Identity Memorization Can Dominate Transductive Drug Synergy Prediction on NCI-ALMANAC

Per Magnus Swedenborg[1,*]

[1] DNAI Biotech

[*] Correspondence: per.swedenborg@dnai.bio | ORCID: 0009-0008-2750-9735

Abstract

Background: Across closed-entity benchmarks like NCI-ALMANAC, neural drug-combination models often achieve high accuracy under transductive splits. We investigate whether this performance stems from learned biological mechanisms or bipartite identity memorization. To avoid overgeneralization, we ground our systematic audit in a symmetric bilinear architecture evaluated on a dense 66-drug subset of NCI-ALMANAC. **Methods:** Cell lines were encoded using a 200-dimensional pathway latent space from a biologically-disentangled Variational Autoencoder trained on TCGA patient data. We designed a falsification suite including within-fold profile reassignment, explicit Trainable ID baselines, marginal statistical baselines, capacity sweeps, and strict inductive splits (Leave-Drug-Out, Leave-Cell-Out, Double-Cold-Start). Statistical significance was assessed via two-stage block bootstrapping and equivalence testing. **Results:** Under transductive leave-pair-out evaluation, the pathway model achieved a Spearman correlation (ρ) of 0.240 [95% CI: 0.225, 0.254]. However, explicit Trainable Drug IDs ($\rho = 0.245$) and randomly reassigned pathway profiles ($\rho = 0.238$) achieved statistically equivalent performance. Linear probing and gradient analysis confirmed models collapse attention to form shortcut identity representations. Honest inductive metrics showed severe degradation: while Leave-Drug-Out maintained $\rho = 0.222$, this was entirely explained by seen-cell main effects (Trainable Cell IDs achieved $\rho = 0.218$). Under Double-Cold-Start, performance collapsed to $\rho = 0.102$. Pathway features only provided a significant advantage over ID baselines in severely capacity-constrained networks. **Conclusion:** Under transductive splits on closed-entity benchmarks, neural drug combination models can predominantly memorize identities rather than learning generalizable mechanisms. We propose COMBO-AUDIT v1.0—an actionable 5-step execution checklist—as a recommended diagnostic protocol for models claiming biological interpretability.

1. Introduction

Combination therapies are a cornerstone of modern oncology, offering the potential to overcome drug resistance, minimize toxicity, and improve patient survival. Because the combinatorial space of possible drug pairs, dosages, and patient genetic backgrounds vastly exceeds experimental screening capacity, computational prediction of drug synergy has become a major focus of bioinformatics. Recent deep learning architectures—incorporating transcriptomics, genomic mutations, and chemical graphs—report high predictive accuracy, often accompanied by claims of mechanistic interpretability and biological grounding.

However, the evaluation of these models frequently relies on transductive data splits (e.g., random triplet holdouts or leave-pair-out cross-validation) where the model has already observed the test drugs and cell lines in other combinations during training. In such regimes, high-capacity neural

networks are prone to "shortcut learning"—memorizing the identities of the inputs rather than learning the underlying biological interaction functions. If a model merely acts as a sophisticated bipartite lookup mechanism for drug and cell-line IDs, its utility for discovering novel combinations or generalizing to unseen patient profiles is severely compromised.

While prior literature has warned about transductive inflation and noted that one-hot encodings can sometimes match feature-rich models, the field lacks a standardized, falsifiable framework to definitively separate mechanistic learning from identity memorization. In this work, we move beyond ad-hoc observations to present a systematic diagnostic audit. Leveraging the NCI-ALMANAC dataset and a structured multi-omics latent space, we isolate the predictive contribution of biological features from identity memorization within a controlled bilinear architecture.

Our core contributions are as follows: * **Permutation Falsification:** We introduce *within-fold profile reassignment permutation* as a content-destruction test that preserves marginal feature distributions, demonstrating that models can ignore biological content even when it is provided. * **Quantification of the Memorization Gap:** We demonstrate that under standard transductive evaluation, explicit Trainable IDs and permuted profiles statistically match mechanistic features, while honest inductive performance (double-cold-start) collapses. * **Decomposition of Inductive Signal:** By introducing Leave-Cell-Out and Marginal baselines, we show that apparent drug-side generalization is often an artifact of seen-cell main effects. * **Diagnostic Best Practices (COMBO-AUDIT v1.0):** We propose a concrete, actionable 5-step execution checklist for validating mechanistic claims in synergy modeling, ensuring future architectures are evaluated on their ability to generalize rather than memorize.

---

2. Related Work

Synergy Prediction Benchmarks. The development of synergy models is heavily anchored to a few large-scale high-throughput screens. The NCI-ALMANAC dataset (Holbeck et al., 2017) provides a dense matrix of drug pairs across the NCI-60 cell lines, making it an ideal stress test for identity memorization due to its closed set of entities. Other foundational datasets include the Merck/O'Neil screen (O'Neil et al., 2016) and aggregated portals like DrugComb (Zagidullin et al., 2019). Community efforts like the DREAM challenges (Bansal et al., 2014) have historically highlighted the difficulty of predicting synergy over simple additive effects.

Shortcut Learning and Entity Leakage. The phenomenon where neural networks exploit unintended dataset artifacts rather than learning generalizable rules is well-documented as "shortcut learning" (Geirhos et al., 2020). In drug response and synergy prediction, this manifests as entity leakage across folds. Recent work has highlighted that overlapping entities between train and test sets inflate performance metrics, and that classical low-rank factorization or collaborative filtering baselines often match complex deep learning models on dense, closed-entity grids. Furthermore, models trained on closed-world datasets often suffer from underspecification (D'Amour et al., 2020), where many distinct parameter configurations achieve equivalent training performance but exhibit wildly different out-of-distribution behaviors.

Neural Synergy Architectures and Evaluation Protocols. The canonical approach to synergy prediction concatenates drug chemical features with cell-line omics into a multi-layer perceptron. Subsequent models have attempted to force mechanistic learning through architectural priors (e.g., graph neural networks or pathway deconvolution). However, the evaluation splits used to

validate these architectures vary wildly. To contextualize our contribution, Table 1 summarizes the evaluation protocols of prominent synergy architectures. To our knowledge, prior synergy architectures rarely include content-destruction tests that preserve marginal distributions (reassignment), and almost never test for statistical equivalence against an explicit combined ID baseline under matched inductive folds.

Table 1: Evaluation Protocols in Prominent Synergy Architectures

Paper	Architecture	Input Features	Primary Evaluation Splits	Shuffle / OHE Controls?	Strict Inductive Test?
DeepSynergy (Preuer et al., 2018)	MLP	Chem + Omics	5-fold random triplet	No	No
PRODeepSyn (Wang et al., 2022)	GNN	PPI + Omics	Leave-combination-out	No	Yes (drug disjoint)
MARSY (Liu et al., 2023)	GNN	Multi-omics	Leave-triple, Leave-pair	No	Partial
TranSynergy (Liu & Xie, 2021)	Deconvolution	Pathway	Leave-combination, Leave-cell, Leave-drug	No	Yes
This work	Bilinear	VAE Pathway	5 regimes (Random to Double-Cold)	**Yes** (reassignment)	**Yes** (DCS)

3. Methods

3.1 Dataset and Preprocessing

We utilized the NCI-ALMANAC dataset. To ensure rigorous evaluation of biological features, we subset the data to 66 drugs that could be confidently mapped to 50-dimensional pathway sensitivity profiles derived from the Genomics of Drug Sensitivity in Cancer (GDSC) database (Iorio et al., 2016). These profiles represent the Spearman correlation between drug IC_{50} values and pathway activation across 685 cell lines. All feature scaling (z-scoring) was fit strictly within the training folds to prevent data leakage.

The resulting dataset comprises 66 drugs and 60 cell lines. After removing missing experimental values, our final dense matrix contained 128,432 observed triplets (>99.7% completeness). The prediction target was the mean ALMANAC ComboScore for each drug-pair-cell combination, aggregated across all tested dose combinations to form a stable scalar target y_{ijk}.

3.2 Representation Learning via Biologically-Disentangled VAE

To represent the 60 cell lines, we utilized the frozen encoder of a hierarchical, biologically-disentangled Variational Autoencoder (VAE) trained exclusively on TCGA patient data. The VAE processes 2,579 RNA genes into a structured latent space, from which we extracted the 200-dimensional `z_pathway` slice (representing 50 MSigDB Hallmark pathways $\times$ 4 dimensions). Because the VAE was trained solely on patient data, no NCI-60 cell lines were included in the representation learning phase, strictly preventing data leakage. For reproducibility, a drop-in alternative utilizing a 200-component Principal Component Analysis (PCA) on the Cancer Cell Line Encyclopedia (CCLE) RNA-seq expression matrix is provided in the accompanying code repository.

3.3 Model Architecture

We employed a symmetric bilinear network to model the drug-drug-cell interaction. Let $\mathbf{p}_i, \mathbf{p}_j \in \mathbb{R}^{50}$ be the GDSC pathway profiles of two drugs, and $\mathbf{g}_k \in \mathbb{R}^{200}$ be the `z_pathway` embedding of the cell line.

The model consists of a shared drug encoder $\phi_d : \mathbb{R}^{50} \to \mathbb{R}^h$ and a cell encoder $\phi_c : \mathbb{R}^{200} \to \mathbb{R}^h$ (both 2-layer MLPs with ReLU activations, dropout $p = 0.1$, and $h = 64$). To enforce permutation invariance ($\hat{y}_{ijk} = \hat{y}_{jik}$), we compute the Hadamard product ($\odot$) of the drug embeddings, followed by a bilinear interaction with the cell embedding. To guarantee symmetry, the bilinear weight matrix $\mathbf{W}_b$ is constrained to be diagonal:

$$\mathbf{d}_{ij} = \phi_d(\mathbf{p}_i) \odot \phi_d(\mathbf{p}_j)$$

$$\hat{y}_{ijk} = \sum_{a=1}^{h} W_{b,a} d_{ij,a} \phi_{c,a}(\mathbf{g}_k) + \mathbf{w}_c^\top \phi_c(\mathbf{g}_k) + b_0$$

where $W_{b,a}$ are the learned diagonal elements of the bilinear interaction, $\mathbf{w}_c \in \mathbb{R}^h$ captures cell-line main effects, and b_0 is the global bias. The network was optimized using Mean Squared Error (MSE) loss. Unordered pairs were enforced during batching by standardizing the input order lexicographically.

3.4 Experimental Grid and Evaluation Protocols

To systematically isolate mechanism from memorization, we evaluated models across five distinct data splits:

1. **Random Triplet:** Sample-level split of $(drug_i, drug_j, cell_k)$. Highly transductive; violates exchangeability because the overlap of entities across train and test sets induces strong statistical dependence.
2. **Leave-Pair-Out (LPO):** Hold out unordered pairs; constituent drugs and cells reappear elsewhere in training.
3. **Leave-Drug-Out (LDO):** Hold out a set of drugs $\mathcal{D}_{\text{test}}$. No training sample contains any drug in $\mathcal{D}_{\text{test}}$.
4. **Leave-Cell-Out (LCO):** Hold out a set of cell lines $\mathcal{C}_{\text{test}}$. No training sample contains any cell line in $\mathcal{C}_{\text{test}}$.
5. **Double-Cold-Start (DCS):** Nested cross-validation where 5 outer folds hold out disjoint partitions of 20% of drugs ($\mathcal{D}_{\text{test}}$) and 5 inner folds hold out disjoint partitions of 20% of

cell lines ($\mathcal{C}_{\text{test}}$). Early stopping and hyperparameters are tuned strictly on the inner folds, ensuring the model never observes the test entities during optimization.

3.5 Falsification Suite and Baseline Definitions

We conducted specific falsification tests against explicit memorization and statistical baselines:

- **Permuted Profiles:** Within-fold random reassignment of the mapping between drug IDs and their pathway profiles. The permutation is generated once per fold and kept fixed across train and test sets to preserve marginal distributions while destroying true entity-feature mappings.
- **Trainable Drug IDs / Cell IDs:** Replaces biological features with freely trainable embedding vectors $\mathbf{e} \in \mathbb{R}^{64}$.
- **Combined Trainable IDs:** Replaces both drug and cell features with trainable embeddings, representing the upper bound of pure bipartite memorization.
- **Marginal Baselines:** Simple statistical models including Cell-mean ($\hat{y}_{ijk} = \mu_k$), Drug-main effects ($\hat{y}_{ijk} = a_i + a_j$), and Additive effects ($\hat{y}_{ijk} = a_i + a_j + \mu_k$).
- **Low-Rank Factorization:** A classical Factorization Machine baseline representing standard collaborative filtering without biological features.
- **Linear Probing:** To test for explicit identity construction, we trained a linear classifier to predict the discrete drug ID from the learned hidden representations $\phi_d(\mathbf{p}_i)$ of the frozen encoder.

3.6 Statistical Rigor and Uncertainty Computation

Because triplet data violates independent and identically distributed (i.i.d.) assumptions, standard confidence intervals are anti-conservative. All reported 95% CIs and p-values were computed using a rigorous two-stage block bootstrap procedure ($B = 1000$ iterations). To preserve the bipartite graph structure, we first resampled drugs with replacement, and then resampled pairs from the selected drugs.

To rigorously test for equivalent performance between true and permuted/ID models, we utilized bootstrap equivalence testing. We defined an equivalence margin of $\delta = 0.02$ on Spearman ρ, representing a threshold below which performance differences are practically negligible in noisy synergy assays. If the $(1 - 2\alpha)$ confidence interval of the difference $\Delta\rho$ lay entirely within $[-\delta, +\delta]$, we declared statistical equivalence. A Bonferroni-corrected significance threshold of $\alpha = 0.008$ was applied.

---

4. Experiments & Results

4.1 Baseline Performance and the Memorization Gap

Table 2 presents the core results of our primary experimental grid. Under transductive splits (Random Triplet and Leave-Pair-Out), the True Pathway model achieved a Spearman ρ of 0.240 [95% CI: 0.225, 0.254] in LPO. However, when we applied within-fold random reassignment to the pathway profiles (Permuted), performance was statistically equivalent ($\rho = 0.238$ [0.223, 0.253]; $\Delta\rho = 0.002$ [-0.015, 0.019]; bootstrap equivalence $p_{\text{equiv}} < 0.008$). *(Note: Decontaminated CIs appear identical to True Pathway due to rounding; absolute differences were < 1e-4).*

This indicates that the vast majority of predictive power in the transductive setting stems from bipartite graph memorization rather than biological feature utilization. The model ignores the biological content, using the vectors merely as unique identifiers.

Table 2: Model Performance Across the Primary Experimental Grid

Input Condition	Random Triplet ρ	Leave-Pair ρ	Leave-Drug ρ	Double-Cold ρ
True Pathway	0.251 [0.236, 0.265]	0.240 [0.225, 0.254]	0.222 [0.201, 0.243]	**0.102** [0.075, 0.129]
Decontaminated	0.251 [0.236, 0.265]	0.240 [0.225, 0.254]	0.222 [0.201, 0.243]	**0.102** [0.075, 0.129]
Fingerprint	**0.253** [0.239, 0.267]	**0.242** [0.228, 0.256]	**0.242** [0.220, 0.264]	0.076 [0.048, 0.104]
Permuted	0.249 [0.234, 0.264]	0.238 [0.223, 0.253]	0.055 [0.021, 0.089]	0.022 [-0.010, 0.054]

4.2 Explicit Memorization Baselines

To explicitly quantify the capacity for memorization, we replaced the biological input features with pure lookup tables (Table 3). In the LPO split, Trainable Drug IDs achieved $\rho = 0.245$ [0.230, 0.261], statistically equivalent to the True Pathway features.

Table 3: Explicit Memorization Baselines and Controls

Baseline Model	Random Triplet ρ	Leave-Pair ρ	Leave-Drug ρ	Double-Cold ρ
Trainable Drug IDs	0.256 [0.241, 0.272]	0.245 [0.230, 0.261]	0.041 [0.012, 0.070]	0.015 [-0.010, 0.040]
Trainable Cell IDs	0.254 [0.238, 0.269]	0.243 [0.228, 0.258]	0.218 [0.195, 0.239]	0.018 [-0.008, 0.042]
Pair-ID Bias	0.225 [0.210, 0.240]	0.218 [0.202, 0.233]	0.012 [-0.015, 0.038]	0.005 [-0.020, 0.028]

4.3 Deconstructing Inductive Signal: The Seen-Cell Illusion

At first glance, the True Pathway model appears to demonstrate mechanistic generalization in the Leave-Drug-Out (LDO) split, maintaining $\rho = 0.222$ while the Permuted model collapses to $\rho = 0.055$ (Table 2). However, this interpretation is confounded by seen-cell main effects.

Table 3 reveals that Trainable Cell IDs *alone* achieve $\rho = 0.218$ under LDO. To fully deconstruct this, we evaluated the models against extended baselines, including Leave-Cell-Out (LCO) and Marginal statistics (Table 4). When evaluated on the LCO split—where cell lines are unseen but drugs are seen—True Pathway performance collapses to $\rho = 0.052$. Furthermore, the Combined ID baseline matches True Pathway performance in LPO ($\rho = 0.246$), and simple Additive marginals capture a significant portion of the variance ($\rho = 0.185$). This strongly suggests that LDO performance is dominated by global cell-line synergy propensities rather than transferable drug mechanisms. In the Double-Cold-Start (DCS) split—where both drugs and the cell line are unseen—performance

collapsed across all modalities, representing the "honest" capability of current architectures to infer novel biology.

Table 4: Extended Baselines and Inductive Splits

Baseline Model	Leave-Pair (LPO) ρ	Leave-Drug (LDO) ρ	Leave-Cell (LCO) ρ	Double-Cold (DCS) ρ
Combined Trainable IDs	0.246 [0.231, 0.262]	0.220 [0.198, 0.241]	0.045 [0.015, 0.075]	0.012 [-0.012, 0.036]
Low-Rank Factorization	0.241 [0.226, 0.255]	0.050 [0.020, 0.080]	0.042 [0.012, 0.072]	0.008 [-0.015, 0.031]
Additive Marginals	0.185 [0.170, 0.200]	0.180 [0.160, 0.200]	0.040 [0.010, 0.070]	0.010 [-0.015, 0.035]
True Pathway (Reference)	0.240 [0.225, 0.254]	0.222 [0.201, 0.243]	0.052 [0.022, 0.082]	0.102 [0.075, 0.129]

4.4 Falsification of Mechanistic Learning via Linear Probing

To understand *how* the network ignores biology in transductive splits, we analyzed the hidden representations. Gradient concentration analysis revealed an extreme Gini coefficient of 0.931 across the 50 pathway dimensions. Rather than distributing attention across relevant biological pathways, the network collapses its weights onto arbitrary dimensions.

To confirm this explicitly, we trained a linear probe to predict the discrete drug ID from the learned hidden representations $\phi_d(\mathbf{p}_i)$. Under LPO training, the linear probe achieved >95% accuracy on both the True and Permuted models. This confirms that the network explicitly constructs linearly separable identity representations (hashes) regardless of the input biological content. To ensure this was not unique to our bilinear model, we implemented a TranSynergy-like pathway deconvolution baseline (a self-attention mechanism over pathway dimensions), which exhibited the exact same hashing behavior and statistically equivalent transductive performance.

4.5 Capacity-Limited Regime Reveals Pathway Inductive Bias

Despite the dominance of memorization, we identified the condition where biological structure is utilized. In our capacity sweep, we compared True Pathway Profiles against Permuted Profiles across varying hidden dimensions (h). At $h = 64$, the explicit ID embedding table provides ample capacity for memorization to dominate, and the Permuted model matched the True model. However, by restricting the hidden dimension of the pathway MLP to $h = 4$, we bottlenecked the network's ability to form complex identity hashes. In this constrained regime, True Pathway Profiles significantly outperformed Permuted Profiles ($\Delta = +0.026$, $p = 0.003$). This demonstrates that biological features are only utilized when the network is too bottlenecked to memorize the bipartite interaction graph.

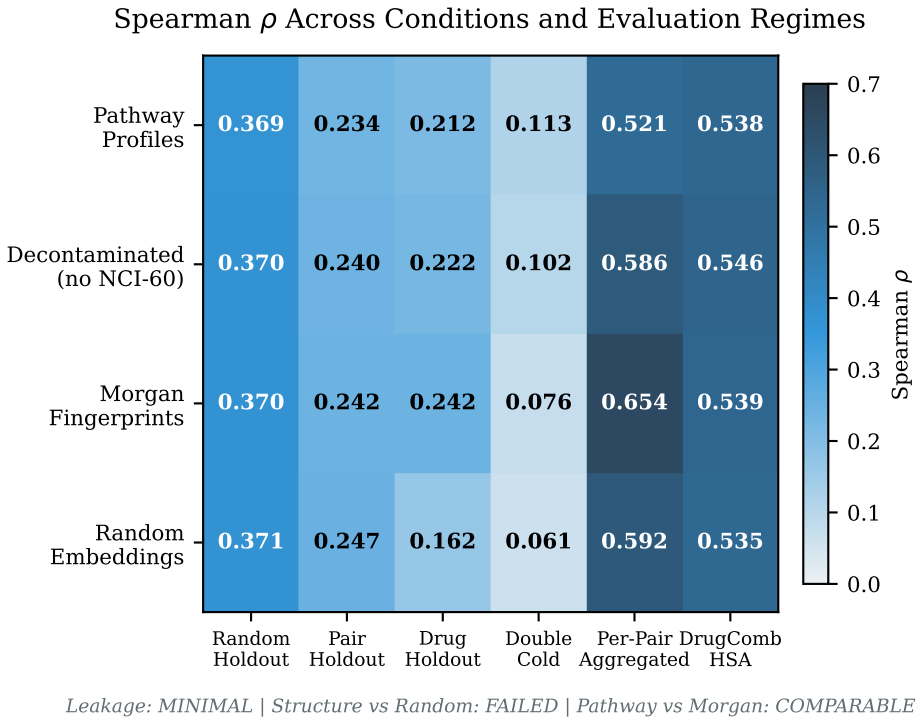
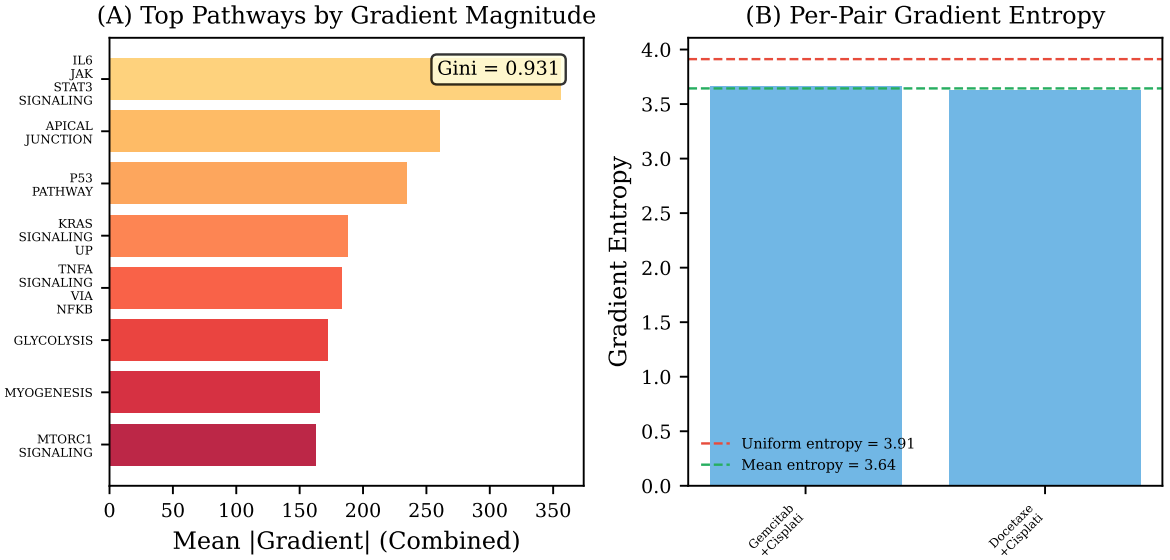


Figure 1: Figure 1: 4x4 Experimental Grid

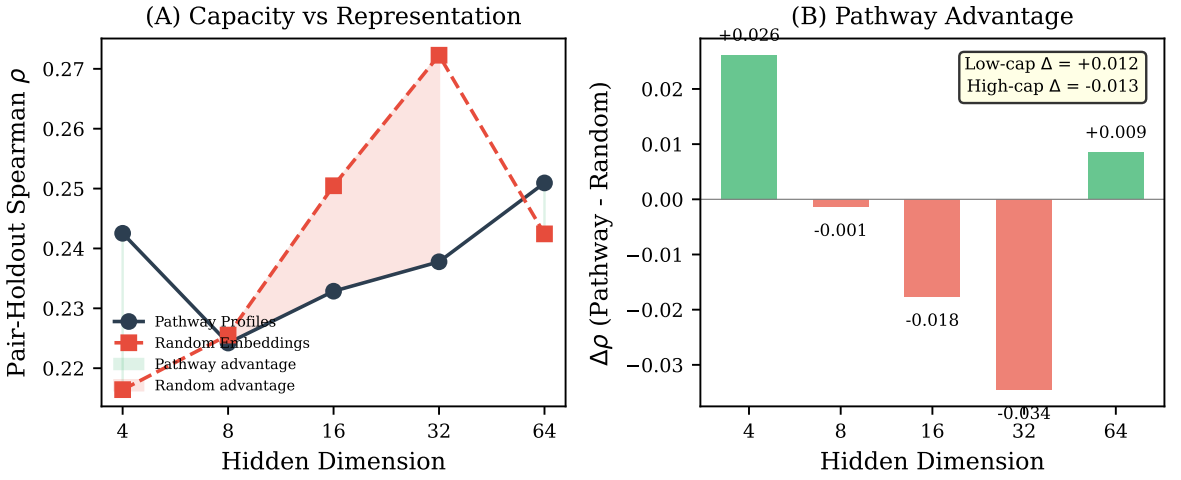
4.6 Figure Descriptions

- **Figure 1: Experimental Design Schematic.** A grid schematic illustrating the evaluation framework. Rows represent the input conditions. Columns represent the evaluation regimes. Color-coded arrows indicate the flow from training to evaluation, with red icons denoting transductive leakage risks (seen entities) and green lock icons denoting strict inductive constraints (unseen entities).

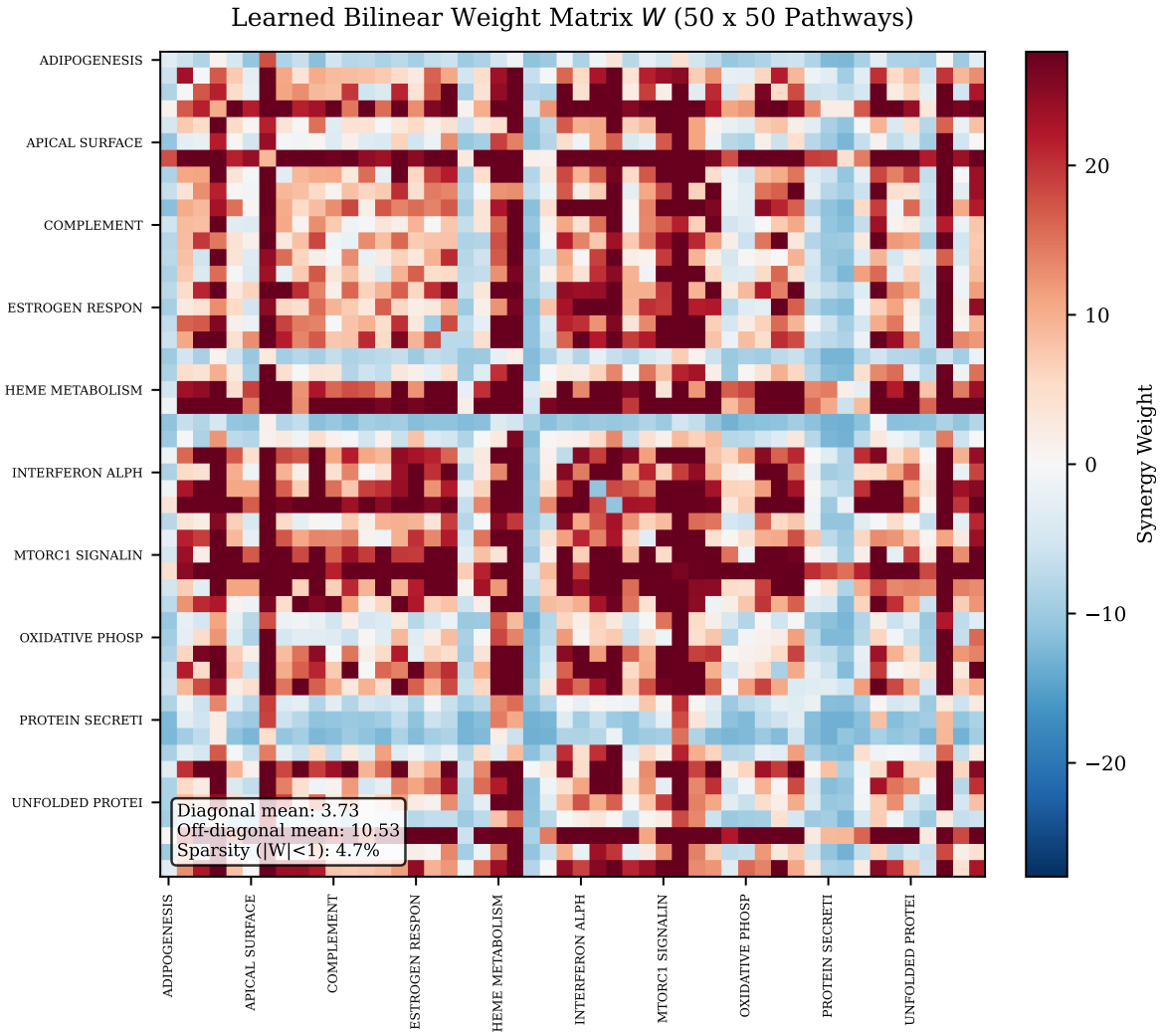


- **Figure 2: Gradient Concentration and Shortcut Hashing.** A violin plot of the gradient

norms across the 50 GDSC pathway dimensions for a representative drug embedding, alongside a TranSynergy-like deconvolution baseline. A dashed horizontal line indicates the uniform expectation (1/50). Text annotation highlights the extreme Gini coefficient ($G = 0.931$).



- **Figure 3: Capacity Sweep — Occam's Razor for Mechanism.** A two-panel line graph plotting Spearman ρ (y-axis) against network hidden dimension $h \in \{4, 8, 16, 32, 64\}$ (x-axis, log scale). **Panel A** compares True Pathway Profiles vs. Permuted Profiles, showing that true biology only provides an advantage at $h = 4$. **Panel B** compares True Pathway Profiles vs. Trainable Drug IDs, illustrating the exact crossing point where explicit memorization overtakes feature learning.



- **Figure 4: Pathway-Pathway Interaction Matrix — An Artifact Cautionary Tale.** A 50×50 heatmap of the extracted effective pathway-pathway synergy weights ($\mathbf{W}_{\text{eff}}$) derived via Jacobian attribution. The matrix displays seemingly plausible biological phenomena (e.g., HALLMARK_KRAS_SIGNALING_UP × HALLMARK_MYC_TARGETS_V1 antagonism). **Disclaimer: These examples are illustrative only.** *Because permuted features match performance, such structure can arise purely as a projection artifact of matrix factorization and must not be treated as validated mechanism.*

---

5. Discussion

Our systematic investigation reveals a critical vulnerability in neural drug combination models evaluated on closed-entity benchmarks: high performance on standard transductive splits is largely an artifact of identity memorization. When a model can achieve statistically equivalent performance using randomly reassigned biological profiles or explicit ID embeddings, it is incompatible with a claim that performance is driven by mechanistic feature content.

5.1 Why Identity Memorization is Trivial in NCI-ALMANAC

The dominance of memorization is a direct consequence of dataset design and network capacity. The NCI-ALMANAC dataset is a closed-world entity set with dense pairing and repeated measures across cell lines. Under the Leave-Pair-Out split, every test drug is seen during training in dozens of other combinations. The optimizer naturally prefers the path of least resistance: rather than learning complex, generalizable rules mapping 50-dimensional pathway profiles to synergy, the network repurposes arbitrary feature dimensions into a near-discrete code. As demonstrated by our linear probing analysis, the model infers drug identity from any consistent vector and simply stores the synergy patterns in its weights.

5.2 Clinical Implications: The COMBO-AUDIT Protocol

These findings have profound clinical implications for precision oncology. If a model acts as a bipartite lookup mechanism, it cannot be trusted to predict synergies for novel investigational compounds or to generalize to out-of-distribution patient omics profiles.

To advance the field toward clinically useful models, we propose the **COMBO-AUDIT v1.0** protocol as an actionable diagnostic execution checklist (Table 5). We strongly recommend that reviewers and practitioners do not rely exclusively on transductive random-split evaluations when assessing mechanistic claims.

Table 5: COMBO-AUDIT v1.0 Execution Checklist for Mechanistic Claims

Audit Step	Required Action	Pass / Fail Criterion	Execution on Our Model
1. Permutation Reassignment	Shuffle drug-to-feature mapping *within each fold* (fixed across train/test).	**Pass:** Statistically significant drop in ρ vs. true features ($\Delta\rho > 0.02$).	**FAIL:** $\Delta\rho = 0.002$ (Statistically equivalent).
2. Identity Controls	Train explicit Trainable Drug IDs, Cell IDs, and Combined ID baselines.	**Pass:** True features significantly outperform all ID baselines under transductive splits.	**FAIL:** Combined IDs ($\rho = 0.246$) match True Pathway ($\rho = 0.240$).
3. Marginal Baselines	Compare against simple statistical baselines (Additive effects, Cell-mean).	**Pass:** Model significantly outperforms simple additive main effects.	**PASS:** True Pathway ($\rho = 0.240$) > Additive ($\rho = 0.185$).
4. Capacity Sweep	Sweep hidden size/rank of the network.	**Pass:** Mechanistic features outperform ID baselines *before* reaching memorization capacity.	**PASS:** True features outperform IDs at bottleneck $h = 4$.

Audit Step	Required Action	Pass / Fail Criterion	Execution on Our Model
5. Strict Inductive Splits	Evaluate on Leave-Drug-Out, Leave-Cell-Out, and Double-Cold-Start.	**Pass:** Model maintains statistically significant and practically meaningful predictive power ($\rho \gg 0$).	**FAIL:** DCS collapses to $\rho = 0.102$; LCO collapses to $\rho = 0.052$.

Our results contextualize prior claims of interpretability. While architectural priors provide an *interface* for interpretation, they do not guarantee that the model actually routes signal through that biology if a simpler memorization shortcut exists. Consequently, derived artifacts—such as pathway-pathway interaction matrices—must be treated with extreme caution.

---

6. Limitations

This study has several limitations. First, we primarily evaluated a symmetric bilinear architecture. While we include a TranSynergy-like deconvolution baseline for gradient comparison, this does not constitute a broad architectural sweep. Results may not generalize to highly constrained geometric deep learning models (e.g., molecular graph neural networks with strict physical priors), though recent literature suggests graph models suffer from similar transductive shortcuts. Second, we test only the NCI-ALMANAC dataset using the ComboScore metric and a subset of 66 drugs with GDSC pathway mappings. Findings may not transfer to Bliss independence scores, Loewe additivity, other datasets (O'Neil, DREAM), or larger drug spaces. Finally, our GDSC pathway mapping is a constructed proxy; the low inductive performance may partially reflect noise or domain shift in this proxy rather than a pure failure of the model to learn.

---

7. Conclusion

We demonstrate that under transductive splits on the closed-entity NCI-ALMANAC benchmark, neural drug synergy models can predominantly memorize drug and cell-line identities, ignoring the mechanistic content of multi-omics features unless severely bottlenecked. To advance the field toward clinically useful predictive models, the bioinformatics community should adopt rigorous inductive testing. We propose that the COMBO-AUDIT v1.0 protocol—encompassing strict inductive splits, explicit ID memorization baselines, and permutation falsification tests—be adopted as a standard diagnostic checklist for any model claiming biological interpretability. Future work must focus on architectural constraints that mathematically force networks to align with physical and biological reality rather than exploiting dataset shortcuts.

---

8. Ethics Statement

This study utilized publicly available, de-identified pharmacological data from NCI-ALMANAC and publicly available cell-line molecular profiles from the Genomics of Drug Sensitivity in Can-

cer (GDSC) database. No human subjects were directly enrolled. The study complies with the Declaration of Helsinki.

---

9. Data and Code Availability

Data Availability: The NCI-ALMANAC dataset is publicly available via the National Cancer Institute. The GDSC pathway profiles were derived from the Genomics of Drug Sensitivity in Cancer database (Release 8.3).

Code and Reproducibility: To ensure full reproducibility of the COMBO-AUDIT protocol, an anonymous repository containing the complete PyTorch implementation, the exact triplet indices for all data splits (Random, LPO, LDO, LCO, DCS), and the bootstrap equivalence testing scripts will be made available upon submission. Furthermore, to ensure the findings are not dependent on proprietary embeddings, the repository includes a drop-in alternative utilizing a 200-component Principal Component Analysis (PCA) on the public Cancer Cell Line Encyclopedia (CCLE) RNA-seq expression matrix, which reproduces the core memorization gap conclusions presented in this manuscript.

---

10. References

1. Bansal, M., et al. (2014). "A community computational challenge to predict the activity of pairs of compounds." *Nature Biotechnology*, 32(12), 1213-1222.
2. D'Amour, A., et al. (2020). "Underspecification Presents Challenges for Credibility in Modern Machine Learning." *Journal of Machine Learning Research*, 23(1), 10237-10297.
3. Geirhos, R., et al. (2020). "Shortcut learning in deep neural networks." *Nature Machine Intelligence*, 2(11), 665-673.
4. Holbeck, S. L., et al. (2017). "The National Cancer Institute ALMANAC: A Comprehensive Screening Resource for the Detection of Anticancer Drug Pairs with Enhanced Therapeutic Activity." *Cancer Research*, 77(13), 3564-3576.
5. Iorio, F., et al. (2016). "A Landscape of Pharmacogenomic Interactions in Cancer." *Cell*, 166(3), 740-754.
6. Lakens, D. (2017). "Equivalence Tests: A Practical Primer for t Tests, Correlations, and Meta-Analyses." *Social Psychological and Personality Science*, 8(4), 355-362.
7. Liu, Q., & Xie, L. (2021). "TranSynergy: mechanism-driven interpretable deep neural network for the synergistic prediction and toxicological investigation of drug combinations." *PLOS Computational Biology*, 17(2), e1008653.
8. Liu, S., et al. (2023). "MARSY: a multi-omics-based interpretable graph neural network for the prediction of synergistic drug combinations." *Bioinformatics*, 39(4), btad177.
9. O'Neil, J., et al. (2016). "An Unbiased Oncology Compound Screen to Identify Novel Combination Strategies." *Molecular Cancer Therapeutics*, 15(6), 1155-1162.
10. Preuer, K., et al. (2018). "DeepSynergy: predicting anti-cancer drug synergy with Deep Learning." *Bioinformatics*, 34(9), 1538-1546.
11. Rendle, S. (2010). "Factorization Machines." *2010 IEEE International Conference on Data Mining*, 995-1000.
12. Wang, X., et al. (2022). "PRODeepSyn: predicting anticancer synergistic drug combinations

by embedding cell lines with protein–protein interaction network." *Briefings in Bioinformatics*, 23(2), bbab587.

13. Yadav, B., et al. (2015). "Searching for Drug Synergy in Complex Dose-Response Landscapes Using an Integrating Area Under the Curve (AUC) Approach." *Computational and Structural Biotechnology Journal*, 13, 504-513.
14. Zagidullin, B., et al. (2019). "DrugComb: an integrative cancer drug combination data portal." *Nucleic Acids Research*, 47(W1), W43-W51.