Combination Overlap in Drug Synergy Benchmarks: An Audit of Cell-Line Holdout Protocols

Kusal Debnath and Preetam Ghosh†

Department of Computer Science, Virginia Commonwealth University, Richmond, Virginia 23284, USA

†*E-mail: pghosh@vcu.edu*

Drug-combination synergy models are generally benchmarked on combinatorial screens, in which a fixed panel of drug pairs is assayed across a fixed panel of cell lines. Random splits over measurements are permissive in nature, as the same drug pair may show up in both the train and test set, albeit alongside different cell lines, allowing the model to interpolate rather than generalize. As a possible mitigation, the field has adopted cold-start protocols, namely, leave-drug-out, leave-cell-line-out, leave-tissue-out, and leave-combination-out. Nevertheless, these protocols differ in what they actually withhold. In this work, we show that holding out a cell line does not hold out a combination. Whenever a screen assays its pair matrix across its entire cell panel, leave-cell-line-out retains every test combination in training by construction. This holds exactly on the fully-crossed ALMANAC, FRIEDMAN, and O'Neil screens, and reaches 92% on the sparser AstraZeneca screen. Additionally, a harmonised DrugComb corpus still retains 72.8%, suggesting that pooling across screens dilutes the effect instead of removing it. To further quantify the consequence, we compare a null model using no molecular information, comprised of a per-cell-line mean along with a per-pair mean, against a symmetric neural baseline and a larger pathway-aware model under matched splits. Under the leave-cell-line-out setup, the null model outperforms both, and collapses when combinations are held out instead. Between leave-combination-out and leave-cell-line-out, the sign of structure's contribution flips, while leave-drug-out yields no resolvable difference and leave-tissue-out is too variable across folds to adjudicate. Moreover, for this one pathway model, ablation locates the advantage in the drug-target branch rather than in pathway topology, although the branch's contribution cannot be separated from the capacity and auxiliary supervision it also supplies. Our work thus emphasizes that predicting a known combination in an uncharacterized context is legitimate and clinically relevant, but distinct from combination generalization; retained overlap is a property of a protocol to be reported alongside the claim it supports, rather than a defect to be corrected.

Keywords: Drug combination synergy; evaluation methodology; cold-start evaluation; train–test overlap; benchmark design; combinatorial drug screens

1. Introduction

In cancer treatment, combination therapy has emerged as the standard protocol. Nevertheless, exhaustive screening of compounds is tedious due to the large candidate space. In silico adoption for screening, pioneered by Preuer et al.,[1] aids in this scenario.

© 2026 The Authors. Open Access chapter published by World Scientific Publishing Company and distributed under the terms of the Creative Commons Attribution Non-Commercial (CC BY-NC) 4.0 License.

Evolution of research in this field has been majorly measured against benchmarks derived from combinatorial screens, namely, O'Neil,[2] NCI-ALMANAC,[3] and their aggregation in DrugComb.[4,5] These screens share a design property that is a virtue experimentally but a hazard statistically: a fixed panel of drug pairs is assayed across a fixed panel of cell lines. As a result, random shuffling and splitting causes the same cell lines to appear in both the train and test sets, just with different pairs. During testing, the model encounters drug pairs and cell lines it has already seen, but with a new intersection. Therefore, the credibility of such test results is a matter of debate due to the inherent bias in the predictions.

Evaluating on novel combinations instead of random measurements, Preuer et al. reported that every method they compared degraded when extrapolating to unexplored drugs or cell lines. Later, Dewulf et al. formalized the cold-start subtasks for drug-drug interaction prediction and argued that matched validation schemes are critical to correct assessment.[6] Recent methods routinely reported leave-drug-out, leave-combination-out and leave-cell-line-out alongside headline numbers.[7–9]

As an instance of a general problem, Kapoor et al. surveyed seventeen scientific fields and observed leakage affecting 294 papers, often producing extremely overoptimistic conclusions. In their reproducibility study of civil-war prediction, correcting leakage eliminated the entire advantage of complex models over vanilla logistic regression.[10,11] Bernett et al. emphasized that leakage is especially hard to detect in biological data due to the complex dependencies between samples,[12] while Rosenblatt et al. showed that repeated subjects drastically inflate performance in connectome models.[13]

In this work, we seek an answer to a question that, to our knowledge, has not been addressed: within each cold-start protocol, how much combination overlap actually survives? Holding out a cell line removes a context, not the combination, as the same pair is measured in the cell lines that remain. The importance of such a scenario is an empirical question related to the redundancy of the screens.

We treat these protocols as definitions of distinct tasks rather than as defective instruments. Leave-cell-line-out defines a well-posed and clinically meaningful task, predicting a profiled combination in an uncharacterized context, and retaining that combination in training is precisely what the task licenses. What we audit is the distance between that definition and the generalization claims commonly attached to the numbers it produces.

In summary, this work is best placed as a method to characterize what each evaluation protocol withholds and what follows for the interpretation of reported performance. We quantify how much combination overlap each cold-start protocol actually leaves, protocol by protocol and screen by screen. Specifically, our work thus contributes the following:

- An audit across six protocols to measure the fraction of test combinations also present in training,
- A null model quantifying the consequence, using a per-cell plus per-pair mean,
- A demonstration that the protocol determines the verdict on held-out combinations, and,
- A component attribution to biological priors that might partially reflect database completion.

2. Background and Related Work

2.1. *Synergy Quantification*

Synergy is defined relative to a reference model of non-interaction. Standard references are Loewe additivity, Bliss independence, Highest Single Agent(HSA), and Zero Interaction Potency(ZIP). These references depict different assumptions and disagree in practice,[14–16] to the point that the choice of dose-response fitting software can flip a combination between synergistic and antagonistic.[17] In this work, we use Loewe as the reference model for quantifying synergy.

2.2. *Cold-Start Protocols*

As one of the earliest approaches, Dewulf et al. identified the distinct subtasks surfaced when drugs, pairs or contexts are novel, and proposed matched validation schemes.[6] In their review, Abbasi et al. concluded with a classic cold-start issue by stating that the methods they implemented performed well for known drugs and cell lines but fall short for novel ones.[18]Yousefabadi et al. showcased that permissive cross-validation allowing drug pairs across folds inflates performance, and thus adopted pair-held-out validation for antibacterial synergy.[19] In a more recent work, Agboola et al. reported a random forest outperforming their deep model under cold-drug validation.[20]

2.3. *Structure-informed Synergy Models*

Incorporating pathway or network structure is an active direction, spanning network-proximity approaches,[21] pathway-activation scores,[22] interpretable subsystem models,[23] graph and transformer architectures,[9,24,25] multi-task formulations,[26,27] multi-omics ensembles[28,29] and proteomics-based models.[8]

2.4. *Leakage in ML-based Science*

Non-independence between train and test samples is one of eight leakage types discussed by Kapoor et al. in their taxonomy,[10,11] addressed by reporting checklists[12,30] and quantified in neuroimaging[13] and other applied fields.[31,32] Cold-start protocols have been already adopted in the subfield as remedy, however, the shortcomings of such a remedy can only be seen by scrutinizing pairs.

Combination overlap is a measurable property of a partition: the fraction of test combinations that also occur in training, computable without fitting a model. Task mismatch is a gap between the entity a protocol withholds and the entity a claim generalizes over, as when leave-cell-line-out, which withholds biological contexts, is used to support a claim about unseen combinations. Leakage is the case in which that gap invalidates the performance estimate for the claim being made. Overlap alone is therefore neither necessary nor sufficient for leakage: a protocol that retains most of its test combinations is performing exactly as intended when the claim concerns a known combination in a new context, and the same retention becomes leakage only under a claim of combination generalization.

In a recent work, Çandır et al.[33] replaced the drug and cell-line representations of five published synergy models with one-hot identifiers and obtained comparable or slightly better performance, concluding that these models exploit identity-linked covariation in the synergy labels rather than the chemical and molecular features they are supplied.

3. Methods

3.1. *Corpus*

We harmonized DrugComb[4] to 634,631 measurements after removing self-pairs and rows lacking a Loewe score, resulting in 73,108 unique unordered pairs across 288 cell lines. Eliminating instances of incomplete drug and cell features drastically shrinks the measurements to 180,746 over 1,945 drugs, 159 cell lines and 15 tissues, drawn from ALMANAC, FRIEDMAN, O'NEIL and ASTRAZENECA. Every protocol, split, overlap figure and performance metrics reported in this study is computed on this feature-complete subset of 180,746 measurements.

3.2. *Protocols*

Here, six protocols are implemented over the common corpus: random, leave-combo-out (unseen combination, both drugs seen), leave-drug-out (one/both variants), leave-cell-out, leave-tissue-out, leave-study-out. Each of the protocols persisted as row indices with a corpus fingerprint so a split replays exactly without redistributing data.

The drug degree distribution is heavy-tailed: a small number of drugs participate in a large share of the pairs and therefore carry a disproportionate share of the rows. Partitioning entities uniformly is consequently unstable, as a single high-degree entity landing on the held-out side moves the realized test fraction far from its nominal value. Instead of uniform partitioning, we therefore assigned entities by mass-aware greedy packing. The packer orders groups by row mass, assigns each to the bin with the largest remaining deficit, and asserts the achieved test fraction before a split is accepted. Every split reported in this paper is produced by this packer on the 180,746-row feature-complete corpus.

The assertion applies to the mass-packed protocols only. The leave-tissue-out and leave-study-out partition on an explicit categorical assignment. As a result, their test fraction is essentially what the held-out block weighs, hence the check is waived consciously. The leave-tissue-out is restricted to the eight tissues with at least 3,500 covered rows, making the distribution severely skewed (skin 71,638 rows, 39.6% of the 180,746-row feature-complete corpus; pancreas 6 rows).

3.3. *Baselines*

We denote the null model as B1. For a test row carrying unordered pair p in cell line c, B1 predicts $y(p, c) = m_c + d_p$, where m_c is the mean Loewe score over the training rows of the cell line c and d_p is the mean of the residual $y_i - m_{c(i)}$ over the training rows i and pair p, both estimated on training rows only. Next, the unseen entities back off: m_c is replaced by the global training mean when c contributes no training row, and d_p is set to zero when p contributes none, so a pair unseen in a cell line unseen receives the global training mean as a constant prediction.

B1 therefore has no fitted weights, no regularization and no hyperparameters, and its score under a given protocol is a property of the split rather than of a training procedure. It uses no structure, targets, pathways or omics since the objective here is diagnosing a benchmark on which B1 is competitive. Signal from such a benchmark might be largely recoverable from pair and context identity alone, without chemistry.

Two additional baselines are a gradient boosting on PCA-reduced embeddings (B2) and a symmetric MLP over drug embeddings and cell features with no ontology, graph or target projection (B3). Both consume the drug embeddings and the cell context specified in Section 3.4, encoded symmetrically as the sum and the absolute difference of the two drug vectors concatenated with the cell vector, so that drug order cannot affect a prediction. B2 reduces the drug embeddings with a 64-component PCA fitted on the stacked training embeddings of both drugs and fits a histogram gradient-boosting regressor with 300 boosting iterations and otherwise default settings. B3 applies the same encoding to the unreduced embeddings and passes it through a 512-unit layer with layer normalization, GELU activation and dropout 0.1, a 256-unit GELU layer and a scalar head, trained for 30 epochs with AdamW at learning rate 0.001, weight decay 0.01 and batch size 512, minimizing mean squared error on winsorized and standardized Loewe scores with no early stopping. B2 is a learned but non-neural comparator that sits between B1 and the neural models, while B3 acts as the reference for "does structure help".

3.4. *Features and Model Configuration*

Cell context is a 160-dimension entity, comprising of 14 PROGENy activities,[34] 64 principal components of protein-coding expression, 64 of CRISPR gene effect over the ontology's gene universe, and 18 genotype flags, all from the DepMap portal.[35] Drug embeddings are generated using MolFormer-XL.[36] The drug-target profiles come from ChEMBL[37] over a 72-gene panel, where only 24.7% of trainable rows have both drugs annotated. Hence, a target-imputation head predicts profiles from structure.

Pathway vocabulary, gene membership, graph topology and mechanism taxonomy are a versioned specification with no assumption about relevant biology is embedded in the architecture. Here, we use MSigDB Hallmark (50 gene sets)[38] and aggregate a pathway-pathway graph from STRING v12 physical interactions at combined score at least 700[39] with no hand-specified edges. The panel genes are absent from every Hallmark set acting as the druggable kinome, while Hallmark sets are expression-response signatures and are placed by STRING guilt-by-association. Finally, the drug target profiles project onto pathways, compose across the pair over the graph, and read out through a synergy head.

3.5. *Training and Model Selection*

Every split is persisted as disjoint training, validation and test bins. The mass-packed protocols divide the row mass in proportions of 0.8, 0.1 and 0.1, while leave-tissue-out and leave-study-out validate on a second tissue and on FRIEDMAN, and the tissue cross-validation serves evaluation only. Because the protocol is the object of comparison, each model keeps one configuration, fixed in advance and shared across protocols, folds and seeds, and no hyper-

parameter search is performed. The pathway model is a two-layer, two-head GAT with node dimension 16, hidden width 256, encoder width 512, dropout 0.1 and max projection, auxiliary target, node and sparsity losses weighted 0.1, 0.05 and 0.01, and an imputation head frozen for the first ten epochs. It is trained with AdamW at learning rate 3×10^{-4}, weight decay 0.01, batch size 512 and gradient clipping at 1.0. B1 has nothing to select, and the learned models select only through stopping rules that never see the test bin. The pathway model keeps its best validation-Spearman checkpoint within 60 epochs (patience 8), B2 applies the scikit-learn default early stopping to a tenth of the training rows, and B3 trains for a fixed 30 epochs, which can only favor the pathway model in Table 5. All preprocessing is fitted on training rows, each test bin is scored once, and the untuned scores are therefore conservative.

4. Results

4.1. *Combination overlap survives cold-start protocols*

As shown in Table 1, leave-study-out retains only 9.8% of test combinations, whereas leave-cell-line-out and leave-tissue-out retain 72.8% and 85.4%, respectively. Replicating this audit within individual contributing screens (Table 2) reveals that pooling the corpus actually dilutes the true overlap. For fully crossed screens, such as ALMANAC, FRIEDMAN, and O'Neil, holding out a cell line removes absolutely no combinations, resulting in 100% overlap, while the partially crossed AstraZeneca screen yields 92% retention. When evaluating leave-one-tissue-out across all eight eligible tissues (Table 3), the retained overlap varies drastically depending on the held-out tissue, ranging from 56.4% for skin to 100.0% for lung, ovary, and kidney. Across these tissue folds, the performance of the context-plus-pair null model correlates almost perfectly with the degree of combination overlap ($r = 0.977$). This association is not an artifact if fold size, as the retained overlap is uncorrelated with the number of test rows in a fold ($r = 0.081$ against log test rows, $p = 0.85$). In addition, the overlap - B1 association is undiminished when fold size is partialled out (partial $r = 0.984$, $p < 0.001$).

Retention under biological partitioning follows from the design of fully crossed screens, which assay every pair across every available cell line, so withholding a context masks the context-specific measurements without removing the pair from training.The correlation between baseline performance and retained overlap is consistent with simple models recovering the mean effect of a specific pair wherever retention is high. We report it as an association rather than a causal estimate: folds differ in tissue biology and in size as well as in overlap, and it is the analytic argument of Section 4.2, not this correlation, that isolates the contribution of pair retention.

4.2. *The overlap is large enough to invert model rankings*

Across the evaluated protocols (Table 4, Fig. 1), the performance of the null model broadly tracks the degree of combination overlap ($r = 0.97$). Under leave-cell-line-out (72.8% overlap), the null model outperforms both the symmetric multilayer perceptron and the pathway-informed model, by 0.055 and 0.121 Pearson respectively; paired over the six seeds, both margins are significant ($t = 7.71$, $p = 0.0006$ and $t = 8.81$, $p = 0.0003$). The same ordering holds under leave-tissue-out, where the null model exceeds the multilayer perceptron in all

Table 1. Fraction of test combinations overlap with training.

Protocol	Test Rows	Test Pairs Present in Train Set	Held Out Combinations
Random	18,076	94.0% ± 0.3%	no
Leave-tissue-out	4,372–71,638	85.4% ± 18.6%	no
Leave-cell-out	18,076	72.8% ± 0.1%	no
Leave-study-out	15,279	9.8% (fixed)	mostly
Leave-drug-out	31,728	0.0%	yes
Leave-combo-out	18,074	0.0%	yes

Table 2. Per-study Replication Statistics for leave-cell-out and leave-combo-out protocols.

Study	Rows	Pairs	Rows Per Pair	Random	Leave Cell Out	Leave Combo Out
ALMANAC	109,897	2,992	36.7	100% (0.677)	100% (0.653)	0% (0.057)
FRIEDMAN	50,107	2,926	17.1	100% (0.380)	100% (0.400)	0% (0.100)
ONEIL	9,369	315	29.7	100% (0.682)	100% (0.600)	0% (0.135)
ASTRAZENECA	5,910	407	14.5	100% (0.421)	92% (0.298)	0% (0.323)

Table 3. Statistics for leave-one-tissue-out on 8-fold cross-validation.

Held Out Tissue	Test Rows	Overlap	B1	B3	Pathway Model
Lung	25,612	100.0%	0.668	0.574	0.795
Ovary	16,912	100.0%	0.677	0.578	0.631
Kidney	11,845	100.0%	0.628	0.582	0.829
Large intestine	16,736	99.9%	0.642	0.544	0.561
Breast	23,644	93.1%	0.609	0.489	0.508
Brain	4,372	67.9%	0.488	0.471	0.795
Haemato./Lymphoid	4,693	66.3%	0.418	0.349	0.313
Skin	71,638	56.4%	0.325	0.210	0.192
Mean over folds		85.4%	0.557	0.475	0.578
Weighted by test rows		79.6%	0.505	0.403	0.461

eight cross-validation folds of Table 3 by a mean of 0.082 ($t = 6.52$, $p = 0.0003$); against the pathway model, however, the fold-to-fold spread is too wide to resolve a difference (-0.021, $p = 0.73$), which is the same limitation recorded for that protocol in Table 5. An off-the-shelf gradient-boosting regressor exceeds both neural baselines under leave-cell-line-out, by 0.005 and 0.071. The inversion is specific to this regime rather than to overlap in general: under random partitioning, where overlap is highest at 94.0%, the null model does not invert the ranking but instead recovers 79% of the multilayer perceptron's correlation (0.588 against 0.745) without using any molecular information. Leave-cell-line-out is the more consequential case precisely because it is the protocol adopted to control for the permissiveness of random splitting, and it is there that a model with no chemistry wins outright. Conversely, under protocols that strictly enforce combination novelty, such as leave-study-out (9.8% overlap) and leave-combo-out (0.0% overlap), the performance of all models decreases precipitously, and the null model performs poorly.

Table 4. Summary of Pearson Correlation on Loewe Reference Model.

Protocol	Overlap	B1	B2	B3	Pathway Model
Random	94.0%	0.588	0.630	0.745	0.744
Leave-tissue-out	85.4%	0.557	0.499	0.475	0.578
Leave-cell-out	72.8%	0.425	0.375	0.370	0.304
Leave-study-out	9.8%	0.111	0.146	0.139	0.123
Leave-drug-out	0.0%	0.125	0.216	0.245	0.224
Leave-combo-out	0.0%	0.200	0.378	0.594	0.618

This inversion of model rankings is consistent with high combination overlap permitting baseline models to exploit pair-specific memorization rather than learning transferable representations. Complex neural models, constrained by their architecture and regularization, may fail to memorize these pair-specific effects as efficiently as simpler models or explicit baselines, leading to lower apparent performance on the overlapping rows. It is a mismatch rather than a leakage when the resulting number is read as evidence of generalization to unseen combinations. When combination novelty is enforced instead, the memorization advantage vanishes, and all models are forced to rely on genuine generalization, resulting in uniformly lower performance that reflects the actual complexity of predicting unseen drug combinations.

Under leave-cell-line-out every test cell line is unseen, so the per-cell-line term of B1 backs off to the global mean on every test row (Section 3.3) and B1 reduces to the global mean plus the per-pair deviation. On test rows whose pair is absent from training, the pair term backs off as well, leaving an identical constant prediction on every such row and no predicted variance to correlate against. By construction, the leave-cell-line-out correlation of 0.425 is therefore carried entirely by the 72.8% of test pairs retained in training. This is an isolation of pair overlap at a fixed split, as it holds the protocol, the training data and the novelty of the held-out cell line constant, and varies only whether a test row's pair was seen. It applies to B1 analytically and not to the learned models, for which the corresponding control, holding the test rows fixed and deleting from training every observation that involves a test pair, requires retraining on a purpose-built partition; we did not run it and record it in Section 6 as the experiment that would complete this argument. Across protocols, the comparison is correspondingly associative, since the held-out entity, the training and test sizes and the intrinsic difficulty of the task all move together with overlap.

4.3. *The protocol determines the verdict*

When comparing the pathway-informed model to a standard neural baseline across different protocols (Table 5), the inclusion of biological structure yields a statistically significant performance improvement of 0.025 only when combinations are strictly held out ($p < 0.0001$). Under random partitioning and strict drug hold-out, the structural prior provides no significant benefit, while under cell-line hold-out, the inclusion of the pathway structure actually results in a statistically significant performance degradation of 0.067 ($p = 0.0018$). Leave-tissue-out carries the largest nominal effect of any protocol, +0.103, yet it is the one protocol under

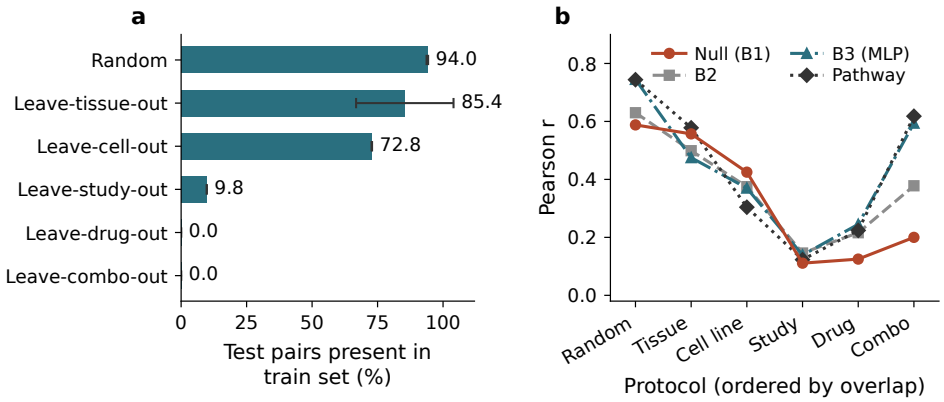


Fig. 1. (a) Combination overlap retained by each protocol. (b) Pearson correlation of each method under the same protocols, ordered by overlap. The null model B1 tracks overlap across the six protocols ($r = 0.97$) and across the eight leave-one-tissue-out folds of Table 3, in which only the held-out block varies ($r = 0.977$); protocols differ in more than overlap, so these are associations.

Table 5. Statistics for Pathway model minus MLP.

Protocol	Pearson Delta	T	P	Conclusion Supported
random	-0.001 ± 0.005	-0.58	0.59	structure is irrelevant
leave-combo-out	+0.025 ± 0.002	29.7	< 0.0001	structure helps
leave-drug-out	-0.021 ± 0.036	-1.44	0.21	structure is irrelevant
leave-cell-out	-0.067 ± 0.027	-6.04	0.0018	structure hurts
leave-tissue-out	+0.103 ± 0.139	2.11	0.073	not resolvable

which no verdict can be reached: the fold-to-fold spread is 0.139, wider than the effect itself, giving $p = 0.073$ over the eight folds. The obstacle is heterogeneity between tissues rather than a small effect, and it is consistent with the fold-level variation documented in Table 3, where retained overlap ranges from 56.4% to 100.0%. A protocol that partitions on a coarse categorical block therefore cannot adjudicate architecture at this sample size, which is itself a reason to prefer protocols whose held-out unit matches the entity under test. The magnitude of these architecture-driven effects is remarkably small, particularly when contrasted with the massive performance fluctuations induced simply by changing the evaluation protocol, where the pathway model's performance ranges from 0.123 to 0.744.

One reading is that where overlap is high the memorized pair effect is available and a structural prior is largely redundant, whereas enforcing combination novelty removes that shortcut and leaves the prior something to contribute. Whatever the mechanism, a single-protocol evaluation cannot separate a genuine benefit from biological structure from an interaction with the pair coverage that a given split leaves in place.

4.4. *What the pathway model actually relies on*

Component-wise ablation locates this model's performance in the drug-target branch rather than in the topology of the pathway graph. What the ablation does not establish is why that

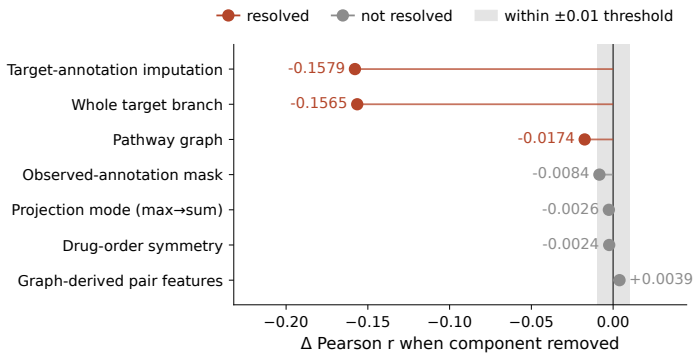


Fig. 2. Ablation deltas on leave-combo-out. Only the pathway graph and the target-annotation imputation exceed the 0.01 resolution threshold, and they differ by an order of magnitude. Removing the imputation head and removing the whole target branch are indistinguishable, which is why the branch's contribution cannot be attributed to annotation completion alone.

branch matters.

Ablation studies conducted under strict combination hold-out (Fig. 2, full model $r = 0.6246$) demonstrate that removing the pathway graph structure results in a negligible decrease in predictive performance of 0.017. In stark contrast, removing the imputation of missing drug-target annotations, or removing the entire target-processing branch, causes a massive drop in performance of approximately 0.158. Furthermore, stratifying model performance by the annotation status of the drug pairs reveals that the model performs better on pairs where neither drug has measured target annotations (0.682) compared to pairs where both drugs are fully annotated (0.582).

Database completion is one reading, under which imputation densifies a sparse annotation matrix and hands the predictor a less noisy profile than the measured annotations alone. It is not the only one, and the ablation cannot select among them: removing the imputation head (-0.1579) and removing the entire target branch (-0.1565) cost the same, so the intervention deletes the branch's learned chemical representation, auxiliary supervision, parameters and optimization path together with the imputed annotations. A capacity-matched control, a head of equal size trained without target supervision, would apportion the drop, and we did not run it (Section 6). The stratified comparison is likewise suggestive rather than decisive, as the advantage on unannotated pairs (0.682 against 0.582) is confounded with the study bias of public databases. What the ablation does support is the contrast that motivates it: the pathway graph is worth -0.0174 here, an order of magnitude less than the target branch, so a reported benefit from biological structure should be resolved into topology and annotation handling before it is attributed to either. Resting on a single architecture, this does not extend to structure-informed models in general.

5. Discussion

Cold-cell-line prediction is pivotal in predicting how a known combination behaves in an uncharacterized cell line. Although such a finding is legitimate and clinically relevant, it cannot be relied upon as a test of combination generalization. A model might perform well on the

former while learning essentially nothing about the underlying chemistry. Hence, a single performance metric cannot adjudicate both capabilities. These are two tasks, not one task and one flaw, and neither is simply the harder version of the other. The symmetric neural baseline scores higher under leave-combo-out than under leave-cell-line-out (0.594 against 0.370), while the null model shows the opposite ordering (0.200 against 0.425), so the protocols differ in the kind of information they leave available rather than along a single axis of difficulty. The corrective we propose is therefore reporting rather than replacement: pair a protocol with the claim it can carry, and publish the combination overlap that tells a reader which claim that is.

Combinatorial screens are dense in biological contexts but sparse in combination space, hence, withholding cell lines barely perturbs the coverage of drug pairs. This gap remains invisible in aggregate model performance and becomes apparent only when pair-level overlap is considered explicitly.

Attributing performance gains to biological structure likewise warrants caution. In the model examined here the advantage sits in the drug-target branch rather than in pathway topology, and studies reporting benefits from biological priors would strengthen the claim by separating knowledge injection from data completion with capacity-matched ablations.

Relative to the shortcut-learning, what the previous work doe not measure is the residue that produces the effect: how many test combinations each cold-start protocol actually leaves in training. Our work quantifies that residue protocol by protocol and screen by screen, then attribute it to full crossing of the pair-by-cell-line matrix, and use it to show that the protocol determines the verdict on structural priors and that the apparent benefit of those priors is traceable to annotation completion.

6. Limitations

DrugComb is the only corpus used in this study. The overlap audit is executed within four contributing screens. Among those, three are fully crossed while one is only partially crossed. Hence, the study addresses the effect of pooling across sources but not that of harmonization across assays.

The ablation study is anchored to a single pathway-informed model. The study does not claim that all such models will behave identically. Extending it would require the same ablations on other structure-injection mechanisms, such as subsystem models, network-proximity scores and graph transformers. Rather, we observe that the attribution of gains to biological structure remains unmeasured in prior work reporting pathway benefits.

We observe that two of the non-neural baselines surpass both neural architectures under leave-cell-line-out. This observation alone is unable to establish that no neural model could recover the per-pair mean from learned embeddings. A hybrid design, comprised of B3 with the per-cell-line and per-pair terms learned by B1 as explicit features would disentangle whether the protocol leaves too much exploitable signal.

In this study, we quantify overlap and demonstrate its consequences using our own baselines. Confirming that specific published results change under an overlap-controlled protocol requires re-training those methods, and we regard this as the most important extension.

Two controls are absent, one per central claim. For pair retention, holding the test rows fixed and deleting from training every observation involving a test pair would vary retention alone, with the held-out cell lines, the test distribution and the model constant; we offer instead the analytic reduction of B1 (Section 4.2) and the size-adjusted association of Section 4.1. For the ablation, a capacity-matched head trained without target supervision would separate annotation completion from the representation and capacity the target branch also contributes. Until both are run, the corresponding statements here are associations and component locations rather than causal attributions.

7. Conclusion

In summary, this study showcased that standard cold-start evaluation protocols for drug-combination synergy prediction retain a substantial and screen-dependent fraction of test combinations in training, and that this residue dominates both the reported numbers and the apparent utility of biological priors. The retention is consistent with the task leave-cell-line-out defines and should not be inferred as leakage, as it becomes an inferential error only when such numbers are taken as evidence of generalization to unseen combinations. Leave-cell-line-out and leave-one-tissue-out evaluations often retain the majority of test combinations within the training set depending on the underlying screen design. Under leave-cell-line-out, the protocol adopted specifically to correct for permissive random splitting, a simple baseline predictor lacking molecular information outperforms both conventional neural architectures outright, and it recovers most of their correlation under random splitting as well. As the choice of evaluation protocol can invert the apparent utility of the structured biological information, we recommend that future studies routinely report combination overlap alongside a context-plus-pair null model. Both are computationally inexpensive and would make the combination coverage a split leaves in place visible at the point where performance is reported.

Acknowledgements

This work was partially supported by the Commonwealth Health Research Board (CHRB-2360623, PI: Das), the National Science Foundation (NSF-2316003, PI: Ghosh), the National Institutes of Health (1S10OD038214-01, PI: Ghosh; 1T32TR006000-01, PI: Cifu; 1R01AR086251-01A1, PI: Boyan), the VCU Breakthrough award (PI: Ghosh), and VCU CCTR Endowment funds (PI: Thangarajh) awarded to P.G.

Code Availability

Scripts to reproduce the results presented in this work can be accessed at https://github.com/debnathk/drugsyn-audit

Use of Large Language Models

A large language model was used as a coding and drafting assistant: for parts of the analysis and plotting code. All experiments, numerical results, and scientific claims were produced and verified by the authors, who take full responsibility for the content.

References

1. K. Preuer, R. P. I. Lewis, S. Hochreiter, A. Bender, K. C. Bulusu, and G. Klambauer. DeepSynergy: predicting anti-cancer drug synergy with Deep Learning. *Bioinformatics*, 34:1538–1546, 2018.
2. J. O'Neil, Y. Benita, I. Feldman, M. Chenard, B. Roberts, Y. Liu, et al. An Unbiased Oncology Compound Screen to Identify Novel Combination Strategies. *Mol. Cancer Ther.*, 15:1155–1162, 2016.
3. S. L. Holbeck, R. Camalier, J. A. Crowell, J. P. Govindharajulu, M. Hollingshead, L. W. Anderson, et al. The National Cancer Institute ALMANAC: A Comprehensive Screening Resource for the Detection of Anticancer Drug Pairs with Enhanced Therapeutic Activity. *Cancer Res.*, 77:3564–3576, 2017.
4. B. Zagidullin, J. Aldahdooh, S. Zheng, W. Wang, Y. Wang, J. Saad, et al. DrugComb: an integrative cancer drug combination data portal. *Nucleic Acids Res.*, 47:W43–W51, 2019.
5. S. Zheng, J. Aldahdooh, T. Shadbahr, Y. Wang, D. Aldahdooh, J. Bao, et al. DrugComb update: a more comprehensive drug sensitivity data repository and analysis portal. *Nucleic Acids Res.*, 49:W174–W184, 2021.
6. P. Dewulf, M. Stock, and B. De Baets. Cold-Start Problems in Data-Driven Prediction of Drug-Drug Interaction Effects. *Pharmaceuticals*, 14:429, 2021.
7. S. Monem et al. A multi-view feature representation for predicting drugs combination synergy based on ensemble and multi-task attention models. *J. Cheminform.*, 2024.
8. B. Boonyarit et al. SynProtX: a large-scale proteomics-based deep learning model for predicting synergistic anticancer drug combinations. *GigaScience*, 2025.
9. W. Alam et al. Unlocking the therapeutic potential of drug combinations through synergy prediction using graph transformer networks. *Comput. Biol. Med.*, 2024.
10. S. Kapoor and A. Narayanan. Leakage and the reproducibility crisis in machine-learning-based science. *Patterns*, 4:100804, 2023.
11. S. Kapoor and A. Narayanan. Leakage and the Reproducibility Crisis in ML-based Science. *arXiv* preprint, 2022.
12. J. Bernett et al. Guiding questions to avoid data leakage in biological machine learning applications. *Nat. Methods*, 2024.
13. M. Rosenblatt et al. Data leakage inflates prediction performance in connectome-based machine learning models. *Nat. Commun.*, 2024.
14. A. H. C. Vlot, N. Aniceto, M. P. Menden, G. Ulrich-Merzenich, and A. Bender. Applying drug synergy metrics to oncology combination screening data: agreements, disagreements and pitfalls. *Drug Discov. Today*, 2019.
15. D. Duarte and N. Vale. Evaluation of synergism in drug combinations and reference models for future orientations in oncology. *Curr. Res. Pharmacol. Drug Discov.*, 2022.
16. K. R. Roell, D. M. Reif, and A. A. Motsinger-Reif. An Introduction to Terminology and Methodology of Chemical Synergy — Perspectives from Across Disciplines. *Front. Pharmacol.*, 2017.
17. O. Fuentealba-Manosalva et al. Mind the Curve: Dose-Response Fitting Biases the Synergy Scores across Software Used for Chemotherapy Combination Studies. *Int. J. Mol. Sci.*, 2023.
18. F. Abbasi and J. Rousu. New methods for drug synergy prediction: a mini-review. *Curr. Opin. Struct. Biol.*, 2024.
19. H. Yousefabadi et al. Bioactivity-Driven Prediction of Antibacterial Synergy Using Machine Learning Models. *bioRxiv* preprint, 2025.
20. O. E. Agboola et al. End-to-end multimodal attention fusion for drug-target interaction prediction with cold-scaffold validation on the Davis kinase benchmark. *J. Mol. Graph. Model.*, 2026.

21. F. Cheng, I. A. Kovacs, and A. L. Barabasi. Network-based prediction of drug combinations. *Nat. Commun.*, 10:1197, 2019.
22. Q. T. Trac et al. Pathway activation model for personalized prediction of drug synergy. *eLife*, 2025.
23. B. M. Kuenzi et al. Predicting Drug Response and Synergy Using a Deep Learning Model of Human Cancer Cells. *Cancer Cell*, 38:672–684.e6, 2020.
24. H. I. Kuru, O. Tastan, and A. E. Cicek. MatchMaker: A Deep Learning Framework for Drug Synergy Prediction. *IEEE/ACM Trans. Comput. Biol. Bioinform.*, 19:2334–2344, 2022.
25. W. B. Sherine et al. Graph Enhanced Transformers for Predicting Drug using Drug Synergies and Antagonistic Interactions. In *ICESC*, 2025.
26. D. Chen et al. Predicting anticancer synergistic drug combinations based on multi-task learning. *BMC Bioinformatics*, 2023.
27. Z. Wang et al. DEML: Drug Synergy and Interaction Prediction Using Ensemble-Based Multi-Task Learning. *Molecules*, 2023.
28. A. J. Preto et al. SYNPRED: prediction of drug combination effects in cancer using different synergy metrics and ensemble learning. *GigaScience*, 2022.
29. J. Lin et al. Pisces: A multi-modal data augmentation approach for drug combination synergy prediction. *Cell Genomics*, 2023.
30. S. Kapoor et al. REFORMS: Consensus-based Recommendations for Machine-learning-based Science. *Sci. Adv.*, 2024.
31. J. Zhu et al. Machine Learning in Environmental Research: Common Pitfalls and Best Practices. *Environ. Sci. Technol.*, 2023.
32. H. Semmelrock et al. Reproducibility in Machine Learning-based Research: Overview, Barriers and Drivers. *AI Magazine*, 2024.
33. Emine Beyza Çandır, Halil İbrahim Kuru, Magnus Rattray, A Ercüment Çiçek, and Oznur Tastan. One-hot news: drug synergy models shortcut molecular features. *Bioinformatics*, 42(3):btag040, 2026.
34. M. Schubert, B. Klinger, M. Klunemann, A. Sieber, F. Uhlitz, S. Sauer, et al. Perturbation-response genes reveal signaling footprints in cancer gene expression. *Nat. Commun.*, 9:20, 2018.
35. A. Tsherniak, F. Vazquez, P. G. Montgomery, B. A. Weir, G. Kryukov, G. S. Cowley, et al. Defining a Cancer Dependency Map. *Cell*, 170:564–576.e16, 2017.
36. J. Ross, B. Belgodere, V. Chenthamarakshan, I. Padhi, Y. Mroueh, and P. Das. Large-scale chemical language representations capture molecular structure and properties. *Nat. Mach. Intell.*, 4:1256–1264, 2022.
37. B. Zdrazil, E. Felix, F. Hunter, E. J. Manners, J. Blackshaw, S. Corbett, et al. The ChEMBL Database in 2023: a drug discovery platform spanning multiple bioactivity data types and time periods. *Nucleic Acids Res.*, 52:D1180–D1192, 2024.
38. A. Liberzon, C. Birger, H. Thorvaldsdottir, M. Ghandi, J. P. Mesirov, and P. Tamayo. The Molecular Signatures Database (MSigDB) hallmark gene set collection. *Cell Syst.*, 1:417–425, 2015.
39. D. Szklarczyk, R. Kirsch, M. Koutrouli, K. Nastou, F. Mehryary, R. Hachilif, et al. The STRING database in 2023: protein-protein association networks and functional enrichment analyses for any sequenced genome of interest. *Nucleic Acids Res.*, 51:D638–D646, 2023.