SOBOL INDICES FOR PARAMETER-CONDITIONED NEURAL NETWORKS

Anonymous authors
Paper under double-blind review

ABSTRACT

Networks conditioned on a continuous parameter, such as surrogates for parametric partial differential equations, replace many trained models with one. Two questions follow: what has such a network learned about its parameter, and which simpler architectures could match it? We address both with the Sobol (functional ANOVA) decomposition of the trained network over inputs and parameter, estimated by standard Monte-Carlo sampling. On a loss-conditional physics-informed network for parametric 2D Helmholtz, the indices match those of the exact solution to 0.002 (0.005 across three training seeds) and show that 43% of the output depends jointly on space and the parameter, which first-order attribution methods cannot express. The indices of one pilot network also predict the lowest error that any restricted architecture can reach. Networks trained by regression reach the bound read from the pilot to within 0.003 on the headline target, and the exact bound to within 0.001 on two further targets. Networks trained with the physics-informed loss stay above it. Finally, a neural HDMR estimator that fits the decomposition directly learns a nearly correct function (three-way index within 0.02), but its standard readout at the default resolution overstates that index by 0.14, six times the seed spread. The reported indices still sum to one, so this error does not appear in them. Computing the indices from the fitted function removes the error; at this problem size Monte-Carlo estimation is also about 500 times cheaper.

1 INTRODUCTION

Parametric partial differential equations describe how a physical system depends on adjustable inputs: spatial coordinates and one or more physical parameters. Loss-conditional PINNs, neural operators (Li et al., 2021; Lu et al., 2021), and physics-informed DeepONets train a single network $u_\theta(x, \lambda)$ across an entire parametric family at one forward pass per query. Such a network is cheap to query. What it has learned about its inputs is less clear. When a parametric network reaches low residual error, what has it learned about the joint (x, λ) structure? Can we tell, before training a simplified architecture, whether it will keep that accuracy?

Our object is the functional ANOVA decomposition of u_θ on the joint input space, classical since Hoeffding (1948) and Sobol (1993) and developed as HDMR by Rabitz & Aliş (1999); the normalised component variances are Sobol indices. The decomposition applies to any trained parametric network; loss-conditional PINNs are the instance we study. The central theorem follows from elementary L^2 geometry. We train one pilot network and compute its Sobol indices. These indices give a lower bound on the error of every restricted architecture in the design space, attained by the L^2 projection onto that class, so architecture classes can be ruled out before any of them is trained. Restricted architectures are worth considering because they are cheaper to evaluate, their additive components can be inspected directly, and a class that cannot reach the required accuracy should not be trained at all. First-order attribution methods, such as Integrated Gradients (Sundararajan et al., 2017), SHAP (Lundberg & Lee, 2017) and permutation importance, cannot make this prediction when high-order interactions dominate. Our main example is in this regime.

The paper makes three contributions.

(1) Error floors for restricted architectures (Theorem 4). If an approximation may use only a given family of input subsets, its relative L^2 error is at least the square root of the Sobol mass on the

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

excluded subsets. Equality holds at the L^2 projection. We compute these floors from a trained pilot network. Six restricted classes trained by regression on the headline target reach their predicted floors to within 0.003 (five seeds), and the floors predict their ordering. The unrestricted model reaches 0.015.

(2) A measurement of what the network learned. Saltelli Monte-Carlo Sobol indices of the trained surrogate agree with the exact indices of the analytic solution to 0.002 on every index, and to 0.005 across three training seeds (Table 1). One of these indices is a three-way interaction, $S_{xyk} = 0.426$. First-order attribution methods cannot represent such interactions.

(3) An analysis of a fast estimator. A Fourier-feature HDMR estimator with Möbius purification (Algorithm 1, Appendix A) returns the ANOVA component functions from samples. Its standard readout overestimates the three-way index by 0.14. The fitted function itself is within 0.02 of the exact value, so the error comes from the readout. Because the readout normalises the indices to sum to one, the error is not visible in the reported values.

The screening idea has precedent. Hooker (2004) scores non-additive structure in a black-box function by the loss of its projection onto additive classes, and Hart & Gremaud (2018) give the approximation-theoretic identity between Sobol indices and best-approximation error; Sobol' et al. (2007) bound the error of fixing unessential factors; work on effective dimension in quasi-Monte Carlo (Caflisch et al., 1997; Kuo et al., 2010) uses the same Sobol indices to measure how much of a function low-order interactions capture; and interpretable ML has used a pilot model's learned interactions to choose restricted additive architectures (Lou et al., 2013; Tsang et al., 2018a; Yang et al., 2021; Agarwal et al., 2021). We add three things. We state the bound and its equality case on the joint (x, λ) space of parametric surrogates, and we use it to select architectures. We test it against analytic projections and against restricted networks trained from scratch. We also show that the readout of a scalable estimator can produce errors that do not appear in its normalised output.

Scope and limitations. Floors are tested on three manufactured 2D Helmholtz targets and illustrated on one tabular case; parabolic and hyperbolic families appear only as qualitative footprints; shocks and full Navier–Stokes are out of scope. References are manufactured analytic or numerical; measured experimental data is future work. Sobol indices are properties of a specific (target, input measure) pair (Remark 6); we quantify this dependence explicitly (Appendix G).

2 Related work

Functional ANOVA and Sobol indices. The decomposition is due to Hoeffding (1948) for U-statistics, developed for variance-based global sensitivity analysis by Sobol (1993); Sobol' (2001), with the HDMR generalisation by Rabitz & Aliş (1999). The Saltelli scheme (Saltelli et al., 2010) is the canonical Monte-Carlo estimator. Hooker (2007) introduces generalised functional ANOVA with identifying constraints for non-product measures, underpinning the weighted extension in Remark 7. For correlated inputs, where the population decomposition already fails to be orthogonal, relaxed component definitions (Li & Rabitz, 2012) and generalised Hoeffding–Sobol indices (Chastaing et al., 2012) exist; our inputs are independent, and the orthogonality failure we document arises in the *fitted* terms, as an estimation artefact. Polynomial chaos (Xiu & Karniadakis, 2002) is the classical surrogate alternative; purification of fitted interaction terms also appears in interpretable ML (Lengerich et al., 2020).

Neural ANOVA, attribution, Fourier features. Additive models with pairwise interactions (Lou et al., 2013) implement low-order ANOVA by architectural restriction. Integrated Gradients (Sundararajan et al., 2017) and SHAP (Lundberg & Lee, 2017) are intrinsically first-order. Tancik et al. (2020) showed Fourier-feature encoding lets MLPs resolve high-frequency content, directly relevant when ANOVA components are oscillatory.

Parametric PINNs and operator learning. PINNs (Raissi et al., 2019) minimise the PDE residual; loss-conditional training (Dosovitskiy & Djolonga, 2020) conditions one network on the coefficients of its own loss, and applying it to a parametric PDE family yields the loss-conditional PINN (LC-PINN) studied here. FNO (Li et al., 2021) and DeepONet (Lu et al., 2021) are operator-learning alternatives. The decomposition is method-agnostic.

3 THEORETICAL FRAMEWORK

This section develops the analytical machinery behind two claims of the paper. The first is that the functional ANOVA decomposition of a parametric neural network in the joint input–parameter space is well-defined and computable. The second is that the resulting Sobol indices give an a-priori lower bound on the relative-L^2 error of any architecturally restricted neural function class on the same target. The mathematics rests on the classical ANOVA orthogonality theorem and the L^2 Pythagorean identity; our contribution is the application to neural-network function classes and the estimator described in Section 4.

3.1 FUNCTIONAL ANOVA ON THE JOINT (x, λ) SPACE

Let $\Omega \subset \mathbb{R}^{d_x}$ be a compact spatial domain with probability measure p_Ω, and let $\Lambda \subset \mathbb{R}^{d_\lambda}$ be a compact parameter domain with probability measure p_Λ. Write the joint domain $Z = \Omega \times \Lambda$ with product measure $p = p_\Omega \otimes p_\Lambda$. We work in the Hilbert space $L^2(Z, p)$ of square-integrable functions on Z with inner product $\langle u, v\rangle = \int_Z u(z)v(z)\, dp(z)$ and norm $\|u\| = \sqrt{\langle u, u\rangle}$.

Let $d = d_x + d_\lambda$, and index the input coordinates by $[d] = \{1, \dots, d\}$. For any subset $S \subseteq [d]$, denote by z_S the corresponding subset of coordinates and by z_{-S} the complement. We assume that the joint measure factors across coordinates: $p = \prod_{i=1}^{d} p_i$. This holds in our setting because p_Ω is a product of uniform marginals on a hyperrectangle and λ is independent of x under p.

Definition 1 (ANOVA decomposition; Hoeffding, Sobol)**.** *For $u \in L^2(Z, p)$, the functional ANOVA decomposition is the unique decomposition*

$$u(z) \;=\; \sum_{S \subseteq [d]} u_S(z_S), \tag{1}$$

in which each component u_S depends only on the coordinates indexed by S and, for $S \neq \varnothing$, integrates to zero in each of its own coordinates, $\int u_S\, dp_i = 0$ for $i \in S$; under the product measure these conditions make the decomposition unique and the components pairwise orthogonal in $L^2(Z, p)$:

$$\langle u_S, u_T\rangle = 0 \quad \textit{for } S \neq T. \tag{2}$$

The components are defined recursively by

$$u_S(z_S) \;=\; \mathbb{E}\big[u \,\big|\, z_S\big] \;-\; \sum_{T \subsetneq S} u_T(z_T), \tag{3}$$

where the conditional expectation integrates over z_{-S} under p. The empty-set component $u_\varnothing = \mathbb{E}[u]$ is the mean.

The decomposition exists and is unique whenever $u \in L^2(Z, p)$ and the measure factorises; the orthogonality follows from the tower property of conditional expectations. Hoeffding (Hoeffding, 1948) established the decomposition in the context of U-statistics; Sobol (Sobol, 1993; Sobol', 2001) reframed it as the foundational object of variance-based global sensitivity analysis. Rabitz and Aliş (Rabitz & Aliş, 1999) provide the multivariate generalisation in the High-Dimensional Model Representation (HDMR) framework that we adapt for the neural setting in Section 4.

Definition 2 (Sobol indices)**.** *Let $V_S = \|u_S\|^2$ be the variance attributable to subset S. By equation 2, the total variance decomposes as $\mathrm{Var}(u) = \sum_{S \neq \varnothing} V_S$. The Sobol index of S is*

$$S_S \;=\; V_S \,/\, \mathrm{Var}(u), \tag{4}$$

with $\sum_{S \neq \varnothing} S_S = 1$.

We apply Definitions 1–2 to the trained parametric network $u_\theta(x, \lambda)$ regarded as a function on the joint space Z, and to manufactured-reference solutions $u_{\mathrm{ref}}(x; \lambda)$ when these are available for cross-validation.

3.2 THE ARCHITECTURAL FLOOR

Sensitivity analysis is usually presented as a tool for interpreting a fixed function. Our central observation is that the Sobol indices of a trained parametric network determine, by an elementary Hilbert-space argument, a lower bound on the relative-L^2 error attainable by any neural function class that excludes specific subsets from the input dependence.

Definition 3 (Subset-restricted function class)**.** *Let $\mathcal{F} \subseteq 2^{[d]}$ be a family of subsets of input indices. The $\mathcal{F}$-restricted function class is*

$$\mathcal{H}_{\mathcal{F}} = \Big\{ v \in L^2(Z,p) : \; v(z) = \sum_{S \in \mathcal{F}} v_S(z_S), \\ v_S \in L^2(Z,p) \text{ depends only on } z_S \Big\}.$$

A neural architecture whose parameter space can realise only functions in $\mathcal{H}_{\mathcal{F}}$ is called $\mathcal{F}$-architecturally restricted*. The complement $\mathcal{F}^c = 2^{[d]} \setminus \mathcal{F}$ is the set of* forbidden *interactions.*

Two examples we use later in the paper, both for the 2D Helmholtz setting where $d_x = 2$, $d_\lambda = 1$, and the coordinates are (x, y, k):

Additive architecture. The class $\mathcal{H}_{\text{add}}$ with $\mathcal{F} = \{\varnothing, \{x\}, \{y\}, \{k\}, \{x, y\}\}$ corresponds to networks of the form $v(x, y, k) = f_{xy}(x, y) + f_k(k)$. The forbidden interactions are $\{x, k\}$, $\{y, k\}$, and $\{x, y, k\}$.

Pair-additive architecture. The class $\mathcal{H}_{\text{pair}}$ admitting all subsets with $|S| \leq 2$ corresponds to networks expressible as a sum of one- and two-variable functions. The only forbidden interaction is the triplet $\{x, y, k\}$.

The central theorem links the forbidden interactions to a tight L^2 floor.

Theorem 4 (Architectural-floor theorem)**.** *Let $p = \otimes_i p_i$ be a product measure, $u \in L^2(Z,p)$ with ANOVA decomposition equation 1, and let $\mathcal{F} \subseteq 2^{[d]}$ be non-empty and downward closed ($S \in \mathcal{F}$ and $T \subset S$ imply $T \in \mathcal{F}$; for a general family apply the theorem to its downward closure, which spans the same class). Assume* $\mathrm{Var}(u) > 0$*. For any $v \in \mathcal{H}_{\mathcal{F}}$, the relative-$L^2$ error satisfies*

$$\frac{\|u - v\|^2}{\|u - \mathbb{E}[u]\|^2} \;\geq\; \sum_{S \in \mathcal{F}^c \setminus \{\varnothing\}} S_S(u), \tag{5}$$

with equality if and only if $v = P_{\mathcal{F}} u$, the L^2-orthogonal projection of u onto $\mathcal{H}_{\mathcal{F}}$.

The proof is elementary Hilbert-space geometry (orthogonal projection plus the Pythagorean identity) and is given in Appendix B.

The mathematics is elementary; the non-triviality in the neural setting is the operational reading. A single pilot training fixes the Sobol indices, which fix the floor for *every* architecturally restricted neural function class one might consider in place of the pilot. Restricted architectures whose forbidden subsets contain large Sobol mass can be ruled out before they are trained; the floor is achievable by the L^2-orthogonal projection $P_{\mathcal{F}} u$, so the bound is tight rather than asymptotic.

Corollary 5 (Floors for 2D Helmholtz)**.** *For the 2D Helmholtz target on (x, y, k), the architectural floors of the additive and pair-additive classes are*

$$\frac{\|u - v\|}{\|u - \mathbb{E}[u]\|} \;\geq\; \begin{cases} \sqrt{S_{xk} + S_{yk} + S_{xyk}} & \text{if } v \in \mathcal{H}_{\text{add}}, \\ \sqrt{S_{xyk}} & \text{if } v \in \mathcal{H}_{\text{pair}}. \end{cases}$$

For the trained network on our manufactured-reference problem, the Saltelli-basis indices of Table 1 column (b) — $S_{xk} = 0.073$, $S_{yk} = 0.073$, $S_{xyk} = 0.426$ — give floors $\sqrt{0.572} \approx 0.756$ (additive) and $\sqrt{0.426} \approx 0.653$ (pair-additive); the verification table (Table 2) evaluates the same floors on its binned-ANOVA basis, where they read 0.7574 and 0.6558 — the two bases differ by 0.004 on the triplet and are labelled wherever they appear.

Remark 6 (Scope; reference-dependence)**.** *The theorem is a statement about L^2 best approximations within a function class, not about gradient-descent training: no optimiser, compute budget, or hyperparameter choice can produce a function in $\mathcal{H}_\mathcal{F}$ with rel-L^2 below the floor, but achieving the floor in practice requires an expressive enough restricted architecture and a training procedure that converges to a point near the projection $P_\mathcal{F}u$. Sobol indices are also properties of a specific (u, p) target – different forcings or different input measures yield different Sobol indices and hence different floors. Sections 5–6 and Appendices F–G report how the indices, and hence the floors, move across five manufactured-reference variants and FD references on three forcings; floor tightness is tested on the headline target only.*

3.3 Finite-sample estimation

Three sources of error enter the operational floor prediction. (1) The Saltelli estimator $\hat{S}_S$ has CLT-rate $O_p(1/\sqrt{N})$ per index; by first-order Taylor expansion of the square root (valid away from a zero floor) the implied floor-prediction error is approximately $(2\sqrt{\sum S_S})^{-1} \sum |\varepsilon_S|$, ≈ 0.007 at the bootstrap run's $N = 2 \times 10^5$ (smaller at the headline $N = 10^6$), with bootstrap 95% half-width ± 0.009 on S_{xyk} (std $= 0.0044$). (2) The trained u_θ differs from the target u. Because $I - P_\mathcal{F}$ is an orthogonal projection, $\big|\, \|u - P_\mathcal{F}u\| - \|u_\theta - P_\mathcal{F}u_\theta\| \,\big| \le \|u - u_\theta\|$, so the unnormalised floor transfers from pilot to target up to the pilot's own error, and the normalised floor up to that error relative to $\|u - \mathbb{E}u\|$ plus the change in variance; this is the same order as the LC-PINN approximation error (10^{-3} to 10^{-2} in our experiments, small relative to the headline problem's architectural floors, all ≥ 0.65). (3) The Fourier HDMR estimator's standard readout overstates the triplet by $+0.36$ before purification and $+0.14$ after it, although the fitted function itself is within 0.02 (Section 7); we use MC-Sobol on the trained network as the headline reference.

3.4 Triplet purification by Möbius inversion

In the neural estimator, the fitted subset MLPs $\hat{f}_S^\theta$ absorb a mix of their target u_S and lower-order leakage. Naive subtraction $\hat{f}_S^\theta - \sum_{T \subsetneq S} \hat{f}_T^\theta$ does not eliminate leakage because the $\hat{f}_T^\theta$ themselves carry approximation error. We instead apply Möbius inversion on the subset lattice to the fitted triplet directly. Let $\hat{E}_T := \widehat{\mathbb{E}}[\hat{f}_S^\theta \mid z_T]$ be the empirical conditional expectation of the fitted triplet on subset $T \subseteq S$ (via B-bin discretisation per coordinate). Then for $S = \{i, j, k\}$,

$$\hat{f}_S^{\text{pure}} = \hat{f}_S^\theta - \hat{E}_{ij} - \hat{E}_{ik} - \hat{E}_{jk} + \hat{E}_i + \hat{E}_j + \hat{E}_k - \hat{E}_\varnothing, \tag{6}$$

with signs $(-1)^{|S \setminus T|}$ for $T \subseteq S$. With exact conditional expectations the purified term is the top-order ANOVA component of the fitted term; with binned estimates the correction is approximate, and its finite-B behaviour is measured rather than assumed. Empirically (Section 7): over six seeded fits on 2D Helmholtz the order-3 estimate is 0.781 ± 0.025 without purification and 0.567 ± 0.023 with it, against the Monte-Carlo reference 0.426. Purification removes about three fifths of the bias; the $+0.14$ residual comes from the binned, self-normalised readout rather than the fit, whose exact ANOVA is within 0.02 of the reference.

Under a non-product measure or a weighted inner product the decomposition is no longer unique (Hooker, 2007); promoting the identifying constraints to hyperparameters yields a weighted-ANOVA family whose indices and floors become functions of the weights. The construction, and an empirical probe along the orthogonality-weight axis, are in Appendix C (Remark 7) and Appendix L.

4 Estimators

This paper uses three estimators, and we report each one separately.

(a) Saltelli MC-Sobol on the analytic reference (when one exists): the ground truth for what the PDE solution's ANOVA structure is.

(b) Saltelli MC-Sobol on the trained network at $N = 10^6$: consistent for the network's own indices (all estimates are consistent, with $O(1/N)$ bias from the subtracted mean and the ratio); the reference for every claim about what the network learned. First-order and closed pair variances use

Table 1: Three estimators on 2D Helmholtz with parametric wavenumber. (a) Exact indices of the analytic reference by tensor-product Gauss–Legendre quadrature (converged to five decimals; a five-seed Saltelli mean at $N = 10^6$ agrees within 0.0012). (b) Saltelli MC at $N = 10^6$ on the trained network (seed-0 checkpoint). They agree to 0.002 on every index (largest gap 0.0024), within the Monte-Carlo error of column (b). (c) Fourier-HDMR, mean of a 50-fit battery purified at $B = 10$ (triplet 0.560 ± 0.019): it agrees on the x, y mains and the mixed pairs (within 0.006) but underestimates S_k and S_{xy}, which compensates its triplet bias under sum-to-one normalisation; the batteries characterised in Section 7 give 0.567 ± 0.023 and 0.569 ± 0.013 (Appendix D). All columns are normalised on the same joint measure; in (b) S_{xyk} is the residual $1 - \sum$ of the other six indices.

Index	(a) analytic MC	(b) network MC	(c) Fourier HDMR
S_x	0.040	0.038	0.041
S_y	0.040	0.039	0.045
S_k	0.074	0.073	0.030
S_{xy}	0.279	0.278	0.171
S_{xk}	0.071	0.073	0.078
S_{yk}	0.071	0.073	0.076
S_{xyk}	**0.427**	**0.426**	0.560
max $\lvert$(a) − (b)$\rvert$		0.002	
max $\lvert$(b) − (c)$\rvert$		0.134	

the correlation estimator; total indices the mean-square-difference estimator; at $d = 3$ the triplet variance is the exact residual.

(c) The Fourier-HDMR estimator: per-subset MLPs $f_S^\theta : \mathbb{R}^{|S|} \to \mathbb{R}$ for $|S| \le L_{\max}$, shared Fourier-feature encoding at $L = 6$ octaves $\omega_\ell = \pi 2^\ell$, hidden width 64, depth 2; fitted jointly to reconstruct the network output on N joint samples by AdamW; the highest-order term is purified by binned inclusion–exclusion (Algorithm 1, Appendix A) at $B = 10$ bins per coordinate (the default, used for every headline estimator number); the resolution sweep of Section 7 deliberately varies B and labels each value. We use $L_{\max} = 3$ for the headline $d = 3$ problem (order ≤ 2 for the qualitative footprints of Table 5) and $L_{\max} = 2$ for $d = 4$; for the tabular cases (California Housing $d = 8$, Sobol-G $d = 20$) $L_{\max} = 2$, chosen as the largest order whose subset count keeps the joint fit within the pilot network's own training budget; Appendix E reports cost scaling with d and $L_{\max}$.

Seeding. Network initialisation and data-loader order are seeded (`torch`, `numpy`); two unseeded refits of the same configuration differ by 0.04 on S_{xyk}. This is larger than several effects discussed below, so we use no unpaired, unseeded comparisons. Estimator intervals are means ± sd over seeded refits; comparisons across settings are paired on the same fitted model unless stated.

5 Sobol indices of the trained network

The 2D Helmholtz family with wavenumber $k \in [1, 10]$ on $[0, 1]^2$ has the manufactured reference $u_{\text{ref}} = \sin(\pi x)\sin(\pi y)\cos(kx)\cos(ky)$. Table 1 compares the three estimators on every Sobol index of the trained loss-conditional PINN.

Columns (a) and (b) give the main result. The trained network reproduces the reference's interaction structure to within Monte-Carlo noise, including the three-way term $S_{xyk} = 0.426$, which carries 43% of the variance. The agreement holds across training — two further independently trained checkpoints give $S_{xyk} = 0.4222$ and 0.4256 at the same $N = 10^6$, and every index of all three lies within 0.0054 of column (a) — and across architectures: on three physics-informed DeepONets for the same family (branch–trunk operator surrogates, mean rel-L^2 0.14–0.30 against 10^{-3}–10^{-2} for the LC-PINN) the decomposition applies unchanged and recovers the same structure, the triplet (0.405–0.432) dominant in all three, with per-index errors up to 0.022 — roughly 4 to 9 times the LC-PINN's, though not monotone in fidelity across the three DeepONets and far smaller than their 14–30% function-space error (Appendix J). Paired on identical sample matrices, the LC-PINN's indices differ from the reference's by at most 0.0006 on every index (ten seeds, 95% intervals of

Table 2: Architectural floors on the 2D Helmholtz analytic reference. Predicted floor $= \sqrt{\sum_{S \in \mathcal{F}^c} S_S}$; measured = rel-$L^2$ of the L^2 projection onto the restricted class. Evaluation basis: this table's indices come from a binned ANOVA decomposition at 40 bins per axis, on which $S_{xyk} = 0.4301$; the Saltelli estimate is 0.426 on the trained network and the exact value 0.427 on the analytic reference (Table 1) — the bases differ by about 0.004 on this index, and basis and object are labelled wherever these values appear. The predicted column uses the 40-bin indices; the measured column evaluates the binned projection on 10^6 fresh samples, so the two differ by up to 0.002 through bin discretisation and sampling.

Restricted class	Predicted floor	Measured rel-L^2
Constant (no inputs)	1.0000	1.0000
Mains only	0.9221	0.9204
Additive $f_{xy} + f_k$	0.7574	0.7554
All pairs (no triplet)	0.6558	0.6558
Full order-3	0.0000	0.0000

width below 10^{-4}), a test whose noise scales with the surrogate error rather than with $\mathrm{Var}(u)$. Column (c) is the fast estimator; we do not use it as evidence for any claim in this section or the next.

Uncertainty. Bootstrap (500 resamples, $N = 2 \times 10^5$): $S_{xyk} = 0.4251$, 95% CI $[0.4159, 0.4337]$. MC convergence: the index moves by 0.0007 from $N = 10^5$ to 10^6 (0.4265, 0.4250, 0.4258). Checkpoint comparisons require matched budgets: at $N = 3 \times 10^4$ seed 1 reads 0.4389, an MC artefact that vanishes at $N = 10^6$ (0.4222).

Robustness to the reference choice. Five manufactured variants of the reference (Appendix F, Table 4) give $S_{xyk} \in [0.171, 0.425]$, the triplet dominant in four of five; the headline variant is the variant-wise maximum and is used because it produces the most demanding floor. Independent finite-difference references on three forcings and three grids (Appendix G) span $S_{xyk} \in [0.04, 0.94]$, mesh-converged and confirmed by a Chebyshev pseudospectral re-solution. Sobol indices are properties of the specific target, as Remark 6 states, and the floor applies pointwise to the target the practitioner intends.

Other families. Table 5 gives qualitative footprints, mostly from the fast estimator. The decomposition recovers different footprints across four PDE families and two tabular datasets (Appendix H): Helmholtz concentrates variance in the spatial–parameter triplet, heat balances mains, wave concentrates in time, California Housing is near-purely additive (the three dominant mains sum to 0.92, all eight to 0.97; each of the 28 pairs is below 0.006).

6 Verification of the predicted floors

Theorem 4 converts the measured indices into predictions. Table 2 checks them against the L^2 projection of the analytic reference, evaluated on 10^6 samples. This check verifies the identity numerically. The measured column is itself an L^2 projection computed from the same decomposition, so the check validates the estimation pipeline, not the operational claim; the operational evidence is the trained restricted networks below, which carry the optimisation gap the theorem does not cover.

Trained restricted networks. Training from scratch adds the optimisation gap the theorem does not cover. We train all seven classes (six restricted, plus the full model) on the headline target — each a sum of per-subset Fourier-feature MLPs over its allowed subsets, fitted by regression on fresh samples of the target (five seeds, 15 000 steps; Appendix I) — and read their floors off the pilot's Saltelli indices (Table 1, column (b)). Every restricted class lands within 0.003 of its predicted floor (largest gap 0.0021); against the floor computed from the target's own indices the signed deviations lie in $[-0.0002, +0.0007]$, inside the Monte-Carlo error of a 10^5-point validation estimate, so values slightly below a predicted floor reflect index estimation, not a violation. The full-order class, floor zero, stops at 0.015. Regression on the target is not available to a PINN practitioner, so we repeat the experiment distilling each class from the pilot network instead (three seeds): every restricted

class reaches the same error to four decimals, and the full class 0.017. The difference is the pilot's own error, as expected for a pilot this accurate, and the setting a practitioner can run. The ordering is reproduced; most pairs are nested and ordered by construction, and the informative comparison is $f_{xy} + f_k$ (0.754) against the two classes $f_{xk} + f_y$, $f_{yk} + f_x$ (0.881), whose own floors are equal by the target's $x \leftrightarrow y$ symmetry. On a half-frequency variant and a near-additive variant with floors 0.085 and 0.073, regression again lands every class within 0.001 of its exact floor (Table 7). Physics-informed training respects every floor but does not approach it, since minimising the residual within a class that cannot represent the solution does not give the L^2 projection: 0.104 and 0.142 against 0.085 and 0.073 in the small-floor case, 0.84–1.00 on the other two targets; floors order these classes with mean Kendall τ of 0.83, 0.87 and 0.41, while the unrestricted model does not converge at this budget.

Comparison with attribution methods. On the same network, normalised first-order shares are: Integrated Gradients $(0.35, 0.32, 0.34)$, gradient saliency $(0.40, 0.38, 0.22)$, permutation importance $(0.36, 0.36, 0.29)$, MC-Sobol first-order $(0.25, 0.26, 0.49)$. These disagree with one another and, being first-order, cannot register that 85% of the variance lies in interactions. Quantities built for interactions can: total-order indices from the same Saltelli run are $S^T = (0.815, 0.813, 0.642)$ against first-order $(0.038, 0.039, 0.073)$, exposing the interaction mass per input, as Shapley effects (Owen, 2014; Song et al., 2016) and interaction attributions (Sundararajan et al., 2020; Janizek et al., 2021) and interaction detection from trained networks (Tsang et al., 2018b) would. The full ANOVA also identifies which subsets carry that mass, and the floor computation needs this information.

A tabular test. On a conditional MLP for California Housing (MedInc as λ), the order-2 decomposition is near-purely additive (pair mass 0.034), which puts an additive surrogate's floor at $\sqrt{0.034} \approx 0.18$ rel-L^2 of the pilot. Trained additive surrogates reach 0.236 ± 0.001 (three seeds), above the floor; an order-2 control at the same budget stops at 0.155 ± 0.012, a fitting error comparable to the floor itself, so the check cannot establish tightness. We report it as an illustration: the features are dependent, the indices come from the fast estimator (order ≤ 2), and the decomposition is fitted on the empirical joint rather than the product measure Theorem 4 assumes (Remark 7; Appendix H).

7 A FAST ESTIMATOR: FITTED FUNCTION AND READOUT

At this scale the Fourier-HDMR estimator is not cheaper than Saltelli MC. On the same hardware and surrogate, Saltelli MC at $N = 10^6$ takes 0.6 s and one HDMR fit 320 s (Appendix E). What the estimator offers is different: it returns the component functions f_S themselves, not only their variances, and it needs only samples of the map rather than a model that can be queried anywhere. Its use case is a surrogate or solver that is expensive to evaluate, or a map known only through data. This section locates, on the headline case and under seeded, paired comparisons, where the estimator's error arises.

Bias of the standard readout. The estimator reports each index as a fitted term's variance divided by the sum of all term variances. Over six seeded fits this readout gives the triplet as 0.781 ± 0.025 before purification and 0.567 ± 0.023 after, against the Monte-Carlo 0.426: purification removes about three fifths of the bias, and the $+0.14$ residual is six times the seed spread.

Where the bias arises. The fitted reconstruction $\hat{u} = \sum_S f_S$ is itself a queryable function, so its exact ANOVA can be computed by Saltelli MC on $\hat{u}$ ($N = 10^6$), with no binning and no self-normalisation. On three seeded fits the readout gives 0.566 ± 0.018 $(+0.140)$ while the exact ANOVA of the same fits gives 0.406 ± 0.016 (-0.020), every index of the fit within 0.06 of the reference. The exact read improves with the fit (0.389, 0.406, 0.422 at validation rel-RMSE 0.52, 0.44, 0.36), while the readout stays 0.12–0.15 high on every seed; tripling the training data $(+0.151)$ or doubling width and epochs $(+0.140)$ leaves it unchanged. Under-fitting explains the small exact-read error. It does not explain the readout's $+0.14$. Exact integration of the fitted function removes this bias.

An orthogonality check. An orthogonal decomposition satisfies $\sum_S \mathrm{Var}(f_S) = \mathrm{Var}(\hat{u})$. The fitted terms violate it by a factor 2.2–4.3 at $B = 10$ (per-seed $\sum_S \mathrm{Var}(f_S)/\mathrm{Var}(u_\theta)$, the purified terms against the network's output variance, min–max over the six fits; Figure 1 plots $\sum_S \mathrm{Var}(f_S)/\mathrm{Var}(\sum_S f_S)$, which averages 4.42 for the unpurified terms); under the MC conven-

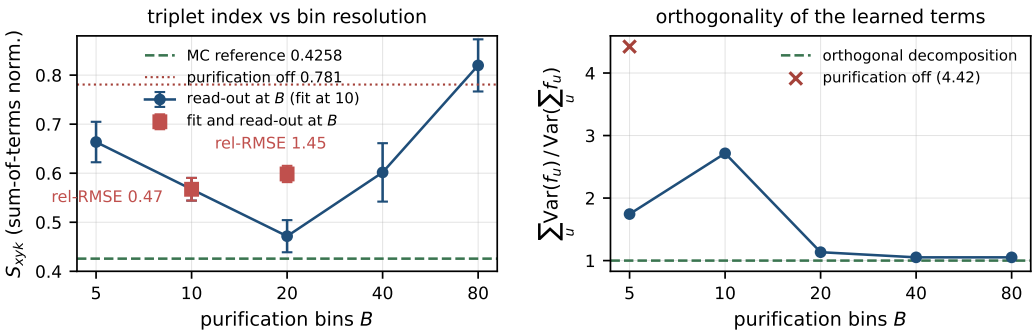

Figure 1: Six seeded fits; error bars are seed sd. Left: S_{xyk} against purification resolution B, read out on models fitted at B=10 (blue), with the MC reference (dashed) and unpurified value (dotted); red squares refit with B threaded through training — at B=20 the index moves to 0.598 ± 0.017 while reconstruction rel-RMSE rises from 0.47 to 1.45. Right: the orthogonality ratio $\sum_S \mathrm{Var}(f_S)/\mathrm{Var}(\sum_S f_S)$, which is 1 for an orthogonal decomposition.

tion (divide by $\mathrm{Var}(\hat{u})$) the triplet term alone scores 1.95 ± 0.49, which exceeds the variance of the function. The reported "captured fraction" is 1.000 for every variant, so the violation is invisible in the normalised output; the readout attributes this non-orthogonality to individual terms. The invariant can be checked with one line of code.

Effect of the binning resolution. Sweeping the purification resolution at read-out traces the U-shaped curve of Figure 1: coarse bins under-correct (B=5: +0.24), fine bins leave too few samples per conditioning cell (B=80: 4.7 samples per 2-D cell), and the minimum at $B = 20$ reads +0.046. Threading $B = 20$ through training returns the bias to +0.17 and raises the validation rel-RMSE from 0.47 to 1.45; per-axis bin edges are slightly worse at every resolution. Neither refit resolution tested ($B = 10, 20$) removes the bias; exact integration does.

Replication. The full battery on a second, independently trained surrogate of the same family (pointwise 1.1% from the first; its own MC reference 0.4248 at $N = 2 \times 10^5$, 0.4256 at $N = 10^6$) reproduces every number within seed spread: bias +0.36 unpurified, +0.14 purified, the same U-curve with minimum +0.048, the same broken refit at $B = 20$ (rel-RMSE 1.38).

Remaining uses of the estimator. Read exactly, the estimator supplies component functions from samples. On Sobol-G ($d = 20$, closed-form indices) its mains match the closed form to 0.007 at worst, though a tanh baseline that reconstructs worse is closer (0.002), so reconstruction quality does not order index accuracy; at $d = 3$ polynomial chaos at degree 12 is within 0.012 of the triplet (Appendices E, K; ablation in Table 3).

8 Discussion

One pilot training and one Sobol decomposition give a lower bound on the error of every restricted architecture. Restricted networks trained by regression reach these bounds; physics-informed training leaves them loose. On the headline problem this predicts the floor of every restricted architecture trained by regression, and their ranking. For the HDMR estimator, the fit is correct, and the error arises because the readout uses non-orthogonal terms. Comparing with an expensive reference on one case, or testing the variance-sum invariant, detects this error.

Limitations. (1) Sobol indices are computed on the trained network, not the true solution; Table 1 controls this gap (0.002, unpaired; 0.0006 paired) for the headline case; DeepONets with far larger error still match to 0.022, so index agreement is necessary but not sufficient evidence of fidelity. (2) Indices are target-specific; forcings move S_{xyk} across $[0.04, 0.94]$ (Appendix G), and every floor is a statement about one target. (3) Restricted networks trained by regression reach their floors on three targets; physics-informed ones respect but do not approach them, and closing that optimisation gap is open. (4) The readout diagnosis covers one estimator on two surrogates.

REFERENCES

Rishabh Agarwal, Levi Melnick, Nicholas Frosst, Xuezhou Zhang, Ben Lengerich, Rich Caruana, and Geoffrey E. Hinton. Neural additive models: Interpretable machine learning with neural nets. In *Advances in Neural Information Processing Systems*, 2021.

R. E. Caflisch, W. Morokoff, and A. B. Owen. Valuation of mortgage-backed securities using Brownian bridges to reduce effective dimension. *Journal of Computational Finance*, 1(1):27–46, 1997.

G. Chastaing, F. Gamboa, and C. Prieur. Generalized Hoeffding-Sobol decomposition for dependent variables — application to sensitivity analysis. *Electronic Journal of Statistics*, 6:2420–2448, 2012.

Alexey Dosovitskiy and Josip Djolonga. You only train once: Loss-conditional training of deep networks. In *International Conference on Learning Representations*, 2020.

Joseph Hart and Pierre A. Gremaud. An approximation theoretic perspective of Sobol' indices with dependent variables. *International Journal for Uncertainty Quantification*, 8(6), 2018.

Wassily Hoeffding. A class of statistics with asymptotically normal distribution. *The Annals of Mathematical Statistics*, 19(3):293–325, 1948.

Giles Hooker. Discovering additive structure in black box functions. In *Proceedings of the Tenth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pp. 575–580, 2004.

Giles Hooker. Generalized functional ANOVA diagnostics for high-dimensional functions of dependent variables. *Journal of Computational and Graphical Statistics*, 16(3):709–732, 2007.

Joseph D. Janizek, Pascal Sturmfels, and Su-In Lee. Explaining explanations: Axiomatic feature interactions for deep networks. *Journal of Machine Learning Research*, 22(104):1–54, 2021.

F. Y. Kuo, I. H. Sloan, G. W. Wasilkowski, and H. Woźniakowski. On decompositions of multivariate functions. *Mathematics of Computation*, 79:953–966, 2010.

B. Lengerich, S. Tan, C.-H. Chang, G. Hooker, and R. Caruana. Purifying interaction effects with the functional ANOVA: an efficient algorithm for recovering identifiable additive models. In *International Conference on Artificial Intelligence and Statistics (AISTATS)*, 2020.

G. Li and H. Rabitz. General formulation of HDMR component functions with independent and correlated variables. *Journal of Mathematical Chemistry*, 50:99–130, 2012.

Zongyi Li, Nikola Kovachki, Kamyar Azizzadenesheli, Burigede Liu, Kaushik Bhattacharya, Andrew Stuart, and Anima Anandkumar. Fourier neural operator for parametric partial differential equations. In *International Conference on Learning Representations*, 2021.

Yin Lou, Rich Caruana, Johannes Gehrke, and Giles Hooker. Accurate intelligible models with pairwise interactions. In *Proc. ACM SIGKDD*, pp. 623–631, 2013.

Lu Lu, Pengzhan Jin, Guofei Pang, Zhongqiang Zhang, and George Em Karniadakis. Learning nonlinear operators via DeepONet. *Nature Machine Intelligence*, 3(3):218–229, 2021.

Scott M. Lundberg and Su-In Lee. A unified approach to interpreting model predictions. In *Advances in Neural Information Processing Systems*, volume 30, pp. 4765–4774, 2017.

Art B. Owen. Sobol' indices and Shapley value. *SIAM/ASA Journal on Uncertainty Quantification*, 2(1):245–251, 2014.

Herschel Rabitz and Ömer F. Aliş. General foundations of high-dimensional model representations. *Journal of Mathematical Chemistry*, 25(2-3):197–233, 1999.

Maziar Raissi, Paris Perdikaris, and George Em Karniadakis. Physics-informed neural networks: A deep learning framework for solving forward and inverse problems involving nonlinear partial differential equations. *Journal of Computational Physics*, 378:686–707, 2019.

Andrea Saltelli, Paola Annoni, Ivano Azzini, Francesca Campolongo, Marco Ratto, and Stefano Tarantola. Variance based sensitivity analysis of model output. *Computer Physics Communications*, 181(2):259–270, 2010.

I. M. Sobol. Sensitivity estimates for nonlinear mathematical models. *Mathematical Modelling and Computational Experiments*, 1(4):407–414, 1993.

I. M. Sobol', S. Tarantola, D. Gatelli, S. Kucherenko, and W. Mauntz. Estimating the approximation error when fixing unessential factors in global sensitivity analysis. *Reliability Engineering & System Safety*, 92:957–960, 2007.

Ilya M. Sobol'. Global sensitivity indices for nonlinear mathematical models and their Monte Carlo estimates. *Mathematics and Computers in Simulation*, 55(1–3):271–280, 2001.

Eunhye Song, Barry L. Nelson, and Jeremy Staum. Shapley effects for global sensitivity analysis: Theory and computation. *SIAM/ASA Journal on Uncertainty Quantification*, 4(1):1060–1083, 2016.

Bruno Sudret. Global sensitivity analysis using polynomial chaos expansions. *Reliability Engineering & System Safety*, 93(7):964–979, 2008.

Mukund Sundararajan, Ankur Taly, and Qiqi Yan. Axiomatic attribution for deep networks. In *International Conference on Machine Learning*, pp. 3319–3328, 2017.

Mukund Sundararajan, Kedar Dhamdhere, and Ashish Agarwal. The Shapley Taylor interaction index. In *International Conference on Machine Learning*, 2020.

Matthew Tancik, Pratul Srinivasan, Ben Mildenhall, Sara Fridovich-Keil, Nithin Raghavan, Utkarsh Singhal, Ravi Ramamoorthi, Jonathan Barron, and Ren Ng. Fourier features let networks learn high frequency functions in low dimensional domains. In *Advances in Neural Information Processing Systems*, volume 33, pp. 7537–7547, 2020.

M. Tsang, H. Liu, S. Purushotham, P. Murali, and Y. Liu. Neural interaction transparency (NIT): disentangling learned interactions for improved interpretability. In *Advances in Neural Information Processing Systems*, volume 31, 2018a.

Michael Tsang, Dehua Cheng, and Yan Liu. Detecting statistical interactions from neural network weights. In *International Conference on Learning Representations*, 2018b.

Dongbin Xiu and George Em Karniadakis. The Wiener–Askey polynomial chaos for stochastic differential equations. *SIAM Journal on Scientific Computing*, 24(2):619–644, 2002.

Zebin Yang, Aijun Zhang, and Agus Sudjianto. GAMI-Net: An explainable neural network based on generalized additive models with structured interactions. *Pattern Recognition*, 120:108192, 2021.

USE OF GENERATIVE AI

In this work, we used generative AI tools (large language models) to polish and edit the writing, to draft and revise sections of the paper, and to search for and summarise related work. The authors reviewed all AI-assisted text and take full responsibility for the contents of the paper.

A THE PURIFICATION ALGORITHM

Algorithm 1 gives the binned inclusion–exclusion (Möbius-inversion) purification applied to the highest-order fitted term; $B = 10$ bins per coordinate is the default, and Section 7 characterises the estimate's dependence on B.

Algorithm 1 Möbius purification of the top-order HDMR term. The inclusion–exclusion step is exact; the conditional expectations it subtracts are estimated by binning, so the correction is exact only in the limit of fine bins with many samples per cell.

Require: fitted top-order subnet $\hat{f}_S^\theta$ with $S = \{1, 2, 3\}$; joint samples $z^{(1)}, \dots, z^{(N)} \in \mathbb{R}^3$; bins per coordinate B

Ensure: purified term $t^{\text{pure}} \in \mathbb{R}^N$ with $\widehat{\mathbb{E}}[\, t^{\text{pure}} \mid z_i, z_j \,] \approx 0$ for all pairs $\{i, j\} \subsetneq S$ (exact when the 3-D cell counts are balanced within every 2-D cell)

$t_n \leftarrow \hat{f}_S^\theta(z^{(n)})$ for $n = 1, \dots, N$
for $i = 1$ **to** 3 **do**
  partition $[\min_n z_i^{(n)}, \max_n z_i^{(n)}]$ into B equal-width bins
  $b_i(n) \leftarrow$ index of the bin containing $z_i^{(n)}$
end for
$\bar{t} \leftarrow \frac{1}{N} \sum_n t_n$ {0-dimensional conditioning}
for $i = 1$ **to** 3 **do**
  $E_i[\beta] \leftarrow \text{mean}\{t_n : b_i(n) = \beta\}$ {1-D, B cells}
end for
for each pair (i, j) with $i < j$ **do**
  $E_{ij}[\beta, \gamma] \leftarrow \text{mean}\{t_n : b_i(n) = \beta,\ b_j(n) = \gamma\}$ {2-D, B^2 cells}
end for
for $n = 1$ **to** N **do**
  $t_n^{\text{pure}} \leftarrow t_n - E_{12}[b_1(n), b_2(n)] - E_{13}[b_1(n), b_3(n)] - E_{23}[b_2(n), b_3(n)] + E_1[b_1(n)] + E_2[b_2(n)] + E_3[b_3(n)] - \bar{t}$
end for
return t^{pure}

B Proof of Theorem 4

Proof. The ANOVA components $\{u_S\}_{S \subseteq [d]}$ form a finite orthogonal family in $L^2(Z, p)$. Set $P_{\mathcal{F}} u = \sum_{S \in \mathcal{F}} u_S \in \mathcal{H}_{\mathcal{F}}$. We show that $r = u - P_{\mathcal{F}} u = \sum_{S \notin \mathcal{F}} u_S$ is orthogonal to every $v = \sum_{T \in \mathcal{F}} v_T \in \mathcal{H}_{\mathcal{F}}$. Take $S \notin \mathcal{F}$ and $T \in \mathcal{F}$. Because $\mathcal{F}$ is downward closed, $S \not\subseteq T$ (otherwise $S \in \mathcal{F}$), so there is an index $i \in S \setminus T$. Under the product measure, Fubini's theorem gives $\langle u_S, v_T \rangle = \int v_T(z_T) \big(\int u_S \, dp_i \big) \, dp_{-i} = 0$, since v_T does not depend on z_i and u_S integrates to zero in z_i (Definition 1); both functions are in L^2, so the integral is well defined. Hence $r \perp \mathcal{H}_{\mathcal{F}}$, and since $\mathcal{H}_{\mathcal{F}}$ is a finite sum of closed subspaces $\{v_T(z_T)\} \cong L^2(p_T)$, the Pythagorean identity gives, for any $v \in \mathcal{H}_{\mathcal{F}}$,

$$\|u - v\|^2 \;=\; \|r\|^2 + \|P_{\mathcal{F}} u - v\|^2 \;\geq\; \|r\|^2,$$

with equality if and only if $v = P_{\mathcal{F}} u$. By orthogonality of the components, $\|r\|^2 = \sum_{S \notin \mathcal{F}} V_S = \text{Var}(u) \sum_{S \in \mathcal{F}^c} S_S$, where $\varnothing \in \mathcal{F}$ because $\mathcal{F}$ is non-empty and downward closed, so the sum runs over non-empty subsets. Dividing by $\|u - \mathbb{E}[u]\|^2 = \text{Var}(u) > 0$ gives equation 5. □

C Weighted-ANOVA construction

Remark 7 (Weight-dependent generalised ANOVA). *Theorem 4 relies on the classical L^2-orthogonality of ANOVA components under the product measure $p = \otimes_i p_i$. Under a non-product measure, or under a weighted inner product $\langle u, v \rangle_w = \int u(z) v(z) w(z) dp(z)$ in which w does not factorise, the recursive decomposition remains well defined but its components are no longer orthogonal; Hooker (2007) restores a unique hierarchically orthogonal variant by imposing identifying constraints, and other constraints yield other admissible decompositions $\{u_S(z_S)\}$, all satisfying $u = \sum_S u_S$. A natural extension promotes those constraints to hyperparameters w and considers the family*

$$L(\{u_S\}; w) = \Big\| u - \sum_S u_S \Big\|^2 + \sum_{S \neq T} w_{ST}^{\text{orth}} |\langle u_S, u_T \rangle|^2 + \sum_S w_S^{\text{grad}} \|\nabla_S u_S\|^2,$$

This functional is not convex, and pairwise orthogonality alone does not identify the classical decomposition without the zero-mean constraints: placing all of u in the top-order term is exactly orthogonal. Appendix L shows this collapse empirically ($S_{xyk} \to 1$ for $w^{\text{orth}} \geq 5$), so the penalty is a probe of stability, not a route to the classical decomposition. The Sobol indices S_S and the architectural floor $\text{floor}(\mathcal{F}; w)$ then become functions of w. Whether the headline indices reported in Section 5 are stable across reasonable w, or whether different w values yield substantively different structural readings, is probed along the w^{orth} axis in Appendix L.

D Reproducibility

Per-subset MLPs: one subnetwork per input subset up to order $L_{\max}$ (seven subnets for $d = 3$), each a two-hidden-layer MLP of width 64 on Fourier features of its inputs; trained by AdamW on mean-squared reconstruction error (weight decay 10^{-6}, staged learning rates, 40 + 80 epochs, order-3 phase 120 where present) on 10^5 joint samples, validated on 3×10^4. Term variances are empirical variances of each subnet's output on validation samples after purification. `torch` and `numpy` are seeded per run: the network initialisation seed is set before model construction and a seeded generator drives the data-loader shuffle; six seeds per configuration in Section 7, five in Table 3. Three estimator batteries exist: an earlier 50-fit battery (Table 1 column (c), 0.560 ± 0.019) and the six- and five-fit batteries under this appendix's staged recipe; the staged recipe is the default configuration and the one Section 7 characterises. Purification: binned conditional-mean subtraction at $B = 10$ per axis unless stated. Wall time per fit $\approx$ 325 s (Apple-silicon backend); the full characterisation study is about 2.5 hours. The MC reference uses two Saltelli matrices with $N = 10^6$ rows; values at $N = 10^5, 3 \times 10^5, 10^6$: 0.4265, 0.4250, 0.4258. The MC convergence run is tagged with its generating checkpoint and pipeline for provenance.

E Cost scaling

Measured on one Apple-silicon machine with no other jobs running, on the seed-0 2D Helmholtz LC-PINN (medians of three runs): Saltelli MC-Sobol with all mains, pairs and the triplet ($8N$ network evaluations) takes 0.07 s at $N = 10^5$, 0.18 s at 3×10^5 and 0.60 s at 10^6; one Fourier-HDMR fit in the default configuration (order 3, $N = 10^5$, phases 40/80/120) takes 320 s, consistent with the 325 s recorded for the Section 7 battery. MC's triplet at $N = 10^5$ is already within 0.01 of its $N = 10^6$ value, against the purified estimator's $+0.14$. At a matched budget of 10^5 network evaluations ($N = 12\,500$, the HDMR fit's training-set size), ten MC runs give $S_{xyk} = 0.432 \pm 0.017$: noisier, but centred on the reference, where the estimator is centred 0.14 above it. The joint fit trains $\sum_{1 \leq \ell \leq L_{\max}} \binom{d}{\ell}$ subnets (7 at $d = 3$, $L_{\max} = 3$; 210 at $d = 20$, $L_{\max} = 2$, where one fit on Sobol-G took 823 s); MC with all pairs needs $N(2 + d + \binom{d}{2})$ evaluations, $212N$ at $d = 20$ — under 2 s at $N = 10^5$ extrapolated from the measured throughput (1.3×10^7 evaluations per second). MC therefore dominates on both cost and accuracy for cheap-to-evaluate surrogates; the estimator's case is an expensive map or one available only as samples (Section 7).

On Sobol-G ($d = 20$, four active inputs) the estimator reaches rel-RMSE 0.046 against 0.065 for a matched tanh baseline; its main-effect indices match the closed-form values to 0.007 at worst (summed absolute error 0.008 over the 20 mains and 0.008 over the 190 pairs), while the tanh baseline's mains are within 0.002 at worst (summed 0.004). Table 3 gives the seeded estimator ablation on the headline surrogate.

F Manufactured-reference variants

$n = 5$ variants is a small designed sample; the claim it supports is structural-qualitative (the triplet-dominated regime is not an artefact of one manufactured choice), not distributional.

G Finite-difference and spectral references

Independent references by second-order FD on three grids (65^2, 129^2, 257^2) and three forcings; Saltelli at $N = 5 \times 10^5$, three seeds. F_1 and F_2 give $S_{xyk} \approx 0.94$; F_3 gives 0.04 with $S_{xy} \approx 0.87$:

Table 3: Estimator ablation on the trained 2D Helmholtz surrogate: mean ± sd over five seeded fits (initialisation and shuffle seeded; data sample fixed), purification at $B = 10$. The order-2 variant has no triplet term by construction. Without Fourier features the estimator is not merely worse — it is unstable: the tanh triplet estimate varies by ± 0.16 across seeds. Evaluation basis: indices normalised by the sum of term variances; validation rel-RMSE against the surrogate output. †Only the mains are trained; the pair and triplet subnets keep their initialisation, so the reconstruction sits at the mains-only floor (0.92) and its triplet index measures untrained output, not structure.

Variant	S_{xyk}	val rel-RMSE
Fourier, order 3 (full)	0.569 ± 0.013	0.479 ± 0.072
tanh, order 3 (no Fourier)	0.225 ± 0.160	0.602 ± 0.053
Fourier, order 2 (no triplet)	—	0.661 ± 0.001
Fourier, mains phase only†	0.131 ± 0.077	0.920 ± 0.001

Table 4: Saltelli MC-Sobol across five manufactured variants of the 2D Helmholtz reference, $N = 10^6$ each. The triplet dominates in four of five; the half-frequency variant shifts mass to the parameter main effect. One MC run per variant; S_y and S_{yk} are omitted for space, so rows do not sum to one. The baseline row is a separate run from Table 1's column (a) and agrees with it within MC error.

Variant	S_x	S_k	S_{xy}	S_{xk}	S_{xyk}
cos-cos baseline	0.039	0.074	0.281	0.073	**0.425**
sin-sin	0.054	0.144	0.226	0.086	**0.349**
cos-cos phase $\pi/4$	0.042	0.060	0.265	0.087	**0.415**
asymmetric cos-sin	0.060	0.085	0.253	0.105	**0.398**
half frequency $k/2$	0.078	**0.263**	0.163	0.124	0.171

mesh-converged (drift < 0.005 between the two finest grids) and confirmed to four decimals by a Chebyshev pseudospectral re-solution at 48 nodes per axis. The spread across forcings is the reference-dependence of Remark 6, quantified.

H Cross-family footprints

Table 5: Dominant Sobol terms across families. Helmholtz 2D (manufactured and FD) and the 3D gold values are Saltelli MC; the remaining rows are from the Fourier-HDMR estimator (order ≤ 2) and are qualitative, subject to Section 7's caveats (means over three seeds, headline forcing per family).

Target	d	Dominant terms
Schrödinger 1D	2	$S_x = 0.99$
Helmholtz 1D	2	$S_{xk} = 0.65$, $S_k = 0.20$, $S_x = 0.13$
Helmholtz 2D (manuf.)	3	$S_{xyk} = 0.43$, $S_{xy} = 0.28$
Helmholtz 2D (FD F_1)	3	$S_{xyk} = 0.94$
Helmholtz 2D (FD F_3)	3	$S_{xy} = 0.87$, $S_y = 0.09$
Helmholtz 3D	4	$S_k = 0.21$, $S_{xyzk} = 0.14$, $S_{xyz} = 0.12$
Heat 1D	3	$S_t = 0.35$, $S_\alpha = 0.31$, $S_x = 0.14$
Wave 1D	3	$S_t = 0.51$, $S_{tc} = 0.21$, $S_{xt} = 0.12$
California Housing	8	$S_{\mathrm{Lat}} = 0.33$, $S_{\mathrm{Lon}} = 0.30$, $S_{\mathrm{MedInc}} = 0.29$

On 3D Helmholtz the gold standard identifies a four-way interaction $S_{xyzk} = 0.137$ that order-2 truncation cannot represent, inflating its S_k to 0.35 against the gold 0.21 — a caution about truncation order, not about the decomposition. California Housing details: conditional MLP with MedInc

as λ; the order-2 Fourier-HDMR fit on the MLP output reaches val rel-RMSE 0.10; the three dominant mains are $S_{\mathrm{Lat}} = 0.33$, $S_{\mathrm{Lon}} = 0.30$, $S_{\mathrm{MedInc}} = 0.29$, all eight mains sum to 0.97, the 28 pairs to 0.03 with each below 0.006 (indices are normalised over the retained order-≤ 2 terms). Two qualifications: the features are dependent, so the fit is on the empirical joint distribution rather than the product measure of Definition 1 — these are variance shares of the fitted decomposition, and Theorem 4 applies only heuristically (the principled route for dependent inputs is the weighted construction of Remark 7 and Chastaing et al. (2012)); and the additive-surrogate prediction of Section 6 was tested by retraining the pilot with a fixed seed (its pair mass reproduces at 0.034) and fitting additive and order-2 surrogates to it (sums of per-subset MLPs, two hidden layers of width 64, Adam, 300 epochs, three seeds): additive 0.236 ± 0.001, order-2 0.155 ± 0.012 rel-L^2 against the pilot on held-out rows. Sklearn-diabetes (bmi as λ) recovers an interaction-richer footprint ($S_{s5} = 0.24$, $S_{\mathrm{bmi}} = 0.22$, $S_{\mathrm{bp}} = 0.13$) under the same qualifications.

I FLOOR-RANKING EXPERIMENT

Each restricted class is a bias plus one Fourier-feature MLP per allowed subset (6 frequency octaves per axis, two hidden layers of width 64, tanh), trained by Adam (learning rate 2×10^{-3}, cosine schedule, 15 000 steps, batch 4096) on fresh uniform samples of the analytic target at every step, on the same joint measure as the MC indices. Error is the centred rel-L^2 of Theorem 4 on 10^5 fixed validation points. Floors come from the pilot network's Saltelli indices (column (b) of Table 1); the target's exact indices (column (a)) move each floor by at most 0.0022.

Table 6: Predicted floors and trained errors on 2D Helmholtz: regression on the analytic target (five seeds) and distillation from the pilot network (three seeds), both scored against the analytic target.

Class	Predicted floor	Trained (mean)	sd	Distilled from pilot
constant	1.0000	1.0000	0	1.0000
mains only	0.9218	0.9210	$< 10^{-4}$	0.9210
$f_{xy} + f_k$	0.7560	0.7539	$< 10^{-4}$	0.7539
$f_{xk} + f_y$	0.8812	0.8812	$< 10^{-4}$	0.8812
$f_{yk} + f_x$	0.8815	0.8814	$< 10^{-4}$	0.8814
all pairs	0.6525	0.6531	0.0001	0.6531
full order 3	0.0000	0.0149	0.0006	0.0167

Table 7: Exact floors (Gauss–Legendre quadrature) and mean trained centred rel-L^2 on three manufactured targets, under regression on the target and physics-informed training (residual plus boundary loss, forcing by automatic differentiation, 15 000 Adam steps); three seeds (five for headline regression). Second column: mean Kendall τ between floors and trained errors over the six non-constant classes.

Target / training	τ	mains	$f_{xy}+f_k$	$f_{xk}+f_y$	$f_{yk}+f_x$	pairs	full
headline floor		0.920	0.754	0.881	0.881	0.653	0.000
regression	0.97	0.921	0.754	0.881	0.881	0.653	0.015
phys.-inf.	0.83	0.998	0.837	0.997	0.997	0.849	0.619
half-freq. floor		0.762	0.645	0.675	0.675	0.409	0.000
regression	0.97	0.762	0.645	0.676	0.675	0.410	0.008
phys.-inf.	0.87	0.972	0.936	0.965	0.966	0.951	0.906
near-add. floor		0.351	0.085	0.350	0.350	0.073	0.000
regression	0.97	0.351	0.085	0.350	0.350	0.073	0.005
phys.-inf.	0.41	0.760	0.104	0.792	0.791	0.142	0.113

J Operator-learning surrogate: physics-informed DeepONet

The decomposition requires only evaluations of a trained map $(x, \lambda) \mapsto u$, so it applies unchanged to operator-learning surrogates. Three physics-informed DeepONets (Lu et al., 2021) for the 2D Helmholtz family (branch network on k, trunk network on (x, y), 160 basis functions, 4×64 hidden layers each, trained on the residual loss with the same k sampling as the LC-PINN) were evaluated by Saltelli MC at $N = 10^6$ on the joint measure of Section 5. Each checkpoint reproduces its own logged per-k rel-L^2 exactly before its indices are reported. All three recover the reference's

Table 8: Saltelli indices of three physics-informed DeepONets (seeds 0–2) against the analytic reference (column (a) of Table 1). Mean rel-L^2 is the checkpoint's own error against the reference over 20 wavenumbers; $|\Delta|$ is computed from unrounded values.

Index	(a) analytic	seed 0	seed 1	seed 2
S_x	0.040	0.046	0.017	0.020
S_y	0.040	0.048	0.020	0.038
S_k	0.074	0.080	0.073	0.077
S_{xy}	0.279	0.270	0.276	0.266
S_{xk}	0.071	0.071	0.091	0.088
S_{yk}	0.071	0.079	0.091	0.090
S_{xyk}	0.427	0.405	0.432	0.421
mean rel-L^2	—	0.14	0.30	0.28
max $\|\Delta\|$ vs (a)	—	0.022	0.022	0.019

structure — the triplet is the largest index in each, followed by S_{xy}. Their per-index errors (0.019–0.022) are roughly 4 to 9 times the LC-PINN's (0.0024 on seed 0, 0.0054 across its three seeds) but do not order the three DeepONets by fidelity, and they are far smaller than the 14–30% function-space errors. Sobol indices are therefore a coarse fidelity signal: they certify which interactions a surrogate carries, and a function-space error metric remains necessary.

K Polynomial chaos comparison

Legendre PCE, from which Sobol indices follow in closed form (Sudret, 2008), on the trained-network output ($N_{\text{tr}} = N_{\text{va}} = 3 \times 10^4$): $S_{xyk} = 0.010$ at $P = 4$ (35 coefficients, 30.6% variance captured), 0.260 at $P = 8$ (165, 73%), 0.414 at $P = 12$ (455, 97%). LASSO-sparse PCE at $P = 12$ keeps ~ 280 active terms without improving the triplet.

L Weighted-ANOVA sensitivity

Table 9: Unpurified Fourier-HDMR indices versus orthogonality-penalty weight w^{orth} (3 seeds, 200 epochs, $N = 2 \times 10^5$).

w^{orth}	S_{xyk}	S_{xy}	rel-RMSE
0	0.582 ± 0.030	0.139 ± 0.021	0.31 ± 0.03
0.1	0.569 ± 0.018	0.158 ± 0.010	0.27 ± 0.01
1	0.629 ± 0.172	0.109 ± 0.059	2.8 ± 4.1
≥ 5	$\to 1.00$	$\to 0$	diverged

In the stable regime the indices are not strongly weight-dependent; large penalties break the fit rather than smoothly drifting it. The 2.8 ± 4.1 rel-RMSE at $w^{\text{orth}} = 1$ is driven by one diverged seed (per-seed: 0.39, 7.5, 0.38). The $w^{\text{orth}} = 0$ row is the unpurified estimator under this appendix's configuration (200 epochs, $N = 2 \times 10^5$, rel-RMSE 0.31), where the unpurified triplet reads $0.582 \pm$

0.030; the main text's 0.781 ± 0.025 comes from the staged recipe of Appendix D at rel-RMSE 0.47. The unpurified bias depends on the fit configuration — both values are far above the MC reference, and only the purified, guarded numbers of Section 7 are used for claims.

M RESTRICTED-NETWORK AND DIAGNOSTIC SETTINGS

Parametric 1D Schrödinger: $-u'' + V(x;\alpha)u = f$ on $(0,1)$, $V = \alpha(x - \frac{1}{2})^2$, $\alpha \sim \mathcal{U}[0.5, 10]$, manufactured $u^\star = \sin(\pi x)e^{-\alpha(x-1/2)^2/2}$. Parametric 1D Helmholtz: $u'' + k^2 u = f$ on $(0,1)$, $k \sim \mathcal{U}[1, 10]$, $u^\star = \sin(\pi x)\cos(kx)$. Restricted networks: one subnetwork per admitted subset, 26,114 parameters, 60,000 Adam steps, seed 0. Errors are uncentred rel-L^2; single training seed. Failure-mode probe: residual-input correlations $|\rho| < 0.021$ at $k \in \{1, 3.5, 7, 10\}$; per-k rel-$L^2 \in [0.006, 0.018]$.