SPARSE DEEP ADDITIVE MODEL WITH INTERACTIONS: ENHANCING INTERPRETABILITY AND PREDICTABILITY

Anonymous authors
Paper under double-blind review

ABSTRACT

Recent advances in deep learning highlight the need for personalized models that can learn from small or moderate samples, handle high-dimensional features, and remain interpretable. To address this challenge, we propose the Sparse Deep Additive Model with Interactions (SDAMI), a framework that combines sparsity-driven feature selection with deep subnetworks for flexible function approximation. Unlike conventional deep learning models, which often function as black boxes, SDAMI explicitly disentangles main effects and interaction effects to enhance interpretability. At the same time, its deep additive structure achieves higher predictive accuracy than classical additive models. Central to SDAMI is the concept of an Effect Footprint, which assumes that higher-order interactions project marginally onto main effects. Guided by this principle, SDAMI adopts a two-stage strategy: first, identify strong main effects that implicitly carry information about important interactions; second, exploit this information—through structured regularization such as group lasso—to distinguish genuine main effects from interaction effects. For each selected main effect, SDAMI constructs a dedicated subnetwork, enabling nonlinear function approximation while preserving interpretability and providing a structured foundation for modeling interactions. Extensive simulations with comparisons confirm SDAMI's ability to recover effect structures across diverse scenarios, while applications in reliability analysis, neuroscience, and medical diagnostics further demonstrate its versatility in addressing real-world high-dimensional modeling challenges.

1 INTRODUCTION

Deep learning regression now underpins applications across science, engineering, and biomedicine (Cesario et al., 2024; Collins et al., 2024). Yet most architectures are tuned to data-rich regimes with large sample sizes (He et al., 2020). In many emerging settings—especially personalized AI—the reality is the opposite: modest numbers of samples paired with extremely high feature counts. Such small-n, large-p problems are increasingly common as measurement technologies extract thousands of variables from limited observations (Jain, 2002; Stefanicka-Wojtas & Kurpas, 2023; Zhou et al., 2015). Our motivating example comes from neuroscience, where we analyze single-cell activity with roughly $n = 500$ observations and over $p = 11{,}000$ candidate features. This regime creates a basic tension. Classical deep models risk overfitting because the effective sample size per parameter is tiny, while aggressive dimensionality reduction can discard meaningful biological signal. Addressing this trade-off requires models that scale to high dimensions, remain stable in small samples, and preserve interpretability for scientific discovery.

When data are abundant, conventional deep models can achieve high predictive accuracy but typically operate as "black boxes," obscuring how individual variables and their interactions drive predictions (Wang & Lin, 2021). That can suffice for tasks like image classification or speech recognition, but scientific studies need insight into which effects matter and why (Molnar, 2020). In small-n, large-p settings, the stakes are higher: high variance and spurious correlations are easy to create, making transparency essential for reliability (Hastie et al., 2009). These considerations motivate structured architectures that explicitly encode regression effects. By modeling main effects

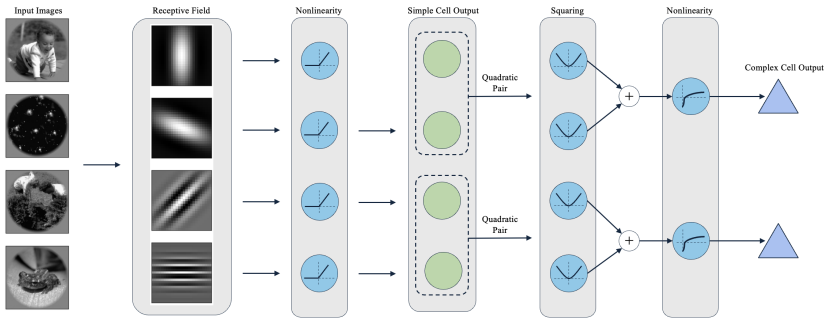

Figure 1: The formation of complex cells arises from nonlinear activation of quadratic pairs of simple cells generated by Gabor-wavelet filters applied to the input.

and interactions in a constrained, interpretable form, one can deliver components that aid inference, support diagnostics, and tie predictions back to hypotheses, even when data are limited.

A concrete illustration comes from modeling visual cortex responses. Images are first passed through localized, orientation- and phase-sensitive Gabor filters to mimic simple-cell receptive fields; outputs then undergo nonlinear transforms to produce single-cell responses. Complex cells are formed by pooling quadrature-phase pairs (square–sum–nonlinearity), yielding phase-invariant responses. This pipeline generates thousands of features from only a few hundred image–response pairs, creating a prototypical small-n, large-p scenario. Classical sparse additive models treat simple and complex cell terms as independent main effects, offering flexibility but ignoring biologically plausible higher-order associations (Kay et al., 2008; Vu et al., 2008). These limitations motivate us to propose the Sparse Deep Additive Model with Interactions (SDAMI), a structured deep additive framework that preserves interpretability while facilitating the discovery of nonlinear effects and interactions.

The Sparse Deep Additive Model with Interactions (SDAMI) is motivated by a new principle we introduce in this study, the notion of the Effect Footprint. The Effect Footprint posits that higher-order interactions leave detectable marginal signatures on main effects, thereby providing a pathway for discovering interactions even when direct estimation is statistically expensive in small-sample regimes. While related ideas have appeared in the statistical literature under hierarchical sparsity and the heredity principle (Bien et al., 2013; Lim & Hastie, 2015), the Effect Footprint is novel in explicitly formalizing this marginal-to-interaction connection and operationalizing it in a deep additive framework. SDAMI leverages this idea in a two-stage procedure. In the first stage, it identifies strong main effects whose marginal signals implicitly carry information about potential interactions. In the second stage, it employs structured regularization—such as group penalties or hierarchical sparsity (Simon et al., 2013; Yuan et al., 2009; Zhao et al., 2009)—to disentangle true main effects from interaction effects and to introduce nonlinear interaction subnetworks only when justified by the data. For each selected main effect, SDAMI constructs a dedicated subnetwork, enabling nonlinear function approximation while retaining interpretability at the effect level. In this way, SDAMI achieves a balance between flexibility and transparency: it adapts deep subnetworks to capture complex nonlinearities while organizing them in an additive structure that preserves clarity. Extensive simulations show that the proposed approach recovers effect structure across diverse scenarios and avoids the pitfalls of either underfitting main effects or overfitting interactions.

Related Work and Differences. To clarify SDAMI's distinct role, we compare it with two established model families: deep models with interactions and additive models with deep structure. Conventional deep learning with entangled architectures can represent complex interactions implicitly (He et al., 2020), but typically offers little insight into which variables matter or how they contribute. Many studies attempt to compensate by adding external attribution modules or post-hoc diagnostics, yet these approaches often rest on restrictive assumptions (e.g., linearized effects), miss general nonlinearities, and are prone to instability when samples are scarce or signals are weak (Molnar, 2020). In contrast, SDAMI is designed to be additive at the top layer: each selected variable receives a dedicated subnetwork, and interaction blocks are introduced only when warranted by estimated footprints. This design provides a direct, training-time pathway that aligns feature at-

tributions with the generative structure, rather than relying on surrogate explanations after the fact. The result is effect-level interpretability, reduced variance, and targeted capacity allocation. Neural additive models likewise attach subnetworks to inputs but usually treat interactions as optional or probe them in an unstructured manner (Agarwal et al., 2021; Xu et al., 2023). SDAMI goes further by embedding interaction discovery into the objective via structured penalties that promote sparsity and respect grouping, allowing statistically disciplined selection of both main effects and interactions (Patel et al., 2020; Shah, 2016). Overall, SDAMI combines interpretable decomposition with expressive subnetworks and data-driven interaction modules, yielding a principled framework for identifying and estimating effects without sacrificing stability in small-n, large-p regimes. It also preserves sample efficiency by allocating capacity only where evidence supports complexity and robustness.

Our Contribution. SDAMI connects three strands of literature. First, *sparse additive modeling* (e.g., SpAM [SpAM]) attains interpretability via main-effect decompositions with sparsity, but struggles with rich nonlinear interactions (e.g., Fan et al. (2011a); Ravikumar et al. (2009); Fan et al. (2011b)). Second, *neural additive frameworks* replace basis expansions with subnetworks, enhancing flexibility yet typically handling interactions ad hoc (e.g., Agarwal et al. (2021); Vaughan et al. (2018); Yang et al. (2021)). Third, *structured sparsity in deep learning* selects parameter groups via group/hierarchical penalties (e.g., Yuan & Lin (2006); Scardapane et al. (2017); Wen et al. (2016)). SDAMI builds on these by tailoring structured sparsity to the hierarchy of regression effects and by operationalizing the *Effect Footprint* principle so evidence in main effects guides interaction discovery, shrinking search, improving stability, and yielding interpretable recovery under small-n, large-p. Our main contributions are summarized below:

- Introduce the *effect footprint* and a response-guided framework for nonlinear model-structure selection in deep models.
- Propose a structured deep additive model with interactions: each f_j and $f(\cdot)$ is a dedicated subnetwork; norm-based input constraints (3) gate connections (pruning when $\|f_j\|{=}0$ or $\|f_{\mathcal{I}}\|{=}0$), yielding sparse, effect-level interpretability.
- Provide theory: (i) conditions under which footprints vanish (Hoeffding–Sobol first-order projections), (ii) *effect-level selection consistency*, and (iii) *prediction convergence in probability*.
- Present extensive $p \gg n$ simulations and applications showing improved *predictability and interpretability* over baselines, with high true positive rate (TPR)/low false positive rate (FPR) and informative component-function visualizations.

2 Problem Setup and Response-Guided Structured Deep Framework

We observe regression data $\{(\mathbf{X}_i, Y_i)\}_{i=1}^n$, where $\mathbf{X}_i = (X_{i1}, \ldots, X_{ik})^\top \in \mathbb{R}^k$ denotes the predictors and $Y_i \in \mathbb{R}$ is the response. The true regression function is assumed to follow a sparse additive-plus-interaction structure of the form

$$Y_i = \sum_{j \in \mathcal{M}} f_j(X_{ij}) + f(\mathbf{X}_{i,\mathcal{I}}) + \epsilon_i, \tag{1}$$

where $\mathcal{M} \subseteq \{1, \ldots, k\}$ is the index set of important main effects, $\mathcal{I} \subseteq \{1, \ldots, k\}$ is the index set of variables entering the interaction component, and ϵ_i is a random error with $\mathbb{E}[\epsilon_i] = 0$ and $\mathrm{Var}(\epsilon_i) = \sigma^2$. We assume $|\mathcal{M}| = p \ll k$, so that only a small fraction of predictors directly contribute as main effects. In addition, we define $|\mathcal{I} \setminus \mathcal{M}| = q$, capturing variables that contribute exclusively through interactions but not as main effects. These *footprint variables* are essential for identifying higher-order dependencies that cannot be explained by additive contributions alone.

To estimate model (1), each main-effect function f_j and the interaction function $f(\cdot)$ are represented by dedicated neural subnetworks. Let θ_j and $\theta_{\mathcal{I}}$ denote their respective parameters. Denote by $W^{(1)}_{\mathcal{M},j}$ the weight vector in the first hidden layer connecting input X_j to its main-effect subnetwork, and by $W^{(1)}_{\mathcal{I},j}$ the weight vector connecting X_j to the interaction subnetwork. The estimation problem

is then formulated as

$$\min_{\theta} \frac{1}{n}\sum_{i=1}^{n}\left(Y_i - \sum_{j\in\mathcal{M}} \mathrm{NN}^{(j)}(X_{ij};\theta_j) - \mathrm{NN}^{(\mathcal{I})}(\mathbf{X}_{i,\mathcal{I}};\theta_{\mathcal{I}})\right)^2, \tag{2}$$

$$\text{subject to} \quad \|W^{(1)}_{\mathcal{M},j}\|_\infty \le \kappa_{\mathcal{M}}\|f_j\|, \quad j=1,\dots,k, \qquad \|W^{(1)}_{\mathcal{I},j}\|_\infty \le \kappa_{\mathcal{I}}\|f_{\mathcal{I}}\|, \quad j\in\mathcal{I}. \tag{3}$$

Here, $\mathrm{NN}^{(j)}(X_{ij};\theta_j)$ denotes a *neural network (NN) submodule* dedicated to the j-th main effect, parameterized by weights θ_j, while $\mathrm{NN}^{(\mathcal{I})}(\mathbf{X}_{i,\mathcal{I}};\theta_{\mathcal{I}})$ denotes a subnetwork for the interaction set $\mathcal{I}$, parameterized by $\theta_{\mathcal{I}}$. Each NN is a standard feedforward network with hidden layers and nonlinear activations, serving as a flexible nonlinear approximator. The reference functions f_j and $f_{\mathcal{I}}$ represent the true main-effect and interaction-effect components of the regression function $f^\star$. The constraints in equation 3 regulate the first-layer weights $W^{(1)}$ relative to $\|f_j\|$ and $\|f_{\mathcal{I}}\|$, ensuring that each subnetwork remains aligned with the magnitude of its corresponding effect and thereby preserving hierarchical structure and interpretability. If $\|f_j\| = 0$, the outgoing weights $W^{(1)}_{\mathcal{M},j}$ vanish, excluding X_j from its subnetwork. Similarly, if $\|f_{\mathcal{I}}\| = 0$, connections into the interaction subnetwork are eliminated, removing the interaction term. Thus sparsity and interpretability are achieved not through explicit penalties, but through norm-based constraints that prune irrelevant effects, while the loss in equation 2 enforces predictive accuracy.

The constrained optimization problem (2) and (3) determines which subnetworks remain active for prediction. However, direct optimization without additional structure becomes infeasible in high dimensions, since it is difficult to distinguish relevant main effects from irrelevant variables or latent contributors to interactions. To overcome this challenge, we introduce the principle of an *effect footprint*, which provides a mechanism for linking variable screening directly to the objective function and guiding the activation of subnetworks in a statistically coherent manner.

3 Fitting Sparse Deep Additive Models with Interactions (SDAMI)

The constrained optimization problem (2) and (3) highlights how sparsity and interpretability can be enforced by linking input-layer weights to functional norms. However, to determine which subnetworks should remain active in high-dimensional settings, we require a principled way to identify variables that may only contribute through latent interactions. This motivates the notion of an *effect footprint*.

Formally, an effect footprint is defined as the marginal influence of a variable that arises solely from its participation in the interaction subnetwork. If $X_j \notin \mathcal{M}$ but $j \in \mathcal{I}$, then

$$m_j(x) = \mathbb{E}\big[f(\mathbf{X}_{\mathcal{I}}) \mid X_j = x\big]$$

may still vary with x, leaving a detectable signal in the marginal regression of Y on X_j. Thus, although f_j is absent in the true model, X_j exhibits a footprint through $f(\mathbf{X}_{\mathcal{I}})$. This principle implies that the regression function can be approximated as

$$Y_i = \sum_{j=1}^{p+q} f_j(X_{ij}) + \epsilon_i,$$

where $\{1,\dots,p\} = \mathcal{M}$ correspond to true main effects and $\{p+1,\dots,p+q\} = \mathcal{I}\setminus\mathcal{M}$ represent footprint variables. Let $\mathcal{S} = \{1,\dots,p+q\}$ denote the union of main and footprint variables. Recovering $\mathcal{S}$ is therefore the first step toward solving objective function (2) and (3).

Motivated by model (3), we can apply a sparse additive screening procedure to identify an estimated active set $\widehat{\mathcal{S}}$ (Ravikumar et al., 2009). This step retains variables with either genuine main effects or non-negligible footprints, while shrinking all others to zero. Once $\widehat{\mathcal{S}}$ is obtained, we refine the decomposition by partitioning $\widehat{\mathcal{S}}$ into $\widehat{\mathcal{M}}$ (main effects) and $\widehat{\mathcal{I}}$ (interaction effects) using group lasso with an orthogonal basis expansion (Yuan & Lin, 2006). The sets $\widehat{\mathcal{M}}$ and $\widehat{\mathcal{I}}$ are associated with penalty parameters λ_1 and λ_2, which are selected via Mallow's C_p and cross-validation, respectively.

These subsets then guide the fitting of deep regression model defined in model (1), implemented in PyTorch. Figure 2 illustrates the SDAMI architecture and how structured constraints impose sparsity on the network. A detailed description of the algorithm is provided in Appendix A of the supplementary material.

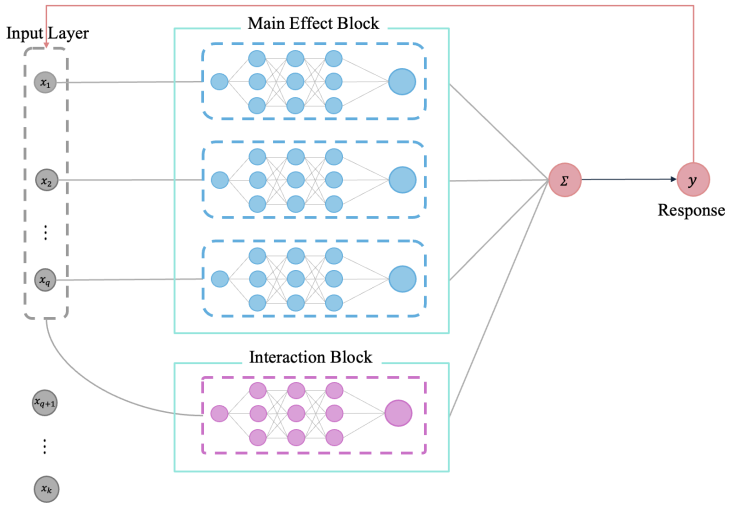

Figure 2: The SDAMI architecture. Screening identifies both main and footprint variables, which guide the activation of subnetworks and enforce biologically and statistically meaningful structure.

4 Theoretical Analysis: The Role of Effect Footprint, Selection Consistency, Model Convergence

We present the theoretical foundation of SDAMI in three parts. First, we formalize the concept of *effect footprint*, which justifies feasible high-dimensional screening. Second, we show that SDAMI attains *effect-level selection consistency*, recovering both the true main effects and the interaction structure. Finally, we establish predictive validity by proving that the fitted predictor converges in probability to the true model (1).

Theorem 4.1 (When effect footprints vanish)**.** *Let* $\mathbf{X}_{\mathcal{I}} = (X_j, \mathbf{Z})$ *be the variables in an interaction* $f(\mathbf{X}_{\mathcal{I}})$ *with* $\mathbb{E}[f(\mathbf{X}_{\mathcal{I}})] = 0$*. Define*

$$m_j(x) = \mathbb{E}[f(\mathbf{X}_{\mathcal{I}}) \mid X_j = x].$$

Then $m_j(x)$ *is constant (no footprint) iff the first–order projection of* f *onto functions of* X_j *vanishes in the Hoeffding–Sobol decomposition (Sobol', 1990; Sobol, 2001). In this case,* f *contains only higher–order components involving* X_j*.*

This characterization isolates the exceptional cases in which footprints fail: a variable leaves no detectable footprint precisely when its influence appears solely through higher–order interactions that vanish after averaging over the remaining inputs. Such a variable may still be essential via interactions, but univariate screening cannot detect it. Two canonical settings illustrate this: (i) independence with centering (e.g., bilinear forms of independent, mean-centered inputs), and (ii) perfect symmetry with antisymmetric interactions (e.g., the XOR rule for binary data or odd functions under symmetric continuous inputs). These conditions are stringent; in practice predictors are correlated, distributions seldom perfectly symmetric, and noise disrupts exact cancellations. Consequently, footprints typically exist, providing a robust signal for screening. A detailed proof is given in Appendix B of the supplementary material.

Theorem 4.2 (Effect-level selection consistency of SDAMI)**.** *Under assumptions (A1)–(A7),*

$$\mathbb{P}\Big(\{j : \widehat{f}_j \neq 0\} = \mathcal{M} \text{ and } (\widehat{f}_{\mathcal{I}} \neq 0 \Leftrightarrow f_{\mathcal{I}} \neq 0)\Big) \longrightarrow 1 \quad \text{as } n \to \infty.$$

Case	Functional Form	Conceptual Description
1	$y = f_1(x_1) + f_2(x_2) + f_3(x_3) + f_4(x_4)$	Only strong main effects, no interactions
2	$y = f_1(x_1) + f_2(x_2) + f_3(x_3) + 0.01 f_4(x_4)$	Main effects with weak signals
3	$y = f_1(x_1) + f_2(x_2) + f_3(x_3) + f_5(x_4, x_5)$	Main effects plus one interaction block with no overlap
4	$y = f_1(x_1) + f_2(x_2) + f_3(x_3) + f_5(x_3, x_4)$	Main effects + 1 interaction block with some overlapping variables
5	$y = f_1(x_1) + f_2(x_2) + f_3(x_3) + f_5(x_2, x_3)$	Main effects plus one interaction block with all variables overlapping
6	$y = f_5(x_1, x_2) + f_5(x_3, x_4)$	Only interaction effects, no main effects

Table 1: The summary table for numerical simulation models.

Thus SDAMI does not merely exploit footprints heuristically; it achieves a rigorous form of oracle recovery. As n grows, SDAMI selects exactly the true set of main effects and correctly detects the interaction with probability tending to one, ensuring that the discovered structure reflects the underlying generative mechanism. The proof (Appendix C of the supplementary material) employs a block-wise primal–dual witness argument for the group-lasso formulation, leveraging footprint-induced group signals and oracle inequalities for group sparsity (Lounici et al., 2011; Negahban et al., 2009).

Theorem 4.3 (Prediction convergence in probability for SDAMI)**.** *Let $\widehat{A}_n$ be the SDAMI-selected index set and let $\widehat{f}_n$ be the SDAMI estimator. Suppose* (B1)–(B6) *hold. Then, for every fixed $\varepsilon > 0$,*

$$\mathbb{P}\Big(|\widehat{f}_n(\mathbf{X}) - f^\star(\mathbf{X})| \ \geq \ \varepsilon \Big) \ \longrightarrow \ 0 \quad \textit{as } n \to \infty,$$

The key idea is to combine sieve approximation with uniform generalization. Selection consistency concentrates learning on the correct coordinates; empirical risk minimization up to a vanishing tolerance, together with a uniform law of large numbers for squared loss (via Rademacher and covering bounds for norm–constrained networks), transfers empirical to population L_2–risk (Bartlett & Mendelson, 2002; Mohri et al., 2018; van de Geer, 2000). In parallel, ReLU approximation theory ensures the sieve approximates the oracle regression under a suitable growth schedule (Barron, 1993; Yarotsky, 2017; Schmidt-Hieber, 2020; Suzuki, 2019). A uniform L_2 envelope (implied by norm constraints and square-integrability) guarantees uniform integrability, so vanishing population risk implies vanishing misfit probability via a Markov-type bound. Full details appear in Appendix D of the supplementary material.

5 Numerical Experiments

We conduct comprehensive numerical simulations to evaluate SDAMI's ability to recover effect structures and achieve predictive accuracy across diverse scenarios. Data are generated under six distinct settings summarized in Table 1, each defined by different functional forms involving only main effects or combinations of main and interaction effects, with varying overlaps among interaction variables.

For each setting:

- Sample sizes n vary across $150, 300$, and 450.
- The feature dimension is fixed at $p = 150$ in a high-dimensional regime, with only a few features having substantive main or interaction effects.
- Responses are generated as $Y_i = \sum_{j \in \mathcal{M}} f_j(X_{ij}) + f(X_{i,\mathcal{I}}) + \epsilon_i$, where $X_i \sim \text{Uniform}(-2.5, 2.5)$ independently and $\epsilon_i \sim N(0, \sigma^2)$.
- True functions are drawn from representative nonlinear forms: $f_1(x) = -2\sin(2x)$, $f_2(x) = \frac{x^2}{2} + 1$, $f_3(x) = x - \frac{1}{2}$, $f_4(x) = e^{-x} + e^{-1} - 1$, and $f_5(x_1, x_2) = e^{\sin(x_1) + \cos(x_2) - 1}$. Detailed formulations for the six experimental cases are provided in Table 1.

We benchmark SDAMI against three alternatives: deep neural networks (DNN, modeling full interactions), fast sparse additive models (fSpAM, a high-dimensional main effect model), and LASSO (a high-dimensional linear model). The architecture of SDAMI is determined by cross-validation and the detail can be found in Appendix E.1 of the supplementary material. Across all simulation settings and sample sizes, SDAMI consistently achieves the lowest mean squared error (MSE)

Method	SDAMI		DNN		fSpAM		LASSO	
	MSE↓	STD↑	MSE↓	STD↑	MSE↓	STD↑	MSE↓	STD↑
Case 1	**0.68**	0.59	14.37	1.03	5.48	0.52	4.77	0.94
Case 2	**0.77**	0.41	5.39	0.42	3.22	0.25	3.02	0.37
Case 3	**0.70**	0.58	5.78	0.43	3.55	0.25	3.61	0.39
Case 4	**0.84**	0.27	7.11	0.51	3.53	0.27	3.34	0.40
Case 5	**0.85**	0.53	5.90	0.43	3.65	0.27	3.59	0.41
Case 6	**0.27**	0.25	1.05	0.11	0.63	0.05	0.61	0.10

Table 2: Performance comparison on SDAMI, DNN, fSpAM, and LASSO by case type when $n = 150$; ↓ means the lowest the better while ↑ means the highest the better.

Method	SDAMI		LASSONET		SODA	
	TPR ↑	FPR ↓	TPR ↑	FPR ↓	TPR ↑	FPR ↓
Case 1	1.0000 (-)	1.1×10^{-5} (-)	0.4900 (0.0490)	0.0037 (0.0028)	0.0175 (0.0641)	6×10^{-4} (2×10^{-4})
Case 2	1.0000 (-)	1.1×10^{-5} (-)	0.2550 (0.0350)	0.0099 (0.0138)	0.0475 (0.1048)	1×10^{-3} (2×10^{-4})
Case 3	0.7500 (-)	10^{-4} (10^{-5})	0.1400 (0.3007)	0.1724 (0.2810)	0.025 (0.0754)	6×10^{-4} (3×10^{-4})
Case 4	0.7600 (-)	10^{-4} (10^{-5})	0.1300 (0.3051)	0.1621 (0.2701)	0.040 (0.1049)	7×10^{-4} (3×10^{-4})
Case 5	0.7550 (0.0249)	10^{-4} (10^{-5})	0.1250 (0.2947)	0.1629 (0.2814)	0.055 (0.1100)	9×10^{-4} (2×10^{-4})
Case 6	0.6000 (-)	10^{-4} (10^{-5})	0.1100 (0.2700)	0.1432 (0.2334)	- (-)	6×10^{-4} (2×10^{-4})

Table 3: Mean (standard deviation) of TPR and FPR over 100 simulations from SDAMI, LASSONET, SODA when $n = 150$ where $(-)$ indicates value $< 1e^{-5}$.

(Tables 2), confirming its capacity to flexibly capture nonlinear main and interaction effects while maintaining interpretability. Increasing sample size improves all methods' performance; however, SDAMI preserves a clear margin of advantage, underscoring its robustness and scalability. Case-specific comparisons further illustrate these findings. In Case 1, which involves only strong main effects, SDAMI attains the best accuracy without introducing spurious interactions, demonstrating parsimony (Tables 2). In Case 2, where true signals are weak, SDAMI continues to outperform benchmarks, reflecting robustness to small effect sizes. In Cases 3–5, which include both main and interaction effects with varying degrees of overlap, fSpAM and LASSO show limited capacity to recover the true structures, while SDAMI consistently models both overlapping and non-overlapping interactions, achieving markedly lower errors across all sample sizes. In Case 6, where effects arise solely from interactions, SDAMI retains strong predictive performance, while fSpAM and LASSO deteriorate substantially and DNN suffers from instability. Taken together, the results across Tables 2 demonstrate that SDAMI provides a balanced combination of flexibility, interpretability, and accuracy. By leveraging effect footprints, it adapts to diverse data-generating mechanisms and consistently outperforms existing approaches, validating its utility as a powerful framework for sparse high-dimensional regression in the presence of complex effect structures. The additional numerical experiments with respect to different sample size is displayed in Appendix E.2 of the supplementary material.

We further evaluate the feature selection performance of SDAMI, focusing on its ability to recover true main and interaction effects while minimizing false discoveries. Table 3 summarizes the TPR and FPR when $n = 150$, averaged over 100 simulations and compared with LASSONET(Lemhadri et al., 2021), and sodavis (SODA)(Li, 2015). TPR measures the proportion of correctly identified active variables, while FPR reflects the rate of spurious selections. SDAMI achieves near-perfect TPRs of 1.0 in Cases 1 and 2, dominated by main effects, showing it reliably identifies relevant signals without omission. In more complex settings with overlapping and non-overlapping interactions (Cases 3–6), SDAMI maintains substantially higher TPRs than LASSONET and SODA, which experience steep sensitivity drops. Concurrently, SDAMI obtains extremely low FPRs, often on the order of 10^{-4}, whereas competitors select irrelevant features at much higher rates. This result indicates that SDAMI strikes a favorable balance between sensitivity and specificity, crucial for high-dimensional regression where false discoveries can obscure interpretation. Stability across 100 replications affirms robustness , while improvements from $n = 150$ to $n = 300$ confirm scalability. Overall, SDAMI demonstrates reliable, precise feature recovery in sparse, high-dimensional problems with complex effect structures. The additional experiment of feature selection for $n = 300$ is shown in Appendix E.2 of the supplementary material.

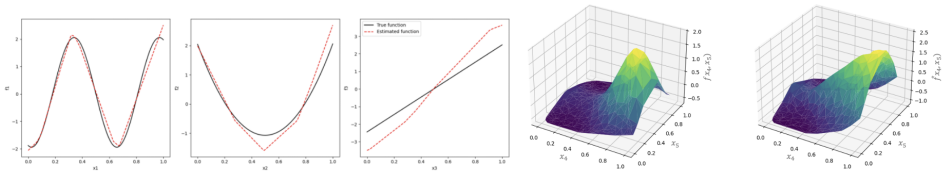

Figure 3: (Case 3) The three figures on the left: Estimated (red dashed lines) versus true additive component functions (solid black lines) for three main effects; the two figures on the far right: the first shows the true response surface for interaction, and the second shows its estimated response surface.

In our simulation studies, a key advantage of SDAMI over other machine learning models is its interpretability through visualization. Unlike black-box methods, SDAMI enables visualization of individual component functions, allowing researchers to inspect each selected main or interaction effect's contribution to predictions. This layered interpretability enhances transparency and offers scientific insight into the modeled relationships. Figure 3 illustrates Case 3 results, where the black solid line shows the true function and the red dashed line shows SDAMI estimates, demonstrating accurate recovery of complex nonlinear patterns. Additionally, visualizations for all simulation cases are provided in Appendix E.3 of the supplementary material, underscoring SDAMI's value for interpretable modeling in high-dimensional regression.

6 Revisit Real Datasets for Better Understanding Practical Use of SDAMI

The V1 fMRI dataset (Kay et al., 2008) records voxel responses from human primary visual cortex at 2 mm×2 mm×2.5 mm resolution on a 4 T scanner while subjects viewed grayscale natural images through a circular aperture. Stimuli are flashed three times per second with interleaved blanks, and signals are preprocessed to reduce noise and nonstationarity. Prior work shows interaction effects among complex cells (Kay et al., 2008; Vu et al., 2008), but how to model such interactions while preserving biological meaning remains underexplored. To foreground the neuroimaging challenge—small n, high p—experiments use 300 unique natural images, each summarized by 1,800 Gabor-filter features derived from complex-cell processing; each voxel reflects pooled, rectified activity organized by a receptive-field hierarchy over space, frequency, and phase. Figure 1 sketches the pipeline producing simple-cell and complex-cell predictors (and Figure 4 shows the SDAMI linkage to voxel responses).

Applying SDAMI to the V1 dataset yields strong predictive gains relative to baselines (Table 4). Modeling only main effects (SDAMI (ME)) already improves MSE over several competitors, and adding explicit interactions (SDAMI (IN)) further boosts performance, underscoring the importance of capturing higher-order structure among complex-cell features. Across datasets in Table 4, SDAMI (IN) attains the best or near-best scores (e.g., lowest MSE on Chip, best root mean squared error (RMSE) and R-squared (R^2) on Diabetes, and lowest MSE / highest R^2 on V1), demonstrating that interaction modules materially enhance accuracy beyond additive structure alone.

Beyond accuracy, SDAMI provides effect-level interpretability via component visualizations. Figure 5 displays estimated main effects from selected Gabor-filter features (highlighting positions, orientations, and scales linked to activity) and interaction surfaces for key feature pairs, revealing synergistic patterns consistent with cortical pooling. SDAMI can also be configured to select coherent pools of complex cells—e.g., grouping features that share spatial location and frequency while varying orientation or phase—thereby aligning selected ensembles with neurophysiological hypotheses rather than ad hoc combinations.

Finally, SDAMI generalizes beyond V1. On the Chip and Diabetes datasets (Table 4), SDAMI (IN) achieves the lowest error and highest R^2 among fSpAM, LASSO, DNN, and LASSONET, while maintaining clear effect decompositions. Together, these results show that SDAMI delivers competitive or superior prediction and biologically grounded interpretability in small-n, large-p regimes, establishing a principled framework for response modeling in neuroscience and other high-

Dataset	Chip		Diabete		V1 Cell	
	MSE↓	R^2 ↑	RMSE↓	R^2 ↑	MSE↓	R^2 ↑
DNN	0.927	0.071	58.375	0.403	0.493	0.074
fSpAM	0.753	0.237	58.797	0.395	0.629	0.258
LASSO	0.276	0.716	59.514	0.380	0.624	0.268
LASSONET	0.904	0.056	56.612	0.439	0.627	0.262
SDAMI (ME)	0.736	0.244	62.052	0.326	0.388	0.272
SDAMI (IN)	0.236	0.758	52.774	0.512	0.372	0.302

Table 4: MSE and R^2 for three real datasets.

dimensional domains. Due to the page limitation, the details of other two dataset analyses are given in Appendix F.1 and Appendix F.2 of the supplementary material.

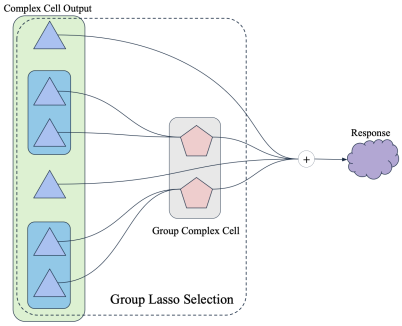

Figure 4: The formation of response arises from complex cells and group complex cells selected by group lasso applied to the input.

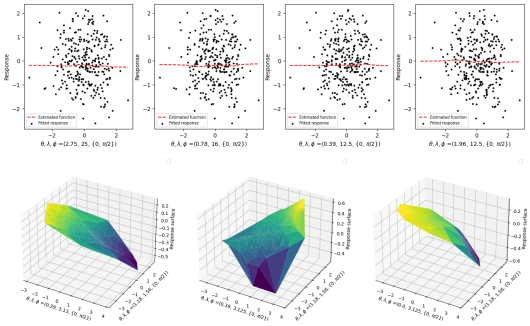

Figure 5: (V1 Cell Dataset) Upper panel: the predicted Marginal main effects (solid black dots); lower panel: the estimated response surface for interactions.

7 Conclusion

This paper introduced the Sparse Deep Additive Model with Interactions (SDAMI), a structured deep learning framework tailored for small-n, large-p regression problems. By leveraging the principle of *effect footprints*, SDAMI offers a systematic approach to detecting and modeling higher-order interactions while retaining effect-level interpretability. The method enforces sparsity through norm-based constraints that prune irrelevant variables and subnetworks, ensuring both statistical stability and interpretability. Theoretical analysis established effect-level selection consistency and prediction convergence in probability, providing rigorous guarantees beyond heuristic interpretability. Simulation studies demonstrated that SDAMI reliably recovers both main and interaction effects, outperforming classical additive models and black-box neural networks. Applications to neuroscience and reliability analysis further illustrated the model's versatility and its ability to bridge deep learning with domain-specific interpretability requirements.

Limitations and Future Directions. While SDAMI achieves both interpretability alongside statistical guarantees, several limitations remain. First, the current two-stage fitting procedure relies on estimating function norms via a sparse additive model (SpAM) step, which can be computationally demanding. This step could be accelerated by employing screening method such as Sure Independence Screening (SIS Fan & Lv (2008); Fan et al. (2011a)) to directly identify important predictors, prioritizing variable selection over full function estimation. This alternative motivates new theoretical development of SDAMI under SIS-style screening, potentially enhancing scalability. Second, the current theoretical results focus on effect-level consistency but do not provide convergence rates. Incorporating recent advances in high-dimensional regression and nonparametric learning provide sharper tools for establishing minimax rates and finite-sample guarantees. Extending SDAMI's theory to include convergence rates would deepen understanding of its performance in finite-sample regimes. Together, these directions offer promising avenues for improving both the computational efficiency and theoretical rigor of SDAMI for large-scale scientific applications.

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. *Advances in neural information processing systems*, 34:4699–4711, 2021.

Andrew R. Barron. Universal approximation bounds for superpositions of a sigmoidal function. *IEEE Transactions on Information Theory*, 39(3):930–945, 1993. doi: 10.1109/18.256500.

Peter L. Bartlett and Shahar Mendelson. Rademacher and gaussian complexities: Risk bounds and structural results. *Journal of Machine Learning Research*, 3:463–482, 2002.

Jacob Bien, Jonathan Taylor, and Robert Tibshirani. A lasso for hierarchical interactions. *Annals of statistics*, 41(3):1111, 2013.

Eugenio Cesario, Carmela Comito, and Ester Zumpano. A survey of the recent trends in deep learning for literature based discovery in the biomedical domain. *Neurocomputing*, 568:127079, 2024.

Gary S Collins, Karel GM Moons, Paula Dhiman, Richard D Riley, Andrew L Beam, Ben Van Calster, Marzyeh Ghassemi, Xiaoxuan Liu, Johannes B Reitsma, Maarten Van Smeden, et al. Tripod+ ai statement: updated guidance for reporting clinical prediction models that use regression or machine learning methods. *bmj*, 385, 2024.

Jianqing Fan and Jinchi Lv. Sure independence screening for ultrahigh dimensional feature space. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 70(5):849–911, 2008.

Jianqing Fan, Yang Feng, and Rui Song. Nonparametric independence screening in sparse ultra-high-dimensional additive models. *Journal of the American Statistical Association*, 106(494): 544–557, 2011a.

Jianqing Fan, Jinchi Lv, and Lei Qi. Sparse high-dimensional models in economics. *Annu. Rev. Econ.*, 3(1):291–317, 2011b.

Trevor Hastie, Robert Tibshirani, and Jerome Friedman. *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*. Springer, New York, 2nd edition, 2009. doi: 10.1007/978-0-387-84858-7.

Tong He, Ru Kong, Avram J Holmes, Minh Nguyen, Mert R Sabuncu, Simon B Eickhoff, Danilo Bzdok, Jiashi Feng, and BT Thomas Yeo. Deep neural networks and kernel regression achieve comparable accuracies for functional connectivity prediction of behavior and demographics. *NeuroImage*, 206:116276, 2020.

Shu-Han Hsu, Ying-Yuan Huang, Yi-Da Wu, Kexin Yang, Li-Hsiang Lin, and Linda Milor. Extraction of wearout model parameters using on-line test of an sram. *Microelectronics Reliability*, 114: 113756, 2020.

Kewal K Jain. Personalized medicine. *Current opinion in molecular therapeutics*, 4(6):548–558, 2002.

V Roshan Joseph, Evren Gul, and Shan Ba. Maximum projection designs for computer experiments. *Biometrika*, 102:371–380, 2015.

Kendrick N Kay, Thomas Naselaris, Ryan J Prenger, and Jack L Gallant. Identifying natural images from human brain activity. *Nature*, 452(7185):352–355, 2008.

Ismael Lemhadri, Feng Ruan, Louis Abraham, and Robert Tibshirani. Lassonet: A neural network with feature sparsity. *Journal of Machine Learning Research*, 22(127):1–29, 2021.

Yang Li. *sodavis: SODA: Main and Interaction Effects Selection for Logistic Regression, Quadratic Discriminant and General Index Models*. R Foundation for Statistical Computing, 2015. URL `https://cran.r-project.org/web/packages/sodavis/`.

Michael Lim and Trevor Hastie. Learning interactions via hierarchical group-lasso regularization. *Journal of Computational and Graphical Statistics*, 24(3):627–654, 2015.

Karim Lounici, Massimiliano Pontil, Sara van de Geer, and Alexandre B. Tsybakov. Oracle inequalities and optimal inference under group sparsity. *Annals of Statistics*, 39(4):2164–2204, 2011. URL https://doi.org/10.1214/11-AOS896.

Mehryar Mohri, Afshin Rostamizadeh, and Ameet Talwalkar. *Foundations of Machine Learning*. MIT Press, 2nd edition, 2018.

Christoph Molnar. *Interpretable machine learning*. Lulu. com, 2020.

Sahand Negahban, Bin Yu, Martin J Wainwright, and Pradeep Ravikumar. A unified framework for high-dimensional analysis of m-estimators with decomposable regularizers. *Advances in neural information processing systems*, 22, 2009.

Lauv Patel, Tripti Shukla, Xiuzhen Huang, David W Ussery, and Shanzhi Wang. Machine learning methods in drug discovery. *Molecules*, 25(22):5277, 2020.

Pradeep Ravikumar, John Lafferty, Han Liu, and Larry Wasserman. Sparse additive models. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 71(5):1009–1030, 2009.

Thomas J Santner, Brian J Williams, and William I Notz. *The Design and Analysis of Computer Experiments (2nd Edition)*. New York, NW: Springer, 2019.

Simone Scardapane, Danilo Comminiello, Amir Hussain, and Aurelio Uncini. Group sparse regularization for deep neural networks. *Neurocomputing*, 241:81–89, 2017.

Johannes Schmidt-Hieber. Nonparametric regression using deep neural networks with relu activation function. *Annals of Statistics*, 48(4):1875–1897, 2020. doi: 10.1214/19-AOS1875.

Rajen D Shah. Modelling interactions in high-dimensional data with backtracking. *Journal of Machine Learning Research*, 17(207):1–31, 2016.

Noah Simon, Jerome Friedman, Trevor Hastie, and Robert Tibshirani. A sparse-group lasso. *Journal of computational and graphical statistics*, 22(2):231–245, 2013.

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.

Il'ya Meerovich Sobol'. On sensitivity estimation for nonlinear mathematical models. *Matematicheskoe modelirovanie*, 2(1):112–118, 1990.

Dorota Stefanicka-Wojtas and Donata Kurpas. Personalised medicine—implementation to the healthcare system in europe (focus group discussions). *Journal of personalized medicine*, 13 (3):380, 2023.

Taiji Suzuki. Adaptivity of deep relu network for learning in besov and mixed smooth besov spaces. In *Proceedings of the 36th International Conference on Machine Learning (ICML)*, volume 97 of *Proceedings of Machine Learning Research*, pp. 11692–11702. PMLR, 2019.

Sara A. van de Geer. *Empirical Processes in M-Estimation*. Cambridge Series in Statistical and Probabilistic Mathematics. Cambridge University Press, 2000.

Joel Vaughan, Agus Sudjianto, Erind Brahimi, Jie Chen, and Vijayan N Nair. Explainable neural networks based on additive index models. *arXiv preprint arXiv:1806.01933*, 2018.

Vincent Q Vu, Bin Yu, Thomas Naselaris, Kendrick Kay, Jack Gallant, and Pradeep Ravikumar. Nonparametric sparse hierarchical models describe v1 fmri responses to natural images. *Advances in Neural Information Processing Systems*, 21, 2008.

Tong Wang and Qihang Lin. Hybrid predictive models: When an interpretable model collaborates with a black-box model. *Journal of Machine Learning Research*, 22(137):1–38, 2021.

Wei Wen, Chunpeng Wu, Yandan Wang, Yiran Chen, and Hai Li. Learning structured sparsity in deep neural networks. *Advances in neural information processing systems*, 29, 2016.

CF Jeff Wu and Michael S Hamada. *Experiments: planning, analysis, and optimization*. John Wiley and Sons, 2011.

Shiyun Xu, Zhiqi Bu, Pratik Chaudhari, and Ian J Barnett. Sparse neural additive model: Interpretable deep learning with feature selection via group sparsity. In *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*, pp. 343–359. Springer, 2023.

Kexin Yang, Taizhi Liu, Rui Zhang, Dae-Hyun Kim, and Linda Milor. Front-end of line and middle-of-line time-dependent dielectric breakdown reliability simulator for logic circuits. *Microelectronics Reliability*, 76:81–86, 2017.

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.

Dmitry Yarotsky. Error bounds for approximations with deep relu networks. *Neural Networks*, 94: 103–114, 2017. doi: 10.1016/j.neunet.2017.07.005.

Ming Yuan and Yi Lin. Model selection and estimation in regression with grouped variables. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 68(1):49–67, 2006.

Ming Yuan, V Roshan Joseph, and Hui Zou. Structured variable selection and estimation. *The Annals of Applied Statistics*, pp. 1738–1757, 2009.

Peng Zhao, Guilherme Rocha, and Bin Yu. The composite absolute penalties family for grouped and hierarchical variable selection. 2009.

Luping Zhou, Lei Wang, Lingqiao Liu, Philip Ogunbona, and Dinggang Shen. Learning discriminative bayesian networks from high-dimensional continuous neuroimaging data. *IEEE transactions on pattern analysis and machine intelligence*, 38(11):2269–2283, 2015.

Supplementary Material for Sparse Deep Additive Model with Interactions: Enhancing Interpretability and Predictability

A SDAMI ALGORITHM

This section describes the detail of the SDAMI algorithm and how the model fitting works.

Algorithm 1 SDAMI Fitting

Require: Data $\{(X_i, Y_i)\}_{i=1}^n$, tuning parameters λ_1, λ_2

1: **Step 1: Effect Footprint Screening (SpAM).**

- Fit the sparse additive model

$$Y_i = \sum_{j=1}^{p+q} f_j(X_{ij}) + \epsilon_i$$

  using SpAM with penalty λ_1.
- Obtain estimated active set $\widehat{\mathcal{S}} \subseteq \{1, \ldots, p+q\}$ containing both true main effects and footprint variables.

2: **Step 2: Partition Active Set (Group Lasso).**

- Apply group lasso with orthogonal basis expansion on $\widehat{\mathcal{S}}$.
- Partition into $\widehat{\mathcal{M}}$ (main effects) and $\widehat{\mathcal{I}}$ (interaction effects).
- Select penalty λ_2 via cross-validation (with λ_1 selected by Mallow's C_p).

3: **Step 3: SDAMI Model Fitting.**

- Fit the constrained deep regression model using $\widehat{\mathcal{M}}$ and $\widehat{\mathcal{I}}$.
- Implement subnetworks in PyTorch, with sparsity imposed via norm-based constraints.

Ensure: Estimated main-effect subnetworks $\{\mathrm{NN}^{(j)}\}_{j\in\widehat{\mathcal{M}}}$ and interaction subnetworks $\{\mathrm{NN}^{(\mathcal{I})}\}_{\mathcal{I}\in\widehat{\mathcal{I}}}$.

Regularization Parameter Selection. The regularization parameters λ_1, λ_2 are selected by minimizing the estimated risk and by cross-validation, respectively. The effective degree of freedom is defined as $\mathrm{df}(\lambda) = \sum_j \nu_j I(\|\hat{f}_j\| \neq 0)$, where $\nu_i = \mathrm{trace}(S_j)$ and S_j denotes the smoothing matrix for the j-th dimension. The estimate is given by

$$C_p = \frac{1}{n}\sum_{i=1}^{n}\left(Y_i - \sum_{j=1}^{p}\hat{f}_j(X_j)\right)^2 + \frac{2\hat{\sigma}^2}{n}\mathrm{df}(\lambda).$$

B PROOF OF THEOREM 4.1

This section provides the detailed proof of Theorem 4.1, which establishes the equivalence between vanishing effect footprints and the disappearance of the first–order projection in the Hoeffding–Sobol decomposition. The result clarifies when a variable contributes only through higher–order interactions and thus leaves no detectable marginal footprint.

We begin with the Hoeffding–Sobol decomposition. Let $f(\mathbf{X}_\mathcal{I})$ be a centered function, i.e., $\mathbb{E}[f(\mathbf{X}_\mathcal{I})] = 0$. Then f admits the unique expansion

$$f(\mathbf{X}_\mathcal{I}) = f_{\{j\}}(X_j) + \sum_{S\subseteq\mathcal{I},\, j\in S,\, |S|\geq 2} f_S(\mathbf{X}_S) + \sum_{S\subseteq\mathcal{I},\, j\notin S,\, |S|\geq 1} f_S(\mathbf{X}_S),$$

where the components f_S are mutually orthogonal in L^2, each has mean zero, and $f_{\{j\}}(X_j)$ represents the unique first–order contribution of X_j. The remaining terms correspond either to higher–order interactions involving X_j or to effects of variables not involving X_j.

Conditional expectation with respect to X_j is the orthogonal projection of f onto the subspace of L^2 functions of X_j, as ensured by the Doob–Dynkin lemma and the Hilbert projection theorem. Hence the footprint $m_j(X_j) = \mathbb{E}[f(\mathbf{X}_\mathcal{I}) \mid X_j]$ coincides with this projection. By uniqueness of the Hoeffding–Sobol components, this projection is exactly $f_{\{j\}}(X_j)$. The two directions now follow. If $f_{\{j\}}$ vanishes identically, then conditioning the decomposition on X_j eliminates all other terms: for S not containing j, centeredness of f_S implies $\mathbb{E}[f_S(\mathbf{X}_S) \mid X_j] = 0$, while for S containing j with $|S| \geq 2$, orthogonality ensures $\mathbb{E}[f_S(\mathbf{X}_S) \mid X_j] = 0$. Thus $m_j(X_j) = 0$, which is constant, so X_j leaves no footprint. Conversely, if $m_j(X_j)$ is constant almost surely, then $\mathbb{E}[f(\mathbf{X}_\mathcal{I}) \mid X_j]$ is identically zero because f is centered. Since this conditional expectation is the projection of f onto the space of functions of X_j, it follows that $f_{\{j\}}(X_j) \equiv 0$.

Therefore, the footprint $m_j(x)$ is constant if and only if the first–order projection $f_{\{j\}}(X_j)$ vanishes. In this case, the variable X_j contributes only through higher–order interactions, and its marginal influence disappears in expectation, thereby proving Theorem 4.1.

C Conditions and Proof of Theorem 4.2

This section establishes the effect-level selection consistency of SDAMI. We begin by introducing the technical assumptions that govern the noise, design structure, signal strength, and basis expansion. These conditions provide the foundation for analyzing the group-lasso estimator used in SDAMI and for verifying the primal–dual witness construction that guarantees selection consistency.

Assumption C.1 (Conditions for effect-level selection)**.**

(A1) (*Noise*) The errors ϵ_i in the true function (1) of the main paper are sub-Gaussian with mean zero and variance proxy σ^2.

(A2) (*Within-group orthonormality*) For each main effect j,

$$\frac{1}{n}\Phi_j^\top \Phi_j = I,$$

and for the interaction block $\Phi_\mathcal{I}$,

$$\frac{1}{n}\Phi_\mathcal{I}^\top \Phi_\mathcal{I} = I, \qquad \frac{1}{n}\Phi_\mathcal{I}^\top \Phi_j = 0 \ \ (j \in \mathcal{I}).$$

(A3) (*Block coherence*) For $g \neq g'$,

$$\left\| \tfrac{1}{n} X_g^\top X_{g'} \right\|_{\text{op}} \leq \mu < 1,$$

where X_g denotes the block of design columns for group g.

(A4) (*Restricted eigenvalue*) The Gram matrix on the active set

$$\Sigma_{A^\star A^\star} = \tfrac{1}{n} X_{A^\star}^\top X_{A^\star}, \qquad A^\star = \mathcal{M} \cup \{\mathcal{I}\},$$

satisfies $\lambda_{\min}(\Sigma_{A^\star A^\star}) \geq \kappa_{\min} > 0$ and the method for constructing the Gram matrix is defined in assumption (A7).

(A5) (*Irrepresentability*) There exists $\eta > 0$ such that

$$\|\Sigma_{A^{\star c} A^\star} \Sigma_{A^\star A^\star}^{-1}\|_{2,\infty} \leq 1 - \eta.$$

(A6) (*Signal strength*) With group weights $w_g \in [1, C_w]$ and tuning parameter $\lambda_n \asymp \sigma\sqrt{\frac{\log G}{n}}$ (where G is the number of candidate groups),

$$\min_{j \in \mathcal{M}} \|f_j\| \geq c_0 \lambda_n, \qquad \|f_\mathcal{I}\| \geq c_0 \lambda_n \ \text{ if the interaction is present,}$$

for some $c_0 > 2/\eta$.

(A7) (*Finite orthonormal basis representation*) Each function f_j and the interaction $f_\mathcal{I}$ is represented in an orthonormal basis expansion of finite dimension (at most quadratic order), with corresponding design blocks Φ_j and $\Phi_\mathcal{I}$.

Having specified the assumptions, we now turn to the proof. The role of (A7) is to provide a finite orthonormal basis representation of all effects, which allows us to formulate the regression problem as a finite-dimensional block group-lasso. Assumptions (A1)–(A6) then control the noise, dependence, eigenstructure, and signal strength needed to verify that the primal–dual witness construction recovers the correct support with probability tending to one.

By (A7), each main effect f_j and the interaction $f_{\mathcal{I}}$ admits a finite-dimensional orthonormal basis representation, say

$$f_j(x_j) = \Phi_j(x_j)^\top \beta_j, \qquad f_{\mathcal{I}}(\mathbf{X}_{\mathcal{I}}) = \Phi_{\mathcal{I}}(\mathbf{X}_{\mathcal{I}})^\top \gamma,$$

where $\Phi_j \in \mathbb{R}^{n\times m_j}$ and $\Phi_{\mathcal{I}} \in \mathbb{R}^{n\times m_{\mathcal{I}}}$ collect the basis evaluations across n samples. Stacking these blocks gives the design matrix

$$X = [X_1, \dots, X_k, X_{\mathcal{I}}], \qquad X_j := \Phi_j, \quad X_{\mathcal{I}} := \Phi_{\mathcal{I}},$$

with block coefficient vector $\theta = (\beta_1, \dots, \beta_k, \gamma)$. The true active set is $A^\star = \mathcal{M} \cup \{\mathcal{I} : f_{\mathcal{I}} \neq 0\}$ and the inactive set is $I^\star = \mathcal{G} \setminus A^\star$, where $\mathcal{G}$ denotes all candidate groups.

The SDAMI estimator solves the block group-lasso problem

$$\widehat{\theta} \in \arg\min_{\theta} \ \frac{1}{2n}\|y - X\theta\|_2^2 + \lambda_n \sum_{g\in\mathcal{G}} w_g \|\theta_g\|_2,$$

with tuning parameter $\lambda_n \asymp \sigma\sqrt{\frac{\log G}{n}}$ and group weights $w_g \in [1, C_w]$. The associated KKT conditions are

$$\frac{1}{n} X_g^\top (y - X\widehat{\theta}) = \lambda_n w_g \widehat{z}_g, \qquad \|\widehat{z}_g\|_2 \le 1, \qquad \widehat{z}_g = \frac{\widehat{\theta}_g}{\|\widehat{\theta}_g\|_2} \ \text{ if } \widehat{\theta}_g \neq 0.$$

Assumption (A1) ensures that the error vector ε is sub-Gaussian. By a union bound over all blocks and coordinates, with probability $1 - o(1)$ the event

$$\max_{g\in\mathcal{G}} \frac{1}{n}\|X_g^\top \varepsilon\|_2 \le \tfrac{1}{2}\lambda_n w_g$$

holds, providing high-probability control of noise terms in the KKT system. Assumptions (A2) and (A3) impose within-block orthonormality and block coherence, ensuring that $\Sigma = X^\top X/n$ has bounded eigenvalues and limited inter-block correlations. Assumption (A4) states a restricted eigenvalue condition, which guarantees that for any deviation vector $\Delta_{A^\star}$ supported on the active set,

$$\frac{1}{n}\|X_{A^\star}\Delta_{A^\star}\|_2^2 \ge \kappa_{\min}\|\Delta_{A^\star}\|_2^2.$$

Assumption (A5) provides the irrepresentability condition, ensuring that inactive blocks cannot mimic active ones in the dual constraints. Finally, assumption (A6) requires minimal signal strength $\|f_g\| \ge c_0\lambda_n$ on all active blocks, so that true coefficients dominate the estimation error.

Under these conditions, the restricted problem on $A^\star$ yields an estimator $\widehat{\theta}_{A^\star}$ with error bound

$$\|\widehat{\theta}_{A^\star} - \theta^\star_{A^\star}\|_2 \le \frac{3\lambda_n}{\kappa_{\min}}\Big(\sum_{g\in A^\star} w_g^2\Big)^{1/2}.$$

Because $c_0 > 2/\eta$, this error is asymptotically smaller than the true signal size, ensuring $\widehat{\theta}_g \neq 0$ for all $g \in A^\star$. Thus, no active block is missed. For inactive groups, the dual feasibility condition requires $\frac{1}{n}\|X_g^\top(y - X_{A^\star}\widehat{\theta}_{A^\star})\|_2 < \lambda_n w_g$. The residual expands as $\widehat{r} = \varepsilon - X_{A^\star}(\widehat{\theta}_{A^\star} - \theta^\star_{A^\star})$. The first term is controlled by (A1), while the second is bounded by (A3) and (A5) together with the error rate above. Consequently, inactive groups satisfy strict dual feasibility, forcing $\widehat{\theta}_g = 0$ for all $g \in I^\star$. This establishes absence of false positives.

For the interaction, if $f_{\mathcal{I}} = 0$, then $\mathcal{I} \in I^\star$ and the dual condition implies $\widehat{f}_{\mathcal{I}} = 0$. If $f_{\mathcal{I}} \neq 0$, then $\mathcal{I} \in A^\star$ and the signal strength bound ensures $\widehat{f}_{\mathcal{I}} \neq 0$. Combining all pieces, with probability tending to one we have

$$\{j : \widehat{f}_j \neq 0\} = \mathcal{M}, \qquad \widehat{f}_{\mathcal{I}} \neq 0 \ \Leftrightarrow \ f_{\mathcal{I}} \neq 0,$$

which proves the effect-level selection consistency of SDAMI as stated in Theorem 4.2.

D CONDITIONS AND PROOF OF THEOREM 4.3

To ground the proof, we first specify the SDAMI function class and estimator used throughout.

Model class of SDAMI. Let $A \subseteq \{1, \dots, p\}$ index a subset of active main effects and interactions. For each main effect $j \in A_{\text{main}}$ and interaction $\mathcal{I} \in A_{\text{int}}$, let $\mathcal{N}_{L,W,B}$ denote the class of feedforward ReLU subnetworks of depth L and maximal width W whose parameters satisfy a norm constraint (e.g., path norm, spectral norm, or ℓ_2 decay) bounded by B. For a growth schedule (L_n, W_n, B_n), define the SDAMI sieve over A by

$$\mathcal{F}_n^{\text{SDAMI}}(A) = \left\{ f(x) = \sum_{j \in A_{\text{main}}} g_j(x_j) \;+\; h_{\mathcal{I}}(x_{\mathcal{I}}) \;:\; g_j \in \mathcal{N}_{L_n,W_n,B_n},\;\; h_{\mathcal{I}} \in \mathcal{N}_{L_n,W_n,B_n} \right\}.$$

Thus SDAMI is an additive model with interactions, where each component is realized by a subnetwork from $\mathcal{N}_{L_n,W_n,B_n}$ restricted to its own argument(s).

Assumptions.

(B1) *Sampling, noise, and approximation.* The data $(\mathbf{X}_i, Y_i)_{i=1}^n$ are i.i.d. from model (1) in the main content with ϵ_i satisfying $E[\epsilon_i] = 0$ and $\text{Var}(\epsilon_i) = \sigma^2 < \infty$. The covariates $\mathbf{X}$ have either bounded support or sub-Gaussian tails, and the true regression function $f^\star \in L_2(P_{\mathbf{X}})$ lies in the $L_2(P_{\mathbf{X}})$-closure of the sieve

$$\bigcup_{n=1}^{\infty} \mathcal{F}_n^{\text{SDAMI}}(A),$$

so that for any $\varepsilon > 0$ there exists n and $f \in \mathcal{F}_n^{\text{SDAMI}}(A)$ with $\|f - f^\star\|_{L_2(P_{\mathbf{X}})} \le \varepsilon$.

(B2) *Effect-level selection consistency (SDAMI).* Let $A^\star$ be the true set of active main effects and interactions. Then $\mathbb{P}(\widehat{A}_n = A^\star) \to 1$.

(B3) *Approximation (DNN sieve over true inputs).* For the restricted DNN class $\mathcal{F}_n^{\text{DNN}}(A^\star)$ with schedule (L_n, W_n, B_n), the sieve approximation error vanishes:

$$\alpha_n \;:=\; \inf_{f \in \mathcal{F}_n^{\text{DNN}}(A^\star)} P\big[(f - f^\star_{A^\star})^2\big] \;\longrightarrow\; 0.$$

(B4) *Empirical risk minimization up to tolerance.* The trained $\widehat{f}_n \in \mathcal{F}_n^{\text{DNN}}(\widehat{A}_n)$ satisfies

$$P_n\big[(\widehat{f}_n - Y)^2\big] \;\le\; \inf_{f \in \mathcal{F}_n^{\text{DNN}}(\widehat{A}_n)} P_n\big[(f - Y)^2\big] \;+\; \delta_n, \qquad \delta_n \downarrow 0.$$

(B5) *Capacity control and uniform generalization.* The norm constraint B_n (and/or width W_n) ensures a vanishing complexity for squared loss:

$$\mathfrak{R}_n(\mathcal{L}_n) \;=\; o(1), \qquad \mathcal{L}_n := \{(f-g)^2 : f \in \mathcal{F}_n^{\text{DNN}}(A),\, g \in \mathcal{F}_n^{\text{DNN}}(A),\, A \subseteq \{1, \dots, p\}\},$$

so that

$$\sup_{h \in \mathcal{L}_n} \big|(P - P_n)h\big| \;=\; o_p(1).$$

(B6) (*Measurability and uniform L_2 envelope*) Each $f \in \mathcal{F}_n^{\text{SDAMI}}(A)$ is measurable, and there exists a constant $M < \infty$ (independent of n, A, and f) such that

$$\sup_{A \subseteq [p]} \sup_{f \in \mathcal{F}_n^{\text{SDAMI}}(A)} Pf^2 \;\le\; M.$$

In particular, for the data–dependent active set $\widehat{A}_n$, the trained $\widehat{f}_n \in \mathcal{F}_n^{\text{SDAMI}}(\widehat{A}_n)$ is measurable and satisfies $P\widehat{f}_n^2 \le M$ almost surely. Hence $\{P\ell(\widehat{f}_n)\}_n$ is uniformly integrable.

With the SDAMI sieve $\mathcal{F}_n^{\text{SDAMI}}(\widehat{A}_n)$ specified and assumptions (B1)–(B6) in place, we now prove Theorem 4.3 by analyzing the empirical minimizer within this class and translating vanishing risk into prediction convergence.

Let P denote expectation with respect to $P_{\mathbf{X}}$ and P_n the empirical average over the training inputs. Write the squared excess prediction loss as $\ell(f) := (f - f^\star)^2$. By the selection consistency of SDAMI (B2), $\mathbb{P}(\widehat{A}_n = A^\star) \to 1$, so it suffices to analyze $\widehat{f}_n \in \mathcal{F}_n^{\mathrm{SDAMI}}(A^\star)$ and the conclusions will then hold unconditionally. Using the empirical-to-population decomposition,

$$P\ell(\widehat{f}_n) \;=\; P_n\ell(\widehat{f}_n) \;+\; (P - P_n)\ell(\widehat{f}_n).$$

To control $P_n\ell(\widehat{f}_n)$, expand the empirical squared loss around $Y = f^\star + \epsilon$:

$$P_n\big[(\widehat{f}_n - Y)^2\big] = P_n\ell(\widehat{f}_n) \;+\; P_n[\epsilon^2] \;+\; 2\,P_n\big[(f^\star - \widehat{f}_n)\epsilon\big].$$

By the empirical optimality up to tolerance (B4), for any $f \in \mathcal{F}_n^{\mathrm{SDAMI}}(A^\star)$,

$$P_n\ell(\widehat{f}_n) \;\le\; P_n\ell(f) \;+\; 2\left|P_n\big[(f^\star - \widehat{f}_n)\epsilon\big]\right| \;+\; 2\left|P_n\big[(f^\star - f)\epsilon\big]\right| \;+\; \delta_n.$$

The noise is centered with bounded conditional variance (B1) and the SDAMI sieve is capacity–controlled (B5), hence the stochastic inner products above are $o_p(1)$ uniformly over $f \in \mathcal{F}_n^{\mathrm{SDAMI}}(A^\star)$ by standard symmetrization/contraction bounds for squared loss. Taking the infimum over $f \in \mathcal{F}_n^{\mathrm{SDAMI}}(A^\star)$ yields

$$P_n\ell(\widehat{f}_n) \;\le\; \inf_{f \in \mathcal{F}_n^{\mathrm{SDAMI}}(A^\star)} P_n\ell(f) \;+\; o_p(1) \;+\; \delta_n.$$

Adding and subtracting population risks and invoking the uniform generalization bound for squared loss from (D5),

$$P\ell(\widehat{f}_n) \;\le\; \inf_{f \in \mathcal{F}_n^{\mathrm{SDAMI}}(A^\star)} P\ell(f) \;+\; o_p(1) \;+\; \delta_n.$$

By the approximation property of the SDAMI sieve on the true inputs (B3), the approximation error $\alpha_n := \inf_{f \in \mathcal{F}_n^{\mathrm{SDAMI}}(A^\star)} P\ell(f)$ satisfies $\alpha_n \to 0$; therefore

$$P\ell(\widehat{f}_n) \;\xrightarrow{p}\; 0. \tag{4}$$

To convert result (4) into prediction convergence, note the inequality

$$\mathbf{1}\Big\{\big|\widehat{f}_n(\mathbf{X}) - f^\star(\mathbf{X})\big| \ge \varepsilon\Big\} \;\le\; \frac{\ell(\widehat{f}_n)(\mathbf{X})}{\varepsilon^2}, \qquad \varepsilon > 0.$$

Taking expectation over $\mathbf{X}$ and then over the training sample gives

$$\mathbb{P}\Big(\big|\widehat{f}_n(\mathbf{X}) - f^\star(\mathbf{X})\big| \ge \varepsilon\Big) \;\le\; \frac{\mathbb{E}\big[P\ell(\widehat{f}_n)\big]}{\varepsilon^2}.$$

The sieve's norm constraints together with (B6) imply a square–integrable envelope on $\mathcal{F}_n^{\mathrm{SDAMI}}(A^\star)$, hence $\{P\ell(\widehat{f}_n)\}_n$ is uniformly integrable; combined with result (4) this yields $\mathbb{E}[P\ell(\widehat{f}_n)] \to 0$. Consequently,

$$\mathbb{P}\Big(\big|\widehat{f}_n(\mathbf{X}) - f^\star(\mathbf{X})\big| \ge \varepsilon\Big) \;\longrightarrow\; 0 \quad \text{for every fixed } \varepsilon > 0,$$

i.e., $\widehat{f}_n(\mathbf{X}) \xrightarrow{p} f^\star(\mathbf{X})$ at the design distribution $P_{\mathbf{X}}$. □

E Supplementary Material for Neural Additive Models

In this section, we summarize the detail of cross validation on architecture selection, additional experiment results, and the visualization of either main effects or interactions effects from the numerical studies.

E.1 SDAMIs on Numerical Studies

We summarize the cross validation on configuration selection for SDAMI and DNN in Table 5 where $(1) = [8, 6, 3]$, $(2) = [12, 10, 6]$, and $(3) = [15, 12, 10]$ represent the hidden layers. The optimal architecture for the Simulation studies is (3) achieving lowest MSE.

Method	SDAMI(1)		SDAMI*(2)		SDAMI(3)		DNN(1)		DNN*(2)		DNN(3)	
	MSE↓	STD↓	MSE↓	STD↓	MSE↓	STD↓	MSE↓	STD↓	MSE↓	STD↓	MSE↓	STD↓
Case 1	2.64	3.23	**0.43**	0.65	0.48	0.71	14.11	0.71	14.10	0.73	13.95	0.68
Case 2	0.94	1.04	0.38	0.62	**0.29**	0.56	5.31	0.40	5.26	0.31	5.23	0.30
Case 3	1.20	1.21	0.46	0.63	**0.29**	0.45	5.76	0.39	5.70	0.32	5.62	0.26
Case 4	0.94	1.03	0.34	0.55	**0.32**	0.58	7.07	0.48	6.98	0.38	6.95	0.36
Case 5	0.72	0.90	**0.35**	0.58	0.37	0.65	5.80	0.38	5.78	0.35	5.74	0.36
Case 6	0.33	0.23	**0.25**	0.21	0.25	0.21	1.03	0.17	0.99	0.20	0.37	0.19

Table 5: Performance of SDAMs and DNNs with respect to different configuration when $n = 300$.

Method	SDAMI		DNN		fSpAM		LASSO	
	MSE↓	STD↓	MSE↓	STD↓	MSE↓	STD↓	MSE↓	STD↓
Case 1	**0.48**	0.71	14.11	0.71	5.57	0.31	3.32	0.24
Case 2	**0.29**	0.56	5.31	0.40	3.04	0.16	2.54	0.17
Case 3	**0.29**	0.45	5.76	0.39	3.37	0.15	2.97	0.20
Case 4	**0.32**	0.58	7.07	0.48	3.32	0.17	2.80	0.16
Case 5	**0.37**	0.65	5.80	0.38	3.45	0.16	2.98	0.20
Case 6	**0.25**	0.21	1.03	0.17	0.60	0.03	0.43	0.032

Table 6: Performance comparison on SDAMI, DNN, fSpAM, and LASSO by case type when $n = 300$.

E.2 Additional experiment results

The performance comparison among different machine learning model is demonstrated in Table 6 7. Also, Table 8 results for additional numerical experiments with different sample size and corresponding TPR/ FPR are demonstrated in the following block.

E.3 Visualization for each case

In the section, we demonstrate the visualization of either main effects or interaction among each cases where the visualization result for Case 3 can be found in Figure 3. In Case 1 - 5, the SDAMI can capture both linearity and nonlinearity underlying the true model. In the interaction-existed cases, we can observes the SDAMI can still depict the response surface to approximate the underlying higher-order effects.

In this section, we demonstrate visualizations of the component functions representing either main effects or interactions across different cases. For Cases 1 through 5, SDAMI successfully captures both the linear and nonlinear structures underlying the true models. In cases involving interactions, we observe that SDAMI effectively depicts the response surfaces, accurately approximating the underlying higher-order effects. These visualizations provide valuable insights into the model's interpretability and can be found in detail in the Figure 6, 7.

F Real Data Analysis

This section illustrates the additional experiment on two other real datasets including the parameter settings and corresponding explanation on the visualization.

F.1 Surrogate Modeling of Product Lifetime Modeling

This subsection showcases the application of SDAMI in evaluating prediction performance, positioning it as an effective surrogate technique— a key approach in the field of computer experiments (Santner et al., 2019; Wu & Hamada, 2011). Surrogate modeling serve as statistical approximations of computationally intensive simulations, facilitating the efficient study of complex system dynamics.

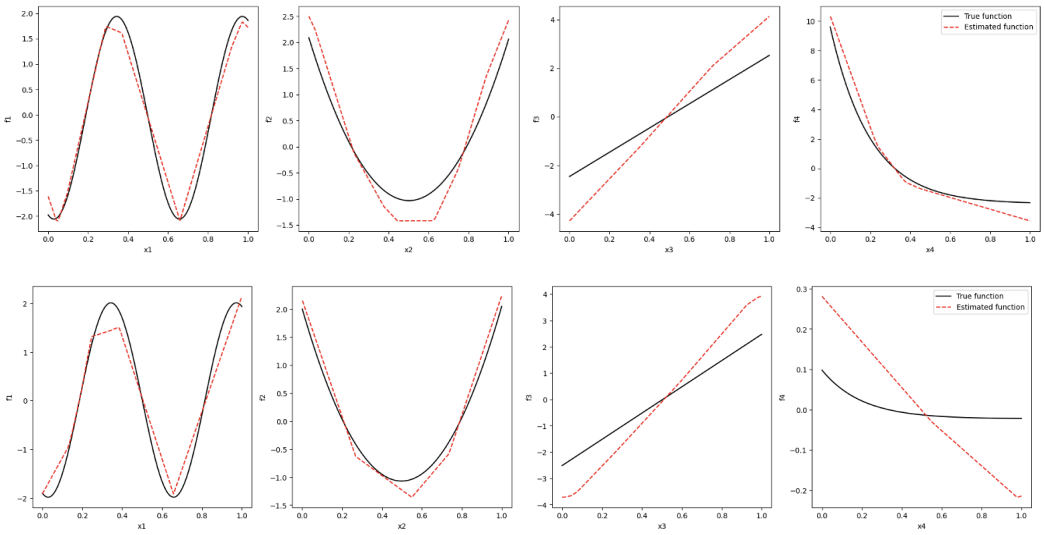

Figure 6: The estimated (red dashed lines) versus true additive component functions (solid black lines) for four main effects for (Upper panel) Case (1) and (Lower panel) Case (2).

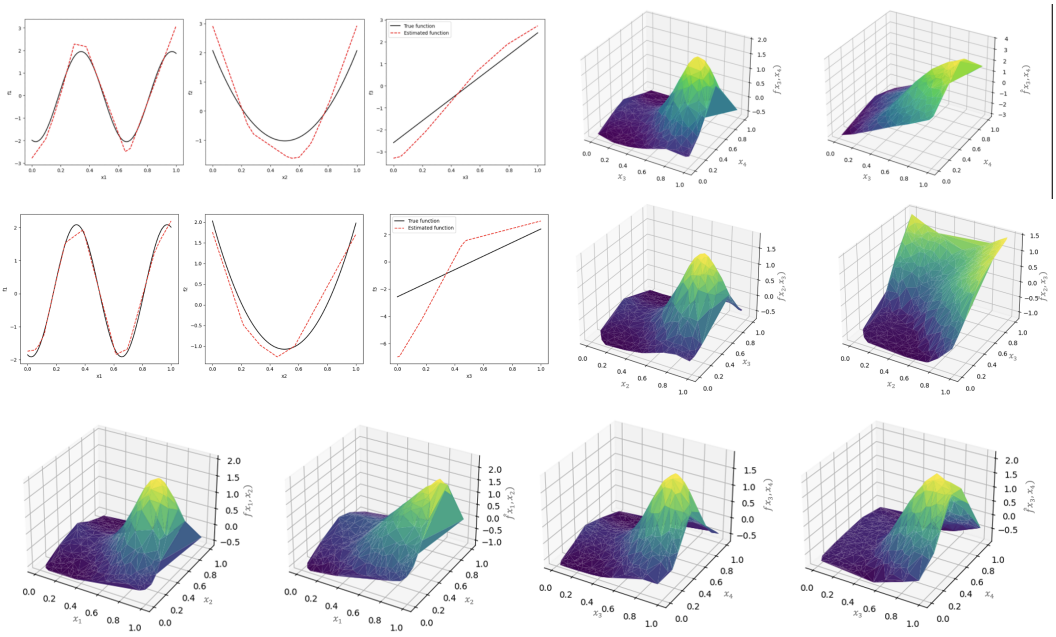

Figure 7: (Upper panel: Case (4); middle panel: Case (5)) The three figures on the left: Estimated (red dashed lines) versus true additive component functions (solid black lines) for three main effects; the two figures on the far right: the first shows the true response surface for interaction, and the second shows its estimated response surface. (Lower panel: Case (6)) The first and third shows the true response surface for interactions, and the second and fourth shows corresponding estimated response surface.

Method	SDAMI		DNN		fSpAM		LASSO	
	MSE↓	STD↓	MSE↓	STD↓	MSE↓	STD↓	MSE↓	STD↓
Case 1	**0.23**	0.63	13.89	0.82	5.43	0.28	3.04	0.17
Case 2	**0.21**	0.46	5.33	0.35	2.98	0.14	2.40	0.11
Case 3	**0.28**	0.37	5.78	0.32	3.33	0.16	2.72	0.14
Case 4	**0.17**	0.21	7.14	0.50	3.24	0.15	2.61	0.13
Case 5	**0.22**	0.19	5.82	0.39	3.41	0.15	2.76	0.13
Case 6	**0.14**	0.18	1.06	0.13	0.59	0.03	0.39	0.02

Table 7: Performance comparison on SDAMI, DNN, fSpAM, and LASSO by case type when $n = 450$.

Method	SDAMI		LASSONET		SODA	
	TPR↑	FPR↓	TPR↑	FPR↓	TPR↑	FPR↓
Case 1	1.0000 (-)	1.1×10^{-5} (-)	0.6100 (0.1241)	0.0129 (0.0091)	0.03 (0.0964)	5×10^{-4} (2×10^{-4})
Case 2	1.0000 (-)	1.1×10^{-5} (-)	0.4550 (0.1083)	0.0168 (0.0083)	0.02 (0.0685)	4×10^{-4} (2×10^{-4})
Case 3	0.7500 (-)	10^{-4} (10^{-5})	0.0200 (0.0980)	0.0451 (0.0418)	0.015 (0.06)	6×10^{-4} (4×10^{-4}
Case 4	0.7600 (0.0490)	10^{-4} (10^{-5})	0.0100 (0.0700)	0.0367 (0.0176)	0.025 (0.0758)	5×10^{-4} (2×10^{-4})
Case 5	0.7525 (0.0249)	10^{-4} (10^{-5})	0.0100 (0.0700)	0.0390 (0.0183)	0.03 (0.0821)	5×10^{-4} (2×10^{-4})
Case 6	0.6100 (0.0436)	10^{-4} (10^{-5})	0.0200 (0.0980)	0.0519 (0.0658)	- (-)	5×10^{-4} (2×10^{-4})

Table 8: Mean (standard deviation) of TPR and FPR over 100 simulations from SDAMI, LASSONET, SODA when $n = 300$ where $(-)$ indicates value $< 1e^{-5}$.

We illustrate this with the analysis of electronic device lifetimes, which can fail due to mechanisms such as front-end fate oxide breakdown (FEOL TDDB) (Yang et al., 2017). This failure occurs when traps accumulate in the gate oxide layer from electrical and thermal stress during operation, eventually creating conductive paths leading to device malfunction. The lifetime distribution for these components is captured by the following function, as characterized in prior work (Hsu et al., 2020):

$$S(t) = \exp\left(-\left(\frac{t}{\mathrm{A_{FEOL}}(\mathrm{WL})^{-\frac{1}{\beta}} e^{-\frac{1}{\beta}} V^{a+bT} \exp\left(\frac{cT+d}{T^2}\right) s^{-1}}\right)^{\beta}\right), \tag{5}$$

where the inputs include process-dependent constants $\mathrm{A_{FEOL}}, a, b, c, d$, voltage V and temperature T, width W and length L of the device, the probability of stress s, and shape parameter β describing failure progression over time.

Although simulating such experiments is straightforward, accurately extracting main and higher-order effects under data sparsity requires sophisticated and interpretable modeling. To that end, we employ the *MaxPro design* (Joseph et al., 2015) to generate space-filling experiments spanning all input factors, with details in Table 9. The dataset includes 100 observations with 9 covariates, augmented by 21 irrelevant noise features randomly sampled uniformly within $[0, 1)$ to test model sparsistency and interaction detection. The log-transformation of the true model is given by

$$\log(\eta) = \log(\mathrm{A}_{FEOL}) - \frac{1}{\beta}\log(\mathrm{WL}) - \frac{1}{\beta} + (a + bT)\log(V) + (\frac{cT+d}{T^2}) - \log(s), \tag{6}$$

where s is constant and η corresponds to a 63% failure quantile from the generalized Wei-bull model (5). This representation admits an additive decomposition involving univariate and bi-variate functions 6,

$$y = \alpha + \sum_i f_i(x_i) + \sum_{i \neq j} f_{ij}(x_i, x_j) + \cdots + \epsilon.$$

allowing comprehensive identification of relevant main and interaction effects. Table 4 presents the comparative performance of various techniques, including SDAMI with and without interactions, DNN, LASSO, LASSONET, and fSpAM, demonstrating SDAMI's prominence in recovering complex dependency structures in sparse, high-dimensional settings.

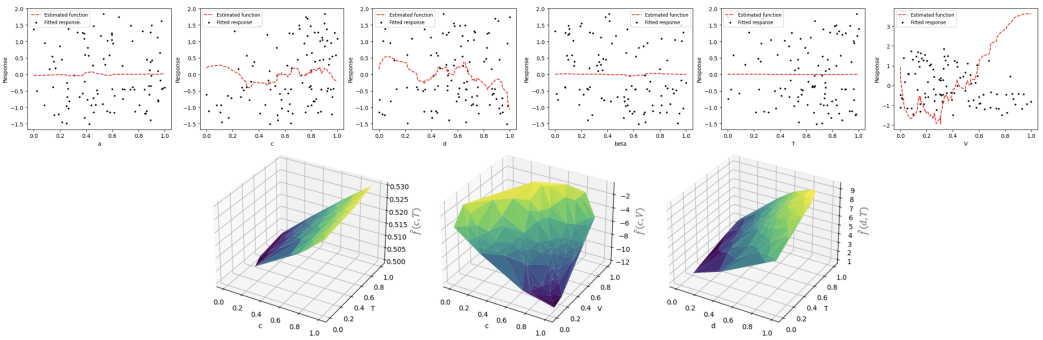

Figure 8: (Upper) Main effects and (Lower) Interaction for Chip Data (Chip Dataset) The six figures on the top panel: Predicted marginal response of target with respect to main effect features (solid black dots) versus estimated additive component functions (red dashed lines) for six main effects; the three figures on the lower panel: Estimated response surface for interactions.

Given the visualization of effects from Figure 8, we can observe that the contribution of main effect is relatively weak. Besides, the interaction have obvious effect on response. To be more specific, when c, T and d, T goes up, the response will increase. when c goes down and V goes up, the response will increase. These phenomenon is predictable because in Equation 6, the higher-order effects are dominant over main effects but the main effects still exist due to its marginal effect on the response.

Parameter	Lower	Upper
a	-81.9	-74.1
b	7.69×10^{-2}	8.51×10^{-2}
c	8.37×10^{3}	9.25×10^{3}
d	-8.14×10^{5}	-7.33×10^{5}
β	1.476	1.804
V	1.2	1.3
T	120	180
WL	4×10^{-4}	6×10^{-4}
A_{FEOL}	4.75×10^{-7}	5.25×10^{-7}
s	1	1

Table 9: Parameter table for generating space-filling experiment on MOSFET device

F.2 Diabetes Response Prediction

For this analysis, we utilize the well-known diabetes dataset from the scikit-learn library, which contains 442 observations and ten baseline covariates. These features capture key demographic and physiological measurements, such age (in years), sex (0: female, 1: male), body mass index (BMI), mean arterial blood pressure, and six standardized blood serum variables known to be relevant for diabetes progression. The target variable is a quantitative measure of disease progression observed one year after baseline, making the dataset suitable for regression modeling and biomarker analysis.

To thoroughly evaluate sparse additive modeling methods under high-dimensional constraints, we purposefully restrict the sample size to $n = 200$ and augment the original dataset with 40 synthetic covariates, each drawn independently from a uniform distribution on the interval $[0, 1)$ distribution. These additional features are explicitly designed to act as non-informative noise, challenging each model's ability to discern relevant predictors. Thus, the expanded dataset includes 50 covariates in total, with the genuine signal confined to the original ten baseline measurements. Standard preprocessing, including normalization and scaling of all features, is performed to ensure comparability and numerical robustness in downstream modeling. This controlled, high-dimensional experimental

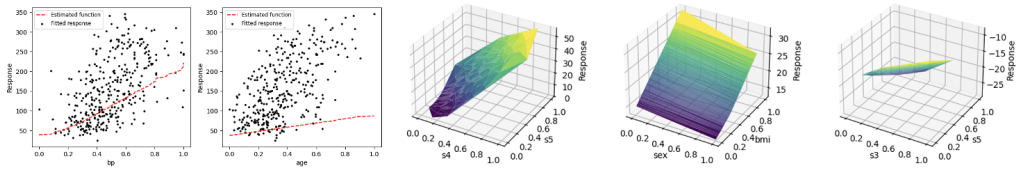

Figure 9: (Diabetes Dataset) The two figures on the left: Predicted marginal response of target with respect to main effect features (solid black dots) versus estimated additive component functions (red dashed lines) for two main effects; the three figures on the far right: Estimated response surface for interactions.

setup provides a rigorous testbed for assessing the sensitivity and variable selection performance of SpAM, and other advanced machine learning algorithms in biomedical contexts.

Visualization of the estimated effects in Figure 9 reveals several interpretable patterns. Both blood pressure and age exhibit a positive association with the disease progression target. Notably, the interaction between total serum cholesterol and the logarithm of serum triglycerides levels further enhances the predictive signal. Across fixed levels of high-density lipoproteins, a higher serum triglycerides value also contributes to increased disease progression. Additionally, the impact of BMI on the response is consistent across both genders, manifesting as a monotonic relationship. The observed relationships align well with clinical expectations and domain knowledge.

Table 4 summarizes model performance for diabetes response prediction. SDAMI with interaction modeling consistently outperform alternative machine learning methods, offering superior predictive accuracy alongside enhanced interpretability thanks to its explicit feature selection and effect visualization capabilities.