and the scores certified are those of the fitted model rather than direct Cas9 activity. Additional investigation is needed to measure the practical looseness of relaxed feature domains, reproduce grouped-tabulation agreement on saved models, assess witness reuse over real guides and verify that all cited data and model archives are versioned and accessible. These studies would make the computational claims easier to reproduce and delimit their biological interpretation.

9 Conclusions

The shifted-block construction gives an auditable family of boosted-tree predictors whose positional interactions admit a path decomposition of width at most $b-1$. An exact count-augmented dynamic program computes the range of the represented model margin over specified nucleotide substitutions while respecting the mismatch feature's functional identity. The existing CHANGE-seq comparison shows partial recovery of predictive performance relative to disjoint blocks, alongside an explicit loss relative to unrestricted boosting. The mathematical guarantees and available numerical checks support the method within the specified model and domain, while independent biological validation, calibrated uncertainty grades and additional computational comparisons remain necessary for stronger applied claims.

List of abbreviations

AP: average precision; CV: cross-validation; DP: dynamic programming; MILP: mixed-integer linear programming; CHANGE-seq: circularization for high-throughput analysis of nuclease genome-wide effects by sequencing.

Declarations

Ethics approval and consent to participate

This secondary computational study used previously published, non-identifiable experimental datasets and did not recruit participants or perform new animal experiments. Ethics approval

and consent to participate are not applicable to the present computational analysis.

Consent for publication

Not applicable: no identifiable personal information is reported.

Availability of data and materials

The original CHANGE-seq data and preprocessing scripts are publicly available through the SysEvalOffTarget project [11]. The supplementary archives associated with this study are identified by the author-provided Zenodo record DOI 10.5281/zenodo.24380698 [12]. The archive upload lists `AMB_Zenodo_validacion_sintetica_v0.1.0.zip` and `AMB_zenodo_deposito.zip`. The previously reported SHA-256 fingerprint of the literal encoded dataset is `ce62ad453b979eaf2a8cfe2c3a18bdf266d789bb23ba0deb72030100b423dce1`. The deposited project files and any regenerated models should be compared with that fingerprint and the historical guide-wise outputs before a claim of exact computational reproduction is made. Synthetic and historical real-model numerical checks are distinguished in the Results section.

Competing interests

The author declares that he has no competing interests.

Funding

No funding was received for this study or preparation of this manuscript.

Authors' contributions

Illych Álvarez Álvarez is the sole author and is responsible for the study design, mathematical analysis, computational evaluation, manuscript drafting and revision. The author has read and approved the final manuscript.

Acknowledgements

Not applicable.

Use of generative AI tools

Generative AI tools were used to assist manuscript drafting and code development. The author is responsible for verification of the mathematical, experimental and bibliographical content and for compliance with journal policies.

References

[1] Lazzarotto CR, Malinin NL, Li Y, Zhang R, Yang Y, Lee G, et al. CHANGE-seq reveals genetic and epigenetic effects on CRISPR–Cas9 genome-wide activity. Nat Biotechnol. 2020;38:1317–1327. doi:10.1038/s41587-020-0555-7.

[2] Hasin F, Minervini M, Mencar C, Ventrella G, Consiglio A, Orro A, et al. Feature representation for explainable CRISPR off-target prediction and base editing efficiency. Front Bioinform. 2026;6:1800237. doi:10.3389/fbinf.2026.1800237.

[3] Devos L, Meert W, Davis J. Versatile verification of tree ensembles. In: Proceedings of the 38th International Conference on Machine Learning. Proc Mach Learn Res. 2021;139:2654–2664. `https://proceedings.mlr.press/v139/devos21a.html`.

[4] Calzavara S, Lucchese C, Marcuzzi F, Orlando S. Feature partitioning for robust tree ensembles and their certification in adversarial scenarios. EURASIP J Inf Secur. 2021;2021:12. doi:10.1186/s13635-021-00127-0.

[5] Calzavara S, Cazzaro L, Pibiri GE, Prezza N. Verifiable learning for robust tree ensembles. In: Proceedings of the ACM SIGSAC Conference on Computer and Communications Security. 2023. doi:10.1145/3576915.3623100.

[6] Calzavara S, Cazzaro L, Lucchese C, Pibiri GE. Verifiable boosted tree ensembles. In: Proceedings of the IEEE Symposium on Security and Privacy. 2025. doi:10.1109/SP61157.2025.00022.

[7] Bauer A, Nakajima S, Müller K-R. Polynomial-time constrained message passing for exact MAP inference on discrete models with global dependencies. Mathematics. 2023;11:2628. doi:10.3390/math11122628.

[8] Zadeh LA. Fuzzy sets. Inf Control. 1965;8:338–353. doi:10.1016/S0019-9958(65)90241-X.

[9] Nguyen XT, Kowalczyk R. On solving distributed fuzzy constraint satisfaction problems with agents. In: Proceedings of the IEEE/WIC/ACM International Conference on Intelligent Agent Technology. 2007. p. 387–390. doi:10.1109/IAT.2007.81.

[10] Chen T, Guestrin C. XGBoost: a scalable tree boosting system. In: Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. 2016. p. 785–794. doi:10.1145/2939672.2939785.

[11] OrensteinLab. SysEvalOffTarget: data preparation and predictive evaluation code. GitHub. `https://github.com/OrensteinLab/SysEvalOffTarget`. Accessed 20 Nov 2026.

[12] Supporting research materials associated with the present manuscript. Zenodo. 2026. `https://doi.org/10.5281/zenodo.24380698`.

Figure legends and illustrations

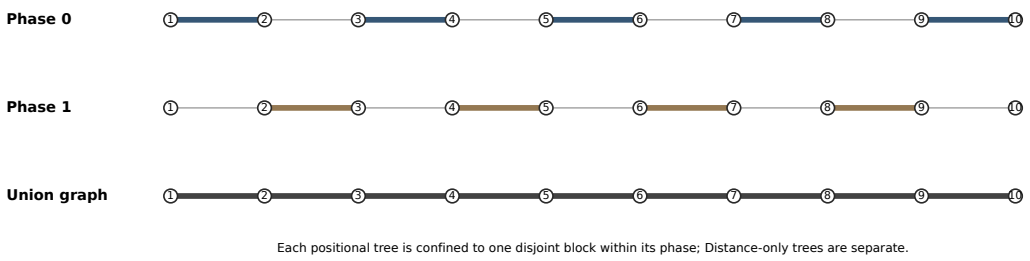


Figure 1: Shifted pair partitions for $b = 2$.

Phase-specific blocks overlap through the union; count-only trees remain separate. This is a structural schematic, not a learned interaction graph.

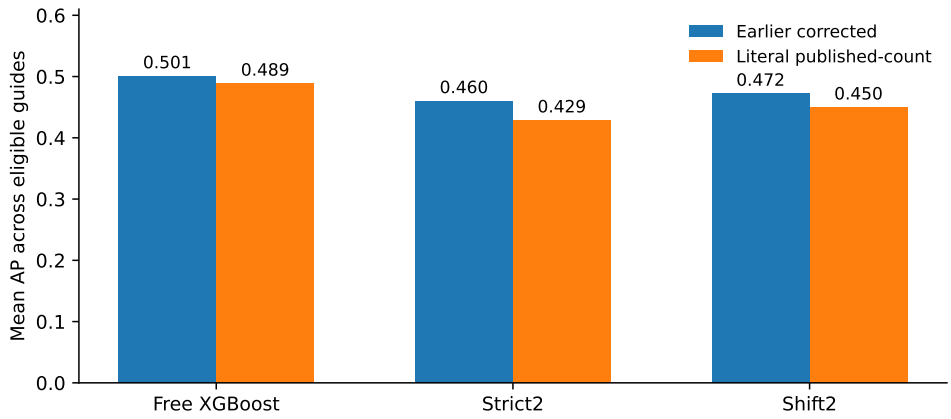


Figure 2: Guide-macro average precision across preprocessing protocols.

The earlier corrected reconstruction and the literal published-count benchmark are not matched nested-CV experiments.

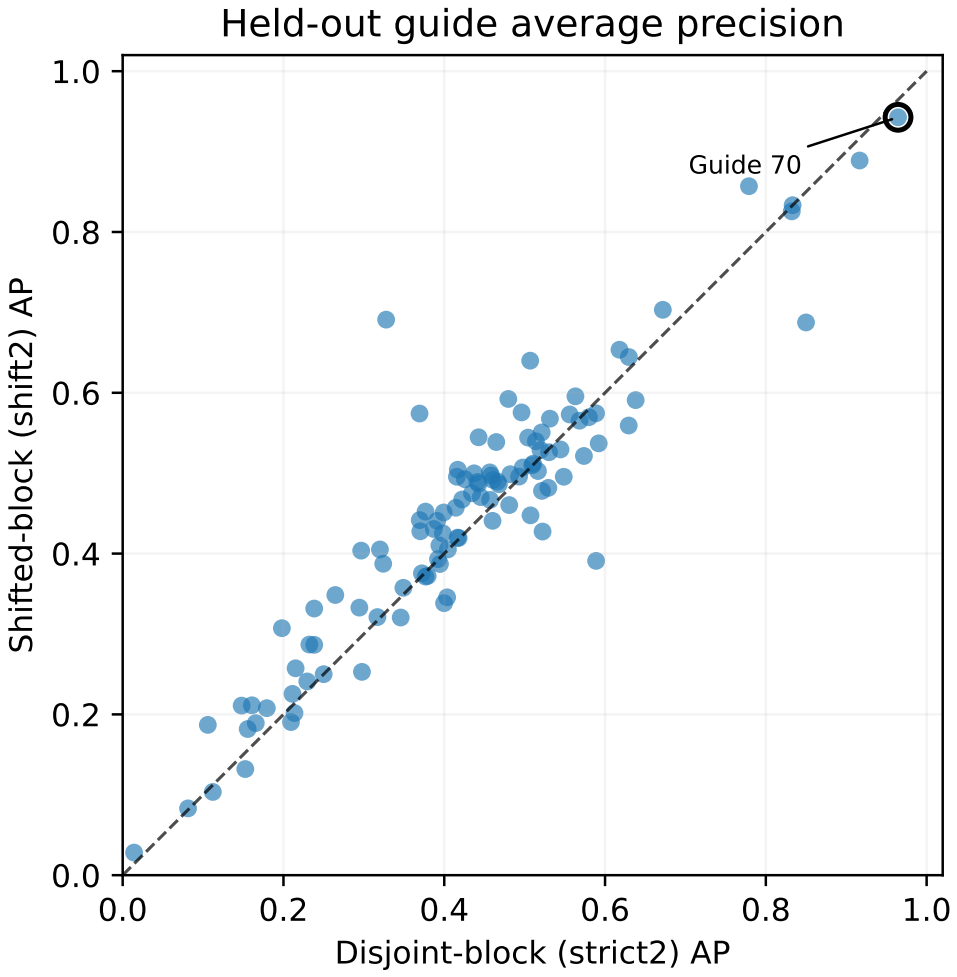


Figure 3: Held-out average precision for two certifiable ensembles.

The diagonal indicates equal average precision. There are 109 evaluable guides; the dominant guide is identified.