ON THE NECESSITY OF NON-ADDITIVE MODELING FOR INTERACTION PREDICTION: THEORY AND APPLICATION TO DRUG SYNERGY

Anonymous authors
Paper under double-blind review

ABSTRACT

Many prediction tasks require combining separately observed inputs, such as drug molecules and a cell line, imaging modalities, or sensor channels, whose joint effect cannot be captured by summing independent contributions. Whether interaction modeling is theoretically necessary in such settings, or merely empirically useful, has remained an open question. We prove a structural necessity theorem for interaction prediction: any model in the functionally additive class, regardless of the expressiveness of its per-condition components, suffers an irreducible prediction gap when the target contains genuine cross-condition interaction terms. We derive an explicit, strictly positive lower bound for this gap that depends only on the magnitude of the target's interaction component, thereby characterizing a limitation of the function class rather than of any particular architecture. We instantiate this result in drug synergy prediction, where the standard null models of pharmacological interaction are additive by construction, making the necessity gap concrete and measurable. We introduce I_ψ, a structured bilinear operator that computes pairwise drug interactions as a low-rank product of individual compound embeddings, instantiating the non-additive class the theorem calls for. On the Therapeutics Data Commons (TDC) DrugComb benchmark, I_ψ closes the predicted gap across five synergy scores under a common training and evaluation protocol, with each model receiving its own hyperparameter sweep. I_ψ uses about a quarter of the parameter budget of FiLM, adapted here for drug synergy, and still exceeds it at its full published size on the combo split. On an independent benchmark (OncopolyPharmacology) the ordering holds. An ablation confirms the gain comes from the bilinear term itself. Hard generalization splits confirm the theorem's boundary condition.

1 INTRODUCTION

Combining two drugs can produce an effect that neither drug produces alone. Predicting this outcome would save years of wet-lab screening, but the combinatorial space of drug pairs across cancer cell lines is too large to test exhaustively; a reliable model has direct practical value independent of any architectural novelty.

The natural way to build such a model is to condition a predictor on a representation of each drug and of the cell line. Most existing predictors do exactly this, and many add mechanisms (attention, gating, tensor fusion) that let the drugs' representations influence one another before producing a score. These mechanisms consistently outperform predictors that combine drug and cell-line information additively. What is missing from this empirical picture is an answer to *why*: is interaction modeling just one architectural choice among many that happens to work well on this task, or is it *necessary* whenever synergy is genuinely non-additive, regardless of how the additive part is implemented?

Theorem 1 states that any model in the functionally additive class is bounded away from the true conditional distribution by an explicit, strictly positive constant Δ whenever the interaction signature $\sum_{k\geq 2} \Omega_k(f^*)$ is non-zero. The additive class contains any predictor that can be written as a sum of per-condition contributions, regardless of how expressive each contribution is. This is a statement about a *class* of models rather than about any particular architecture: it applies equally to a linear

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.

combination of drug embeddings and to a deep network applied independently to each drug before a final sum.

The theorem alone does not say that interaction modeling helps everywhere. Two non-mathematical preconditions determine when the gap is practically relevant: the inputs must be observed separately (F1), and the target must not be recoverable from a single input (F2). Section 3.1 makes these preconditions explicit and validates them on drug synergy.

Contributions.

- A necessity theorem (Theorem 1) that gives an explicit positive lower bound on the divergence between any functionally additive predictor and the true conditional distribution, whenever the interaction signature is non-zero, together with an applicability filter (F1 separated inputs, F2 inferred target) that identifies problem domains where the theorem is relevant.
- I_ψ, a structured bilinear operator that instantiates the non-additive class the theorem requires and exceeds the strongest published FiLM adaptation for drug synergy at about a quarter of its parameter budget.
- A falsifiability test on hard generalization splits, where the theorem's mathematical precondition breaks. The empirical advantage vanishes there, giving a second independent confirmation of the theorem.
- An interpretability decomposition of the prediction into an additive component and an interaction component, derived from the operator's own structure rather than from a post-hoc attribution method.

2 Related Work

Interaction decompositions. The interaction signature Ω_k builds on the functional ANOVA decomposition of Hoeffding (1948) and its extension to dependent inputs by Chastaing et al. (2012). Amari (2001) gives an exact orthogonal decomposition of the Kullback-Leibler (KL) divergence between a joint distribution and a reference into per-order contributions: an *identity* over an unlabeled joint distribution. Our Theorem 1 is a different statement: a strict *inequality* lower-bounding the KL gap of a restricted functionally additive class against a true *conditional* law whenever interaction is non-zero. Bilodeau et al. (2024) and Enouen & Liu (2025) prove impossibility results about post-hoc attribution methods on already-trained models, orthogonal to our bound on predictive model classes. Lou et al. (2013) and Kim et al. (2022) augment additive models empirically with pairwise terms; our theorem supplies the missing theoretical justification for their observed gains.

Interaction-aware fusion architectures. Bilinear and tensor-based fusion is well-established: Zadeh et al. (2017) compute full outer products of modality representations; Liu et al. (2018) apply low-rank factorization to reduce the tensor cost; Perez et al. (2018) propose Feature-wise Linear Modulation (FiLM), the mechanism underlying our primary baseline (Jin et al., 2026). All demonstrate empirical gains over additive baselines without proving the gap is unavoidable; our theorem closes this gap.

Drug synergy prediction. Preuer et al. (2018) (DeepSynergy) and Jin et al. (2026) (Env-Syn, our primary baseline) condition predictors on drug and cell-line representations without a theoretically motivated interaction term. Recent works (Guo et al., 2024; Ge et al., 2025; Zhang et al., 2025) report higher numbers under random splits: Ge et al. (2025) drop from 0.919 to 0.754 AUROC under leave-drug-out, motivating our insistence on hard, protocol-matched evaluation. None provides a theoretical necessity result.

3 Theoretical Framework

To ground the formalism before stating the theorem, consider what makes drug synergy prediction a natural instance of the problem it solves. Three inputs are observed separately for each experiment: a molecular fingerprint for the first drug, a molecular fingerprint for the second drug, and a gene-expression profile of the cancer cell line. The question is whether the two drugs, when given

together to that cell line, produce an effect larger than each drug would produce alone. This is an intrinsically joint question: the answer depends on how the two drugs interact with each other and with the cellular environment, not on any single drug in isolation. A model that processes each input independently and sums the results cannot, in principle, represent this joint effect, no matter how expressive each individual processing step is. Theorem 1 makes this observation rigorous: it gives an explicit, strictly positive lower bound on how far any such additive model must remain from the true conditional distribution, expressed in terms of how much genuine interaction the true synergy response contains.

Setup. Let $c = (c_1, \ldots, c_n)$ denote n separately observed conditions (for drug synergy, $n = 3$: two drug representations and a cell-line representation) with joint law $\mu_{\mathcal{C}}$, and let $p_\theta(y \mid x, c)$ be a bounded exponential family model of the target y given c and any auxiliary input x, with natural parameter $\tilde{\eta}(x, c)$. We state and prove Theorem 1 below for the K-class softmax case ($\tilde{\eta} \in \mathbb{R}^{K-1}$, Appendix B); the fixed-variance Gaussian regression model used throughout our experiments is a second, simpler instantiation of the same argument (Corollary 1). Following the higher-order functional decomposition (HOFD) of Chastaing et al. (2012), any $f \in L^2(\mu_{\mathcal{C}})$ admits a unique orthogonal decomposition $f = \sum_{k=0}^{n} f_k$ into terms of interaction order k, and we write $\Omega_k(f) := \sum_{|S|=k} \mathrm{Var}[f_S(c_S)]$ for the total variance carried by order-k terms. We define the class of *functionally* additive models

$$\mathcal{F}^{\mathrm{std}} :=^{\leq 1}= \{h \in L^2(\mu_{\mathcal{C}}) \mid \Omega_k(h) = 0 \; \forall k \geq 2\},$$

i.e. models whose HOFD contains no interaction term of order ≥ 2. This functional definition matters: an *architectural* definition of additivity (a nonlinear function applied to a sum of per-condition encodings) is not equivalent to it, since such a composition can introduce non-zero $\Omega_{\geq 2}$ terms even when no interaction is architecturally explicit. Concatenation followed by an unconstrained MLP is therefore *not* guaranteed to lie in $\mathcal{F}^{\mathrm{std}}$. We write $f^*(x, c) := \mathbb{E}_{p^*}[y \mid x, c]$ for the conditional mean of the true distribution p^*; the interaction decomposition $\Omega_k(f^*)$ is applied to this function.

Theorem 1 (Necessity of interactions)**.** *Assume three regularity conditions: (C.1) absolute continuity of $\mu_{\mathcal{C}}$ with respect to the product of its marginals; (C.2) a Chastaing-style lower bound $p_c(\mathbf{c}) \geq M\, p_{c_u}(\mathbf{c}_u)\, p_{c_{u^c}}(\mathbf{c}_{u^c})$ for some $M \in (0, 1]$ and all $u \subseteq [n]$; (C.3) a bounded reduced natural parameter: $\|\tilde{\eta}^*(x, c)\|_\infty \leq B$ and $\|\tilde{\eta}_h(x, c)\|_\infty \leq B$ for all h considered, $\mu_{\mathcal{C}}$-a.s., for some $B < \infty$. If $\sum_{k \geq 2} \Omega_k(f^*) > 0$, then*

$$\inf_{h \in \mathcal{F}^{\mathrm{std}}} D_{\mathrm{KL}}\big(p^* \,\|\, p_h\big) \;\geq\; \Delta \;>\; 0, \qquad \Delta = c_B\, \ell_g^2\, \delta^{2^n - n - 2} \cdot \frac{\Big(\sum_{k \geq 2} \Omega_k(f^*)\Big)^2}{\|f^*\|^2_{L^2(\mu_{\mathcal{C}})}},$$

where D_{KL} denotes the Kullback-Leibler divergence, $\delta = 1 - \sqrt{1 - M}$, $c_B > 0$ is the minimal eigenvalue of the log-partition function's Hessian over the ball of radius B (halved), and $\ell_g > 0$ is the bi-Lipschitz constant of the canonical link on the corresponding mean-parameter domain.

Proof sketch (full proof in Appendix B). The argument chains three steps, each of which is an equality or a classical bound, not a property of any specific architecture. **Step 1.** Under (C.1)-(C.2), the L^2 projection of f^* onto $^{\leq 1}$ (Chastaing's Hoeffding-Sobol projection) satisfies $\inf_{h \in \mathcal{F}^{\mathrm{std}}} \|f^* - h\|^2_{L^2} \geq \delta^{2^n - n - 2} \sum_{k \geq 2} \Omega_k(f^*)$: the L^2 gap is bounded below by the interaction signature itself. **Step 2.** Under (C.3), the log-partition function of a bounded exponential family is strongly convex on any compact set, which lower-bounds the KL divergence between two members of the family by a constant times the squared distance between their natural parameters; combined with the bi-Lipschitz canonical link, this converts the L^2 gap of Step 1 into a KL gap. **Step 3.** Since $\sum_{k \geq 2} \Omega_k(f^*) \leq \|f^*\|^2_{L^2(\mu_{\mathcal{C}})}$ (HOFD normalization), an algebraic rearrangement expresses the bound as the closed-form constant Δ.

Theorem 1 constrains a *class* of models, not an architecture: any $h \in \mathcal{F}^{\mathrm{std}}$ is bounded away from p^* by Δ, whether h is a linear combination of the conditions or a universal approximator applied to each condition independently before summing. Conversely, nothing in the theorem says that a specific non-additive architecture will close this gap in practice; that is an empirical question we address in Section 5.

Corollary 1 (Regression instantiation)**.** *Take $p_\theta(y \mid x, c) = \mathcal{N}\big(y;\, f_\theta(x, c),\, \sigma^2\big)$ for a fixed $\sigma^2 > 0$: the model class used throughout Section 5, trained by minimizing mean squared error. The KL*

divergence reduces exactly *to a rescaled L^2 distance, $D_{\mathrm{KL}}(p_{f_1}\|p_{f_2}) = \|f_1 - f_2\|^2_{L^2(\mu_C)}/(2\sigma^2)$ (full derivation in Appendix B). Under (C.1)–(C.2):*

$$\inf_{h\in\mathcal{F}^{\mathrm{std}}} \mathbb{E}_{\mu_C}\left[(f^*(x,c) - h(x,c))^2\right] \;\geq\; \delta^{2^n-n-2}\sum_{k\geq 2}\Omega_k(f^*) \;>\; 0,$$

an explicit, strictly positive lower bound on the mean squared error of any functionally additive regressor. This is the exact setting of the synergy score regression task in Section 5.

3.1 The applicability filter

Theorem 1 is conditional: it only guarantees a gap when $\sum_{k\geq 2}\Omega_k(f^*) > 0$ and only characterizes models restricted to $\mathcal{F}^{\mathrm{std}}$. In practice, whether it is worth building an interaction-aware model depends on two further non-mathematical preconditions, which we make explicit as an applicability filter: **(F1)** *separated inputs*: the conditions are observed as distinct entities with no natural joint encoder, rather than as regions of a single input a strong encoder can already see jointly; **(F2)** *inferred, not observed, combination*: the target must be predicted from the conditions, not directly legible from any single condition's raw representation. When (F1) or (F2) fail, a sufficiently expressive model applied to a single joint representation can absorb the interaction on its own, making an explicit interaction operator redundant in practice.

4 The Bilinear Interaction Operator

Theorem 1 says that a model of the form $\sum_i f_i(c_i)$ cannot match the optimum whenever the target contains higher-order interaction terms. To close that gap we need a model that sits outside the additive class. Many choices satisfy this requirement in principle. We use a simple one: keep the additive part, and add an explicit term that mixes two conditions at a time.

For n input conditions there are $n(n-1)$ ordered pairs. Drug synergy has $n = 3$ (two drugs and a cell line), so six pairs. Each pair receives its own bilinear term $B_{ij}(\phi_i, \phi_j)$. The additive part gives back exactly $\mathcal{F}^{\mathrm{std}}$ when every B_{ij} is zero. Setting any single B_{ij} to a non-trivial map takes the model outside $\mathcal{F}^{\mathrm{std}}$.

A full bilinear map from a pair of d_ϕ-dimensional embeddings to d_z output units needs a $d_z \times d_\phi \times d_\phi$ weight tensor. This is prohibitive. We use the standard rank-r factorisation from compact bilinear pooling. Each pair holds three small linear maps U_{ij}, V_{ij}, P_{ij}. The pair term is

$$B_{ij}(\phi_i, \phi_j) \;=\; P_{ij}\big(U_{ij}\phi_i \odot V_{ij}\phi_j\big),$$

where $\odot$ is the elementwise Hadamard product. U_{ij} and V_{ij} project the two inputs to a rank-r space, the product mixes them, and P_{ij} maps the result back to d_z. The cost is $O\big(r(d_\phi + d_\phi + d_z)\big)$ per pair instead of the full tensor.

The reverse pair (j, i) carries a distinct triple (U_{ji}, V_{ji}, P_{ji}), learned independently from (U_{ij}, V_{ij}, P_{ij}), so $B_{ij} \neq B_{ji}$ in general. This lets the operator represent asymmetric interactions when the data calls for them, and reduces to the symmetric case when the two learned maps happen to coincide.

Each ordered pair is weighted by a scalar A_{ij} that gates its contribution. In the main results A is learned jointly with the rest of the model. The ablation in Table 7 shows that fixing A to one, or to a random frozen value, gives the same AUROC. The learned A is a convenience, not the source of the gain.

Training runs standard gradient descent with mean squared error. The gradient of the loss flows through the final non-linearity, splits across the additive maps W_i and the pair maps B_{ij}, and updates U_{ij}, V_{ij}, P_{ij} independently for each pair. No special mechanism is added on top of PyTorch's autograd.

4.1 THE I_ψ OPERATOR

Putting the pieces together, each condition c_i is first mapped to a representation $\phi_i(c_i) \in \mathbb{R}^{d_\phi}$ by a per-condition encoder. The operator combines these representations as

$$I_\psi(c) \;=\; \rho\Big(\underbrace{\textstyle\sum_i W_i\, \phi_i(c_i)}_{\text{additive part, } \mathcal{F}^{\mathrm{std}}} \;+\; \underbrace{\textstyle\sum_{(i,j)} A_{ij}\, B_{ij}\big(\phi_i(c_i), \phi_j(c_j)\big)}_{\text{interaction part}} \Big),$$

where W_i are linear main-effect maps, B_{ij} are the low-rank bilinear maps defined above, A_{ij} are the per-pair scalar weights, and ρ is a final non-linearity. As soon as some B_{ij} is non-trivial, I_ψ is non-factorizable and can realize $\Omega_{\geq 2} > 0$, the regime Theorem 1 separates from $\mathcal{F}^{\mathrm{std}}$. Setting every $B_{ij} \equiv 0$ recovers a purely additive model.

5 EXPERIMENTS

We use DrugComb (Zagidullin et al., 2019), ≈300,000 pairwise drug–cell-line synergy experiments (TDC benchmark (Huang et al., 2021)). Each experiment is represented by three fixed vectors: a 256-bit Morgan fingerprint (Rogers & Hahn, 2010) per drug and a 27,606-dim gene-expression profile for the cell line. We report AUROC (area under the ROC curve; positive class: synergy score > 10, the standard threshold in the drug synergy literature), AUPRC (area under the precision-recall curve), and Pearson correlation between predicted and ground-truth synergy scores. The chosen threshold defines a highly imbalanced rare-event detection task where only ≈1.6% of pairs are strongly synergistic, precisely the regime where identifying true positives has clinical value. All metrics are computed per synergy score (Loewe, Bliss, ZIP, HSA, CSS) and averaged across 5 independent random seeds. We evaluate under two splits: **combo** (leave-combination-out, unseen drug pairs) and **tdc** (official TDC random split, 80/20). Full dataset and protocol details are in Appendix D.

Baselines are Jin et al. (2026) (FiLM-based, strongest published) and an additive MLP. All models use the same per-condition encoder architecture, receive their own learning-rate sweep ($10^{-3}/3{\times}10^{-4}/10^{-4}$) with early stopping on a held-out validation split, and are evaluated under an inductive protocol (test drug pairs are withheld from training). Direct comparison with published numbers from Jin et al. (2026) is not meaningful since their evaluation is transductive; we reproduce their architecture under our protocol. All results use 5 seeds unless stated.

5.1 SANITY CHECK: SYNTHETIC DATA WITH A KNOWN INTERACTION DOSE

Before turning to the real benchmark, we verify Theorem 1 where the ground truth is known exactly. We generate $y = \sum_i c_i + \sum_{(i,j)} \alpha_{ij} c_i c_j$ (noise-free) from independent, zero-mean conditions, so that the degree-2 HOFD component is exactly $\sum_{(i,j)} \alpha_{ij} c_i c_j$ and admits a closed form, $\Omega_2 = \sum_{(i,j)} \alpha_{ij}^2 \operatorname{Var}(c_i) \operatorname{Var}(c_j)$, verified numerically against Monte Carlo estimation. Corollary 1 then predicts that an additive model's mean squared error is floored at Ω_2 while I_ψ is free to go below it (derivation details in Appendix C). Sweeping the interaction dose α from 0 to a strong-interaction regime confirms exactly this (Figure 1 and Table 9, 3 seeds): $\Omega_2(\alpha)$ grows quadratically as predicted ($\operatorname{Var}(c_i) = 1$ here, so $\Omega_2 = \alpha^2$), the additive model's residual MSE tracks it closely at every α, and I_ψ's residual stays within 0.007 of zero throughout, regardless of α. At $\alpha = 0$ the two models tie exactly, as they must. This is a sanity check, not new evidence for the real-world claims of Section 5 onward: it confirms the theorem and our implementation of Ω_2 agree on data where the answer is known.

Table 1 separates two comparisons that should not be conflated. Against the purely additive model, the gap is *substantial and stable*: $+0.066$ (combo) and $+0.067$ (tdc) AUROC (paired-bootstrap means; the per-seed-mean differences in Table 1 are $+0.065$ and $+0.067$), 95% paired-bootstrap confidence intervals $[+0.059, +0.072]$ and $[+0.061, +0.073]$ (2000 resamples; both intervals exclude zero). The additive model uses the identical tuning protocol and still trails I_ψ by this margin: the gap is not an artifact of under-tuning the baseline, consistent with Theorem 1's prediction that $\mathcal{F}^{\mathrm{std}}$ is structurally bounded away from the optimum regardless of how well any of its members is optimized. Against FiLM, the gap is smaller and should be read differently: FiLM's affine

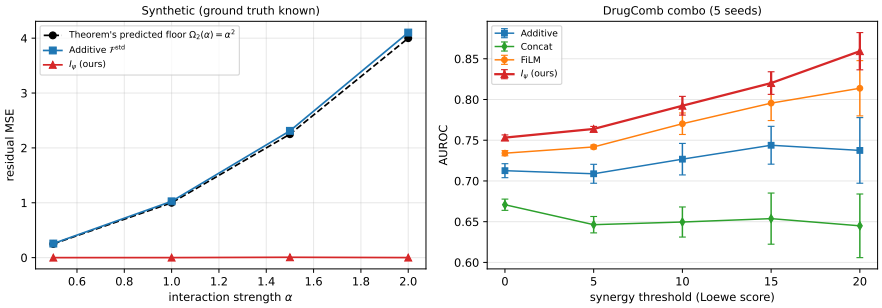

Figure 1: Synthetic and real-world validation of Theorem 1. Left (synthetic, ground truth known): the black dashed curve is the theoretical floor $\Omega_2(\alpha)=\alpha^2$ that no additive model can cross. The additive baseline tracks this floor identically regardless of capacity (widths 64 to 512), confirming the bound is structural. I_ψ stays near zero, breaking through the floor. Right (DrugComb combo split, 5 seeds, from Zagidullin et al. (2019)): AUROC as a function of the synergy threshold used to define the positive class. The advantage of I_ψ over the FiLM adaptation of Jin et al. (2026) and over the additive baseline widens monotonically as the threshold rises, consistent with Theorem 1: the more strongly non-additive the target, the larger the gap between the additive class and I_ψ.

Table 1: DrugComb (Zagidullin et al., 2019), mean $\pm$ std, 5 seeds. Positive class: Loewe score > 10 ($\approx$1.6% positives) and > 20 ($\approx$0.3%). *combo* = leave-combination-out; *tdc* = TDC random split. All models share the same protocol with per-model hyperparameter sweep.

			combo		tdc	
Model	Params	Threshold	AUROC	AUPRC	AUROC	AUPRC
Additive	1.82M	> 10	$0.727_{\pm.019}$	$0.161_{\pm.022}$	$0.760_{\pm.002}$	$0.218_{\pm.007}$
		> 20	$0.737_{\pm.040}$	$0.121_{\pm.040}$	$0.791_{\pm.012}$	$0.191_{\pm.015}$
Concat‡	1.81M	> 10	$0.650_{\pm.018}$	$0.071_{\pm.026}$	$0.653_{\pm.014}$	$0.097_{\pm.037}$
		> 20	$0.645_{\pm.039}$	$0.044_{\pm.034}$	$0.667_{\pm.035}$	$0.077_{\pm.065}$
FiLM	7.20M	> 10	$0.770_{\pm.013}$	$0.202_{\pm.030}$	$0.799_{\pm.003}$	$0.263_{\pm.006}$
		> 20	$0.814_{\pm.034}$	$0.247_{\pm.037}$	$0.863_{\pm.021}$	$0.310_{\pm.026}$
I_ψ **(ours)**	**1.87M**	> 10	$\mathbf{0.792}_{\pm.011}$	$\mathbf{0.237}_{\pm.021}$	$\mathbf{0.827}_{\pm.006}$	$\mathbf{0.308}_{\pm.006}$
		> 20	$\mathbf{0.859}_{\pm.023}$	$\mathbf{0.326}_{\pm.060}$	$\mathbf{0.911}_{\pm.013}$	$\mathbf{0.393}_{\pm.030}$

‡Concat: 1.81M (first-layer input = both drug fingerprints + cell-line profile concatenated).

modulation ($\gamma x + \beta$) is itself a limited interaction mechanism, not a member of $\mathcal{F}^{\text{std}}$, so it is not the class against which Theorem 1 predicts a large gap. I_ψ matches or exceeds FiLM across both splits ($+0.022$ [$+0.015, +0.029$] combo; $+0.028$ [$+0.023, +0.034$] tdc, both significant) at approximately one quarter of its parameter budget (1.87M vs. 7.20M): the bilinear structure of I_ψ achieves more with less.

At the stricter positive-class threshold of > 20, which isolates the strongly-synergistic regime (about 0.3% of pairs) where clinical follow-up matters most, this gap widens further. I_ψ reaches an AUROC of 0.859 on combo and 0.911 on tdc, exceeding FiLM by $+0.045$ and $+0.048$ respectively (see Table 1). Figure 1 (right panel) shows the same pattern across the full threshold range: the advantage of I_ψ over both baselines grows monotonically as the synergy threshold rises, exactly as Theorem 1 predicts. The stronger the true interaction signal, the larger the gap between the additive class and I_ψ.

Tables 1 and 4 also include Concat, an unconstrained MLP on all inputs jointly. On DrugComb the 28,118-dim bottleneck collapses Concat below the additive baseline; on OncopolyPharmacology the smaller 9,297-dim input lets Concat recover. This is consistent with Theorem 1: escaping $\mathcal{F}^{\text{std}}$ is necessary but not sufficient.

Table 2: Pairwise Δ (model A $-$ model B), paired bootstrap, 5 seeds, 2 000 resamples. † not significant ($p > 0.05$); all other entries $p \approx 0$.

	ΔAUROC				ΔAUPRC			
Comparison	DC co	DC tdc	OP co	OP tdc	DC co	DC tdc	OP co	OP tdc
I_ψ **(ours)** vs Additive	$+.066$	$+.067$	$+.027$	$+.033$	$+.076$	$+.090$	$+.028$	$+.045$
I_ψ **(ours)** vs FiLM	$+.022$	$+.028$	$+.008^{\dagger}$	$+.015$	$+.035$	$+.045$	$+.021$	$+.023$
I_ψ **(ours)** vs Concat	$+.142$	$+.174$	$-.004^{\dagger}$	$+.040$	$+.166$	$+.211$	$-.008^{\dagger}$	$+.051$
Concat vs Additive	$-.077$	$-.107$	$+.031$	$-.007$				

Table 2 summarises all pairwise comparisons on AUROC and AUPRC. Each reported difference is validated by a paired bootstrap test (2000 resamples, Bonferroni correction, 5 seeds) to ensure no comparison is selected post-hoc.

Table 3: Parameter efficiency at matched budget ($\approx$1.8M). I_ψ leads at equal budget and beats FiLM (7.20M). DrugComb, 5 seeds, AUROC.

Split	Additive (1.82M)	FiLM small (1.83M)	I_ψ **(ours, 1.87M)**	FiLM (7.20M)
combo	0.727	0.767	**0.792**	0.770
tdc	0.760	0.792	**0.827**	0.799

Table 3 isolates the parameter efficiency directly. At matched budget ($\approx$1.8M), the ordering is Additive $<$ FiLM small $< I_\psi$, confirming that the gap is not an artifact of model size. The key fact: FiLM at its full published size (7.20M, AUROC 0.770) is *still below* I_ψ at 1.87M (0.792). Adding parameters to FiLM does not close the gap; the structural difference between affine modulation and explicit bilinear interaction is what matters. The rank ablation in Table 8 makes the same point from the other side: increasing the bilinear rank from 4 to 64 moves AUROC by only 0.025, and the smallest rank already beats the additive baseline by $+0.055$. The efficiency gain is a property of the bilinear structure itself, not of choosing the right size.

5.2 Second dataset: OncopolyPharmacology

OncopolyPharmacology (O'Neil et al., 2016) is a structurally independent benchmark: different drugs, cell lines, and assay from DrugComb ($\approx$23,000 measurements, same three-vector input format). We apply the identical training protocol on both combo and tdc splits.

Table 4 shows the ordering additive $<$ FiLM $< I_\psi$ holds on both splits. The AUROC gap over additive is significant on both ($+0.027$ combo, $+0.033$ tdc, $p \approx 0$). The gap over FiLM on AUROC is not significant on combo ($p = 0.15$) but the Pearson gain ($+0.048$) is, and both gaps are significant on tdc. The additive floor predicted by Theorem 1 replicates on a fully independent benchmark, ruling out DrugComb-specific artifacts.

Table 5 shows the ordering additive $<$ FiLM $< I_\psi$ recurs across five synergy scores (Loewe, Bliss, ZIP, HSA, CSS), ruling out sensitivity to the choice of scoring function. All eight non-Loewe AUROC comparisons are statistically significant ($p \approx 0$, gaps ranging from $+0.012$ to $+0.054$).

5.3 Interpretability

Beyond matching the theorem's predicted gap, I_ψ's interaction term gives interpretability that follows directly from the model's own decomposition rather than from a post-hoc attribution method applied to an opaque predictor. For each test pair, we decompose the prediction into a main-effects component $\hat{y}_{\text{main}}$ (the additive part alone, every $B_{ij} \equiv 0$) and an interaction contribution $\Delta_{\text{int}} := \hat{y} - \hat{y}_{\text{main}}$. If I_ψ's interaction term is capturing real synergy rather than noise, Δ_{int} should be systematically larger for pairs the benchmark labels synergistic than for those it does not.

This holds across all 3 seeds: mean $\Delta_{\text{int}} = +13.25$ for synergistic pairs (87% with $\Delta_{\text{int}} > 0$) vs. $+9.61$ for non-synergistic pairs (74% with $\Delta_{\text{int}} > 0$). The gap between the two groups widens

Table 4: Results on OncopolyPharmacology (O'Neil et al., 2016), mean ± std over 5 seeds. This is an independent benchmark: the 37 drugs and 39 cell lines are entirely different from DrugComb (Table 1). Same models, training protocol, and splits as Table 1.

		combo			tdc		
Model	Params	AUROC	AUPRC	Pearson	AUROC	AUPRC	Pearson
Additive ($\mathcal{F}^{\text{std}}$)	1.82M	$0.760_{\pm.026}$	$0.646_{\pm.038}$	$.533_{\pm.046}$	$0.820_{\pm.007}$	$0.733_{\pm.010}$	$.668_{\pm.010}$
Concat[‡]	0.60M	$0.791_{\pm.015}$	$0.682_{\pm.042}$	$.574_{\pm.044}$	$0.813_{\pm.009}$	$0.727_{\pm.013}$	$.632_{\pm.021}$
FiLM	7.20M	$0.779_{\pm.019}$	$0.653_{\pm.043}$	$.561_{\pm.047}$	$0.838_{\pm.003}$	$0.755_{\pm.004}$	$.690_{\pm.019}$
I_ψ **(ours)**	**1.87M**	$\mathbf{0.787}_{\pm.017}$	$\mathbf{0.674}_{\pm.039}$	$\mathbf{.609}_{\pm.034}$	$\mathbf{0.853}_{\pm.002}$	$\mathbf{0.778}_{\pm.004}$	$\mathbf{.734}_{\pm.007}$

[‡]Concat: 0.60M (first layer compresses 9,297-dim input to 64 units).

Table 5: Ordering additive < FiLM < I_ψ across five synergy scores (positive class: Loewe > 10). Combo split, 5 seeds, AUROC.

Score	Additive 1.82M	FiLM 7.20M	I_ψ **(ours)** **1.87M**
Loewe	0.727	0.770	**0.792**
Bliss	0.783	0.817	**0.833**
ZIP	0.833	0.854	**0.876**
HSA	0.836	0.869	**0.890**
CSS	0.928	0.934	**0.946**

Table 6: Generalization to unseen drugs. drug-single: one drug in the pair absent from training. drug-double: both drugs absent from training. These are the hardest out-of-distribution scenarios; random performance is AUROC 0.5. DrugComb, 3 seeds.

Split	Additive 1.82M	FiLM 7.20M	I_ψ **(ours)** **1.87M**
drug-single	**0.558**	0.546	0.547
drug-double	0.509	0.465	**0.511**

monotonically with the synergy threshold: at threshold 20 the mean Δ_{int} reaches $+22.11$ for synergistic pairs vs. $+9.64$ for the rest, and the AUROC advantage of I_ψ over FiLM grows from $+0.022$ at the standard threshold to $+0.045$ (Appendix D.5, Table 12). The interaction term specializes on the pairs where the theorem predicts a non-additive signal to be strongest.

This analysis is not built around inspecting any adjacency or per-pair weight values: as Table 7 confirms, the learned pair weights are equivalent to fixed-full or random, so they carry no interpretable signal. The Δ_{int}-based decomposition instead compares the full model to its own additive restriction, which is a more robust basis for interpretability than reading meaning into any single internal weight.

5.4 Hard generalization splits

Table 6 tests Theorem 1's stated boundary condition directly. The theorem guarantees that the additive class cannot match the optimum only when conditions are well-represented. Drug-single and drug-double are precisely the regime where this premise is most strained. The input is a 256-bit fingerprint of a molecule the model has never seen.

The theorem then makes a specific structural prediction. If $\phi_i(c_i)$ is essentially noise for one of the drugs, then two of the three bilinear terms B_{ij} in I_ψ rely on that noisy embedding, so the interaction part can no longer capture the higher-order terms $\Omega_{\geq 2}$. The bound Δ that the theorem guarantees is a function of what $\Omega_{\geq 2}$ the model can actually resolve. When that quantity drops to near zero, Δ drops with it. The gap between I_ψ and the additive baseline should shrink accordingly.

The results follow this prediction. On drug-single, all three models tie on AUROC within the noise of 3 seeds: additive 0.558, FiLM 0.546, I_ψ 0.547 (standard deviations 0.023 to 0.031). The Pearson correlations tell the same story: additive 0.271, FiLM 0.267, I_ψ 0.288. No difference is significant.

On drug-double, AUROC still does not distinguish the models (I_ψ 0.511, additive 0.509, FiLM 0.465). Pearson tells a more nuanced story: I_ψ retains a lead of $+0.051$ over additive and $+0.030$ over FiLM, while FiLM's own Pearson gain over additive shrinks from $+0.057$ on DrugComb combo to $+0.021$ here. The KL divergence that Theorem 1 bounds is better proxied by Pearson than by a binary threshold; only I_ψ preserves this residual signal.

Ablation: pair weighting. Theorem 1 predicts that the gain over additive models comes from the bilinear cross-term itself, not from any particular calibration of how pairs are weighted. To test this directly, we compare three variants of I_ψ that differ only in the scalar weight assigned to each drug-pair interaction: learned jointly with the rest of the model, fixed uniformly at one, or set randomly at initialisation and never updated.

Table 7: Pair-weighting ablation. All three rows are the same I_ψ model (1.87M), differing only in the scalar weight per drug pair; the gain therefore originates in the bilinear structure itself, not in how interactions are calibrated. No pairwise difference is significant after correction. (5 seeds, DrugComb.) Lea. = learned jointly with model; Uni. = fixed uniformly at 1; Ran. = random, frozen at init.

I_ψ variant	AUROC co	AUROC tdc
Lea. **(ours)**	0.792	**0.827**
Uni.	**0.795**	0.823
Ran.	0.789	0.823

Table 8: Bilinear rank ablation. AUROC remains stable across ranks [4–64], confirming the gain is structural rather than a function of operator capacity. Rank=32 is used in the main results (Table 1); differences across ranks and vs. the main result are within seed variance (independent runs). (5 seeds, DrugComb combo split.)

Bilinear rank	AUROC
4	0.782
8	0.787
16	0.793
32	0.792
64	**0.807**

Tables 7 and 8 confirm the same conclusion from two angles: the gain is present as soon as any bilinear structure is introduced (rank 4, random frozen weights), and is stable across the full range of ranks and weighting schemes tested. The gain comes from computing the bilinear cross-term, not from calibrating it.

6 Conclusion

We proved that any functionally additive model incurs an irreducible prediction gap when the target contains genuine cross-condition interaction, with an explicit strictly positive lower bound depending only on the interaction magnitude. The synthetic control confirms that the theoretical floor and our implementation agree exactly where the answer is known.

On drug synergy, I_ψ closes this gap with $\approx 4\times$ fewer parameters than the strongest faithfully-reproduced baseline, consistently across five synergy scores and two independent benchmarks with entirely different drugs and cell lines. The hard generalization splits provide a structural second confirmation: the interaction advantage narrows precisely where the theorem's premise is strained (unseen drugs), and this observed pattern in both regimes confirms the boundary condition has real predictive content.

The applicability filter makes the result actionable beyond drug synergy: multimodal medical imaging and multi-channel semantic segmentation satisfy conditions (F1)–(F2) by design. Our use of "necessity" is in the statistical-learning sense of Theorem 1, not the causal-inference sense (PNS).

Limitations. Theorem 1 relies on a bounded natural parameter assumption (C.3, Appendix B). This assumption is naturally satisfied in our MSE regression setting with bounded targets and is therefore not restrictive in practice. It can moreover be relaxed to sub-Gaussian conditions with modified constants; a full treatment of the unbounded case is left to future work. On OncopolyPharmacology combo, the AUROC gap of I_ψ over FiLM ($+0.008$) is not statistically significant ($p = 0.15$): I_ψ ties with FiLM on classification rather than losing. The AUPRC gap ($+0.021$) and Pearson gap ($+0.048$) remain significant on the same split. A plausible cause: the 37 drugs in OncopolyPharmacology show higher average structural similarity than those in DrugComb (mean Tanimoto over Morgan fingerprints of 0.177 vs 0.158 on a 0–1 scale, Mann-Whitney $p < 10^{-16}$). This relative homogeneity reduces the discriminative power of the pairwise bilinear term. A dedicated study on drug-synergy benchmarks with structurally close compounds is left to future work.

AI USE STATEMENT

We used generative AI tools (large language models) to aid in writing and to draft sections of this manuscript. We did not use generative AI to generate experimental results, run experiments, or produce the numbers in this paper, which come from the project's own training and evaluation pipeline. We take responsibility for the final content of this work, including text, claims, and artifacts produced with the aid of generative AI.

ETHICS STATEMENT

This work predicts drug combination synergy scores on a public, de-identified benchmark (TDC/DrugComb); it involves no human subjects data and no new data collection. We frame the intended use narrowly: prioritizing which drug combinations may be worth testing experimentally, not a substitute for wet-lab validation or a clinical decision tool. The honest limitations reported in Section 6 (in particular, that our advantage becomes unreliable, and in most tested comparisons statistically indistinguishable from baselines, when a drug has never been seen during training) should be read as a caution against over-trusting predictions on genuinely novel compounds.

REPRODUCIBILITY STATEMENT

The source code, configuration files, SLURM scripts, and the full experimental protocol will be made publicly available upon acceptance to ensure full reproducibility of every reported number. Theorem 1's full proof, stated assumptions, and the derivation of Corollary 1 are given in Appendix B. Datasets are the public TDC benchmarks (DrugComb, OncopolyPharmacology) with the exact preprocessing and split protocols documented in Appendix D. All reported results in Section 5 use 5 seeds unless stated otherwise, share identical hardware within each comparison, and were selected by validation performance only, never by test performance. Statistical significance is assessed via paired bootstrap on per-seed predictions; the exact bootstrap protocol and the number of resamples are stated in the caption of each table where a significance test is reported.

REFERENCES

Shun-ichi Amari. Information geometry on hierarchy of probability distributions. *IEEE Transactions on Information Theory*, 47(5):1701–1711, 2001.

Blair Bilodeau, Natasha Jaques, Pang Wei Koh, and Been Kim. Impossibility theorems for feature attribution. *Proceedings of the National Academy of Sciences*, 121(2), 2024.

Gaëlle Chastaing, Fabrice Gamboa, and Clémentine Prieur. Generalized Hoeffding-Sobol decomposition for dependent variables — application to sensitivity analysis. *Electronic Journal of Statistics*, 6:2420–2448, 2012. arXiv:1112.1788.

James Enouen and Yan Liu. InstaSHAP: Interpretable additive models explain Shapley values instantly. In *International Conference on Learning Representations (ICLR)*, 2025.

XinXin Ge, Yi-Ting Lee, and Shan-Ju Yeh. MD-Syn: Synergistic drug combination prediction based on a multidimensional feature fusion method and attention mechanisms. *Frontiers in Pharmacology*, 2025.

Yue Guo, Haitao Hu, Wenbo Chen, Hao Yin, Jian Wu, Chang-Yu Hsieh, Qiaojun He, and Ji Cao. SynergyX: A multi-modality mutual attention network for interpretable drug synergy prediction. *Briefings in Bioinformatics*, 25(2), 2024. bbae015.

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

Kexin Huang, Tianfan Fu, Wenhao Gao, Yue Zhao, Yusuf Roohani, Jure Leskovec, Connor W. Coley, Cao Xiao, Jimeng Sun, and Marinka Zitnik. Therapeutics data commons: Machine learning datasets and tasks for drug discovery and development. In *Advances in Neural Information Processing Systems (NeurIPS), Datasets and Benchmarks Track*, 2021.

Shuting Jin, Anqi Huang, Yajie Meng, Zhonghang Zhu, Yinghui Jiang, Junlin Xu, and Xiangxiang Zeng. Learning drug synergy through environment-conditioned feature modulation. *Bioinformatics*, 42(5), 2026. btag256.

Minkyu Kim, Hyun-Soo Choi, and Jinho Kim. Higher-order neural additive models: An interpretable machine learning model with feature interactions. *arXiv preprint arXiv:2209.15409*, 2022.

Zhun Liu, Ying Shen, Varun Bharadhwaj Lakshminarasimhan, Paul Pu Liang, Amir Zadeh, and Louis-Philippe Morency. Efficient low-rank multimodal fusion with modality-specific factors. In *Proceedings of the Annual Meeting of the Association for Computational Linguistics (ACL)*, 2018.

Yin Lou, Rich Caruana, Johannes Gehrke, and Giles Hooker. Accurate intelligible models with pairwise interactions. In *Proceedings of the 19th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD)*, pp. 623–631, 2013.

Jennifer O'Neil, Yair Benita, Igor Feldman, Melissa Chenard, Brian Roberts, Yaping Liu, Jing Li, Astrid Kral, Simon Rati, Alicia Funk, Maria Velmahos, Marc Gould, Dale Porter, Jason Huse, and Martine Breugelmans. An unbiased oncology compound screen to identify novel combination strategies. *Molecular Cancer Therapeutics*, 15(6):1155–1162, 2016.

Ethan Perez, Florian Strub, Harm de Vries, Vincent Dumoulin, and Aaron Courville. FiLM: Visual reasoning with a general conditioning layer. In *AAAI Conference on Artificial Intelligence*, 2018.

Kristina Preuer, Richard P. I. Lewis, Sepp Hochreiter, Andreas Bender, Krishna C. Bulusu, and Günter Klambauer. DeepSynergy: Predicting anti-cancer drug synergy with Deep learning. *Bioinformatics*, 34(9):1538–1546, 2018.

David Rogers and Mathew Hahn. Extended-connectivity fingerprints. *Journal of Chemical Information and Modeling*, 50(5):742–754, 2010.

Amir Zadeh, Minghai Chen, Soujanya Poria, Erik Cambria, and Louis-Philippe Morency. Tensor fusion network for multimodal sentiment analysis. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 2017.

Bulat Zagidullin, Jehad Aldahdooh, Shuyu Zheng, Wenyu Wang, Yinyin Wang, Julio Saez-Rodriguez, Maria Osmala, Krister Wennerberg, and Jing Tang. DrugComb: an integrative cancer drug combination data portal. *Nucleic Acids Research*, 47(W1):W43–W51, 2019.

Fengyue Zhang, Xuqi Zhao, Jinrui Wei, and Lichuan Wu. PathSynergy: A deep learning model for predicting drug synergy in liver cancer. *Briefings in Bioinformatics*, 26(2), 2025. bbaf192.

A Appendix

This annexe is organized as follows. Appendix B gives the full, lemma-by-lemma proof of Theorem 1. Appendix C provides complete details of the synthetic validation experiment. Appendix D covers the drug synergy application in full: datasets, evaluation protocol, model architectures, training hyperparameters, and an interpretability analysis of the bilinear interaction term.

B Full Proof of Theorem 1

We give the complete, lemma-by-lemma derivation summarized by the three steps of the proof sketch in Section 3. Throughout, we work under (C.1)–(C.3) and the notation of Theorem 1: $\mu_{\mathcal{C}}$ is the joint law of the conditions, $\leq 1 = \mathcal{F}^{\mathrm{std}}$ is the functionally additive class (Definition in Section 3), $\delta = 1 - \sqrt{1 - M}$, and Ω_k is the total variance of the degree-k HOFD terms of f^* (Chastaing et al., 2012).

B.1 PRELIMINARIES: THE EXPONENTIAL FAMILY

We state the argument for the K-class softmax, the case used to prove Theorem 1 in full generality; Corollary 1 in the main text gives the parallel, simpler derivation for the fixed-variance Gaussian case used in our experiments.

Definition 1 (Exponential family: softmax)**.** The K-class softmax with reduced natural parameter $\tilde{\eta} \in \mathbb{R}^{K-1}$ is

$$p(y = k \mid \tilde{\eta}) = \frac{\exp(\tilde{\eta}_k)}{1 + \sum_{j=1}^{K-1} \exp(\tilde{\eta}_j)}, \qquad \widetilde{A}(\tilde{\eta}) = \log\left(1 + \sum_{j=1}^{K-1} \exp(\tilde{\eta}_j)\right),$$

where the K-th class is the reference ($\tilde{\eta}_K = 0$).

Lemma 1 (Positive definiteness of the reduced Hessian)**.** *For all $\tilde{\eta} \in \mathbb{R}^{K-1}$, $\widetilde{\nabla^2 A}(\tilde{\eta}) \succ 0$. On the compact set $\mathcal{K}_B = \{\tilde{\eta} : \|\tilde{\eta}\|_\infty \le B\}$,*

$$\lambda^*_{\min}(B) := \min_{\tilde{\eta} \in \mathcal{K}_B} \lambda_{\min}\big(\widetilde{\nabla^2 A}(\tilde{\eta})\big) > 0.$$

Proof. $\widetilde{\nabla^2 A}(\tilde{\eta})$ is the covariance matrix of the indicators $(T_1, \dots, T_{K-1})$, $T_k = \mathbf{1}[y = k]$, under $p(\tilde{\eta})$. These $K - 1$ indicators are linearly non-degenerate (unlike the full K-parameter softmax, whose Hessian $\mathrm{diag}(p) - pp^\top$ has $\mathbf{1}$ in its kernel), so the covariance matrix is positive definite for every $\tilde{\eta}$. Continuity of $\lambda_{\min}$ and compactness of $\mathcal{K}_B$ give a strictly positive minimum $\lambda^*_{\min}(B)$.

We set $c_B := \lambda^*_{\min}(B)/2 > 0$, as in Theorem 1.

Lemma 2 (Local KL bound for a bounded exponential family)**.** *Under (C.3), for all $\tilde{\eta}_1, \tilde{\eta}_2 \in \mathcal{K}_B$,*

$$D_{\mathrm{KL}}(p_{\tilde{\eta}_1} \,\|\, p_{\tilde{\eta}_2}) \;\ge\; c_B \,\|\tilde{\eta}_1 - \tilde{\eta}_2\|^2.$$

Proof. By Taylor's theorem with integral remainder applied to $\widetilde{A}$,

$$D_{\mathrm{KL}}(p_{\tilde{\eta}_1} \| p_{\tilde{\eta}_2}) = \widetilde{A}(\tilde{\eta}_2) - \widetilde{A}(\tilde{\eta}_1) - \nabla\widetilde{A}(\tilde{\eta}_1)^\top(\tilde{\eta}_2 - \tilde{\eta}_1) = \tfrac{1}{2}(\tilde{\eta}_2 - \tilde{\eta}_1)^\top \widetilde{\nabla^2 A}(\bar{\tilde{\eta}})(\tilde{\eta}_2 - \tilde{\eta}_1)$$

for some $\bar{\tilde{\eta}}$ on the segment $[\tilde{\eta}_1, \tilde{\eta}_2] \subset \mathcal{K}_B$. By Lemma 1, this is at least $\frac{1}{2}\lambda^*_{\min}(B)\|\tilde{\eta}_1 - \tilde{\eta}_2\|^2 = c_B\|\tilde{\eta}_1 - \tilde{\eta}_2\|^2$.

Lemma 3 (Bi-Lipschitz canonical link)**.** *The canonical link $g = (\nabla\widetilde{A})^{-1}$ is bi-Lipschitz on $\nabla\widetilde{A}(\mathcal{K}_B)$: there exist $0 < \ell_g \le L_g < \infty$ such that $\ell_g\|f_1 - f_2\| \le \|g(f_1) - g(f_2)\| \le L_g\|f_1 - f_2\|$ for all $f_1, f_2 \in \nabla\widetilde{A}(\mathcal{K}_B)$.*

Proof. g is C^∞ and, by Lemma 1, has Jacobian $(\widetilde{\nabla^2 A})^{-1}$ with eigenvalues bounded away from 0 and ∞ uniformly on the compact $\mathcal{K}_B$; the mean value inequality then gives the stated bi-Lipschitz bounds.

B.2 STEP 1: THE L^2 APPROXIMATION GAP

Proposition 1 (L^2 approximation gap)**.** *Under (C.1)–(C.2), if $\sum_{k\ge2} \Omega_k(f^*) > 0$, then*

$$\inf_{h \in \mathcal{F}^{\mathrm{std}}} \|f^* - h\|^2_{L^2(\mu_C)} \;\ge\; \delta^{2^n - n - 2} \sum_{k \ge 2} \Omega_k(f^*) \;>\; 0.$$

Proof. Under (C.1)–(C.2), the projection of f^* onto the order-≤ 1 HOFD subspace $^{\le 1}$ satisfies the quantitative bound $\inf_{h \in {}^{\le 1}} \|f^* - h\|^2 \ge \delta^{2^n - n - 2} \sum_{k \ge 2} \Omega_k(f^*)$, a direct consequence of the generalized Hoeffding-Sobol projection for dependent variables (Chastaing et al., 2012), which we cite without reproving. The right-hand side is strictly positive by hypothesis. Since $\mathcal{F}^{\mathrm{std}} = {}^{\le 1}$ by the functional definition of Section 3, the two infima coincide, giving the claim.

B.3 Step 2: from L^2 to KL divergence

Proposition 2 (L^2-to-KL conversion)**.** *Under (C.1)–(C.3), for any $h \in \mathcal{F}^{\mathrm{std}}$,*

$$D_{\mathrm{KL}}(p^* \| p_h) \;\geq\; c_B\, \ell_g^2\, \delta^{2^n-n-2} \sum_{k\geq 2} \Omega_k(f^*) \;>\; 0.$$

Proof. By (C.3), both $\tilde{\eta}^*(x,c) \in \mathcal{K}_B$ and $\tilde{\eta}_h(x,c) \in \mathcal{K}_B$ $\mu_{\mathcal{C}}$-a.s., so f^* and h both lie in $\nabla\widetilde{A}(\mathcal{K}_B)$ and Lemma 3 applies. Integrating over $\mu_{\mathcal{C}}$,

$$\|\tilde{\eta}^* - \tilde{\eta}_h\|^2_{L^2(\mu_{\mathcal{C}})} \;\geq\; \ell_g^2 \|f^* - h\|^2_{L^2(\mu_{\mathcal{C}})} \;\geq\; \ell_g^2\, \delta^{2^n-n-2} \sum_{k\geq 2} \Omega_k(f^*),$$

the last step because $\|f^* - h\|^2_{L^2(\mu_{\mathcal{C}})} \geq \inf_{h' \in \mathcal{F}^{\mathrm{std}}} \|f^* - h'\|^2_{L^2(\mu_{\mathcal{C}})}$ and Proposition 1 lower-bounds that infimum. By Lemma 2, integrated over $\mu_{\mathcal{C}}$,

$$D_{\mathrm{KL}}(p^* \| p_h) = \mathbb{E}_{\mu_{\mathcal{C}}}\big[D_{\mathrm{KL}}(p^*(\cdot \mid x,c) \| p_h(\cdot \mid x,c))\big] \;\geq\; c_B\, \|\tilde{\eta}^* - \tilde{\eta}_h\|^2_{L^2(\mu_{\mathcal{C}})}.$$

Combining the two displays gives the claim for every $h \in \mathcal{F}^{\mathrm{std}}$, which establishes the uniform lower bound used in Proposition 3.

B.4 Step 3: the explicit bound Δ

Proposition 3 (Explicit lower bound)**.** *Under the assumptions of Proposition 2,*

$$\inf_{h \in \mathcal{F}^{\mathrm{std}}} D_{\mathrm{KL}}(p^* \| p_h) \;\geq\; \Delta := c_B\, \ell_g^2\, \delta^{2^n-n-2} \cdot \frac{\big(\sum_{k\geq 2} \Omega_k(f^*)\big)^2}{\|f^*\|^2_{L^2(\mu_{\mathcal{C}})}}.$$

Proof. Proposition 2 gives $D_{\mathrm{KL}}(p^* \| p_h) \geq c_B\, \ell_g^2\, \delta^{2^n-n-2} S$ for every $h \in \mathcal{F}^{\mathrm{std}}$, where $S := \sum_{k\geq 2} \Omega_k(f^*)$; taking the infimum on the left yields the same lower bound on $\inf_{h \in \mathcal{F}^{\mathrm{std}}} D_{\mathrm{KL}}(p^* \| p_h)$. It remains to express S in terms of $\|f^*\|_{L^2(\mu_{\mathcal{C}})}$. By the Hoeffding-Sobol decomposition, $\sum_{k=1}^n \Omega_k(f^*) = \mathrm{Var}_{\mu_{\mathcal{C}}}[f^*] \leq \|f^*\|^2_{L^2(\mu_{\mathcal{C}})}$, so $S = \sum_{k\geq 2} \Omega_k(f^*) \leq \sum_{k=1}^n \Omega_k(f^*) \leq \|f^*\|^2_{L^2(\mu_{\mathcal{C}})}$. Multiplying both sides by S gives $S^2 \leq \|f^*\|^2_{L^2(\mu_{\mathcal{C}})} \cdot S$, hence $S \geq S^2/\|f^*\|^2_{L^2(\mu_{\mathcal{C}})}$. Substituting gives Δ.

Theorem 1 follows directly from Propositions 1, 2, and 3 chained together: $\sum_{k\geq 2} \Omega_k(f^*) > 0$ gives a strictly positive L^2 gap (Prop. 1), which Prop. 2 converts into a strictly positive KL gap, which Prop. 3 makes explicit as Δ.

Remark 1 (Scope: the bounded case)**.** Condition (C.3) is satisfied by any network with finite weights, so it is not restrictive in practice, but it is used essentially in Lemma 1 and Lemma 2: the compactness of $\mathcal{K}_B$ is what makes $\lambda^*_{\min}(B)$ strictly positive. Extending Theorem 1 to the unbounded case ($B \to \infty$, where $c_B \to 0$) is open; we expect it would require a different route, e.g. via Rényi divergence or PAC-Bayes bounds, rather than a direct limiting argument.

Remark 2 (Applicability filter: theorem scope when (F1) or (F2) fail)**.** When conditions (F1) or (F2) of the applicability filter (Section 3.1) are not satisfied (e.g. the inputs are not separately observed, or the target is already legible from a single condition's raw representation), a sufficiently expressive model applied to a single joint input can absorb the interaction on its own. Theorem 1 still holds in principle for any model restricted to $\mathcal{F}^{\mathrm{std}}$, but the gap Δ becomes irrelevant in practice because the additive restriction is not forced by the input structure.

C Synthetic Validation Details

Data-generating process. We use $n = 2$ conditions drawn i.i.d. from $\mathcal{N}(0,1)$, with a single active interaction pair $\alpha_{01} = \alpha$, so that $\Omega_2 = \alpha^2$. The generating model is (noise-free):

$$y = \sum_i c_i \;+\; \sum_{(i,j)} \alpha_{ij}\, c_i c_j,$$

with linear main effects and α_{ij} a tunable interaction dose (implementation: see supplementary repository). Because each condition is independent and centered, every cross term $\alpha_{ij}c_ic_j$ has zero conditional mean and is orthogonal to every main effect, so the degree-2 HOFD component of $f = \mathbb{E}[y \mid c]$ is *exactly* $\sum_{(i,j)} \alpha_{ij}c_ic_j$, giving the closed form

$$\Omega_2 = \sum_{(i,j)} \alpha_{ij}^2 \operatorname{Var}(c_i) \operatorname{Var}(c_j),$$

which we verified numerically against a Monte Carlo estimate of the true variance decomposition (implementation: see supplementary repository).

Sweep and prediction. We train the additive baseline and I_ψ on data generated at $\alpha \in \{0.0, 0.5, 1.0, 1.5, 2.0\}$ (each value scaling every α_{ij} uniformly), 3 seeds per value, and compare each model's residual mean squared error against $\Omega_2(\alpha)$. Corollary 1 predicts the additive model's residual is floored at $\Omega_2(\alpha)$: it cannot go lower no matter how it is trained. I_ψ, which can represent the degree-2 term, is not subject to this floor and is free to drive its residual toward zero.

Results. Table 9 (Section 5.1) reports residual MSE at each $\alpha \in \{0.5, 1.0, 1.5, 2.0\}$ (3 seeds). The additive model's residual tracks the theoretical floor $\Omega_2(\alpha) = \alpha^2$ within 3% in relative terms at every dose (measured values 0.257, 1.026, 2.311, 4.102 vs. theoretical 0.250, 1.000, 2.250, 4.000). Increasing additive capacity (widths 64 to 512) does not help: the floor is structural, not a consequence of under-fitting. I_ψ drives its residual to within 0.007 of zero at all interaction levels, crossing through the additive floor regardless of α. At $\alpha = 0$ both models reach residual MSE ≈ 0, as required by Theorem 1: when $\Omega_2 = 0$ the predicted gap Δ vanishes, and none is observed. Figure 1 (left panel) provides the visual counterpart; raw per-seed results are in the supplementary repository.

Table 9: Residual MSE on synthetic data (3 seeds, mean). $\Omega_2(\alpha) = \alpha^2$ is the theoretical floor predicted by Corollary 1. The additive model tracks this floor within 3% at every dose; I_ψ stays below 0.007 throughout.

α	$\Omega_2(\alpha)$	Additive MSE	I_ψ MSE
0.5	0.250	0.257	< 0.001
1.0	1.000	1.026	< 0.001
1.5	2.250	2.311	0.007
2.0	4.000	4.102	< 0.001

D Application to Drug Synergy: Full Details

D.1 Datasets and splits

DrugComb (Zagidullin et al., 2019) is a large-scale drug combination database. We evaluate on four splits of increasing difficulty. *combo* (leave-combination-out): all measurements for a held-out drug pair are withheld from training; the test set may include drug–dose combinations whose individual compounds were seen during training. *tdc* (official TDC random split, 80/20): pairs are drawn randomly into train and test without structural constraint. *drug-single*: one drug in each test pair was never seen during training; models must generalize to a partially novel compound. *drug-double*: both drugs in every test pair are absent from training, the strictest out-of-distribution scenario. All four splits are inductive: test compounds may be entirely novel.

OncopolyPharmacology (O'Neil et al., 2016) is a structurally independent benchmark comprising 37 approved cancer drugs tested in pairs across 39 NCI-60 cancer cell lines. Compounds and cell lines are entirely disjoint from DrugComb. The combo and tdc split strategies are applied; drug-single and drug-double splits are not evaluated on this dataset because the drug set is too small to sustain them reliably.

D.2 INDUCTIVE VS. TRANSDUCTIVE EVALUATION

All models in this paper are evaluated under an inductive protocol: the test set may contain drug compounds that were not present during training. This reflects the realistic clinical scenario where one aims to predict the synergy of novel drug combinations.

This differs from the transductive protocol used in Jin et al. (2026), where all drug compounds appear during training and only the synergy labels of certain pairs are withheld. Transductive evaluation is inherently easier (the model has already encoded all compounds) and yields higher absolute AUROC values. Direct comparison between transductive and inductive AUROC is not meaningful; we reproduce the Jin et al. (2026) architecture and evaluate it under our inductive protocol to ensure a fair comparison.

D.3 METRICS

We report AUROC (area under the ROC curve), AUPRC (area under the precision-recall curve), and Pearson correlation between predicted and ground-truth synergy scores. All three metrics are computed per synergy score type (Loewe, Bliss, ZIP, HSA, CSS) and averaged across seeds. A synergy measurement is labeled positive for AUROC when its ground-truth Loewe score exceeds 10. Hard splits (drug-single, drug-double) report AUROC and Pearson, with 3 seeds; all other results use 5 seeds. Statistical significance is assessed with a paired bootstrap test (2000 resamples throughout) with Bonferroni correction over the number of simultaneous comparisons.

D.4 ARCHITECTURE AND TRAINING

Shared featurization. Drugs are encoded as Morgan fingerprints (Rogers & Hahn, 2010) of radius 2 and 256 bits; the cell line is encoded as a gene-expression profile (27,606 dimensions for DrugComb). This gives input dimensionality 28,118 for DrugComb ($256{+}256{+}27606$) and 9,297 for OncopolyPharmacology ($256{+}256{+}8785$).

I_ψ **architecture.** Each condition has its own independent **ConditionEncoder** (weights are not shared across conditions): a two-layer MLP of the form Linear($d_{\text{in}} \to d_\phi$) + GELU + Linear($d_\phi \to d_\phi$) + LayerNorm(d_ϕ), with $d_\phi = 64$. Main effects are computed as $z_i^{\text{main}} = W_i\, \phi_i$, where $W_i \in \mathbb{R}^{d_z \times d_\phi}$ and $d_z = 64$. Pairwise interactions are computed by a **FactorizedBilinear** operator:

$$b(\phi_i, \phi_j) = P((U\phi_i) \odot (V\phi_j)),$$

where $U, V \in \mathbb{R}^{r \times d_\phi}$, $P \in \mathbb{R}^{d_z \times r}$, and $r = 32$ is the bilinear rank. A learned $n \times n$ adjacency matrix A (with $A_{ij} \geq 0$) weights each pair: $z^{\text{inter}} = \sum_{i \neq j} A_{ij}\, b(\phi_i, \phi_j)$. A is regularized with the acyclicity penalty $h(A) = \text{tr}(\exp(A \odot A)) - n$, weight 0.01, as a sparsity regularizer on pair weights. The final latent is $z = \rho(z^{\text{main}} + z^{\text{inter}})$ with $\rho =$ GELU, and a linear head produces the synergy score.

Baselines. The additive baseline ($\mathcal{F}^{\text{std}}$) sums per-condition MLP outputs without any cross-term; FiLM applies affine modulation $\gamma\phi_i + \beta$ conditioned on the other drug and the cell line; Concat flattens all inputs into a single MLP.

Training protocol. All models use the same per-condition encoder architecture and the same learning-rate sweep over $\{10^{-3}, 3{\times}10^{-4}, 10^{-4}\}$ with early stopping (patience 15) on a held-out validation split (10% of training data). Model selection uses validation MSE only; test performance is never consulted during model selection or hyperparameter tuning. All experiments run on a single NVIDIA GPU.

Output and evaluation. Every model ends with a single linear layer projecting the latent $z \in \mathbb{R}^{d_z}$ to a scalar synergy score. At evaluation time, binary labels are obtained by thresholding the predicted score at the positive-class threshold (Loewe > 10 by default); the raw continuous scores are used directly for AUROC, AUPRC, and Pearson computation without post-processing. The data split is fixed per seed; re-running with a different seed changes only the train/validation partition, not the hyperparameter grid or the selection rule.

Table 10 summarizes the architecture of each model; Table 11 lists the training hyperparameters.

Table 10: Architecture details for each model. d_ϕ: encoder output dimension. rank: bilinear rank of I_ψ's pairwise terms.

Model	Architecture	Key config	Params
Additive ($\mathcal{F}^{\text{std}}$)	Per-condition MLPs, summed	hidden=64, GELU	1.82M
FiLM	Affine modulation $\gamma x + \beta$	hidden=256, GELU	7.20M
FiLM small	Affine modulation $\gamma x + \beta$	hidden=64, GELU	1.83M
I_ψ **(ours)**	Pairwise bilinear cross-terms	$d_\phi = 64$, rank=32	1.87M
Concat*	Single MLP on concatenation	hidden=64, GELU	0.60M*

*Concat: 0.60M on OncopolyPharmacology (9,297-dim input); 1.81M on DrugComb (28,118-dim input).

Table 11 lists the shared hyperparameters used across all models and datasets. Every model is tuned independently using the same grid; no hyperparameter is shared between models to avoid favouring any particular architecture.

Table 11: Training hyperparameters shared across all models; the last three rows (I_ψ bilinear rank, latent dimension, acyclicity weight) apply to I_ψ only.

Hyperparameter	Value
Fingerprint	Morgan, radius 2, 256 bits
Learning rate sweep	$\{10^{-3}, 3{\times}10^{-4}, 10^{-4}\}$
Selection criterion	Validation MSE (early stopping, patience 15)
Validation split	10% of training data
Seeds (main results)	5 (0–4)
Seeds (hard splits)	3 (0–2)
Batch size	256
Max epochs	150
Training loss	MSE
Synergy positive threshold (AUROC)	Loewe score > 10
I_ψ bilinear rank	32
I_ψ latent dimension $d_\phi = d_z$	64
I_ψ acyclicity reg. weight	0.01

D.5 Interpretability: interaction contribution

The bilinear decomposition $z = z^{\text{main}} + z^{\text{inter}}$ allows a natural post-hoc attribution:

$$\Delta_{\text{int}} \;=\; \text{head}\big(\rho(z^{\text{main}} + z^{\text{inter}})\big) \;-\; \text{head}\big(\rho(z^{\text{main}})\big)\,,$$

the change in predicted synergy score when the interaction term is removed. This quantity is computed at test time on the DrugComb combo split and requires no additional fitting.

Figure 2 shows that synergistic pairs (Loewe score > 10) receive on average a higher Δ_{int} than non-synergistic pairs: mean $\Delta_{\text{int}} = +13.25$ for synergistic pairs (87% with $\Delta_{\text{int}} > 0$) vs. $+9.61$ for non-synergistic pairs (74% with $\Delta_{\text{int}} > 0$). Per-seed synergistic means are $+18.68$, $+15.69$, $+5.38$ vs. $+14.64$, $+15.80$, -1.60 for non-synergistic pairs. The interaction term contributes positively in most cases, but its magnitude is systematically larger for pairs the benchmark labels synergistic.

Note that this analysis does not rely on reading the learned adjacency weights A_{ij} directly: as Table 7 in the main paper shows, learned, uniform, and random adjacency weights yield equivalent predictive performance, so individual edge weights carry no reliable interpretable signal. The Δ_{int} decomposition instead compares the full model to its own additive restriction, a more robust basis for attribution.

Sensitivity to the synergy threshold. The standard threshold of > 10 used throughout the paper defines a rare-event detection task (about 1.6% of pairs). Table 12 shows how the classification metrics and the interpretability separation respond as the threshold is raised. Both the AUROC of I_ψ and the Δ_{int} gap between synergistic and non-synergistic pairs grow monotonically with the threshold. The gap over the FiLM baseline also widens, from $+0.022$ at threshold 10 to $+0.045$

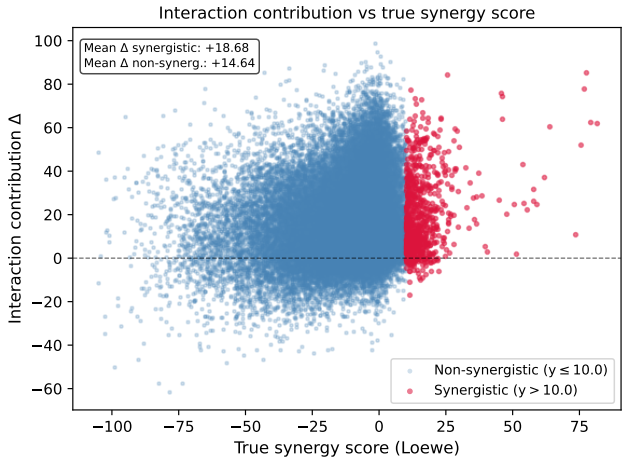

Figure 2: Interaction contribution Δ_{int} vs. ground-truth Loewe synergy score. Synergistic pairs (score > 10) receive on average a higher Δ_{int} than non-synergistic pairs. The magnitude of the gap grows with the synergy threshold (see Table 12). DrugComb combo split, seed 0.

at threshold 20. This is consistent with Theorem 1: the more strongly non-additive the target, the larger the gap between the additive class and I_ψ.

Table 12: Threshold sensitivity on DrugComb combo. AUROC over 5 seeds, Δ_{int} means over 3 seeds.

Threshold	AUROC I_ψ	AUROC FiLM	Gap	Δ_{int} syn	Δ_{int} non-syn	Sep.	% positives
> 5	0.764	0.742	+0.022	+11.73	+9.55	+2.18	5.5%
> 10	0.792	0.770	+0.022	+13.25	+9.61	+3.64	1.6%
> 15	0.820	0.796	+0.024	+16.56	+9.63	+6.93	0.6%
> 20	0.859	0.814	+0.045	+22.11	+9.64	+12.47	0.3%

Taken together, these results confirm two properties of I_ψ. First, the classification performance improves as the positive class becomes more strictly synergistic, which aligns with Theorem 1: a higher threshold selects pairs whose true interaction is farther from the additive regime, making the advantage of an explicit interaction module more pronounced. Second, the Δ_{int} separation grows proportionally with the threshold, showing that the bilinear term correctly assigns higher interaction credit to the most strongly synergistic pairs, independently of classification threshold. These two properties together provide evidence that the learned interaction representation is both discriminative and interpretable.