Supplementary Materials for

Knowledge-Guided Explanations for Tree Ensemble Models via Levels-Structure Values

Hyunwook Koh

Department of Applied Mathematics and Statistics
The State University of New York, Korea
Incheon, South Korea

E-mail: hyunwook.koh@stonybrook.edu

Notation

Let $N = \{1, \dots, p\}$ denote the feature set and let $v_x : 2^N \to \mathbb{R}$ be a fixed finite characteristic function. Let $\mathcal{H}$ be a rooted hierarchy tree whose root is the grand coalition N and whose leaves are the individual features. For a node u in $\mathcal{H}$, let $A(u) \subseteq N$ denote the set of feature leaves in the subtree rooted at u, and let $\mathrm{ch}(u) = \{u_1, \dots, u_d\}$ denote the children of u. A feature-parent node is a node whose children are individual feature leaves; $\mathcal{F}(\mathcal{H})$ denotes the set of feature-parent nodes of $\mathcal{H}$ and $K_1 := |\mathcal{F}(\mathcal{H})|$. It is convenient to regard a bare feature leaf $\{i\}$ as a degenerate feature-parent node with $A(\{i\}) = \{i\}$, so that the recursion below is defined at every node of an arbitrary rooted hierarchy tree of this form, including nodes whose children mix feature leaves and hierarchy nodes. With that convention, Eq. (S2) at such a node reduces to $\phi_i^{\{i\}}(v^u_{x,C}) = v_x(C \cup \{i\}) - v_x(C)$, which is exactly the value obtained by wrapping the leaf in a singleton feature-parent node; no computed value depends on which of the two representations is used. The feature-parent form used in the main text for depth-unbalanced taxonomies is one such tree. In particular, degree-one carry-forward nodes contribute a single child ordering, a factor of 1!, and therefore do not change $\Omega(\mathcal{H})$ or the resulting value. Because feature-parent nodes need not all lie at the same level of a depth-unbalanced hierarchy, all statements below are indexed by nodes of $\mathcal{H}$ rather than by a level index; in particular $\{A(u) : u \in \mathcal{F}(\mathcal{H})\}$ is always a partition of N.

Throughout, v_x is a fixed finite set function, so

$$\Lambda := 2 \max_{S \subseteq N} |v_x(S)| < \infty$$

is a finite constant, and every marginal contribution of v_x satisfies $|v_x(S \cup \{i\}) - v_x(S)| \le \Lambda$. For the tree ensemble application of Section S2, v_x is the path-dependent tree-aware set function defined by Eqs. (14)–(15) of the main text together with conventions (a) and (b) stated there; all exactness statements are relative to that fixed choice.

A permutation of N is $\mathcal{H}$-consistent if $A(u)$ appears as a contiguous block for every node u in $\mathcal{H}$. Let $\Omega(\mathcal{H})$ denote the set of all such permutations. Equivalently, a permutation in $\Omega(\mathcal{H})$ is generated by independently drawing a uniform ordering of the children at every internal node and concatenating the resulting suborderings recursively. Therefore,

$$|\Omega(\mathcal{H})| = \prod_{u \in \mathrm{int}(\mathcal{H})} |\mathrm{ch}(u)|!.$$

For a permutation $\pi = (i_1, \dots, i_p)$, let $\mathrm{pos}_\pi(i)$ denote the position of feature i in π, and define the predecessor set

$$P_i^\pi := \{j \in N : \mathrm{pos}_\pi(j) < \mathrm{pos}_\pi(i)\}.$$

For a node u and an outer context $C \subseteq N \setminus A(u)$, define the restricted game

$$v^u_{x,C}(S) := v_x(C \cup S), \qquad S \subseteq A(u). \tag{S1}$$

When u is a feature-parent node, the local Shapley value (Shapley, 1953) of $i \in A(u)$ under context C is

$$\phi_i^{A(u)}(v^u_{x,C}) = \sum_{S \subseteq A(u) \setminus \{i\}} \frac{|S|!\,(|A(u)| - |S| - 1)!}{|A(u)|!} \{v_x(C \cup S \cup \{i\}) - v_x(C \cup S)\}. \tag{S2}$$

For a non-feature-parent node u with children $u_1, \dots, u_d$ and an outer context $C \subseteq N \setminus A(u)$, the *local quotient game* on the children is the d-player game

$$q_{u,C}(T) := v_x\left(C \cup \bigcup_{t \in T} A(u_t)\right), \qquad T \subseteq \{1, \dots, d\}, \tag{S3}$$

and $\mathrm{Sh}_j(q_{u,C})$ denotes the ordinary Shapley value of player j in $q_{u,C}$. Finally, $\mathcal{R}_u$ denotes the multiset of context–weight pairs (C, w) with which the recursive allocation reaches node u from the root: each pair corresponds to one sequence of sibling-subset choices along the root-to-u path, C being the union of the accumulated preceding sibling blocks and w the product of the corresponding local weights in Eq. (S5). By construction $\mathcal{R}_{\mathrm{root}} = \{(\emptyset, 1)\}$ and $\sum_{(C,w) \in \mathcal{R}_u} w = 1$ for every u.

S1. Proof of Proposition 1

Lemma 1 (Recursive representation)**.** *For any rooted hierarchy tree $\mathcal{H}$ whose leaves are the individual features, and for any finite characteristic function v_x, the permutation-average definition of the levels-structure value is equivalent to the recursive local-Shapley allocation over the hierarchy. The padded feature-parent representation used for depth-unbalanced taxonomies is a special case; padding is a representational convenience and is not required for the argument, provided a bare feature leaf $\{i\}$ is read as a degenerate feature-parent node as in the Notation, so that the base case* Eq. (S4) *applies to it.*

Proof of Lemma 1.

Proof. For any node u and context $C \subseteq N \setminus A(u)$, define recursively an allocation $R_i(u, C)$ for $i \in A(u)$. If u is a feature-parent node, set

$$R_i(u, C) := \phi_i^{A(u)}(v_{x,C}^u), \qquad i \in A(u). \tag{S4}$$

By the Notation convention this covers a bare feature leaf $u = \{i\}$ as well, for which $R_i(\{i\}, C) = v_x(C \cup \{i\}) - v_x(C)$; the recursion is therefore total on an arbitrary rooted hierarchy tree. If u is not a feature-parent node and has children $u_1, \ldots, u_d$, define, for $i \in A(u_j)$,

$$R_i(u, C) := \sum_{T \subseteq \{1,\ldots,d\} \setminus \{j\}} \frac{|T|!\,(d - |T| - 1)!}{d!} R_i\left(u_j,\, C \cup \bigcup_{t \in T} A(u_t)\right). \tag{S5}$$

The case $d = 1$ is included, since the only possible subset is $T = \emptyset$ and the corresponding weight is 1.

We prove by induction on the height of the subtree rooted at u that $R_i(u, C)$ equals the average marginal contribution of i over hierarchy-consistent permutations of the features in $A(u)$, with all features in C treated as an already accumulated outer context. If u is a feature-parent node, the hierarchy-consistent permutations of $A(u)$ are ordinary permutations of the individual features in $A(u)$. The average marginal contribution of i over these permutations is exactly the Shapley value in Eq. (S2), so the claim holds by Eq. (S4).

Now suppose the claim holds for all proper child subtrees of an internal node u with children $u_1, \ldots, u_d$. Fix $i \in A(u_j)$. A uniform hierarchy-consistent permutation of $A(u)$ is obtained by first drawing a uniform ordering of the d children of u and then, independently, drawing uniform hierarchy-consistent permutations within the child subtrees. For a subset $T \subseteq \{1, \ldots, d\} \setminus \{j\}$, the probability that exactly the children indexed by T precede u_j is

$$\frac{|T|!\,(d - |T| - 1)!}{d!}, \tag{S6}$$

because the children in T may be ordered in $|T|!$ ways before u_j, the remaining children may be ordered in $(d - |T| - 1)!$ ways after u_j, and there are $d!$ total child orderings. Conditional on this event, all features in $\cup_{t \in T} A(u_t)$ are added to the outer context before the recursion enters u_j, and the ordering within u_j is uniform over hierarchy-consistent permutations of that child subtree. By the induction hypothesis, the conditional average marginal contribution of i is

$$R_i\left(u_j,\, C \cup \bigcup_{t \in T} A(u_t)\right).$$

Averaging over all possible predecessor-child subsets T gives Eq. (S5). Applying this identity at the root with $C = \emptyset$ proves the equivalence between the recursive representation and the permutation-average definition in the main text. □

Lemma 2 (Block allocation and quotient games)**.** *For every node u of $\mathcal{H}$ and every context $C \subseteq N \setminus A(u)$,*

$$\sum_{i \in A(u)} R_i(u, C) = v_x(C \cup A(u)) - v_x(C). \tag{S7}$$

If, in addition, u is not a feature-parent node and has children $u_1, \ldots, u_d$, then for every $j \in \{1, \ldots, d\}$

$$\sum_{i \in A(u_j)} R_i(u, C) = \mathrm{Sh}_j(q_{u,C}), \tag{S8}$$

with $q_{u,C}$ as in Eq. (S3)*. Finally, $|R_i(u, C)| \le \Lambda$ for every $i \in A(u)$.*

Proof of Lemma 2.

Proof. We prove Eq. (S7) by induction on the height of the subtree rooted at u. If u is a feature-parent node, then by Eq. (S4) and efficiency of the Shapley value for the game $v_{x,C}^u$,

$$\sum_{i \in A(u)} R_i(u, C) = \sum_{i \in A(u)} \phi_i^{A(u)}(v_{x,C}^u) = v_{x,C}^u(A(u)) - v_{x,C}^u(\emptyset) = v_x(C \cup A(u)) - v_x(C).$$

Suppose now that u is not a feature-parent node and that Eq. (S7) holds at every child of u. Write $C_T := C \cup \bigcup_{t \in T} A(u_t)$ and $w_T^{(d)} := |T|!\,(d - |T| - 1)!/d!$. Summing Eq. (S5) over $i \in A(u_j)$ and applying the induction hypothesis at u_j with context C_T gives

$$\begin{aligned}
\sum_{i \in A(u_j)} R_i(u, C) &= \sum_{T \subseteq \{1,\ldots,d\} \setminus \{j\}} w_T^{(d)} \sum_{i \in A(u_j)} R_i(u_j, C_T) \\
&= \sum_{T \subseteq \{1,\ldots,d\} \setminus \{j\}} w_T^{(d)} \left\{v_x(C_T \cup A(u_j)) - v_x(C_T)\right\} \\
&= \sum_{T \subseteq \{1,\ldots,d\} \setminus \{j\}} w_T^{(d)} \left\{q_{u,C}(T \cup \{j\}) - q_{u,C}(T)\right\} = \mathrm{Sh}_j(q_{u,C}),
\end{aligned}$$

which is Eq. (S8). Summing this over $j = 1, \dots, d$ and using efficiency of the Shapley value for the d-player game $q_{u,C}$,

$$\sum_{i \in A(u)} R_i(u, C) = \sum_{j=1}^{d} \mathrm{Sh}_j(q_{u,C}) = q_{u,C}(\{1, \dots, d\}) - q_{u,C}(\emptyset) = v_x(C \cup A(u)) - v_x(C),$$

completing the induction. Note that Eq. (S8) was derived using only the induction hypothesis at the children, so the two statements are established simultaneously and the argument is not circular.

For the bound, induct again on height. At a feature-parent node, $\phi_i^{A(u)}(v^u_{x,C})$ is a convex combination of marginal contributions of v_x, each of absolute value at most Λ, so $|R_i(u, C)| \le \Lambda$. At a non-feature-parent node, Eq. (S5) expresses $R_i(u, C)$ as a convex combination of the quantities $R_i(u_j, C_T)$, because the weights $w_T^{(d)}$ are nonnegative and sum to one over $T \subseteq \{1, \dots, d\} \setminus \{j\}$; the bound is therefore preserved. □

Proposition 1 (Basic properties)**.** *For any hierarchy $\mathcal{H}$, $\Psi^{\mathcal{H}}$ coincides with the levels-structure value (Winter, 1989). Consequently, it satisfies efficiency, linearity, the null-player property, level symmetry, multi-resolution coherence, and the Shapley/Owen reductions stated in Proposition 1 of the main text.*

Proof of Proposition 1.

Proof. By Lemma 1, the permutation definition of $\Psi^{\mathcal{H}}$ agrees with the recursive levels-structure allocation: at each internal node, a Shapley allocation is applied to the local quotient game over the node's children, and the allocated value is then distributed recursively within the selected child block. This is the levels-structure value (Winter, 1989) for the hierarchy.

Efficiency. Let $\pi = (i_1, \dots, i_p)$ be any hierarchy-consistent permutation. Then

$$\sum_{i \in N} \{v_x(P_i^\pi \cup \{i\}) - v_x(P_i^\pi)\} = \sum_{r=1}^{p} \{v_x(\{i_1, \dots, i_r\}) - v_x(\{i_1, \dots, i_{r-1}\})\} \quad \text{[S9]}$$

$$= v_x(N) - v_x(\emptyset), \quad \text{[S10]}$$

where the last equality follows by telescoping. Averaging over $\pi \in \Omega(\mathcal{H})$ gives

$$\sum_{i \in N} \Psi_i^{\mathcal{H}}(v_x) = v_x(N) - v_x(\emptyset).$$

Under the conditional-expectation value function in the main text, this becomes $f(x) - \mathbb{E}[f(X)]$.

Linearity. For games v, w and scalars α, β, each marginal contribution satisfies

$$\begin{aligned}(\alpha v + \beta w)(P_i^\pi \cup \{i\}) - (\alpha v + \beta w)(P_i^\pi) \\ = \alpha\{v(P_i^\pi \cup \{i\}) - v(P_i^\pi)\} + \beta\{w(P_i^\pi \cup \{i\}) - w(P_i^\pi)\}.\end{aligned}$$

Averaging over $\Omega(\mathcal{H})$ yields linearity.

Null-player property. If $v_x(S \cup \{i\}) = v_x(S)$ for every $S \subseteq N \setminus \{i\}$, then every marginal contribution of i in every hierarchy-consistent permutation is zero. Hence $\Psi_i^{\mathcal{H}}(v_x) = 0$.

Level symmetry. Let u be an internal node with children $u_1, \dots, u_d$. By Lemma 2, for every outer context C the aggregate allocation that the recursion sends into child u_j is exactly $\mathrm{Sh}_j(q_{u,C})$, the ordinary Shapley value of player j in the local quotient game Eq. (S3). Two remarks are needed, because the quotient game depends on C and there is in general more than one context reaching u.

First, fix C and suppose that children u_j and $u_{j'}$ are symmetric in $q_{u,C}$, that is, $q_{u,C}(T \cup \{j\}) = q_{u,C}(T \cup \{j'\})$ for every $T \subseteq \{1, \dots, d\} \setminus \{j, j'\}$. The Shapley symmetry axiom then gives $\mathrm{Sh}_j(q_{u,C}) = \mathrm{Sh}_{j'}(q_{u,C})$.

Second, unrolling Eq. (S5) from the root down to u as in Eq. (S11) and summing over $i \in A(u_j)$ gives

$$\sum_{i \in A(u_j)} \Psi_i^{\mathcal{H}}(v_x) = \sum_{(C,w) \in \mathcal{R}_u} w \, \mathrm{Sh}_j(q_{u,C}) = \Gamma_{u_j}(v_x),$$

the last equality holding because $\mathcal{R}_{u_j}$ is obtained from $\mathcal{R}_u$ by refining each pair (C, w) into the pairs $(C_T,\, w\, w_T^{(d)})$ indexed by $T \subseteq \{1, \dots, d\} \setminus \{j\}$, so that

$$\Gamma_{u_j}(v_x) = \sum_{(C,w) \in \mathcal{R}_u} w \sum_{T \subseteq \{1,\dots,d\} \setminus \{j\}} w_T^{(d)} \, \{q_{u,C}(T \cup \{j\}) - q_{u,C}(T)\} = \sum_{(C,w) \in \mathcal{R}_u} w \, \mathrm{Sh}_j(q_{u,C}) \, .$$

Consequently the aggregate allocations of u_j and $u_{j'}$ agree provided the symmetry of the preceding paragraph holds for *every* context C appearing in $\mathcal{R}_u$. This quantifier cannot be dropped: symmetry in $q_{u,C}$ for a single context does not by itself equate

the $\mathcal{R}_u$-weighted averages. A sufficient condition that does not refer to $\mathcal{R}_u$, and which is the form in which the axiom is usually stated, is that $A(u_j)$ and $A(u_{j'})$ be symmetric blocks of v_x, that is,

$$v_x(S \cup A(u_j)) = v_x(S \cup A(u_{j'})) \qquad \text{for every } S \subseteq N \setminus (A(u_j) \cup A(u_{j'}))\,,$$

since this implies symmetry in $q_{u,C}$ for every admissible C. With either hypothesis the levels-structure allocation satisfies level symmetry at u, and the argument applies verbatim at every internal node.

Multi-resolution coherence. Let u be any non-leaf hierarchy node; for a feature leaf, the statement is immediate. Let $\mathcal{R}_u$ denote the multiset of context–weight pairs (C, w) with which the recursive allocation reaches u from the root, that is, the pairs obtained by accumulating, along the path from the root to u, the outer contexts and the products of the local Shapley weights Eq. (S5). Unrolling Eq. (S5) from the root down to u gives, for every $i \in A(u)$,

$$\Psi_i^{\mathcal{H}}(v_x) = R_i(\text{root}, \emptyset) = \sum_{(C,w)\in\mathcal{R}_u} w\, R_i(u, C). \tag{S11}$$

The block-level attribution assigned to $A(u)$ at that level is

$$\Gamma_u(v_x) := \sum_{(C,w)\in\mathcal{R}_u} w\{v_x(C \cup A(u)) - v_x(C)\}. \tag{S12}$$

For each pair $(C, w) \in \mathcal{R}_u$, Eq. (S7) of Lemma 2 gives

$$\sum_{i\in A(u)} R_i(u, C) = v_x(C \cup A(u)) - v_x(C).$$

Multiplying by w and summing over $(C, w) \in \mathcal{R}_u$, and using Eq. (S11), yields

$$\sum_{i\in A(u)} \Psi_i^{\mathcal{H}}(v_x) = \Gamma_u(v_x).$$

Thus the feature-level attributions contained in any hierarchy node sum exactly to the block-level attribution assigned to that node. Two consequences are worth recording. Taking u to be a child of an internal node and combining with Eq. (S8) shows that Γ_u is an $\mathcal{R}$-weighted average of quotient-game Shapley values, so the block-level number carries independent game-theoretic content; and taking $u =$ root recovers efficiency. It also follows that a numerical check of this identity is informative only if Γ_u is computed from Eq. (S12) directly: re-summing the reported feature-level attributions over $A(u)$ reproduces the left-hand side by construction and tests nothing, since the children of u partition $A(u)$. This proves multi-resolution coherence.

Reductions. If $L = 0$, the hierarchy has only the root and singleton leaves. Then $\Omega(\mathcal{H})$ is the set of all $p!$ permutations of N, and the definition of $\Psi_i^{\mathcal{H}}$ reduces to the ordinary Shapley permutation formula (Shapley, 1953).

If $L = 1$, the hierarchy consists of the root, one partition $\mathcal{G} = \{G_1, \ldots, G_K\}$, and feature leaves. For $i \in G_k$, a hierarchy-consistent permutation is generated by first ordering the K groups and then ordering the features within each group. Let $U \subseteq \mathcal{K} \setminus \{k\}$ be the set of groups preceding G_k. The probability of this event is

$$\frac{|U|!\,(K - |U| - 1)!}{K!}.$$

Conditional on this event, the features within G_k are uniformly ordered, so the conditional average marginal contribution of i is the within-group Shapley value $\phi_i^{G_k}(v_{x,C_U})$, where $C_U = \cup_{\ell\in U} G_\ell$. Therefore

$$\Psi_i^{\mathcal{H}}(v_x) = \sum_{U\subseteq\mathcal{K}\setminus\{k\}} \frac{|U|!\,(K - |U| - 1)!}{K!}\, \phi_i^{G_k}(v_{x,C_U}),$$

which is the Owen value for the partition $\mathcal{G}$ (Owen, 1977).

Finally, Winter (1989) establishes an axiomatic characterization for cooperative games equipped with a levels structure: the levels-structure value is the unique value satisfying levels-structure analogues of efficiency, additivity, the null-player property, and symmetry, with symmetry imposed within each level of the structure. The properties verified above are those axioms specialized to the hierarchy $\mathcal{H}$ and restated in the present notation, with the level-symmetry statement understood in the quantified form established above; the original formulation of the axioms and the precise statement of the uniqueness theorem are given by Winter (1989) and are not reproduced here. Since $\Psi^{\mathcal{H}}$ is the levels-structure value of $\mathcal{H}$, it inherits that uniqueness. □

S2. Proof of Proposition 2

Proposition 2 (Exactness of TLSV-EXACT)**.** *Suppose that, whenever the hierarchy recursion reaches a feature-parent node u under an outer context $C \subseteq N \setminus A(u)$, the inner solver returns the exact local Shapley values*

$$\{\phi_i^{A(u)}(v_{x,C}^u) : i \in A(u)\}, \qquad v_{x,C}^u(S) = v_x(C \cup S), \quad S \subseteq A(u).$$

Then the TLSV-EXACT algorithm in the main text returns $\Psi_i^{\mathcal{H}}(v_x)$ for every $i \in N$. For the path-dependent tree-aware cooperative game used for tree ensemble models, the context-conditioned INNERTREEDP solver satisfies this inner exactness condition.

Proof of Proposition 2.

Proof. We prove a stronger invariant for the recursive subroutine AGGREGATE. For any node u, context $C \subseteq N \setminus A(u)$, and scalar weight w, the call AGGREGATE(u, C, w) adds to the output vector, for each $i \in A(u)$, exactly

$$w\, R_i(u, C), \tag{S13}$$

where $R_i(u, C)$ is the recursive allocation defined in Eq. (S4)–Eq. (S5). The proposition follows by applying this invariant at the root with $C = \emptyset$ and $w = 1$. The proof is by induction on the height of the subtree rooted at u.

Base case. Suppose $u \in \mathcal{F}(\mathcal{H})$. The main-text exact algorithm calls an exact inner solver and adds w times its output to $\Psi[A(u)]$. By the inner exactness condition, the solver output for feature $i \in A(u)$ is

$$\phi_i^{A(u)}(v_{x,C}^u) = R_i(u, C),$$

where the equality follows from the base-case definition Eq. (S4). Hence the call contributes $wR_i(u, C)$ to each $i \in A(u)$, proving Eq. (S13) in the base case.

For the path-dependent tree-aware game, the exact inner solver is INNERTREEDP, and the inner exactness condition is supplied by Lemma 3 below.

Lemma 3 (Reduction to TreeSHAP on a restricted evaluator)**.** *Fix disjoint $C, G \subseteq N$ and let $\widetilde{v}_{C,G}(S) := v_x(C \cup S)$ for $S \subseteq G$, where v_x is the path-dependent tree-aware set function of Eq. (14) of the main text. Define a* restricted evaluator *on each tree by traversing it and, at an internal node t splitting on $j(t)$:*

(a) if $j(t) \in C$, contracting the node to the child determined by $x_{j(t)}$ (and, when that child is not determined, to the cover-weighted average of its children);

(b) if $j(t) \notin C \cup G$, replacing the node by the cover-weighted average of its children;

(c) if $j(t) \in G$, retaining the node as an active split.

Then for every $S \subseteq G$ the restricted evaluator returns $v_x(C \cup S)$, and applying the TreeSHAP path recursion to it with player set G returns $\phi_i^G(\widetilde{v}_{C,G})$ exactly for every $i \in G$. INNERTREEDP is precisely this computation, so it satisfies the inner exactness condition of Proposition 2.

Proof of Lemma 3. Fix $S \subseteq G$ and evaluate $v_x(C \cup S)$ by Eq. (14). A split on $j(t) \in C$ has $j(t) \in C \cup S$ for every such S, so Eq. (14) always follows the child determined by $x_{j(t)}$: rule (a) substitutes this constant behavior and is therefore value preserving, uniformly in S. A split on $j(t) \notin C \cup G$ has $j(t) \notin C \cup S$ for every such S, so Eq. (14) always takes the cover-weighted average $\sum_c \omega_{c|t} g_c$; rule (b) substitutes exactly that. A split on $j(t) \in G$ is followed or averaged according to whether $j(t) \in S$, which is the defining behavior of an active TreeSHAP split, and rule (c) leaves it untouched. Hence the restricted evaluator agrees with $v_x(C \cup \cdot)$ on all of 2^G, i.e. it represents the game $\widetilde{v}_{C,G}$.

Because rules (a) and (b) only contract nodes or replace them by fixed convex combinations of their children, the restricted evaluator is again a nonnegative path-weighted sum of leaf values whose only free features are those of G. It is therefore an instance of the object to which the TreeSHAP path recursion applies with player set G, and the standard EXTEND/UNWIND argument returns the Shapley values of $\widetilde{v}_{C,G}$ without enumerating $2^{|G|}$ subsets. The repeated-feature correction is unchanged: if an active $j \in G$ splits twice on a root-to-node path, UNWIND removes the earlier occurrence and returns its stored zero- and one-fractions, which are then carried into the new EXTEND, so each active feature occurs at most once in the path object.

Three conventions must match between the evaluator and the solver for the reduction to be exact, and all three are adopted here. First, the absent-feature weights $\omega_{c|t}$ used in rules (a) and (b) are the same children-sum proportions derived from learner-native node-cover masses and used by Eq. (14), including the stated fallbacks when the denominator is unavailable, non-finite or zero. Second, at an active split whose complementary child is absent from the tree, the entire absent-coalition mass is assigned to the child that is present, matching the corresponding fallback of Eq. (14); assigning it in equal parts would discard half of that mass and break the agreement. Third, at a split whose observed child $c^*(t)$ is undetermined in the sense of convention (a) of the main text—$x_{j(t)}$ missing with no default direction, or the selected child absent with no default child—the solver descends by the same cover-weighted average that Eq. (14) uses, and does *not* extend the path object even when $j(t) \in G$. This is value preserving and is also the only choice consistent with the game: under Eq. (14) the coalitions containing $j(t)$ and

those not containing it receive the identical value at t, so $j(t)$ has no local marginal contribution there and must receive no local Shapley weight. A childless internal node is likewise treated as a leaf carrying its stored value μ_t, matching the first line of Eq. (14). Finally, the split rule and the comparison precision used to determine $c^*(t)$ are those of the fitted implementation, fixed identically in the evaluator and in the solver; a mismatch there would make the solver exact for a different set function than the one being evaluated. □

Inductive step. Suppose u is an internal node with children $u_1, \dots, u_d$ and that the invariant holds for all child subtrees. Fix a child u_j and a feature $i \in A(u_j)$. For every sibling subset $T \subseteq \{1, \dots, d\} \setminus \{j\}$, the algorithm makes the recursive call

$$\textsc{Aggregate}\left(u_j,\, C \cup \bigcup_{t \in T} A(u_t),\, w\, w_T^{(d)}\right), \qquad w_T^{(d)} = \frac{|T|!\,(d - |T| - 1)!}{d!}.$$

By the induction hypothesis, this call contributes

$$w\, w_T^{(d)} R_i\left(u_j,\, C \cup \bigcup_{t \in T} A(u_t)\right)$$

to feature i. Summing over all sibling subsets T gives

$$w \sum_{T \subseteq \{1,\dots,d\} \setminus \{j\}} w_T^{(d)} R_i\left(u_j,\, C \cup \bigcup_{t \in T} A(u_t)\right) = w\, R_i(u, C),$$

where the equality is exactly the recursive definition Eq. (S5). This proves the invariant at u. At the root, the call is $\textsc{Aggregate}(\mathrm{root}(\mathcal{H}), \emptyset, 1)$. By Lemma 1, $R_i(\mathrm{root}(\mathcal{H}), \emptyset) = \Psi_i^{\mathcal{H}}(v_x)$, and therefore the exact TLSV algorithm returns the levels-structure value for every feature. □

S3. Proof of Proposition 3

Proposition 3 (Statistical guarantees for recursive TLSV-MC)**.** *Consider the recursive TLSV-MC algorithm in the main text with predetermined local sampling budgets $\{R_u\}$ at all sampled hierarchy-node calls, with mutually independent draws at distinct node calls, and without adaptive stopping. Suppose that, whenever the recursion reaches a feature-parent node u under an outer context $C \subseteq N \setminus A(u)$,* InnerTreeDP *returns the exact local Shapley values $\{\phi_i^{A(u)}(v^u_{x,C}) : i \in A(u)\}$ for the restricted path-dependent tree-aware game. Then, for every feature $i \in N$, the following hold.*

(i) Unbiasedness. $\mathbb{E}\{\widehat{\Psi}_i^{\mathcal{H}}(v_x)\} = \Psi_i^{\mathcal{H}}(v_x)$ *at every fixed budget configuration.*

(ii) Exact efficiency. *For every realized set of sampled child orderings,* $\sum_{i\in N} \widehat{\Psi}_i^{\mathcal{H}}(v_x) = v_x(N) - v_x(\emptyset)$.

(iii) L^2 consistency. *As* $\min_u R_u \to \infty$, $\mathbb{E}[\{\widehat{\Psi}_i^{\mathcal{H}}(v_x) - \Psi_i^{\mathcal{H}}(v_x)\}^2] \to 0$, *hence also convergence in probability.*

(iv) Rate. *If a common budget $R_u \equiv R$ is used at all sampled node calls, then* $\mathrm{Var}\{\widehat{\Psi}_i^{\mathcal{H}}(v_x)\} = O(R^{-1})$, *so the Monte Carlo standard error is $O(R^{-1/2})$, with constants determined by $\mathcal{H}$ and v_x.*

This is Proposition 3 of the main text. Almost sure convergence is treated separately in Corollary 1 below, which is Corollary 1 of the main text.

Proof of Proposition 3.

Proof. Let $R_i(u, C)$ denote the exact recursive allocation defined in Eq. (S4)–Eq. (S5). For a fixed set of local sampling budgets, let $\widehat{R}_i(u, C)$ denote the random allocation returned by the recursive Monte Carlo subroutine when it is called at node u with outer context C and unit weight. We prove by induction on the height of the subtree rooted at u that

$$\mathbb{E}\{\widehat{R}_i(u, C)\} = R_i(u, C), \qquad i \in A(u). \tag{S14}$$

Base case. Suppose u is a feature-parent node. The recursive TLSV-MC algorithm does not sample feature subsets within $A(u)$. Instead, it calls InnerTreeDP with active set $A(u)$ and fixed-present context C. By the assumed exactness of InnerTreeDP for the restricted path-dependent tree-aware game, its output for feature $i \in A(u)$ is

$$\phi_i^{A(u)}(v^u_{x,C}) = R_i(u, C).$$

Thus $\widehat{R}_i(u, C) = R_i(u, C)$ deterministically in the base case, and Eq. (S14) holds.

Inductive step and unbiasedness. Suppose u is an internal non-feature-parent node with children $u_1, \ldots, u_d$, and assume that Eq. (S14) holds for all child subtrees. In one Monte Carlo draw at u, the algorithm samples a uniform ordering of the d children. Fix a child u_j and a feature $i \in A(u_j)$. The subset $T \subseteq \{1, \ldots, d\} \setminus \{j\}$ of siblings that precedes u_j in the sampled ordering has probability

$$\frac{|T|!\,(d - |T| - 1)!}{d!},$$

because the children in T can be ordered before u_j in $|T|!$ ways and the remaining children can be ordered after u_j in $(d-|T|-1)!$ ways. Conditional on this event, the context passed to u_j is $C \cup \bigcup_{t\in T} A(u_t)$. Therefore, by the induction hypothesis, the expected contribution to feature i from one sampled child ordering is

$$\sum_{T\subseteq\{1,\ldots,d\}\setminus\{j\}} \frac{|T|!\,(d - |T| - 1)!}{d!} R_i\left(u_j, C \cup \bigcup_{t\in T} A(u_t)\right) = R_i(u, C),$$

where the equality is the recursive definition Eq. (S5). Averaging R_u independent child-ordering draws at the call to u preserves this expectation. Applying the result at the root with $C = \emptyset$ gives

$$\mathbb{E}\{\widehat{\Psi}_i^{\mathcal{H}}(v_x)\} = \Psi_i^{\mathcal{H}}(v_x).$$

Consistency and rate. A direct appeal to the strong law at each sampled node call is not by itself sufficient here: as the local budgets grow, the number of instantiated node–context calls also grows and the law of each child estimator changes with the child budgets, so the estimator is a nested triangular array rather than an average of identically distributed summands. We therefore argue through the variance recursion, which yields L^2 consistency and the rate simultaneously. Almost sure convergence is not obtained here; it is treated separately in Corollary 1 by a concentration argument applied at the root call.

Fix a deterministic budget schedule $\{R_u^{(m)}\}$ with $\min_u R_u^{(m)} \to \infty$ as $m \to \infty$. Let u be a sampled internal node with children $u_1, \ldots, u_d$ and fix $i \in A(u_j)$. Writing T for the (random) set of siblings preceding u_j in a sampled ordering, the R_u summands averaged at (u, C) are independent and identically distributed, each with mean $R_i(u, C)$ by the inductive step above, so

$$\mathrm{Var}\{\widehat{R}_i(u, C)\} = \frac{1}{R_u}\Big[\mathrm{Var}_T\{R_i(u_j, C_T)\} + \mathbb{E}_T\,\mathrm{Var}\{\widehat{R}_i(u_j, C_T)\}\Big]. \tag{S15}$$

At a feature-parent node the estimator is deterministic (InnerTreeDP is exact), so its variance is 0. By the last statement of Lemma 2, $|R_i(u_j, C_T)| \le \Lambda$ for every child and every context, so the first term in Eq. (S15) is bounded by Λ^2, a constant depending only on v_x and not on any budget. Induction on subtree height therefore gives $\mathrm{Var}\{\widehat{R}_i(u, C)\} \to 0$ as $\min_u R_u \to \infty$, hence

$$\mathbb{E}\big[\{\widehat{\Psi}_i^{\mathcal{H}}(v_x) - \Psi_i^{\mathcal{H}}(v_x)\}^2\big] \longrightarrow 0,$$

so the estimator is L^2-consistent and, a fortiori, consistent in probability. If a common budget $R_u \equiv R$ is used, unrolling Eq. (S15) over the finitely many levels leaves a leading term of order $1/R$, which is the $O(R^{-1/2})$ standard-error statement.

This proves parts (i), (iii) and (iv). Part (ii) is proved below.

Exact efficiency at every fixed budget. We prove the stronger realized-sample identity

$$\sum_{i \in A(u)} \widehat{R}_i(u, C) = v_x(C \cup A(u)) - v_x(C)$$

for every node u, context C, and realized set of sampled child orderings. If u is a feature-parent node, this follows from efficiency of the exact local Shapley allocation returned by InnerTreeDP. If u is an internal node, consider one sampled ordering $(u_{(1)}, \dots, u_{(d)})$ and let

$$C_0 = C, \qquad C_q = C \cup \bigcup_{r=1}^{q} A(u_{(r)}), \quad q = 1, \dots, d.$$

By the induction hypothesis applied to the child calls, the sum of feature attributions contributed by child $u_{(q)}$ equals

$$v_x(C_q) - v_x(C_{q-1}).$$

Summing over $q = 1, \dots, d$ telescopes to

$$\sum_{q=1}^{d} \{v_x(C_q) - v_x(C_{q-1})\} = v_x(C \cup A(u)) - v_x(C).$$

Averaging over the fixed local budget of sampled child orderings at u leaves the same identity unchanged. Applying this identity at the root gives exact efficiency for every realized fixed-budget recursive Monte Carlo estimate:

$$\sum_{i \in N} \widehat{\Psi}_i^{\mathcal{H}}(v_x) = v_x(N) - v_x(\emptyset).$$

□

Corollary 1 (Concentration and almost sure convergence along summable budget schedules)**.** *Adopt the assumptions of Proposition 3 and set $c_\varepsilon := \varepsilon^2/(2\Lambda^2)$ for $\varepsilon > 0$, with $\Lambda = 2\max_{S \subseteq N} |v_x(S)|$ as in the Notation. Let R_{root} denote the sampling budget used at the root call. Then, for every $\varepsilon > 0$ and every $i \in N$,*

$$\Pr\left\{\left|\widehat{\Psi}_i^{\mathcal{H}}(v_x) - \Psi_i^{\mathcal{H}}(v_x)\right| > \varepsilon\right\} \le 2\exp(-c_\varepsilon R_{\mathrm{root}}), \tag{S16}$$

uniformly in the budgets used at the descendant node calls. Consequently, for any deterministic budget schedule $\{R_u^{(m)}\}_{m \ge 1}$ with $\sum_m \exp(-c_\varepsilon R_{\mathrm{root}}^{(m)}) < \infty$ for every $\varepsilon > 0$—for instance $R_u^{(m)} = m$ for all u—one has $\widehat{\Psi}_i^{\mathcal{H},(m)}(v_x) \to \Psi_i^{\mathcal{H}}(v_x)$ almost surely as $m \to \infty$.

Proof of Corollary 1.

Proof. We first record that every realized recursive estimate is uniformly bounded by Λ, irrespective of the budgets. Indeed, by induction on the height of the subtree rooted at u: at a feature-parent node $\widehat{R}_i(u, C) = \phi_i^{A(u)}(v_{x,C}^u)$ is a convex combination of marginal contributions of v_x, hence $|\widehat{R}_i(u, C)| \le \Lambda$; and at a sampled node the realized value is an average over the R_u draws of the realized child estimates, so the bound is preserved. The same argument applied to $R_i(u, C)$ is the last statement of Lemma 2.

Now write the root call as

$$\widehat{\Psi}_i^{\mathcal{H}}(v_x) = \frac{1}{R_{\mathrm{root}}} \sum_{r=1}^{R_{\mathrm{root}}} Y_r,$$

where Y_r is the contribution to feature i of the rth sampled ordering of the root's children, including all randomness generated in the descendant recursion of that draw. Because draws at distinct node calls are mutually independent and each root draw regenerates the entire descendant recursion, $Y_1, \dots, Y_{R_{\mathrm{root}}}$ are independent and identically distributed. By the inductive step of Proposition 3(i), $\mathbb{E}(Y_r) = \Psi_i^{\mathcal{H}}(v_x)$ for every budget configuration, and by the preceding paragraph $Y_r \in [-\Lambda, \Lambda]$, an interval

whose endpoints do not depend on any budget. Hoeffding's inequality for the mean of independent bounded variables with range 2Λ gives

$$\Pr\left\{\left|\widehat{\Psi}_i^{\mathcal{H}}(v_x) - \Psi_i^{\mathcal{H}}(v_x)\right| > \varepsilon\right\} \le 2\exp\left(-\frac{2R_{\text{root}}\varepsilon^2}{(2\Lambda)^2}\right) = 2\exp(-c_\varepsilon R_{\text{root}})\,,$$

which is Eq. (S16). Note that the inequality is obtained at the root call alone; no union bound over the internal node–context states is required, and the descendant budgets enter only through the distribution of Y_r, not through the bound. Under the stated summability condition, $\sum_m \Pr\{|\widehat{\Psi}_i^{\mathcal{H},(m)} - \Psi_i^{\mathcal{H}}| > \varepsilon\} < \infty$ for every $\varepsilon > 0$, so the Borel–Cantelli lemma gives almost sure convergence along a sequence of $\varepsilon \downarrow 0$.

Some such condition is needed. The inequality Eq. (S16) bounds the deviation probabilities from above, and divergence of $\sum_m \exp(-c_\varepsilon R_{\text{root}}^{(m)})$ does not by itself imply that the deviation events occur infinitely often; conversely, with fresh independent draws at every m and sufficiently slow budget growth there is no coupling forcing the successive estimates to stabilize, so almost sure convergence may fail in nondegenerate examples while Proposition 3(iii) continues to hold. □

Remark 1 (Hybrid exact–Monte Carlo estimator)**.** *The same conditioning argument applies to the hybrid estimator described in the main text. Low-degree hierarchy nodes expanded exactly integrate the corresponding local child-ordering distributions analytically, whereas high-degree hierarchy nodes estimate the remaining conditional expectations by Monte Carlo sampling of child orderings. At feature-parent nodes for tree ensemble models, the default exact solver is* InnerTreeDP*, which computes the local Shapley allocation within the terminal feature group without subset enumeration. By iterated expectation, replacing selected exact hierarchy-level local averages by unbiased Monte Carlo averages preserves the target expectation $\Psi^{\mathcal{H}}(v_x)$. Therefore, under predetermined local sampling budgets and mutually independent draws at distinct node calls, the hybrid estimator remains unbiased for the levels-structure value. The realized-sample efficiency argument and the variance recursion apply verbatim, with the exactly expanded nodes contributing a zero sampling variance term, so parts (ii)–(iv) of Proposition 3 also carry over; if the root itself is expanded exactly, the concentration bound* Eq. (S16) *relies on the root call being sampled, so that its summands are independent and identically distributed. When the root is expanded exactly, that structure is no longer available, and an analogous bound would require a separate argument accounting for the weighted combination of several sampled descendant estimators. No concentration or almost-sure statement is asserted for that configuration.*

S4. Detailed Simulation Settings

This section provides the simulation grids, reference computations, sampling conventions, diagnostic definitions, and run-completeness checks underlying the study summarized in the main text. The experiments evaluate attribution accuracy, finite-budget Monte Carlo behavior, numerical coherence, computational scalability, and sensitivity to the local Monte Carlo budget. Scalability is examined separately with respect to hierarchy depth, feature-parent size, and hierarchy degree because these settings isolate different sources of computational growth. The primary TLSV-MC configuration uses the fixed local child-ordering budget $R_u = 3$ at each sampled non-feature-parent node. A separate paired sensitivity experiment varies this budget while holding the fitted model, hierarchy, explained observations, and exact reference fixed. Child orderings are sampled uniformly and independently at distinct node calls; no adaptive stopping or antithetic pairing is used. Table S1 lists the hierarchy regimes and the role each plays, and Table S2 gives the full grid of data-generating, model-fitting and implementation settings.

Scope of the reference computations. The available same-target references differ substantially in cost. DIRECT-ENUM is used in the small exhaustive-validation settings to evaluate the permutation definition directly. A dedicated inner-solver comparison replaces INNERTREEDP with direct feature-subset enumeration while leaving the hierarchy recursion and coalition-value evaluator unchanged. Multi-resolution coherence is assessed against block values Γ_u computed directly from their context–weight definition rather than from the reported feature-level output. These numerical checks complement the theoretical results in the main text.

Table S1. Hierarchy regimes used in the simulation study. The notation $b_1 \times b_2 \times \cdots \times s$ denotes successive hierarchy-level branching factors followed by feature-parent size s, so that the number of terminal features is $p = s \prod_\ell b_\ell$.

Regime	Structures	Main role
Exhaustive validation	$2 \times 2 \times 2$ ($p = 8$; $\lvert\Omega(\mathcal{H})\rvert = 128$); $2 \times 3 \times 2$ ($p = 12$; $\lvert\Omega(\mathcal{H})\rvert = 4{,}608$)	Same-target comparison of DIRECT-ENUM, TLSV-EXACT, and TLSV-MC at $R_u = 3$; the second setting checks the degree-three recursion on the original multiway hierarchy
Depth scaling	$\underbrace{3 \times 3 \times \cdots \times 3}_{m \text{ factors}}$, $m \in \{3, 4, 5, 6, 7, 8\}$	One-factor-at-a-time assessment of hierarchy-depth effects with every branching factor and feature-parent size fixed at three
Feature-parent-size scaling	$3 \times 3 \times s$, $s \in \{3, 6, 9, 12, 15, 18\}$	Assessment of within-group scaling with the higher-level hierarchy fixed; a dedicated same-target inner-solver comparator uses direct subset enumeration over the same $s \leq 18$ grid
Hierarchy-degree scaling	$d \times 2$, $d \in \{3, 5, 7, 9, 11, 13, 15, 17\}$	One-factor-at-a-time assessment of node degree in the regime where Monte Carlo child-order sampling is most beneficial; DIRECT-ENUM is used for $d \leq 5$
Monte Carlo budget sensitivity	15×2 with $R \in \{1, 3, 5, 10, 20\}$	Paired accuracy–cost analysis in a fixed high-degree hierarchy; all budgets share the same fitted model, explained observations, and TLSV-EXACT reference

Table S2. Detailed simulation grid and implementation settings.

Component	Setting
Data types	Regression and binary classification
Tree models	XGBoost, LightGBM, and ranger
Methods compared	DIRECT-ENUM, TLSV-EXACT, TLSV-MC, and an inner-solver comparator obtained by replacing INNERTREEDP with direct subset enumeration at feature-parent nodes
Training size	$n_{\mathrm{tr}} = 500$ observations per simulation replicate
Held-out explanations	$n = 50$ target observations for attribution accuracy, primary Monte Carlo evaluation, and coherence; $n = 20$ for the depth-, feature-parent-size-, hierarchy-degree, and Monte Carlo budget-sensitivity experiments; and $n = 4$ for the dedicated inner-solver comparator
Simulation replicates	100 replicates for attribution accuracy, primary Monte Carlo evaluation, and coherence; 5 replicates for the scaling, budget-sensitivity, and inner-solver comparison experiments
XGBoost hyperparameters	`eta` $= 0.2$, `max_depth` $= 3$, `colsample_bytree` $= 0.5$, `subsample` $= 1$, `reg_lambda` $= 1$, `nrounds` $= 500$; squared-error objective for regression and logistic objective for classification
LightGBM hyperparameters	`learning_rate` $= 0.2$, `max_depth` $= 3$, `num_leaves` $= 8$, `feature_fraction` $= 0.5$, `lambda_l2` $= 1$, `min_data_in_leaf` $= 10$ for regression and 5 for classification, `nrounds` $= 500$
Ranger hyperparameters	`num.trees` $= 500$, `mtry` $= \lfloor \sqrt{p} \rfloor$, `max.depth` $= 3$, `min.node.size` $= 10$ for regression and 5 for classification, with `probability` $=$ `TRUE` for classification and `node.stats` $=$ `TRUE` so native `treeInfo()$numSamples` node-cover masses are available
Feature-generation model	Multilevel latent-factor model with $a_0 = 0.4$; hierarchy-level loadings increase linearly from 0.5 at the coarsest intermediate level to 0.9 at the finest, with loading 0.7 when only one intermediate level is present
Node aggregates	$B_u^0(X) = \lvert A(u)\rvert^{-1/2} \sum_{i \in A(u)} X_i$; $B_u(X)$ is obtained by standardizing $B_u^0(X)$ using training-sample constants and applying the same constants to held-out observations
Feature-level signals	Approximately 10% active features, with at least two active features when feasible
Node-level signals	Up to four active internal hierarchy nodes selected across multiple depths
Interaction signals	Up to two interactions between non-overlapping hierarchy branches
Coefficient magnitudes	Magnitudes sampled independently from $\mathrm{Unif}[0.5, 1.5]$ with random signs
Regression outcome	$Y = \eta(X) + \varepsilon$, with $\varepsilon \sim \mathcal{N}(0, 1)$
Classification outcome	$\Pr(Y = 1 \mid X) = \mathrm{logit}^{-1}\{\alpha + \eta(X)\}$, where α is chosen within each replicate so that the empirical mean predicted probability in the training sample is approximately 0.5
Classification explanation scale	Raw-margin attributions for XGBoost and LightGBM; probability-scale attributions for ranger probability forests
Monte Carlo budget	Primary fixed local child-ordering budget $R_u = 3$ at each sampled non-feature-parent node; paired sensitivity grid $R \in \{1, 3, 5, 10, 20\}$ at $d = 15$; independent sampling across distinct calls, no adaptive stopping, and no antithetic pairing
Same-target references	DIRECT-ENUM in the exhaustive-validation settings; TLSV-EXACT in the larger scaling settings; direct subset enumeration in the inner-solver comparison; and directly evaluated Γ_u in the block-coherence analysis
Accuracy metrics	Correlation, MAE, RMSE, maximum absolute error, and relative-tolerance agreement at 10^{-3} against the corresponding same-target reference
Monte Carlo diagnostics	Feature- and observation-specific empirical bias, variance, and RMSE across 20 independent runs; pooled relative bias and RMSE normalized by the RMS exact-attribution magnitude; bias-to-Monte-Carlo-standard-error ratio; and runtime/speed-up summaries across budgets
Coherence diagnostics	Δ_{eff}, Δ_{agg}, Δ_{blk}, and the degree-one invariance check, all defined below; Δ_{eff}, Δ_{blk} and the invariance check are substantive, whereas Δ_{agg} is an implementation-level bookkeeping diagnostic
Output-scale tolerance	$v_x(N)$ is required to reproduce the fitted learner's prediction within $\lvert \mathrm{native} - \mathrm{unified} \rvert \le a + r \max\{\lvert \mathrm{native} \rvert, \lvert \mathrm{unified} \rvert\}$, with $r = 10^{-6}$ throughout, $a = 10^{-5}$ for the double-precision learners (LightGBM, ranger) and $a = 10^{-4}$ for the single-precision learner (XGBoost), the latter accommodating the observed numerical reproduction floor associated with XGBoost's single-precision model representation
Computational metrics	Runtime, number of hierarchy contexts, number of context-conditioned inner-solver calls, and number of sampled hierarchy-level child orderings
Runtime implementation	Single-core C++ implementation on a Mac Studio with M1 Ultra and $64\,\mathrm{GB}$ memory

Learner-interface validation and cover provenance. Before the production campaign, each fitted-learner interface underwent an independent small-game preflight that was separate from the timed experiments. At the fixed $2 \times 2 \times 2$ hierarchy ($p = 8$), all $2^8 = 256$ coalitions were checked. For XGBoost, the production JSON representation was compared with an independent table-based parser; for LightGBM, the production `lgb.dump()` JSON representation was compared with an independent `lgb.model.dt.tree()` reconstruction; and for ranger, the unified evaluator was compared with a direct `treeInfo()`-based evaluator while native `type="terminalNodes"` and `predict.all=TRUE` outputs independently verified terminal-node predictions. The native full-coalition prediction additionally anchored every learner. These checks target the learner-to-v_x interface itself and therefore complement the same-target algorithmic comparisons, which intentionally hold v_x fixed.

The absent-feature weights in Eqs. (14)–(15) use learner-native node-cover masses. XGBoost and LightGBM obtain the required cover/count information from their native serialized tree representations, whereas ranger uses `treeInfo()$numSamples` and is therefore fitted with `node.stats=TRUE`. The children-sum normalization and the degenerate-node fallbacks are then applied identically by the coalition evaluator and InnerTreeDP. Learner-specific routing conventions are also held fixed: the XGBoost interface uses its strict split comparison on the single-precision evaluation path, whereas the LightGBM and ranger interfaces use their validated non-strict, double-precision representations.

For XGBoost and LightGBM, model preparation also calibrates one additive constant when the serialized trees do not by themselves reproduce the native raw prediction. The constant is estimated from preparation rows, required to be stable within the model-interface tolerance, appended as a one-leaf tree, and revalidated against native predictions. Because this same constant is added to every coalition value, it cancels from all marginal contributions; its role is to align the unified full-coalition prediction and to keep the baseline fixed across later explanation batches. The preparation rows used for this calibration do not define a SHAP background distribution.

Run completeness and numerical audit. The finalized reporting set contains 1,200 replicate-level cells for exact validation (Experiment A), 1,200 cells for the fixed-budget Monte Carlo analysis (Experiment B), and 1,200 cells for the coherence analysis (Experiment C). Each Experiment B cell contains 20 independent Monte Carlo runs, yielding 24,000 fixed-budget runs in total. The hierarchy-degree experiment contains 240 completed cells, the Monte Carlo budget-sensitivity experiment contains 150 completed cells, and the dedicated inner-solver comparison contains 180 completed cells. The production depth and feature-parent-size sweeps contain 360 design cells in total, each executed under both TLSV-Exact and TLSV-MC for 720 solver runs. An additional 360 production-scaling Direct-Enum jobs are recorded as not enumerable by design because they exceed the permutation threshold described below. All expected design keys are present without duplicates, the production scaling jobs contain no failures, all publication metrics are finite, and the observed logical inner-solve counts agree exactly with the theoretical counts used for the complexity comparisons.

The production implementation uses only exact-preserving preprocessing. Fixed Shapley weights and model-side cover quantities are cached, deterministic present-feature routing is reused within each observation, and feature-parent solves use local workspaces rather than full p-dimensional scratch vectors. In addition, for a feature-parent node u, InnerTreeDP skips any tree containing no split feature in $A(u)$, because that tree is constant in the restricted game on $A(u)$ and therefore contributes exactly zero to its local Shapley allocation. These optimizations do not change the path-dependent coalition-value evaluator, the hierarchy recursion, the logical context counts Q_u and $\widetilde{Q}_u$, or the target TLSV allocation.

Exhaustive hierarchy-consistent enumeration. Direct-Enum averages feature marginal contributions over every hierarchy-consistent permutation in $\Omega(\mathcal{H})$ under the same path-dependent tree-aware value function used by TLSV-Exact and TLSV-MC. Its cost is $O(|\Omega(\mathcal{H})|p)$ per explained observation, where

$$|\Omega(\mathcal{H})| = \prod_{u \in \mathrm{int}(\mathcal{H})} |\mathrm{ch}(u)|!.$$

For the $2 \times 2 \times 2$ hierarchy,

$$|\Omega(\mathcal{H})| = (2!)^{1+2+4} = 128,$$

and for the multiway $2 \times 3 \times 2$ hierarchy,

$$|\Omega(\mathcal{H})| = 2!(3!)^2(2!)^6 = 4{,}608.$$

Both settings therefore permit literal enumeration on the original hierarchy without binary refinement or rebalancing. Detailed learner–task exactness results are reported in Table S7. By contrast, the $3 \times 3 \times 3$ hierarchy already has

$$|\Omega(\mathcal{H})| = (3!)^{1+3+9} = 6^{13},$$

which illustrates why exhaustive permutation averaging is restricted to the smaller validation settings. The restriction is enforced programmatically: a design point is evaluated by Direct-Enum only when $|\Omega(\mathcal{H})|$ does not exceed 2×10^5, and is otherwise recorded as not enumerable. This is why Direct-Enum appears in the two exhaustive-validation hierarchies and, in the hierarchy-degree experiment, only for $d \leq 5$: the single-level hierarchy $d \times 2$ has $|\Omega(\mathcal{H})| = d!\, 2^d$, which equals 3,840 at $d = 5$ but 645,120 at $d = 7$. In the depth- and feature-parent-size sweeps the threshold is exceeded at every design point, so no Direct-Enum value is reported there and TLSV-Exact, itself validated against Direct-Enum in the smaller settings, serves as the reference.

Monte Carlo evaluation. For the primary finite-budget accuracy and stability analysis, the fitted model and target observations are held fixed while 20 independent TLSV-MC runs are performed at $R_u = 3$. Sampling occurs only at non-feature-parent nodes; feature-parent nodes are evaluated exactly by InnerTreeDP. For each feature–observation pair, empirical bias, variance, and RMSE are calculated across the 20 runs relative to the designated exact reference and then summarized across observations and simulation replicates. Feature-specific calculation avoids interpreting cancellation of signed errors across features as evidence of unbiasedness. The repeated runs assess finite-budget behavior; unbiasedness and asymptotic convergence are established theoretically in Proposition 3 of the main text. The separate paired budget-sensitivity experiment below complements this primary analysis by varying only R while holding the fitted model, explained observations, hierarchy, and exact reference fixed.

Monte Carlo budget-sensitivity experiment. The primary simulations deliberately use $R_u = 3$ to stress the low-budget regime. To characterize how this computational choice affects accuracy, an additional paired experiment fixes the hierarchy at 15×2 and varies

$$R \in \{1, 3, 5, 10, 20\}.$$

For each learner–task–replicate cell, the five budgets reuse the same fitted model, the same 20 explained observations, and the same observation-specific TLSV-Exact reference already computed for the corresponding $d = 15$ degree-sweep cell. Twenty independent Monte Carlo runs are generated at each budget. Let Ψ_{ij} denote the exact attribution for feature j of explained observation i, and let $\widehat{\Psi}_{ij}^{(b)}(R)$ be Monte Carlo run b. The pooled relative RMSE is the RMS Monte Carlo error divided by the RMS exact-attribution magnitude over all feature–observation entries within a cell. The analogous pooled relative bias uses the run-averaged Monte Carlo attribution in the numerator. This normalization permits aggregation across learners that operate on different attribution scales. Runtime is measured for one TLSV-MC call, and speed-up is paired against the matching exact runtime.

Table S3. Monte Carlo budget sensitivity at the fixed high-degree hierarchy $d = 15$ (feature-parent size 2). The primary estimator remains $R = 3$; this diagnostic varies only the local child-order budget. Entries are medians over 6 learner–task cells and 5 simulation replicates; each cell uses 20 independent Monte Carlo runs per budget.

R	Contexts/child	Rel. RMS bias	Rel. pooled RMSE	Bias/MCSE	MC runtime (s)	Exact/MC	Δ_{eff}
1	1	0.0447	0.2111	1.00	0.105	2429.9	6.5×10^{-15}
3	3	0.0260	0.1215	0.96	0.138	1872.2	4.4×10^{-15}
5	5	0.0197	0.0961	0.96	0.173	1523.9	4.0×10^{-15}
10	10	0.0142	0.0671	1.02	0.256	1038.9	3.9×10^{-15}
20	20	0.0101	0.0465	0.97	0.420	635.1	3.7×10^{-15}

All budgets are evaluated against the same fitted-model and observation-specific TLSV-Exact reference already computed for the $d = 15$ degree-sweep cell. Relative errors are normalized by the RMS exact attribution magnitude. Bias/MCSE uses an RMS bias numerator and the matching RMS Monte Carlo standard error. The runtime column is the median runtime of one TLSV-MC call.

Coherence diagnostics: efficiency and cross-resolution agreement. Allocation efficiency is measured by

$$\Delta_{\text{eff}} = \frac{\left| \sum_{i \in N} \widehat{\Psi}_i^{\mathcal{H}}(v_x) - \{v_x(N) - v_x(\emptyset)\} \right|}{1 + |v_x(N) - v_x(\emptyset)|}.$$

For an implementation-level aggregation check, define

$$\widehat{\Psi}_u^{\mathcal{H}}(v_x) = \sum_{i \in A(u)} \widehat{\Psi}_i^{\mathcal{H}}(v_x),$$

with $\widehat{\Psi}_{\{i\}}^{\mathcal{H}} = \widehat{\Psi}_i^{\mathcal{H}}$ for a feature leaf. Then

$$\Delta_{\text{agg}} = \max_{u \in \text{int}(\mathcal{H})} \frac{\left| \widehat{\Psi}_u^{\mathcal{H}}(v_x) - \sum_{u' \in \text{ch}(u)} \widehat{\Psi}_{u'}^{\mathcal{H}}(v_x) \right|}{1 + \left| \widehat{\Psi}_u^{\mathcal{H}}(v_x) \right|}.$$

Because the children of u partition $A(u)$, this identity holds for any internally consistent aggregation of a feature-level vector. Thus, Δ_{agg} checks the implementation of reported node summaries but is not an independent test of Proposition 1(v).

The substantive cross-resolution check compares the aggregated feature-level attribution with the block value computed independently from

$$\Gamma_u(v_x) = \sum_{(C,w) \in \mathcal{R}_u} w \left\{ v_x(C \cup A(u)) - v_x(C) \right\}.$$

This calculation uses a separate pass over $\mathcal{R}_u$ and does not read or reaggregate the reported feature-level output. Define

$$\Delta_{\mathrm{blk}} = \max_{u \in \mathrm{int}(\mathcal{H}) \setminus \{\mathrm{root}\}} \frac{\left| \sum_{i \in A(u)} \widehat{\Psi}_i^{\mathcal{H}}(v_x) - \Gamma_u(v_x) \right|}{1 + |\Gamma_u(v_x)|}.$$

The root is excluded because its block identity coincides with global efficiency. For DIRECT-ENUM and TLSV-EXACT, Δ_{eff} and Δ_{blk} should be at numerical precision. A realized TLSV-MC allocation at $R_u = 3$ also preserves global efficiency to numerical precision, whereas the corresponding quantity under TLSV-MC measures finite-budget approximation error relative to the exact block target and is accordingly reported as the *MC block-target approximation error* rather than as Δ_{blk}. It should not be interpreted as a failure of the exact multi-resolution identity. Figure S1 collects the residuals from all of these comparisons on a common scale, so that the exact identities and the finite-budget approximation error can be seen side by side. Finally, the degree-one invariance check reports $\Delta_{\mathrm{deg1}} = \max_{i \in N} |\widehat{\Psi}_i^{\mathcal{H}}(v_x) - \widehat{\Psi}_i^{\mathcal{H}^+}(v_x)|$, where $\mathcal{H}^+$ inserts a degree-one carry-forward node above every feature-parent node of $\mathcal{H}$. Since the two hierarchies induce the same permutation set, this must be exactly zero rather than merely small, and it is the numerical check on the normalization of Section 3.3 of the main text.

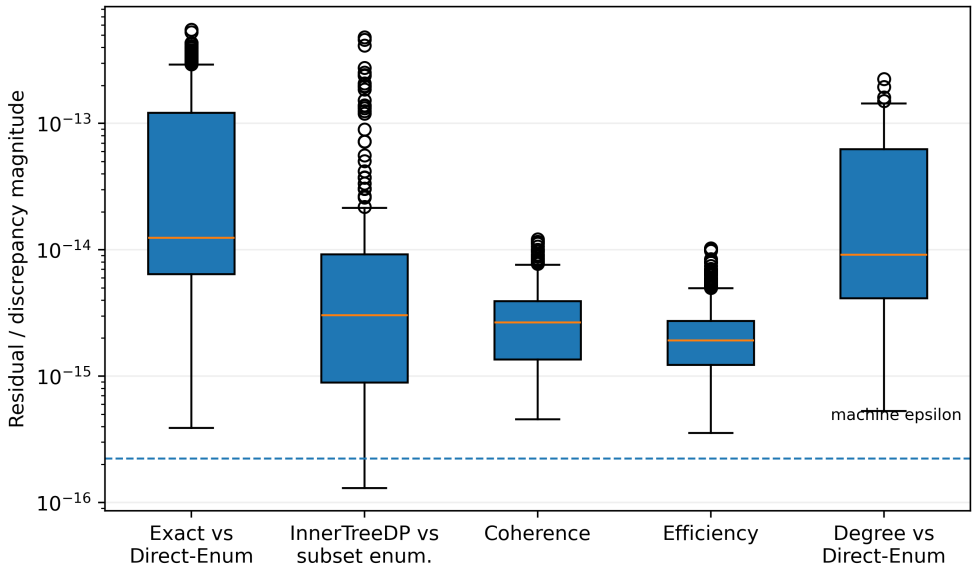


Fig. S1. Numerical residuals from the exact reference comparisons and the allocation identities. From left: TLSV-EXACT against DIRECT-ENUM in the exhaustive-validation hierarchies; INNERTREEDP against direct subset enumeration; the multi-resolution residual Δ_{blk}; the efficiency residual Δ_{eff}; and TLSV-EXACT against DIRECT-ENUM in the hierarchy-degree experiment at $d \leq 5$. The first, second and fifth entries are absolute discrepancies between two computations of the same quantity, whereas Δ_{blk} and Δ_{eff} are the normalized residuals defined in this section; the vertical scale is therefore common only in order of magnitude and the groups are not directly comparable as a single quantity. Boxes span the interquartile range across all design points, learners, tasks and replicates.

Depth-scaling experiment. To isolate hierarchy depth, every branching factor and the feature-parent size are fixed at three, while the number of factors varies as

$$p = 3^m, \qquad m \in \{3, 4, 5, 6, 7, 8\}.$$

There are 3^{m-1} feature-parent nodes, and each is reached through $m - 1$ non-feature-parent nodes of degree three. Because exact recursion considers $2^{3-1} = 4$ sibling subsets at each such node, the number of context-conditioned inner-solver calls per explained observation is

$$3^{m-1} 4^{m-1} = 12^{m-1}.$$

At $m = 8$, this is $12^7 = 35{,}831{,}808$ calls. Each additional factor multiplies the exact count by 12. This experiment therefore characterizes depth scaling of the hierarchy recursion. The complete depth and feature-parent-size runtime summaries are reported in Table S6. Since all nodes have degree three, the exact local expansion is relatively small, and a large Monte Carlo speed advantage is not assumed in this regime.

Feature-parent-size scaling experiment. To isolate feature-parent size, the higher-level hierarchy is fixed at 3×3 and

$$3 \times 3 \times s, \qquad s \in \{3, 6, 9, 12, 15, 18\}.$$

Thus, $p = 9s$, while the higher-level grouping structure remains unchanged. Each of the nine feature-parent nodes is reached through two degree-three nodes, producing

$$Q_u = 2^{3-1} 2^{3-1} = 16$$

outer contexts per feature-parent node and $9 \times 16 = 144$ context-conditioned inner-solver calls per explained observation. This count is constant in s, so runtime growth in this sweep primarily reflects the within-group solver.

Inner-solver comparator. At a feature-parent node u, the restricted game has player set $A(u)$. Direct evaluation of the local Shapley formula Eq. (S2) can reuse coalition values but still requires evaluating all $2^{|A(u)|}$ subsets of $A(u)$. The InnerTreeDP solver defined in the main text computes the same context-conditioned local Shapley values by exploiting the fitted tree structure and the path-dependent recursion in Eqs. (14)–(15), rather than explicitly enumerating these feature subsets.

For a same-target comparison, TLSV-Exact is rerun with direct subset enumeration substituted for InnerTreeDP at every feature-parent node. The fitted model, target observations, hierarchy, outer contexts, coalition-value semantics, model-side branch preprocessing, observation routing, and feature-parent-specific tree-relevance filtering are held fixed. The two implementations therefore differ only in the local feature-subset solver and should agree to numerical precision, while their runtime ratio measures the computational benefit of the tree-aware dynamic-programming solver. This dedicated comparison uses $s \in \{3, 6, 9, 12, 15, 18\}$ and four target observations per fitted model. Extending the comparator through $s = 18$ makes the crossover and subsequent exponential separation directly visible while keeping all other computations identical between the two arms.

Table S4. Effect of the within-group solver, with the higher-level hierarchy fixed at 3×3 and the feature-parent size s varying. Runtimes are medians over 6 learner–task cells and 5 replicates.

s	p	Inner solves (4 obs.)	2^s	InnerTreeDP (s)	Subset enum. (s)	Subset/DP	Max abs. diff.
3	27	576	8	0.142	0.09	0.6	5.3×10^{-15}
6	54	576	64	0.144	0.31	2.1	5.8×10^{-15}
9	81	576	512	0.144	2.01	14.0	7.1×10^{-15}
12	108	576	4,096	0.145	15.23	103.2	9.8×10^{-15}
15	135	576	32,768	0.145	122.73	829.6	5.6×10^{-14}
18	162	576	262,144	0.146	962.85	6475.5	4.8×10^{-13}

Both arms run Algorithm 1 of the main text on the same fitted models and the same hierarchies and differ only in the solver used at feature-parent nodes, so the runtime ratio isolates that solver. The number of context-conditioned inner solves is constant in s by construction, and the two arms agree to numerical precision.

Hierarchy-degree experiment. The depth and feature-parent-size sweeps use only degree-three hierarchy nodes. If a non-feature-parent node has d children, exact local integration for a particular child considers every subset of its remaining $d - 1$ siblings and therefore uses 2^{d-1} hierarchy contexts per child. At $d = 3$, this is only four contexts, compared with the Monte Carlo budget $R_u = 3$.

To examine the regime in which child-order sampling is most advantageous, a separate experiment uses single-level hierarchies

$$d \times 2, \qquad d \in \{3, 5, 7, 9, 11, 13, 15, 17\},$$

where the root has d feature-parent children, each containing two features. Hence the root is the only non-feature-parent node and $p = 2d$. For each root child, TLSV-Exact evaluates 2^{d-1} sibling subsets, whereas TLSV-MC uses $R_u = 3$ sampled child orderings. The ratio of local context counts is therefore

$$\frac{2^{d-1}}{R_u},$$

which ranges from $4/3$ at $d = 3$ to $65536/3$ at $d = 17$. The total number of hierarchy-consistent feature permutations is $d!2^d$, so Direct-Enum is used for $d \leq 5$, with TLSV-Exact serving as the same-target reference for larger d.

Table S5. Hierarchy-degree scaling in the single-level hierarchies $d \times 2$, at the fixed budget $R_u = 3$. Runtimes are medians over 6 learner–task cells and 5 replicates.

d	p	2^{d-1}	Solves (exact)	Solves (MC)	Exact (s)	MC (s)	Exact/MC	Rel. RMSE	Max abs. diff. vs DIRECT-ENUM
3	6	4	240	180	0.15	0.137	1.1	0.150	1.5×10^{-14}
5	10	16	1,600	300	0.35	0.139	2.5	0.135	2.2×10^{-13}
7	14	64	8,960	420	1.11	0.139	7.9	0.140	—
9	18	256	46,080	540	4.25	0.141	30	0.130	—
11	22	1,024	225,280	660	16.89	0.139	117	0.105	—
13	26	4,096	1,064,960	780	68.10	0.142	469	0.129	—
15	30	16,384	4,915,200	900	266.34	0.138	1880	0.115	—
17	34	65,536	22,282,240	1,020	1,065.62	0.140	7533	0.119	—

Solve counts are totals over the 20 explained observations used in each runtime job. The corresponding per-observation counts are $d2^{d-1}$ for TLSV-EXACT and dR_u for TLSV-MC, and the observed counts match Q_u and $\widetilde{Q}_u$ exactly. Relative RMSE divides the Monte Carlo error by the root-mean-square attribution magnitude. Raw errors are not directly comparable across d because both dimensionality and the scale of individual attributions change with the design. The final column is the maximum absolute discrepancy between TLSV-EXACT and DIRECT-ENUM over all explained observations and features in the cell; DIRECT-ENUM is feasible only for $d \leq 5$, so the entry is omitted for larger d.

Table S6. Production scaling of the hierarchy-level recursion. Contexts are the context-conditioned inner solves per explained observation predicted by Q_u and $\widetilde{Q}_u$; runtimes are medians over six learner–task settings and five simulation replicates at $R_u = 3$.

Hierarchy	p	Contexts (exact)	Contexts (MC)	Exact (s)	MC (s)	Exact/MC	Predicted ratio
Hierarchy depth, 3^m							
$3 \times 3 \times 3$	27	144	81	2.2	2.0	1.07	1.78
$3 \times 3 \times 3 \times 3$	81	1,728	729	3.3	2.5	1.33	2.37
$3 \times 3 \times 3 \times 3 \times 3$	243	20,736	6,561	7.8	3.8	2.05	3.16
$3 \times 3 \times 3 \times 3 \times 3 \times 3$	729	248,832	59,049	26.3	7.9	3.31	4.21
$3 \times 3 \times 3 \times 3 \times 3 \times 3 \times 3$	2,187	2,985,984	531,441	109.2	21.6	5.08	5.62
$3 \times 3 \times 3 \times 3 \times 3 \times 3 \times 3 \times 3$	6,561	35,831,808	4,782,969	509.8	74.7	6.85	7.49
Feature-parent size, $3 \times 3 \times s$							
$3 \times 3 \times 3$	27	144	81	2.2	2.0	1.08	1.78
$3 \times 3 \times 6$	54	144	81	2.2	2.0	1.08	1.78
$3 \times 3 \times 9$	81	144	81	2.2	2.1	1.08	1.78
$3 \times 3 \times 12$	108	144	81	2.3	2.1	1.08	1.78
$3 \times 3 \times 15$	135	144	81	2.2	2.1	1.07	1.78
$3 \times 3 \times 18$	162	144	81	2.3	2.1	1.08	1.78

The measured runtime ratio approaches the ratio predicted from the context counts as the fixed per-observation overhead becomes negligible. Under the feature-parent-size sweep the context count is constant by construction, so the runtime remains nearly flat; the dedicated same-target within-group solver comparator is reported separately in the main text. The hierarchy $3 \times 3 \times 3$ is the smallest design point of both sweeps and was executed as a separate job in each; the two rows therefore report independent runs of the same configuration, and the small difference in the measured ratio (1.07 versus 1.08) reflects run-to-run timing variation rather than a difference in the computation. The Exact/MC column is the median of paired per-cell ratios, not the ratio of the two median runtimes.

Table S6 gives the corresponding counts and runtimes, and Figure S2 shows how the measured exact-to-Monte-Carlo runtime ratio approaches the ratio predicted by those counts as depth grows.

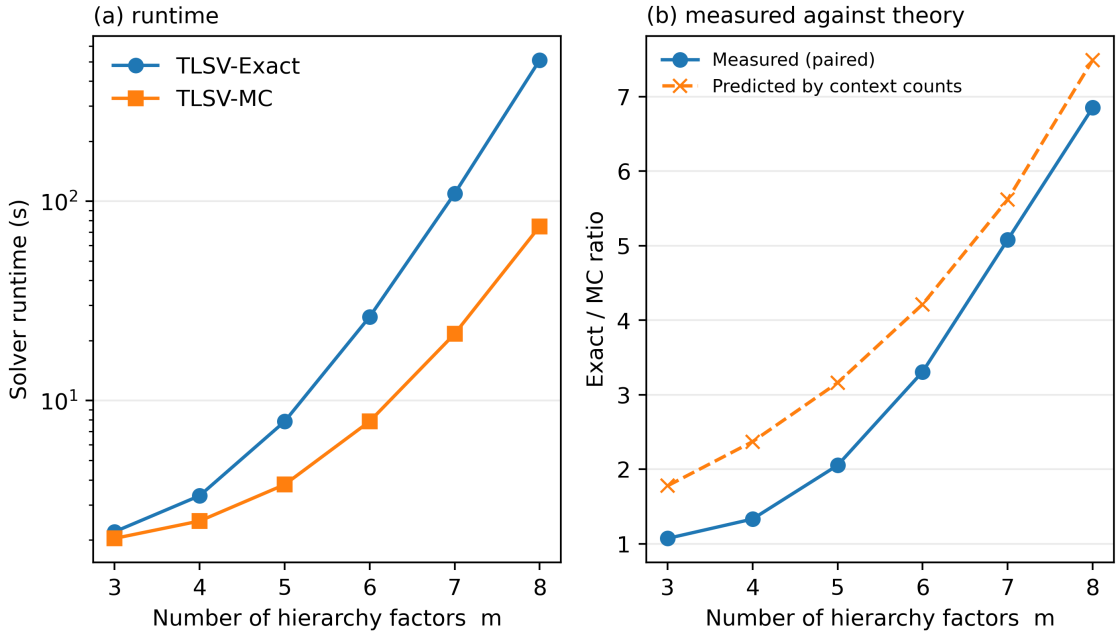


Fig. S2. Hierarchy-depth scaling for the production sweep $3 \times 3 \times \cdots \times 3$, with $m \in \{3, 4, 5, 6, 7, 8\}$ factors and the primary budget $R_u = 3$. (a) Median single-core solver runtimes for TLSV-EXACT and TLSV-MC. (b) The paired median exact-to-Monte-Carlo runtime ratio compared with the ratio predicted by the corresponding context-conditioned inner-solve counts. Points summarize six learner–task cells and five simulation replicates. The complete numeric summaries are given in Table S6; the measured ratio approaches the count-based prediction as depth increases and fixed implementation overhead becomes less dominant.

Table S7. Detailed same-target numerical agreement of TLSV-EXACT with DIRECT-ENUM across the 12 exhaustive-validation hierarchy–learner–task cells. Correlation is the minimum and the error columns are maxima over 100 simulation replicates per cell.

Hierarchy	$\lvert\Omega(\mathcal{H})\rvert$	Learner	Task	Corr.	RMSE	Max abs.	Agree (%)	Δ_{eff}
$2 \times 2 \times 2$	128	XGBoost	Regression	1.000	2.0×10^{-15}	1.5×10^{-14}	100.0	6.0×10^{-15}
		XGBoost	Classification	1.000	3.0×10^{-15}	2.2×10^{-14}	100.0	9.1×10^{-15}
		LightGBM	Regression	1.000	2.1×10^{-15}	1.8×10^{-14}	100.0	6.9×10^{-15}
		LightGBM	Classification	1.000	4.0×10^{-15}	3.1×10^{-14}	100.0	8.8×10^{-15}
		Ranger	Regression	1.000	9.5×10^{-16}	6.7×10^{-15}	100.0	3.0×10^{-15}
		Ranger	Classification	1.000	2.6×10^{-16}	1.3×10^{-15}	100.0	1.6×10^{-15}
$2 \times 3 \times 2$	4,608	XGBoost	Regression	1.000	2.9×10^{-14}	2.6×10^{-13}	100.0	3.9×10^{-15}
		XGBoost	Classification	1.000	4.8×10^{-14}	4.3×10^{-13}	100.0	6.4×10^{-15}
		LightGBM	Regression	1.000	2.9×10^{-14}	2.9×10^{-13}	100.0	5.2×10^{-15}
		LightGBM	Classification	1.000	5.5×10^{-14}	5.5×10^{-13}	100.0	1.3×10^{-14}
		Ranger	Regression	1.000	1.5×10^{-14}	1.8×10^{-13}	100.0	2.1×10^{-15}
		Ranger	Classification	1.000	1.6×10^{-15}	1.2×10^{-14}	100.0	1.9×10^{-15}

DIRECT-ENUM averages marginal contributions over all $|\Omega(\mathcal{H})|$ hierarchy-consistent permutations and is therefore a same-target reference for the definition itself. "Agree" is the percentage satisfying $|\widehat{\Psi} - \Psi_{\text{ref}}| \leq 10^{-3} \max(|\Psi_{\text{ref}}|, 10^{-8})$. Δ_{eff} is the efficiency residual of the realized allocation.

S5. Real Data Application: Full Hierarchical Attributions

Exact computation and validation. The real-data attributions of Section 4.3 of the main text were computed with TLSV-EXACT rather than by sampling. This is a demanding setting for exact computation: the six-level taxonomy has 115 terminal feature groups, a maximum node degree of 11 and depth 5, so the context count of the main text gives 1,549,737,472 context-conditioned inner solves per patient. For comparison, TLSV-MC at $R_u = 3$ would require 27,945 and the hybrid estimator with degree threshold 2 would require 6,033. The exact run completed in 79,149 seconds, about 22 hours, on twelve cores. It is feasible at all because the within-group solver removes the $2^{|A(u)|}$ factor that would otherwise multiply this count; only the hierarchy-level factor remains.

Two safeguards were checked before any interpretation. The full-coalition value reproduced the fitted LightGBM margin with a maximum absolute gap of zero and no patient outside tolerance, and the maximum efficiency residual was 3.4×10^{-11} across all 257 patients. Because the attributions are exact rather than sampled, no seed-stability analysis applies.

Table S8 lists every taxon that receives a non-zero levels-structure attribution in the immune checkpoint inhibitor cohort of Section 4.3 of the main text, together with its mean absolute node attribution and the share of patients receiving a positive contribution. Entries follow the nesting of the taxonomy rather than being grouped by level: each phylum is followed by the classes it contains, each class by its orders, and so on down to species, with indentation marking depth and siblings ordered by decreasing importance. Reading the table from top to bottom therefore traverses the same tree that Figure 2 of the main text draws, so a block and the taxa composing it stay adjacent. The "Level" column abbreviates phylum, class, order, family, genus and species as P, C, O, F, G and S. The "No." column gives the label used in that figure: taxa are numbered by decreasing mean absolute attribution across all six levels together, so the number gives a taxon's position in that ordering, and the 45 largest are labelled.

Of the 218 internal nodes and 379 species in the six-level taxonomy, 116 internal nodes and 116 species receive a non-zero attribution. The remainder are omitted because the fitted LightGBM ensemble never splits on the corresponding features, so their levels-structure value is exactly zero rather than merely small; this is a property of the fitted model, not a thresholding choice. The mean absolute node attribution is the quantity I_u^{node} defined in Eq. (22) of the main text, and the positive share is the fraction of the 257 patients for whom the node attribution is positive. The two are reported together because they answer different questions: the first measures how much a taxon contributes, and the second in which direction it acts. All values are on the raw-margin scale of the fitted classifier, and none of them should be read as a causal effect.

Table S8. Levels-structure attributions for every taxon receiving a non-zero attribution, arranged so that the nesting of the taxonomy is preserved: each phylum is followed by the classes it contains, each class by its orders, and so on to species. Indentation indicates level. Numbered entries are the ones labeled in Figure 2 of the main text.

Taxon	No.	Level	Mean $\lvert\Psi_u\rvert$	Positive share
Firmicutes	1	P	1.1313	0.424
Clostridia	3	C	0.7632	0.416
Clostridiales	4	O	0.7632	0.416
Clostridiaceae	10	F	0.6121	0.362
Clostridium	11	G	0.5824	0.342
Clostridium sp. CAG:433	15	S	0.4665	0.292
Clostridium sp. CAG:217		S	0.1051	0.175
Clostridium sp. 27_14		S	0.0718	0.245

continued on the next page

Table S8, continued from the previous page.

Taxon	No.	Level	Mean $\lvert\Psi_u\rvert$	Positive share
Clostridium sp. CAG:798		S	0.0591	0.700
Clostridium sp. CAG:343		S	0.0287	0.362
Clostridium sp. CAG:440		S	0.0121	0.475
Clostridium sp. AT4		S	0.0115	0.802
Clostridium sp. CAG:127		S	0.0106	0.486
Clostridium sp. KLE 1755		S	0.0081	0.424
Clostridium sp. CAG:448		S	0.0037	0.475
Clostridium sp. L2-50		S	0.0007	0.362
Clostridium sp. CAG:138		S	0.0006	0.412
Clostridium sp. CAG:302		S	0.0001	0.708
Hungatella	35	G	0.1564	0.665
Hungatella hathewayi	36	S	0.1564	0.665
Ruminococcaceae	24	F	0.1839	0.700
Ruminococcus	26	G	0.1820	0.700
Ruminococcus sp. CAG:177	25	S	0.1836	0.700
Ruminococcus bromii		S	0.0152	0.288
Ruminococcus bicirculans		S	0.0085	0.405
Ruminococcus sp. CAG:254		S	0.0085	0.249
Ruminococcus callidus		S	0.0034	0.619
Anaerotruncus		G	0.0319	0.494
Anaerotruncus colihominis		S	0.0319	0.494
Ruminiclostridium		G	0.0226	0.409
[Clostridium] leptum		S	0.0219	0.381
[Eubacterium] siraeum		S	0.0038	0.720
Faecalibacterium		G	0.0099	0.428
Faecalibacterium prausnitzii		S	0.0099	0.428
Lachnospiraceae	27	F	0.1745	0.377
Blautia	32	G	0.1576	0.339
Blautia wexlerae	39	S	0.1390	0.284
Blautia sp. KLE 1732		S	0.0365	0.218
[Ruminococcus] torques [C]		S	0.0117	0.642
Blautia obeum		S	0.0102	0.109
[Ruminococcus] gnavus		S	0.0071	0.521
Coprococcus		G	0.0430	0.230
Coprococcus comes		S	0.0430	0.230
Lachnoclostridium		G	0.0323	0.529
[Clostridium] symbiosum [C]		S	0.0245	0.521
[Clostridium] asparagiforme		S	0.0171	0.272
[Clostridium] sp. [C boltae/clostridioforme]		S	0.0014	0.720
Anaerostipes		G	0.0322	0.222
Anaerostipes hadrus [C]		S	0.0322	0.222
Roseburia		G	0.0194	0.494
Roseburia inulinivorans		S	0.0159	0.440
Roseburia sp. 40_7		S	0.0090	0.374
Roseburia intestinalis		S	0.0031	0.319
Roseburia sp. CAG:309		S	0.0013	0.416
Tyzzerella		G	0.0102	0.206
Tyzzerella nexilis		S	0.0102	0.206
Dorea		G	0.0026	0.420
Dorea formicigenerans [C]		S	0.0025	0.436
Dorea longicatena		S	0.0004	0.778
unresolved family within Clostridiales		F	0.0881	0.358
Flavonifractor		G	0.0677	0.222
Flavonifractor plautii [C]		S	0.0677	0.222
Pseudoflavonifractor		G	0.0438	0.179
Pseudoflavonifractor capillosus		S	0.0438	0.179
Intestinimonas		G	0.0068	0.385
Intestinimonas butyriciproducens		S	0.0068	0.385
Oscillospiraceae		F	0.0218	0.366
Oscillibacter		G	0.0218	0.366
Oscillibacter sp. CAG:155		S	0.0197	0.257
Oscillibacter sp. KLE 1728		S	0.0082	0.171
Eubacteriaceae		F	0.0087	0.409
Eubacterium		G	0.0087	0.409
Eubacterium sp. CAG:156		S	0.0088	0.362
[Eubacterium] hallii		S	0.0008	0.607
Negativicutes	16	C	0.3410	0.549
Selenomonadales	20	O	0.2672	0.712
Selenomonadaceae	21	F	0.2672	0.712
Mitsuokella	22	G	0.2672	0.712
Mitsuokella jalaludinii	23	S	0.2672	0.712
Veillonellales	37	O	0.1472	0.510
Veillonellaceae	38	F	0.1472	0.510
Megasphaera		G	0.1245	0.339
Megasphaera elsdenii		S	0.1245	0.339
Dialister		G	0.1185	0.284
Dialister invisus		S	0.1066	0.284
Dialister succinatiphilus		S	0.0199	0.467
Veillonella		G	0.0267	0.724
Veillonella parvula [C]		S	0.0302	0.724

continued on the next page

Table S8, continued from the previous page.

Taxon	No.	Level	Mean $\|\Psi_u\|$	Positive share
Veillonella sp. oral taxon 158		S	0.0132	0.089
Veillonella atypica [C]		S	0.0121	0.195
Acidaminococcales		O	0.0183	0.529
Acidaminococcaceae		F	0.0183	0.529
Phascolarctobacterium		G	0.0170	0.533
Phascolarctobacterium sp. CAG:266		S	0.0155	0.428
Phascolarctobacterium sp. CAG:207		S	0.0078	0.700
Phascolarctobacterium succinatutens		S	0.0026	0.673
Acidaminococcus		G	0.0046	0.362
Acidaminococcus fermentans		S	0.0046	0.362
Bacilli		C	0.1357	0.490
Lactobacillales		O	0.1264	0.556
Streptococcaceae		F	0.0894	0.385
Lactococcus		G	0.0695	0.323
Lactococcus lactis		S	0.0695	0.323
Streptococcus		G	0.0360	0.210
Streptococcus mitis		S	0.0356	0.148
Streptococcus thermophilus		S	0.0084	0.440
Enterococcaceae		F	0.0677	0.319
Enterococcus		G	0.0677	0.319
Enterococcus faecium [C]		S	0.0677	0.319
Lactobacillaceae		F	0.0352	0.140
Lactobacillus		G	0.0352	0.140
Lactobacillus mucosae		S	0.0352	0.140
Bacillales		O	0.0658	0.529
Staphylococcaceae		F	0.0658	0.529
Staphylococcus		G	0.0658	0.529
Staphylococcus sp. CAG:324		S	0.0658	0.529
Erysipelotrichia		C	0.0664	0.198
Erysipelotrichales		O	0.0664	0.198
Erysipelotrichaceae		F	0.0664	0.198
Merdibacter		G	0.0632	0.144
Merdibacter massiliensis		S	0.0632	0.144
Holdemania		G	0.0145	0.420
Holdemania massiliensis		S	0.0103	0.253
Holdemania filiformis		S	0.0092	0.611
Erysipelatoclostridium		G	0.0060	0.288
[Clostridium] innocuum [C Clostridioides difficile]		S	0.0060	0.288
Proteobacteria	2	P	0.8044	0.268
Gammaproteobacteria	7	C	0.7400	0.268
Enterobacteriales	8	O	0.7398	0.268
Enterobacteriaceae	9	F	0.7398	0.268
Citrobacter	5	G	0.7419	0.268
Citrobacter koseri	6	S	0.7419	0.268
Hafnia		G	0.0169	0.424
Hafnia alvei		S	0.0169	0.424
Pasteurellales		O	0.0029	0.315
Pasteurellaceae		F	0.0029	0.315
Haemophilus		G	0.0029	0.315
Haemophilus parainfluenzae		S	0.0029	0.315
Alphaproteobacteria		C	0.1038	0.619
Rhodospirillales		O	0.1038	0.619
Rhodospirillaceae		F	0.1038	0.619
Azospirillum		G	0.1038	0.619
Azospirillum sp. CAG:239		S	0.0952	0.782
Azospirillum sp. 51_20		S	0.0442	0.553
Betaproteobacteria		C	0.0237	0.623
Burkholderiales		O	0.0237	0.623
Sutterellaceae		F	0.0237	0.623
Parasutterella		G	0.0200	0.732
Parasutterella excrementihominis [C]		S	0.0200	0.732
Sutterella		G	0.0125	0.428
Sutterella wadsworthensis		S	0.0079	0.529
Sutterella sp. CAG:351		S	0.0077	0.179
Sutterella sp. CAG:521		S	0.0023	0.307
Deltaproteobacteria		C	0.0063	0.510
Desulfovibrionales		O	0.0063	0.510
Desulfovibrionaceae		F	0.0063	0.510
Mailhella		G	0.0063	0.510
Mailhella massiliensis		S	0.0063	0.510
Bacteroidetes	12	P	0.5226	0.354
Bacteroidia	13	C	0.5226	0.354
Bacteroidales	14	O	0.5226	0.354
Rikenellaceae	17	F	0.3409	0.280
Alistipes	18	G	0.3409	0.280
Alistipes sp. An31A	19	S	0.3316	0.276
Alistipes inops		S	0.0475	0.444
Alistipes senegalensis		S	0.0370	0.541
Alistipes onderdonkii [C]		S	0.0229	0.451

continued on the next page

Table S8, continued from the previous page.

Taxon	No.	Level	Mean $\lvert\Psi_u\rvert$	Positive share
Alistipes obesi		S	0.0209	0.506
candidatus Alistipes marseilloanorexicus		S	0.0105	0.198
Alistipes shahii		S	0.0090	0.311
Alistipes finegoldii [C]		S	0.0012	0.642
Bacteroidaceae		F	0.1348	0.514
Bacteroides		G	0.1348	0.514
Bacteroides fragilis [C]		S	0.0666	0.510
Bacteroides sp. [C fragilis/ovatus]		S	0.0587	0.595
Bacteroides sp. [C rodentium/uniformis]		S	0.0524	0.654
Bacteroides coprophilus		S	0.0516	0.518
Bacteroides sp. [C dorei/vulgatus]		S	0.0357	0.187
Bacteroides xylanisolvens [C]		S	0.0194	0.553
Bacteroides thetaiotaomicron [C]		S	0.0162	0.786
Bacteroides salyersiae		S	0.0154	0.377
Bacteroides stercoris		S	0.0126	0.654
Bacteroides caccae		S	0.0125	0.580
Bacteroides clarus		S	0.0067	0.467
Bacteroides finegoldii		S	0.0066	0.514
Bacteroides coprocola		S	0.0040	0.265
Bacteroides salanitronis		S	0.0019	0.475
Prevotellaceae		F	0.1339	0.284
Paraprevotella		G	0.1320	0.280
Paraprevotella xylaniphila		S	0.1286	0.276
Paraprevotella clara		S	0.0098	0.459
Prevotella		G	0.0072	0.249
Prevotella stercorea		S	0.0072	0.249
Porphyromonadaceae		F	0.0177	0.638
Parabacteroides		G	0.0164	0.685
Parabacteroides merdae		S	0.0164	0.685
Barnesiella		G	0.0127	0.183
Barnesiella intestinihominis		S	0.0127	0.183
Actinobacteria	33	P	0.1574	0.195
Actinobacteria	29	C	0.1580	0.187
Bifidobacteriales	30	O	0.1580	0.187
Bifidobacteriaceae	31	F	0.1580	0.187
Bifidobacterium	28	G	0.1580	0.187
Bifidobacterium sp. [C gallinarum/pullorum/saeculare]	34	S	0.1572	0.167
Bifidobacterium adolescentis		S	0.0166	0.195
Bifidobacterium longum [C]		S	0.0065	0.716
Bifidobacterium pseudocatenulatum		S	0.0047	0.245
Bifidobacterium dentium [C]		S	0.0035	0.346
Coriobacteriia		C	0.0125	0.537
Eggerthellales		O	0.0117	0.556
Eggerthellaceae		F	0.0117	0.556
Eggerthella		G	0.0114	0.537
Eggerthella lenta [C Clostridioides difficile]		S	0.0114	0.537
Adlercreutzia		G	0.0013	0.533
Adlercreutzia equolifaciens		S	0.0013	0.533
Coriobacteriales		O	0.0049	0.288
Coriobacteriaceae		F	0.0049	0.288
Collinsella		G	0.0049	0.288
Collinsella aerofaciens [C]		S	0.0049	0.288
Enorma		G	4.6×10^{-5}	0.642
Enorma massiliensis		S	4.6×10^{-5}	0.642
Verrucomicrobia	40	P	0.1388	0.280
Verrucomicrobiae	41	C	0.1382	0.268
Verrucomicrobiales	42	O	0.1382	0.268
Akkermansiaceae	43	F	0.1382	0.268
Akkermansia	44	G	0.1382	0.268
Akkermansia muciniphila	45	S	0.1382	0.268
Opitutae		C	0.0144	0.327
Puniceicoccales		O	0.0144	0.327
Puniceicoccaceae		F	0.0144	0.327
Coraliomargarita		G	0.0144	0.327
Coraliomargarita sp. CAG:312		S	0.0144	0.327