GENESYN-HGNN: STRUCTURED CONTEXT-CONDITIONED INTERACTION LEARNING FOR DRUG SYNERGY PREDICTION

Anonymous authors
Paper under double-blind review

ABSTRACT

Learning the effect of an entity pair under a high-dimensional context requires structured entity representations and a scorer that captures pair-context dependence. Drug synergy prediction exemplifies this setting, as the same drug pair can exhibit different effects across cellular states. Directly modeling this dependence with an unrestricted interaction over two drug embeddings and raw gene expression, however, requires a large number of decoder coefficients. We introduce the Gene-aware Drug Synergy Hypergraph Neural Network (GeneSyn-HGNN), which integrates motif-guided multi-relational hypergraph encoding and a structured cell-conditioned decoder. Its central Gene-Synergy Kernel (GSK) models a parameter-efficient interaction between the gated drug pair and the encoded cell representation. Two complementary pathways support this interaction: Cell-Conditioned Gating (CCG) modulates both drug representations according to the encoded cell state, and Gene-Selective Attention (GSA) applies a gated-pair query to expression-scaled gene keys to produce a pair-dependent additive score. We analyze the decoder's expressiveness and coefficient complexity and derive a conditional decoder-level capacity bound under fixed, sample-independent upstream feature maps. Experiments on O'Neil and NCI-ALMANAC show that GeneSyn-HGNN achieves the highest listed AUC and AUPR values and maintains strong performance under reduced training fractions. These findings support structured cell conditioning as an effective and parameter-efficient approach to modeling drug-pair-cell interactions.

1 INTRODUCTION

Context-dependent interaction prediction concerns how the effect of an entity pair varies under a high-dimensional context. Drug synergy prediction exemplifies this setting: a combination can be synergistic in one cell line and inactive or antagonistic in another (Kalkan et al., 2025). This variation calls for structured representations of the two drugs and a predictor that relates their joint effect to cellular state. Such a predictor must also determine which cellular features are informative for the current drug pair. Computational models can narrow the vast combination space explored by costly experimental screens (Liu et al., 2022; Cheng et al., 2026), yet their predictions depend on how molecular structure and cellular state enter the scoring function.

Existing approaches differ in how they represent drugs and cell lines. DeepSynergy fuses chemical descriptors with cell-line expression for context-specific prediction (Preuer et al., 2018). SynergyX uses mutual attention to align drug and cell modalities, while JointSyn learns complementary views for personalized prediction (Guo et al., 2024; Li et al., 2024). Hypergraph models extend pairwise graphs by encoding higher-order associations among drugs, targets and cell lines. Recent hypergraph-based methods emphasize different structural aspects. MRHGNN focuses on multi-relational structure, HypergraphSynergy on multi-way interactions and HypertranSynergy on cross-granularity modeling, while MHGSynergy integrates multimodal structure (Chen et al., 2025; Liu et al., 2022; Wang et al., 2024; Zhang et al., 2025c).

Methods also differ in where cellular information enters the predictor. Feature-fusion and attention models integrate expression with drug features before producing the final score (Preuer et al.,

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.

2018; Guo et al., 2024). MRHGNN instead scores a pair through the cell-indexed diagonal relation $z_a^\top D(w_c)WD(w_c)z_b$, where w_c is a free parameter associated with an observed cell line (Chen et al., 2025). HypergraphSynergy and MHGSynergy feed learned drug and cell representations into a joint scorer (Liu et al., 2022; Zhang et al., 2025c), whereas HIG-Syn constructs fine-grained drug-cell interactions across multiple granularities (Gu et al., 2025).

The unresolved issue is how to expose the influence of cellular state within the final drug-pair score while preserving pair specificity and coefficient efficiency. Our distinction from MRHGNN is therefore not cell-specific diagonal conditioning itself, but an expression-conditioned decomposition of its roles: CCG generates a shared-mapping gate that modulates both drug representations, GSK adds a latent drug-drug-cell interaction whose coefficients are generated from the encoded cell state, and GSA performs pair-specific gene selection. A direct interaction implemented as a fully parameterized tensor over two d-dimensional drug embeddings and a G-dimensional expression vector would make this dependence explicit through $O(d^2G)$ decoder coefficients, a parameterization that is large relative to the available sample size. These considerations motivate a decoder that assigns the effects of cellular state to distinct components whose parameters are shared across drug pairs and cell lines.

Motivated by this gap, we introduce the Gene-aware Drug Synergy Hypergraph Neural Network (GeneSyn-HGNN), which integrates motif-guided multi-relational hypergraph encoding and a structured context-conditioned decoder. The encoder constructs complementary drug representations from six drug-target motif relations and labeled drug-cell interactions. At the decoder, Cell-Conditioned Gating (CCG) first modulates both drug representations using the expression-derived cell embedding. The gated pair then enters a base bilinear scorer and two parallel cell-conditioned components. The Gene-Synergy Kernel (GSK) models a diagonal drug-drug-cell interaction, while Gene-Selective Attention (GSA) constructs a query from the gated drug pair to attend to expression-scaled gene keys. With the upstream representations held fixed, GSK requires $O(dd_c)$ coefficients rather than the $O(d^2G)$ coefficients of an unrestricted raw-expression interaction. Our decoder-level analysis establishes an expressiveness result for CCG and characterizes the coefficient complexity and representability of GSK. It also examines pair selectivity in GSA. The covering bound is conditional on sample-independent feature maps under explicit assumptions. Under the reported transductive within-cell-line protocol, GeneSyn-HGNN achieves AUC/AUPR values of 0.97/0.97 on O'Neil and 0.93/0.94 on NCI-ALMANAC. Component ablations show that removing GSK produces the largest observed ranking degradation, and the full model reaches a test AUC of 0.90 using approximately 21% of the O'Neil training data compared with 48% without GSK. Appendix D gives the formal analysis, while Appendix C reports cell-line structure, cross-cell attribution, and a case study supported by a non-exhaustive retrospective literature search. Our key contributions are as follows:

- We integrate six motif-defined hyperedge relations with a decoder that exposes explicit cell-conditioned scoring mechanisms.
- We introduce GSK as the central structured latent drug-pair-cell interaction, reducing the coefficient count of the three-way decoder term. CCG and GSA provide complementary drug-level modulation and pair-dependent gene selection, respectively.
- We analyze how CCG, GSK and GSA alter the family of scoring functions expressible by the decoder and the number of coefficients required to represent these functions. The associated capacity upper bound is conditional on feature maps fixed independently of the sample used to fit the decoder and is not a bound for end-to-end training.
- We evaluate GeneSyn-HGNN on O'Neil and NCI-ALMANAC. The experiments compare predictive performance, examine the contributions of individual components, and test the robustness of the encoder. We further study reduced-data behavior and analyze cell-dependent model behavior.

2 Related Work

Cellular context in drug-pair scoring. Drug-synergy predictors differ in how cell state enters the final score. DeepSynergy early-fuses drug descriptors and expression, whereas SynergyX and JointSyn use attention or dual-view learning (Preuer et al., 2018; Guo et al., 2024; Li et al., 2024). HypergraphSynergy, MHGSynergy and HIG-Syn instead pass learned drug and cell representations

to joint scorers (Liu et al., 2022; Zhang et al., 2025c; Gu et al., 2025). MRHGNN makes the relation explicit as $s^{\mathrm{MR}}_{a,b,c} = z_a^\top D(w_c) W D(w_c) z_b$, with a free vector w_c for each observed cell line (Chen et al., 2025). Our decoder retains this relation in s^{base} but uses the decomposition $s^{\mathrm{base}} + (\widetilde{z}_a \odot \widetilde{z}_b)^\top W_{\mathrm{gs}}^\top x_c + s^{\mathrm{gsa}}$, separating expression-derived drug modulation, latent pair-cell interaction and pair-specific gene selection.

Conditional modulation. FiLM provides a general precedent for feature-wise conditioning through the affine transformation $\gamma(x) \odot h + \beta(x)$ (Perez et al., 2018). CCG uses a more restricted construction: the encoded cell state generates a sigmoid-constrained multiplicative gate that is shared by both drugs and applied before symmetric pair scoring. We therefore do not claim CCG as a new generic conditioning operator. Its role is to separate expression-derived drug modulation from the cell-indexed base relation, the latent pair-cell interaction modeled by GSK and the pair-specific gene selection performed by GSA.

Structured higher-order interaction. Full bilinear and tensor fusion exposes cross-modal interactions but can be coefficient intensive. MUTAN controls bilinear fusion with Tucker and low-rank decompositions, while TuckER uses a shared Tucker core for relational scoring (Ben-Younes et al., 2017; Balazevic et al., 2019). Within drug synergy, TensoGraph also combines graph representations with Tucker decomposition (Wang et al., 2026). Unlike general Tucker-style parameterizations, GSK imposes the drug-mode diagonal structure $T^{\mathrm{GSK}}_{rsm} = \delta_{rs}(W_{\mathrm{gs}})_{mr}$, requiring dd_c coefficients, while the base bilinear term retains cross-dimensional drug interactions.

Hypergraph encoding and evaluation scope. MRHGNN's multi-relational drug-protein hypergraph informs our encoder, while HypergraphSynergy, HypertranSynergy, HIG-Syn and MHGSynergy provide alternative high-order or multimodal encoders (Chen et al., 2025; Liu et al., 2022; Wang et al., 2024; Gu et al., 2025; Zhang et al., 2025c). These works motivate structured encoding, whereas our main departure is at the scorer. Evaluation settings also vary: cross-study comparisons depend on dataset scale, data quality and synergy metric (Cheng et al., 2026). Gene-level context models further motivate fine-grained conditioning (Huang et al., 2024; Zhang et al., 2026), but the present evaluation remains transductive.

3 METHOD

3.1 PROBLEM SETUP

Given a drug pair (a, b) and a cell line c, let $y_{a,b,c} \in \{0, 1\}$ denote the observed binary synergy label. We learn a predictor

$$\widehat{y}_{a,b,c} = f_\theta(X_a, X_b, X_c) \in [0, 1], \tag{1}$$

where X_a and X_b denote the molecular features of the two drugs and $X_c \in \mathbb{R}^G$ denotes the genome-wide expression profile of cell line c. Hereafter, cellular context refers to the expression-defined state of cell line c, represented by the raw expression profile X_c or its encoded feature x_c, depending on the decoder component. A sample is labeled positive when its synergy score reaches the dataset-specific threshold τ_{pos} (30 for O'Neil and 10 for NCI-ALMANAC) and negative when its score is below zero. Samples in the interval $[0, \tau_{\mathrm{pos}})$ are excluded. Drug combinations are treated as unordered, so the predictor is required to satisfy $f_\theta(X_a, X_b, X_c) = f_\theta(X_b, X_a, X_c)$.

Let d denote the drug-embedding dimension and let $G = 977$ denote the gene dimension. A naively and fully parameterized Drug $\times$ Drug $\times$ Gene interaction tensor would contain d^2G coefficients. With $d = 256$, this amounts to approximately 6.4×10^7 coefficients, which is large relative to the available training data.

3.2 OVERVIEW

GeneSyn-HGNN consists of a motif-guided multi-relational hypergraph encoder followed by a structured cell-conditioned decoder. Figure 1 traces the model through four stages. First, the BioEncoder maps protein, drug and cell-line features to x_p, x_d and x_c, respectively. These features then enter two complementary hypergraph branches. RHGNN1 propagates drug-protein information over the motif-defined hypergraph $\mathcal{G}_1$. RHGNN2 propagates training-set combination evidence over the cell-specific hypergraph $\mathcal{G}_2$. Their aligned drug outputs are averaged into the fused embeddings z_a and z_b. In addition to its indirect contribution through RHGNN2, cellular information

enters the decoder directly through mechanisms that bypass drug-view fusion. The encoded feature x_c conditions CCG and GSK, while the raw expression profile X_c supplies GSA. After CCG gates both fused drug embeddings, the base scorer, GSK and GSA produce parallel terms. Their sum is mapped through a sigmoid to the synergy probability. The resulting score is

$$s_{a,b,c} = s^{\text{base}}_{a,b,c} + s^{\text{gs}}_{a,b,c} + s^{\text{gsa}}_{a,b,c}, \qquad \hat{y}_{a,b,c} = \sigma(s_{a,b,c})\,. \tag{2}$$

This score decomposition allows the decoder to represent complementary effects of cellular information without introducing an unrestricted drug-drug-gene interaction tensor. All three terms are constructed to be invariant to exchanging drugs a and b.

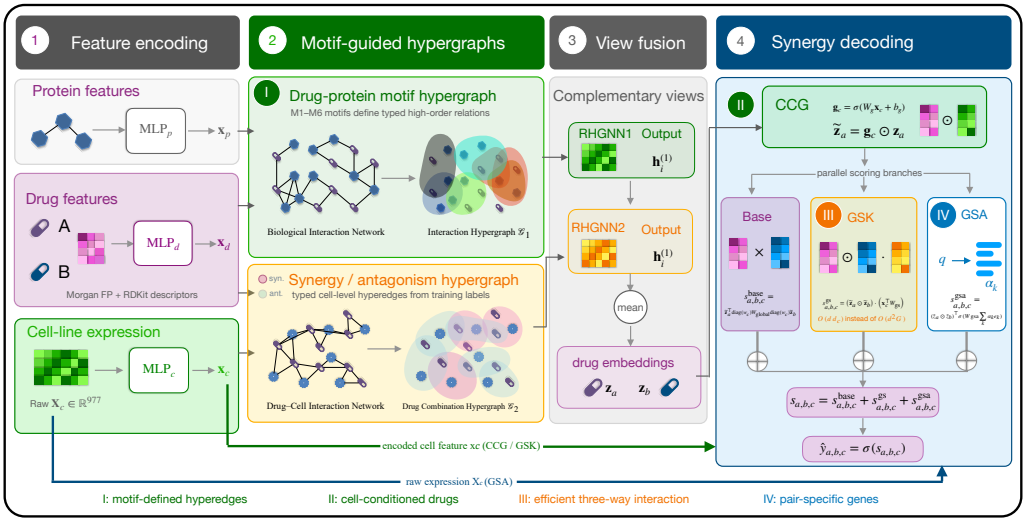


Figure 1: Overview of GeneSyn-HGNN. (1) BioEncoder maps protein features, molecular fingerprints and descriptors, and cell-line expression into modality-specific latent features x_p, x_d, and x_c. (2) Motifs M1-M6 define typed high-order relations in the drug-protein hypergraph $\mathcal{G}_1$ for RHGNN1. Training-set synergy and antagonism labels define the drug-combination hypergraph $\mathcal{G}_2$ for RHGNN2. (3) The complementary RHGNN outputs are aligned in the shared drug space and averaged to obtain the fused drug embeddings z_a and z_b. (4) The encoded cell feature x_c and raw expression X_c bypass view fusion and are routed to CCG/GSK and GSA, respectively. CCG uses x_c to gate both drug embeddings. Base, GSK and GSA then compute parallel scores. Base models a cell-scaled bilinear pair interaction. GSK adds a structured drug-drug-cell interaction using x_c, and GSA applies pair-specific gene attention to raw expression X_c. The three scores are summed and mapped through a sigmoid to the synergy probability. Labels I-IV identify motif construction, CCG, GSK and GSA, respectively.

3.3 Multi-relational Hypergraph Encoding

A hypergraph $\mathcal{G} = (V, E)$ represents high-order associations, with each hyperedge $e \in E$ connecting a subset of nodes. Building on the relational-hypergraph framework of MRHGNN (Chen et al., 2025), RHGNN1 operates on the structural hypergraph $\mathcal{G}_1 = (\mathcal{V}_1, \mathcal{E}_1, \phi_1)$. Its node set is $\mathcal{V}_1 = \mathcal{V}_D \cup \mathcal{V}_P$, where $\mathcal{V}_D$ and $\mathcal{V}_P$ contain drug and protein nodes. Each hyperedge in $\mathcal{E}_1$ is an instance of a drug-protein motif, and the relation mapping $\phi_1 : \mathcal{E}_1 \to \{\text{M1}, \ldots, \text{M6}\}$ assigns one of six motif types. These motifs extend the four M1-M4 motifs defined in the original MRHGNN paper. M1-M3 characterize target structures associated with a single drug. M1 represents two interacting targets. M2 represents three targets forming a protein-protein interaction triangle, and M3 represents two non-interacting targets. M4-M6 characterize drug-pair structures distinguished by DDI status and target relations. M4 represents two drugs with a known DDI, each connected to a distinct target, where the two targets have a recorded PPI. M5 represents two drugs without a recorded DDI whose distinct targets have a recorded PPI. M6 represents two drugs with a known

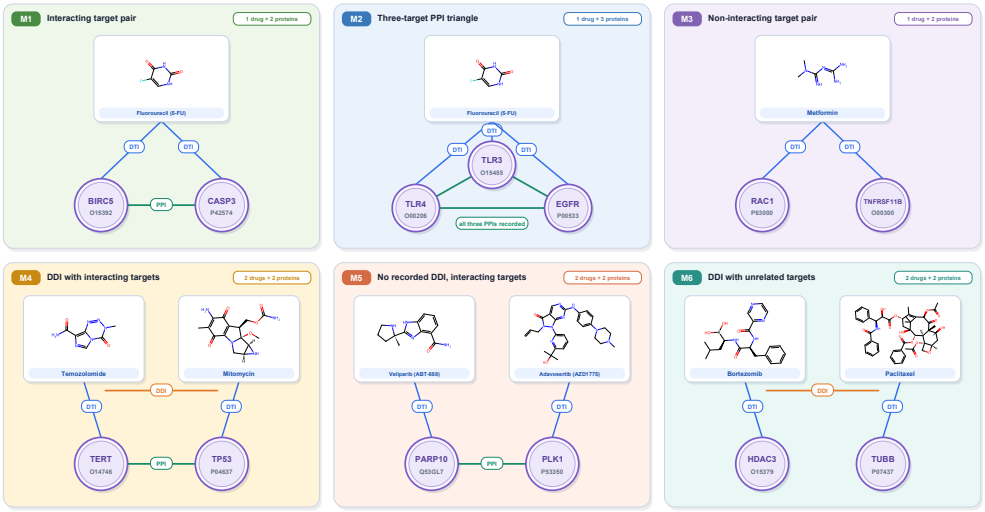


Figure 2: Examples of the six drug-protein motif types used to construct the structural hypergraph $\mathcal{G}_1$. Blue edges denote drug-target interactions, green edges denote protein-protein interactions and orange edges denote drug-drug interactions. The absence of an edge denotes that the corresponding interaction is not recorded in the input network.

DDI whose distinct targets have no recorded PPI. Figure 2 gives concrete instances of all six patterns from the O'Neil input interaction network.

RHGNN2 operates on the cell-specific interaction hypergraph $\mathcal{G}_2 = (\mathcal{V}_2, \mathcal{E}_2, \phi_2)$. Its node set is $\mathcal{V}_2 = \mathcal{V}_D \cup \mathcal{V}_C$, where $\mathcal{V}_C$ contains cell-line nodes. For every eligible labeled training triple (a, b, c), $\mathcal{E}_2$ contains the three-node hyperedge $e_{a,b,c} = \{a, b, c\}$. The relation mapping $\phi_2 : \mathcal{E}_2 \to \{\text{synergistic}, \text{antagonistic}\}$ assigns the synergistic type when $y_{a,b,c} = 1$ and the antagonistic type when $y_{a,b,c} = 0$. The model rebuilds $\mathcal{G}_2$ independently for every fold using training triples only. Validation and test labels therefore do not enter the encoder hypergraph. This construction allows RHGNN2 to propagate cell-specific combination evidence between drug and cell-line nodes.

Each hypergraph encoder performs attention-based bidirectional message passing. For edge type t, the node-to-hyperedge stage aggregates node features into a hyperedge representation using learned attention coefficients:

$$\begin{aligned} h_e^{(t)} &= \sum_{v \in e} \alpha_{e,v}^{(t)} \, \mathbf{W}_t x_v, \\ \alpha_{e,v}^{(t)} &= \frac{\exp\big(\mathrm{LeakyReLU}(a_t^\top \mathbf{W}_t x_v)\big)}{\sum_{v' \in e} \exp\big(\mathrm{LeakyReLU}(a_t^\top \mathbf{W}_t x_{v'})\big)}. \end{aligned} \tag{3}$$

The hyperedge-to-node stage then aggregates incident hyperedge representations. Here, $\beta_{v,e}$ denotes the normalized attention assigned to incident hyperedge e when updating node v. A residual connection, ℓ_2 normalization and a nonlinear activation produce the updated node representation:

$$x'_v = \mathrm{LeakyReLU}\Big(\mathrm{norm}\Big(\mathbf{W}_0 x_v + \sum_{e \ni v} \beta_{v,e} \, \mathbf{W} h_e\Big)\Big). \tag{4}$$

Let the BioEncoder produce drug, cell-line and protein embeddings x_d, x_c and x_p, respectively. RHGNN1 takes $\mathrm{cat}(x_d, x_p)$ as input, while RHGNN2 takes $\mathrm{cat}(x_d, x_c)$. The final representation of drug node i is the average of the two encoder outputs:

$$z_i = \frac{1}{2}\Big(\big[\mathrm{RHGNN1}(\mathrm{cat}(x_d, x_p))\big]_i + \big[\mathrm{RHGNN2}(\mathrm{cat}(x_d, x_c))\big]_i\Big). \tag{5}$$

3.4 CELL-CONDITIONED GATING

MRHGNN uses learned cell-specific diagonal scaling in its bilinear decoder (Chen et al., 2025). CCG adds a distinct, expression-derived dimension-wise gate to the drug representations. It computes this gate from the cell embedding $x_c \in \mathbb{R}^{d_c}$ using a linear transformation followed by a sigmoid:

$$g_c = \sigma(W_g\, x_c + b_g), \qquad \widetilde{z}_a = g_c \odot z_a, \quad \widetilde{z}_b = g_c \odot z_b. \tag{6}$$

Each entry of $g_c \in (0,1)^d$ controls the contribution of the corresponding drug-embedding dimension in cell line c. Although W_g and b_g are shared across cell lines, the gate depends on the expression-derived embedding x_c. Cell lines with similar expression profiles can therefore receive similar gates. The modulated representations enter a bilinear term with cell-specific diagonal scaling. To respect the unordered nature of drug combinations, we parameterize the shared interaction matrix as $W_{\text{global}} = (A + A^\top)/2$:

$$s^{\text{base}}_{a,b,c} = \widetilde{z}_a^\top \,\text{diag}(w_c)\, W_{\text{global}}\, \text{diag}(w_c)\, \widetilde{z}_b. \tag{7}$$

Here, $w_c \in \mathbb{R}^d$ denotes the cell-specific scaling vector and $W_{\text{global}} \in \mathbb{R}^{d\times d}$ is shared across cell lines. Unlike g_c, which is generated from x_c by parameters shared across cell lines, w_c is a learned cell-line-indexed vector inherited from the base bilinear decoder. The symmetry of W_{global} makes the base score invariant to exchanging a and b, while GSK and GSA are invariant by their elementwise-product and sum constructions. Since w_c is indexed by observed cell lines, the present decoder does not directly support an unseen cell line without estimating a new w_c. CCG and w_c also provide partially overlapping diagonal modulation, so their individual roles are interpreted through ablation rather than parameter identifiability. For the observed cell lines, CCG is therefore an expression-linked parameter-sharing bias, not a claim of strict expressiveness over the unrestricted cell-indexed scaling in the complete base decoder.

3.5 GENE-SYNERGY KERNEL

An unrestricted, fully parameterized interaction between two d-dimensional drug embeddings and the raw G-dimensional expression profile requires $O(d^2G)$ decoder coefficients. With the upstream representations held fixed, GSK instead uses the expression-derived latent cell embedding x_c in a structured diagonal three-way term with $O(dd_c)$ coefficients:

$$s^{\text{gs}}_{a,b,c} = (\widetilde{z}_a \odot \widetilde{z}_b) \cdot (x_c^\top W_{\text{gs}}), \tag{8}$$

where $W_{\text{gs}} \in \mathbb{R}^{d_c\times d}$ maps the cell embedding to a d-dimensional cell-conditioned interaction vector. Unlike MRHGNN's cell-indexed relation, GSK generates this separate additive interaction vector from x_c through coefficients shared across cell lines rather than retrieving it solely by cell identity. In our implementation, this matrix contains $512 \times 256 = 131{,}072$ coefficients. This count isolates the GSK decoder term and excludes the upstream BioEncoder and hypergraph encoders. Before adding GSA, the decoder score decomposes into two terms:

$$s^{\text{base}}_{a,b,c} + s^{\text{gs}}_{a,b,c} = \underbrace{\widetilde{z}_a^\top W_{\text{eff},c} \widetilde{z}_b}_{\text{cell-scaled pair interaction}} + \underbrace{(\widetilde{z}_a \odot \widetilde{z}_b) \cdot (x_c^\top W_{\text{gs}})}_{\text{latent cell interaction}}. \tag{9}$$

Here, $W_{\text{eff},c} = \text{diag}(w_c) W_{\text{global}} \text{diag}(w_c)$ depends on the cell line through its learned diagonal scaling. Both terms share one synergy-classification loss rather than separate objectives. The base term captures cell-scaled cross-dimensional pair interactions, whereas GSK captures expression-derived modulation along shared latent dimensions. This lower-dimensional functional split motivates the reduced-data experiment in Section 4.4. The conditional capacity analysis, however, does not predict the empirical learning-curve difference of the jointly trained model.

3.6 GENE-SELECTIVE ATTENTION

GSA uses the gated drug pair as a query to reweight genes dynamically. Each gene k has a learnable embedding $e_k \in \mathbb{R}^{d_a}$, where d_a denotes the query-key dimension and equals 64 in our implementation. Multiplying this embedding by the raw expression value $X_{c,k}$ forms a gene-specific key, while W_q projects the sum of the two gated drug representations into the same attention space. Cross-attention assigns a weight to each gene:

$$q = W_q(\widetilde{z}_a + \widetilde{z}_b) \in \mathbb{R}^{d_a}, \qquad \alpha_k = \mathrm{softmax}_k\Big(\frac{q^\top (e_k X_{c,k})}{\sqrt{d_a}}\Big), \tag{10}$$

$$s^{\mathrm{gsa}}_{a,b,c} = (\widetilde{z}_a \odot \widetilde{z}_b) \cdot \sigma\Big(W_{\mathrm{gsa}}\Big(\textstyle\sum_k \alpha_k e_k\Big)\Big). \tag{11}$$

The attention weight α_k is the model-assigned normalized weight for gene k under the current drug-pair query and cell-line expression profile. Softmax gives relatively lower weights to genes with lower query-key compatibility, so different drug pairs in the same cell line can attend to different gene subsets. These weights alone do not establish biological importance or causal gene effects.

The final logit follows Equation 2. Its three terms are added directly without an additional concatenation-based predictor.

4 EXPERIMENTS

Our evaluation addresses four questions. How well does the model perform under the reported within-cell-line protocol? Is GSK the principal contributor among the added decoder components? Does its advantage persist when fewer training records are retained? What cell-dependent behavior emerges in the learned representations and attributions? These questions concern model behavior under the observed benchmarks and evaluation splits. The first three analyses evaluate predictive performance, the central contribution of GSK, and reduced-data behavior. CCG and GSA are treated as complementary pathways rather than equally supported empirical contributions. As exploratory diagnostics, we examine cell-line structure, compare attributions across cell lines and present a case study supported by a non-exhaustive retrospective literature search in Appendix C.

4.1 EXPERIMENTAL SETUP

The raw O'Neil screen contains 22,737 drug-pair-cell-line measurements involving 38 drugs and 39 cell lines. Matching the available expression features retains 21,571 measurements from 37 cell lines before label-based exclusion. NCI-ALMANAC contains 299,548 raw measurements involving 103 drugs and 60 cell lines. Feature matching retains 239,286 measurements from 48 cell lines before label-based exclusion. These pre-exclusion counts do not represent the effective fold sizes. Recovering the eligible positive and negative counts after thresholding requires the released fold indices. The available run record does not specify an additional duplicate-record aggregation rule, so we describe the split unit as a retained measurement record rather than asserting that every record is a unique drug-pair-cell triple. In the implemented labeling pipeline, samples are assigned positive labels for scores of at least 30 on O'Neil and 10 on NCI-ALMANAC, negative labels for scores below zero, and excluded status for scores between zero and the corresponding positive threshold. Each drug is represented by a 300-bit Morgan fingerprint with radius 6 concatenated with RDKit descriptors. We remove features with zero variance or missing values and standardize the remaining drug-feature matrix. Each cell line is represented by a standardized 977-dimensional gene-expression profile. These unsupervised transformations are applied to the available entity-feature matrices before interaction splitting, consistent with the transductive within-cell-line evaluation setting.

All methods are evaluated using the same record-level, within-cell-line ten-fold cross-validation protocol. Within each cell line, positive and negative examples are shuffled separately. In each fold, one positive partition and an equally sized negative partition form the test set, the next partitions form the validation set, and the remaining eligible examples form the training set. The construction does not enforce drug-disjoint or drug-pair-disjoint folds: drugs, and potentially the same pair in

another cell line, can occur in both training and evaluation records. Together with entity-level preprocessing and the training-label hypergraph, this defines a transductive evaluation over observed drugs and cell lines. For GeneSyn-HGNN, Adam is used with learning rate in $\{10^{-2}, 10^{-3}, 10^{-4}\}$ and weight decay in $\{0, 10^{-2}, 10^{-4}, 10^{-6}\}$. Following the recorded implementation, the epoch with the highest validation ACC is retained. AUPR, computed with `average_precision_score`, is the primary cross-method metric. ACC and F1 are retained as recorded descriptive outputs but are not compared because the archived baseline outputs do not establish a common operating point. We compare against DeepSynergy (Preuer et al., 2018), MRHGNN (Chen et al., 2025), HypergraphSynergy (Liu et al., 2022), HypertranSynergy (Wang et al., 2024) and MHGSynergy (Zhang et al., 2025c). Appendix A reports architectural settings, metric procedures, tuning-scope limitations and the absence of a numeric cross-fold pair-overlap rate.

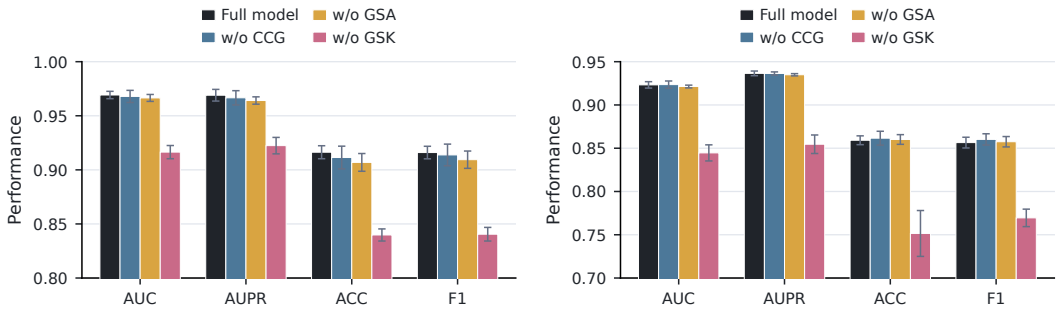
4.2 Predictive Performance under the Reported Protocol

Table 1 reports ten-fold results. GeneSyn-HGNN attains the highest listed AUC/AUPR, 0.97/0.97 on O'Neil and 0.93/0.94 on NCI-ALMANAC, improving AUPR over the strongest listed baseline by 0.02 and 0.06. Because evaluation folds are balanced, AUPR describes the controlled 1:1 distribution and not precision at the lower prevalence of an unfiltered screen.

Table 1: Ten-fold performance (mean $\pm$ standard deviation). Bold marks the best AUC/AUPR. ACC and F1 are retained as method-specific descriptive outputs and are not compared across methods.

	O'Neil Dataset				NCI-ALMANAC Dataset			
Methods	AUC	AUPR	ACC	F1	AUC	AUPR	ACC	F1
MRHGNN	0.94 ± 0.02	0.95 ± 0.01	0.87 ± 0.02	0.87 ± 0.02	0.88 ± 0.01	0.88 ± 0.02	0.80 ± 0.01	0.81 ± 0.01
MHGSynergy	0.93 ± 0.01	0.67 ± 0.01	$\mathbf{0.92 \pm 0.01}$	$\mathbf{0.92 \pm 0.01}$	0.79 ± 0.01	0.57 ± 0.01	0.78 ± 0.02	0.78 ± 0.02
HypergraphSynergy	0.93 ± 0.01	0.68 ± 0.01	$\mathbf{0.92 \pm 0.01}$	$\mathbf{0.92 \pm 0.01}$	0.80 ± 0.01	0.59 ± 0.01	0.80 ± 0.01	0.80 ± 0.01
DeepSynergy	0.85 ± 0.02	0.82 ± 0.02	0.75 ± 0.03	0.73 ± 0.03	0.83 ± 0.03	0.83 ± 0.04	0.75 ± 0.03	0.76 ± 0.02
HypertranSynergy	0.87 ± 0.01	0.49 ± 0.02	0.90 ± 0.01	0.90 ± 0.01	0.86 ± 0.01	0.88 ± 0.02	0.81 ± 0.01	0.80 ± 0.01
GeneSyn-HGNN	$\mathbf{0.97 \pm 0.01}$	$\mathbf{0.97 \pm 0.01}$	$\mathbf{0.92 \pm 0.01}$	$\mathbf{0.92 \pm 0.01}$	$\mathbf{0.93 \pm 0.01}$	$\mathbf{0.94 \pm 0.01}$	$\mathbf{0.86 \pm 0.01}$	$\mathbf{0.86 \pm 0.01}$

4.3 Contribution of the Structured Interaction



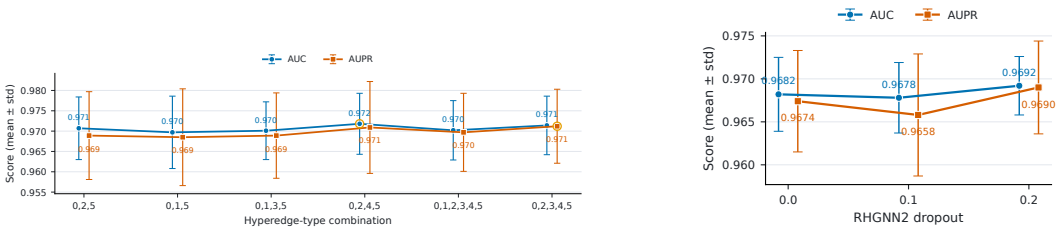
(a) O'Neil (b) NCI-ALMANAC

Figure 3: Component ablations (ten-fold mean $\pm$ standard deviation). Removing GSK produces the largest AUC/AUPR degradation.

Component ablations. Removing GSK reduces AUC from 0.9692 to 0.9164 on O'Neil and from 0.9232 to 0.8446 on NCI-ALMANAC, the largest observed degradation in Figure 3. CCG and GSA have smaller effects, with CCG inconclusive on NCI-ALMANAC, and are therefore treated as complementary pathways. Because each variant is retrained, the ablation shows component-level association rather than an isolated causal effect.

Hyperedge-type combinations. The six configurations in Figure 4(a) select subsets of the RHGNN1 drug-target hyperedge types. Mean AUC and AUPR vary within 0.0021 and 0.0027, respectively. No subset leads on both metrics, and performance does not increase monotonically with the number of included types. This suggests partial redundancy but does not identify any motif

as dispensable: types 0 and 5 occur in every configuration, and the sweep is neither factorial nor a no-motif control. Appendix B analyzes Figure 4(b).



(a) Hyperedge-type combinations (b) RHGNN2 dropout

Figure 4: Encoder robustness analyses on O'Neil. Error bars denote one standard deviation over ten folds.

4.4 Reduced-Data Learning Curves

Figure 5 examines how GSK affects performance as the amount of available training data is reduced. We construct seeded, nested subsets of the O'Neil training split while keeping the validation and test partitions fixed. The full model consistently outperforms the variant without GSK across all retained-data ratios and degrades more gradually as the training set becomes smaller. Its test AUC reaches 0.90 at an interpolated 21% of the training data, whereas the model without GSK requires approximately 48%. Notably, the full model trained with only 40% of the data already exceeds the test AUC obtained by the GSK-ablated model using the entire training set. These results suggest that GSK helps the model extract useful pair–context interaction structure from fewer labeled examples. Because the subsets are nested and the 0.90 crossings are obtained by interpolation on a fixed split, however, this experiment should be interpreted as evidence of improved reduced-data efficiency under this evaluation setting rather than as a general estimate of sample complexity.

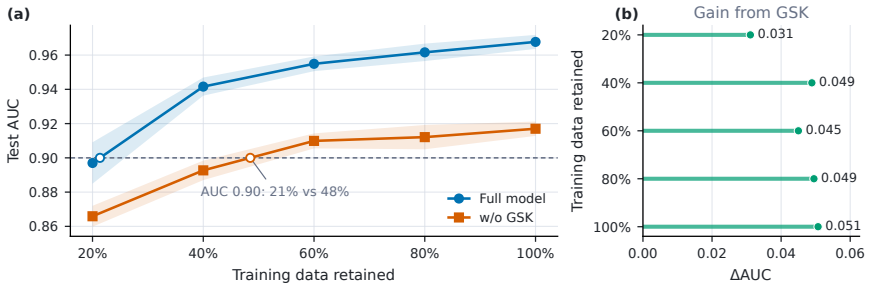


Figure 5: O'Neil reduced-data curves: test AUC and the difference associated with retaining GSK (ten-fold mean ± standard deviation).

5 Conclusion

GeneSyn-HGNN is built on motif-guided hypergraph representations and a structured cell-conditioned decoder that separates drug modulation, pair-context interaction and gene-level scoring. GSK provides the structured interaction term, while CCG and GSA introduce complementary representation-level and gene-level conditioning. Under the reported protocol, GeneSyn-HGNN achieves the strongest listed AUC and AUPR results on both benchmarks. Removing GSK produces the largest observed reduction in ranking performance, and the learning curves show that the full model retains an advantage over the variant without GSK under reduced training fractions on O'Neil. The decoder-level analysis provides a capacity-based interpretation of this structured interaction without establishing the observed empirical gain as a theoretical consequence. The attribution results serve as model diagnostics and require independent biological validation.

AI USE STATEMENT

AI assistance was used to aid and polish the writing and to support the development of mathematical proofs. All AI-assisted material was reviewed and validated by the authors, who remain fully responsible for the paper's content, claims, and conclusions.

REPRODUCIBILITY STATEMENT

The main text specifies the data representation, labeling rule, split construction, model-selection procedure, and evaluation metrics. Appendix A reports model parameters and training details. Sections 4.3 and 4.4 specify the ablation, robustness and learning-curve protocols next to their corresponding figures. Appendix C describes the case-study and retrospective-search procedures and reports the complete results. Appendix D states the assumptions and provides complete proofs for the theoretical claims. The anonymized implementation and supporting materials are available at `https://anonymous.4open.science/status/GeneSyn-HGNN-2F45`.

REFERENCES

Obada T. Alhalabi, Mona Göttmann, Maxwell P. Gold, Silja Schlue, Thomas Hielscher, Murat Iskar, Tobias Kessler, Ling Hai, Tolga Lokumcu, Clara C. Cousins, et al. Integration of transcriptomics, proteomics and loss-of-function screening reveals WEE1 as a target for combination with dasatinib against proneural glioblastoma. *Cancer Letters*, 605:217265, 2024. doi: 10.1016/j.canlet.2024.217265.

Ivana Balazevic, Carl Allen, and Timothy M. Hospedales. TuckER: Tensor factorization for knowledge graph completion. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing*, pp. 5185–5194, 2019. doi: 10.18653/v1/D19-1522.

Hedi Ben-Younes, Remi Cadene, Matthieu Cord, and Nicolas Thome. MUTAN: Multimodal tucker fusion for visual question answering. In *Proceedings of the IEEE International Conference on Computer Vision*, pp. 2612–2620, 2017.

M. Chen et al. MRHGNN: Enhanced multimodal relational hypergraph neural network for synergistic drug combination forecasting. *IEEE Transactions on Neural Networks and Learning Systems*, 2025.

Yingjuan Cheng, Qing Ye, Linlong Jiang, Yu Kang, Chang-Yu Hsieh, and Tingjun Hou. Unified heterogeneity-aware benchmark of drug synergy prediction: A cross-study analysis of traditional machine learning and graph deep learning models. *Journal of Cheminformatics*, 2026. doi: 10.1186/s13321-026-01242-1.

Paraskevi Giannakakou, L. Villalba, H. Li, M. Poruchynsky, and T. Fojo. Combinations of paclitaxel and vinblastine and their effects on tubulin polymerization and cellular cytotoxicity: Characterization of a synergistic schedule. *International Journal of Cancer*, 75(1):57–63, 1998.

Y. Gu, J. Zu, Y. Sun, and L. Zhang. HIG-Syn: A hypergraph and interaction-aware multigranularity network for predicting synergistic drug combinations. *Bioinformatics*, 41:i86–i95, 2025.

Amy D. Guertin, Melissa M. Martin, Brian Roberts, Melissa Hurd, Xiaoli Qu, Nathan R. Miselis, Ying Liu, Jing Li, Ilana Feldman, Yair Benita, Amy Bloecher, Carlo Toniatti, and Stuart D. Shumway. Unique functions of CHK1 and WEE1 underlie synergistic anti-tumor activity upon pharmacologic inhibition. *Cancer Cell International*, 12:45, 2012. doi: 10.1186/1475-2867-12-45.

Y. Guo et al. SynergyX: A multi-modality mutual attention network for interpretable drug synergy prediction. *Briefings in Bioinformatics*, 25(2):bbae015, 2024.

T. Huang, Z. Song, R. Ying, and W. Jin. Protein-nucleic acid complex modeling with frame averaging transformer. In *Advances in Neural Information Processing Systems*, 2024.

Fatma Neslihan Kalkan, Muhammed Sadik Yildiz, N. Ezgi Wood, Michael Farid, Melissa McCoy, Milo M. Lin, Chengcheng Zhang, Bruce A. Posner, Stephen S. Chung, and Erdal Toprak. Synergistic and antagonistic drug interactions are prevalent but not conserved across acute myeloid leukemia cell lines. *Scientific Reports*, 15:19431, 2025. doi: 10.1038/s41598-025-03242-x.

X. Li et al. Dual-view jointly learning improves personalized drug synergy prediction. *Bioinformatics*, 40(10):btae604, 2024.

X. Liu, C. Song, S. Liu, M. Li, X. Zhou, and W. Zhang. Multi-way relation-enhanced hypergraph representation learning for anti-cancer drug synergy prediction. *Bioinformatics*, 38(20):4782–4789, 2022.

Ethan Perez, Florian Strub, Harm de Vries, Vincent Dumoulin, and Aaron Courville. FiLM: Visual reasoning with a general conditioning layer. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 32, 2018. doi: 10.1609/aaai.v32i1.11671.

K. Preuer, R. P. I. Lewis, S. van Hoorn, R. P. Smith, T. Ezaki, and W. Huber. DeepSynergy: Predicting anti-cancer drug synergy with deep learning. *Bioinformatics*, 34(9):1538–1546, 2018.

Cristiana Sessa, Daniela Tosi, Laura Viganò, Joan Albanell, Daniel Hess, Massimo Maur, Stefania Cresta, Andrea Locatelli, Roger Angst, Federico Rojo, et al. Phase Ib study of weekly mammalian target of rapamycin inhibitor ridaforolimus (AP23573; MK-8669) with weekly paclitaxel. *Annals of Oncology*, 21(6):1315–1322, 2010. doi: 10.1093/annonc/mdp504.

Ni Shi, Hao Yu, and Tong Chen. Inhibition of esophageal cancer growth through the suppression of the PI3K/AKT/mTOR signaling pathway. *OncoTargets and Therapy*, 12:7637–7647, 2019. doi: 10.2147/OTT.S205457.

Hussein Tawbi, John M. Kirkwood, Ying-Chih Lin, Shama C. Buch, Yan Lin, Lyubov Kublo, Jan H. Beumer, Merrill J. Egorin, Edward Chu, and Pamela A. Hershberger. Abstract 5499: Temozolomide cytotoxicity in human melanoma cell lines is enhanced by concomitant treatment with vorinostat and the PARP inhibitor MK-4827. *Cancer Research*, 71(8 Supplement):5499, 2011. doi: 10.1158/1538-7445.AM2011-5499. Conference abstract.

Steven Paul Tinsley. *To Establish the Role of Mutations in c-KIT Tyrosine Kinase in the Pathogenesis and Therapy of Core-Binding Factor-Related Acute Myeloid Leukaemia*. PhD thesis, University College London, 2014. URL `https://discovery.ucl.ac.uk/id/eprint/1433249/`.

W. Wang et al. A granularity-level information fusion strategy on hypergraph transformer for predicting synergistic effects of anticancer drugs. *Briefings in Bioinformatics*, 25(1):bbad522, 2024.

X. Wang, D. Zhang, and L. Li. TensoGraph: A tensor-transformer framework for global-local drug synergy prediction on heterogeneous graphs. *Briefings in Bioinformatics*, 27(2):bbag157, 2026.

Wei Wen, Ernest S. Han, Thanh H. Dellinger, Leander X. Lu, Jun Wu, Richard Jove, and John H. Yim. Synergistic anti-tumor activity by targeting multiple signaling pathways in ovarian cancer. *Cancers*, 12(9):2586, 2020. doi: 10.3390/cancers12092586.

Yang Xiao, Judi Ramiscal, Kaska Kowanetz, Christopher Del Nagro, Shiva Malek, Marie Evangelista, Elizabeth Blackwood, Peter K. Jackson, and Thomas O'Brien. Identification of preferred chemotherapeutics for combining with a CHK1 inhibitor. *Molecular Cancer Therapeutics*, 12 (11):2285–2295, 2013. doi: 10.1158/1535-7163.MCT-13-0404.

Huaiwu Zhang, Xinliang Sun, Jianxin Wang, Min Li, and Jing Tang. A causal inference framework for identifying essential genes to enhance drug synergy prediction. *Bioinformatics*, 42(5):btag010, 2026. doi: 10.1093/bioinformatics/btag010.

Z. Zhang, T. Luo, X.-G. Chen, X. Yang, and J. Gong. Multimodal hypergraph representation learning for drug synergy prediction. *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, 23(1):518–527, 2025c.

APPENDIX

A EVALUATION PROTOCOL AND MODEL PARAMETERS

The shared evaluation uses positive thresholds of 30 for O'Neil and 10 for NCI-ALMANAC. Scores below zero are negative, while samples in the interval between zero and the positive threshold are excluded. Positive records are split by cell line into training, validation, and test sets in an $8{:}1{:}1$ ratio over ten folds. Validation and test negatives match the number of positives, and every remaining eligible negative is assigned to training. This record-level construction does not impose drug- or pair-disjoint folds and therefore evaluates transductive prediction over observed drugs and cell lines. The same fold construction and label definitions are used for all reported methods. AUPR is the primary evaluation metric and is computed from continuous prediction scores using `average_precision_score` for every method. Results are summarized as mean $\pm$ standard deviation over ten folds. For GeneSyn-HGNN, training records the epoch with the highest validation ACC, and the reporting code computes an F1-maximizing classification threshold separately for each evaluated split. Its resulting test ACC and F1 are descriptive rather than validation-threshold estimates. The archived baseline ACC and F1 values are retained as recorded, but a common operating-point procedure is not established. Accordingly, these metrics are therefore not used for cross-method claims. GeneSyn-HGNN uses PyTorch and Adam. The learning-rate candidates are $\{10^{-2}, 10^{-3}, 10^{-4}\}$ and the weight-decay candidates are $\{0, 10^{-2}, 10^{-4}, 10^{-6}\}$.

GeneSyn-HGNN. GeneSyn-HGNN uses seed 5 and is trained for at most 1,500 epochs per fold. It minimizes cell-wise binary cross-entropy with logits. The positive-class weight is $\sqrt{N_-/N_+}$. Gradient norms are clipped to 1.0. Adam learning rate and weight decay are configurable over the shared candidate sets, and the learning rate follows cosine annealing with $T_{\max} = 1500$ and a minimum value of 10^{-6}.

The BioEncoder maps the drug input to 512 dimensions with one linear ReLU layer, maps the 977-dimensional cell-line profile through a $977 \to 488 \to 512$ ReLU MLP, and represents each protein with a learned 512-dimensional embedding. $\mathrm{RHGNN1}$ uses dimensions $512 \to 256 \to 256$, six relation types, and dropout 0.2. $\mathrm{RHGNN2}$ uses $512 \to 128 \to 256$, two relation types, and dropout 0. The two 256-dimensional drug representations are averaged. The decoder parameterizes its 256×256 global bilinear matrix as $W_{\mathrm{global}} = (A + A^\top)/2$ and learns a 256-dimensional local vector for every cell line. CCG projects the 512-dimensional cell embedding to a 256-dimensional sigmoid gate. GSK uses $W_{\mathrm{gs}} \in \mathbb{R}^{512\times256}$. GSA uses 64-dimensional gene embeddings, a $256 \to 64$ drug-pair query, and a $64 \to 256$ gene gate. All three components are enabled in the full model.

Baseline architectures and method-specific parameters. The following settings document the recorded implementations. The available run records do not establish that every method received an identical hyperparameter-search budget, so the results should not be read as a controlled comparison of tuning effort. DeepSynergy concatenates the two drug representations and the cell-line profile and applies a $2048 \to 1024 \to 1$ MLP with ReLU hidden activations, input dropout 0.2, hidden dropout 0.5, and He initialization.

MRHGNN uses a 512-dimensional BioEncoder, RHGNN branches of $512 \to 256 \to 256$ and $512 \to 128 \to 256$, and branch dropout rates of 0.2 and 0. Its drug-protein and synergy hypergraphs use four and two relation types, respectively. The self-supervised edge-masking loss has weight 0.2 and mask ratio 0.2.

HypergraphSynergy encodes 75-dimensional atom features with a $75 \to 128 \to 100$ GCN and the cell profile with a $G \to 128 \to 100$ MLP. Three hypergraph-convolution layers map $100 \to 256 \to 256 \to 256$. The decoder maps the concatenated 768-dimensional drug-drug-cell representation through $768 \to 384 \to 192 \to 1$ with dropout 0.4. Its drug- and cell-reconstruction losses have weight 0.4.

HypertranSynergy uses a two-layer 128-dimensional drug GNN, a $G \to 128 \to 128$ cell encoder, two-head self-attention, a 256-dimensional feed-forward block, and a $128 \to 512$ hypergraph convolution. Attention and feed-forward dropout are 0.5, while the drug GNN, hypergraph layer, and decoder use dropout 0.3. The decoder is $1536 \to 768 \to 384 \to 1$, and the reconstruction loss has weight 0.4.

MHGSynergy maps molecular structure, chemical descriptors, drug targets, and cell-line expression to 128-dimensional embeddings. Its three modality-specific hypergraph branches use $128 \to 128 \to 128$ transformations with dropout 0.3, followed by attention-based modality fusion. The decoder maps the 1152-dimensional concatenated drug-drug-cell representation through $1152 \to 576 \to 288 \to 1$ with dropout 0.4.

B Encoder Robustness Analyses

Protocol. These analyses vary two encoder choices while retaining the O'Neil evaluation protocol in Appendix A. They address whether the observed ranking performance requires one narrowly specified hyperedge subset or RHGNN2 dropout rate. They do not replace the decoder-component ablations in Figure 3. Each plotted point is a ten-fold mean with a between-fold standard deviation. We compare the magnitudes and consistency of changes rather than treating overlapping standard-deviation bars as a formal significance test.

RHGNN2 **dropout.** RHGNN2 encodes the cell-level synergy and antagonism hypergraph assembled from training labels, making its regularization a separate question from RHGNN1 edge selection. Figure 4(b) evaluates dropout rates of 0, 0.1, and 0.2 under the O'Neil protocol described in Appendix A. Each point is a ten-fold mean, and each error bar denotes the corresponding between-fold standard deviation. The maximum changes in mean AUC and AUPR are 0.0014 and 0.0032, respectively. Both changes are smaller than the between-fold variation, and the response is non-monotonic. The sweep therefore indicates local robustness within the tested range rather than a statistically reliable advantage for a particular dropout rate.

Scope of the robustness claim. Taken together, the analyses support stability within the six tested RHGNN1 edge subsets and three RHGNN2 dropout settings on O'Neil. They do not establish that motif construction is unnecessary, that any single edge type has no effect, or that the same sensitivity pattern holds on NCI-ALMANAC. The dropout result also does not establish insensitivity to stronger regularization. Fold-level paired differences or additional seeds would be needed to assess small configuration contrasts more precisely. A no-motif or alternative-hypergraph control would be needed to isolate the contribution of motif-guided construction itself.

C Cell-Dependent Behavior and Case Study

Cell-line structure in the gene-synergy space. Figure 6 visualizes the 37 O'Neil cell lines using $x_c^\top W_{\mathrm{gs}}$. The first two principal components explain 29.4% and 20.8% of the variance. Clustering and the silhouette calculation use the original GSK space rather than the two-dimensional projection. Clustering with $K = 5$ yields a silhouette score of 0.198, indicating partial structure rather than a definitive tissue taxonomy. Cluster 1 contains EFM192B, KPL1, LNCAP, T47D, and ZR751. Cluster 2 contains HT29, NCIH2122, OV90, RKO, UWB1289, and VCAP. Cluster 3 contains A2058, A2780, A375, A427, ES2, HCT116, HT144, MDAMB436, NCIH23, NCIH460, PA1, RPMI7951, SKMEL30, and UACC62. Cluster 4 contains COLO320DM, DLD1, LOVO, MSTO, NCIH520, and SW620. Cluster 5 contains CAOV3, NCIH1650, OVCAR3, SKMES1, SKOV3, and SW837.

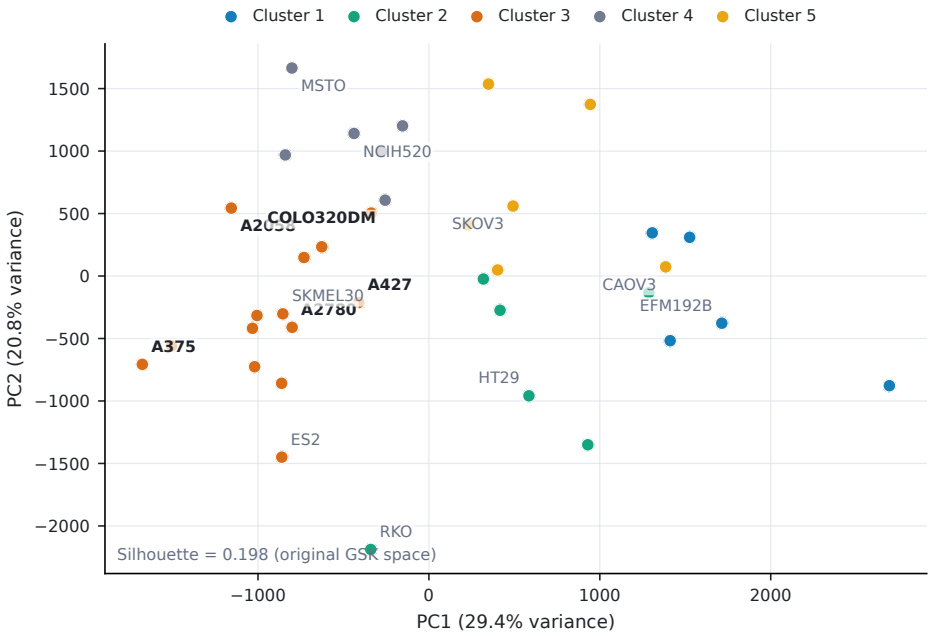


Figure 6: Cell-line structure learned by GSK on O'Neil. Each point represents $x_c^\top W_{\text{gs}}$ projected onto the first two principal components. The axis labels report explained variance. Colors denote five clusters ($K = 5$, silhouette score $= 0.198$), and selected cell-line names are shown for readability.

Cell-line-specific gene attribution. We rank genes using the absolute input gradient of the GSK contribution, $\text{Imp}_{a,b,c,k} = |\partial s^{\text{gs}}_{a,b,c} / \partial X_{c,k}|$. For ABT-888 and DASATINIB (IDs 1 and 7), the predicted probabilities are 0.9980 in A2780 and 0.0012 in COLO320DM. Table 2 summarizes the contrast.

Table 2: Cross-cell-line attribution comparison for ABT-888 and DASATINIB. "Specific genes" occur in one cell line's top 20 and are absent from the other's.

	A2780	COLO320DM
Predicted probability	0.9980	0.0012
Highest-ranked gene	PYCR1 (0.170459)	MEST (0.136277)
Top-20-specific genes	PCM1, SSBP2, DMTF1, KCTD5, TRIB1, PCMT1, SERTAD3, RAP1A	EFCAB14, MBTPS1, COQ3, CD19, TCIRG1, RABGGTA, ABAT, CXCR4

Top-20 overlap: 12 genes (60%). Shared genes can nevertheless occupy different ranks.

Twelve of the top 20 genes overlap between the two cell lines, leaving eight cell-line-specific genes in each list. PYCR1 is highest-ranked in A2780, while MEST is highest-ranked in COLO320DM. Shared genes can also shift priority, as BCL7B moves from rank 17 in A2780 to rank 10 in COLO320DM. These results show that the GSK attribution varies with cellular context. Because this analysis uses one selected pair and one pretrained checkpoint, it is a model diagnostic rather than a population-level stability result. It does not establish the stability of the rankings, and no gene perturbation is performed here. Complete rankings appear in Appendix J.

Case study and retrospective literature search. The analysis script loads the pretrained fold-0 checkpoint, scores every unordered pair among the available drugs for each cell line, sorts pairs within a cell line, and reports the first five for A2058, A2780, A375, and A427. It does not exclude pairs seen during training or require the corresponding drug-drug-cell triple to have been experimentally measured. Consequently, "score" below means the model's retrospective sigmoid output and

should not be interpreted as an experimentally calibrated probability. We conduct a non-exhaustive retrospective literature search to assess whether the selected predictions are supported by previously published evidence. Evidence is graded at the drug-pair level while preserving cell-line and disease context. We use D for an exact pair tested in the same cell line, X for an exact pair tested only in another context, A for a pathway analogue using a different inhibitor, F for feasibility without pairwise synergy validation, and U when no traceable pair-specific source was located. Conference abstracts and a doctoral thesis are identified explicitly rather than treated as peer-reviewed same-context evidence.

The conservative tally is therefore D$=1$, X$=10$, A$=1$, F$=1$, and U$=7$. The strongest observation is the same-cell-line A2058 result for dual WEE1-CHK1 inhibition. Cross-context rows support biological plausibility only. The schedule dependence of paclitaxel-vinblastine and the genotype dependence of BEZ-235-dasatinib also show why drug-pair identity alone is insufficient. Prospective dose-matrix experiments with biological replicates, predefined synergy metrics, and held-out candidate selection are required before these predictions can be considered validated.

Table 3: Supportive literature evidence for the 20 case-study predictions identified through a non-exhaustive retrospective search. A repeated citation indicates that one source supports multiple rows. Evidence in another cell line, tissue, assay, or treatment schedule does not validate the predicted cell-line-specific response.

Cell	Drug pair	Score	Grade	Source type	Evidence and scope
A2058	AZD1775 + MK-8776	0.9999	D	Journal	Exact pair in A2058. At 150 nM, MK-8776 shifted the MK-1775 EC_{50} from 225 to 45 nM (Guertin et al., 2012).
A2058	MK-8776 + temozolomide	0.9992	A	Journal	Temozolomide potentiation in melanoma, including A2058, used GNE-783 rather than MK-8776 (Xiao et al., 2013).
A2058	MK-4827 + temozolomide	0.9989	X	Abstract	Exact pair tested in A375 and FEMX-1, not A2058. Moderate potentiation was dose dependent (Tawbi et al., 2011).
A2058	dasatinib + MK-8776	0.9982	U	-	No traceable pair-specific experimental source located.
A2058	mitomycin + MK-8776	0.9967	U	-	No traceable pair-specific experimental source located.
A2780	dasatinib + MK-8776	1.0000	U	-	No traceable pair-specific experimental source located.
A2780	BEZ-235 + dasatinib	1.0000	X	Thesis	Exact pair was synergistic in c-KIT-mutant Kasumi-1 and HMC1.2 cells, whereas it was antagonistic in several c-KIT-wild-type controls (Tinsley, 2014).
A2780	AZD1775 + dasatinib	1.0000	X	Journal	Exact pair validated in proneural glioblastoma stem-like cells, not A2780 (Alhalabi et al., 2024).
A2780	dasatinib + sunitinib	0.9999	X	Journal	Exact pair was synergistic in MDAH2774 and SKOV3 ovarian cells, not A2780 (Wen et al., 2020).
A2780	dasatinib + MK-2206	0.9999	X	Journal	Exact pair was synergistic in MDAH2774 and SKOV3 ovarian cells, not A2780 (Wen et al., 2020).
A375	BEZ-235 + MK-8669	1.0000	U	-	No traceable pair-specific experimental source located.
A375	BEZ-235 + MK-2206	0.9999	X	Journal	Exact pair showed enhanced antitumor effects in esophageal squamous-cell models, not A375 (Shi et al., 2019).
A375	AZD1775 + MK-8776	0.9999	X	Journal	Exact pair validated in A2058 and other cell lines. The cited study does not establish A375-specific response (Guertin et al., 2012).
A375	paclitaxel + vinblastine	0.9998	X	Journal	Exact pair tested in KB and MCF-7 cells. Synergy occurred with sequential dosing, whereas simultaneous dosing was antagonistic (Giannakakou et al., 1998).
A375	MK-8669 + paclitaxel	0.9998	F	Phase Ib	Although the combination was clinically feasible in solid tumors, the study did not validate A375-specific or pairwise synergy (Sessa et al., 2010).
A427	BEZ-235 + dasatinib	0.9996	X	Thesis	Exact pair was context dependent in c-KIT-mutant and c-KIT-wild-type leukemia models, not A427 (Tinsley, 2014).
A427	dasatinib + MK-8776	0.9995	U	-	No traceable pair-specific experimental source located.
A427	BEZ-235 + L778123	0.9990	U	-	No traceable pair-specific experimental source located.
A427	dasatinib + L778123	0.9980	U	-	No traceable pair-specific experimental source located.
A427	BEZ-235 + MK-2206	0.9980	X	Journal	Exact pair showed enhanced antitumor effects in esophageal squamous-cell models, not A427 (Shi et al., 2019).

D THEORETICAL OVERVIEW AND NOTATION

Notation. The following tables collect the notation used throughout the method and theoretical analysis. Symbols that occur only locally are defined next to the corresponding equation.

Data, indices, and dimensions.

a, b	Indices of the two drugs in a candidate combination.
i, j	Generic drug or drug-node indices.
c	Cell-line index.
k	Gene index ($k \in \{1, \ldots, G\}$).
m	Latent cell-representation coordinate ($m \in \{1, \ldots, d_c\}$).
r, s	Drug-representation coordinates used in the tensor analysis.
d	Drug-representation dimension ($d = 256$ in our implementation).
d_c	BioEncoder cell-embedding dimension ($d_c = 512$).
d_a	Gene-attention subspace dimension ($d_a = 64$).
G	Number of gene-expression features ($G = 977$).
X_a, X_b	Molecular feature vectors of drugs a and b.
$X_c \in \mathbb{R}^G$	Raw gene-expression profile of cell line c.
$x_c = \Phi(X_c)$	BioEncoder representation of cell line c, with $x_c \in \mathbb{R}^{d_c}$.
x_d, x_c, x_p	Initial drug, cell-line, and protein embeddings produced by the BioEncoder.
$z_i, z_j \in \mathbb{R}^d$	Final drug representations obtained by averaging the two RHGNN outputs.
$y_{a,b,c} \in \{0, 1\}$	Synergy label for drugs a, b in cell line c.

Hypergraph encoder.

$\mathcal{G} = (V, E)$	A hypergraph with node set V and hyperedge set E.
v, e, t	A node, a hyperedge, and a hyperedge-relation type, respectively.
x_v, x'_v	Input and updated representations of node v.
$h_e^{(t)}$	Representation of type-t hyperedge e.
$\mathbf{W}_t, a_t$	Type-specific projection matrix and attention vector.
$\alpha_{e,v}^{(t)}$	Attention weight of node v when constructing hyperedge e of type t.
$\beta_{v,e}$	Attention weight of incident hyperedge e when updating node v.
$\mathbf{W}_0, \mathbf{W}$	Residual and hyperedge-message projection matrices.
$\mathrm{RHGNN1}, \mathrm{RHGNN2}$	Drug-protein and synergy/antagonism hypergraph encoders.

Decoder and theoretical analysis.

$g_c = \sigma(W_g x_c + b_g)$	Cell-conditioned gate generated from the encoded cell state.
$D_c = \mathrm{diag}(g_c)$	Diagonal matrix associated with the cell-conditioned gate.
$\widetilde{z}_i = g_c \odot z_i$	Drug representation after cell-conditioned gating.
w_c	Learned cell-specific diagonal scaling vector in the base bilinear decoder.
$W_{\mathrm{global}}, W_{\mathrm{eff},c}$	Symmetric shared drug-drug matrix and its cell-scaled effective matrix.
$W_{\mathrm{gs}} \in \mathbb{R}^{d_c \times d}$	Projection from the latent cell embedding to the drug-interaction space.
$v_c = x_c^\top W_{\mathrm{gs}}$	Cell-conditioned latent interaction vector used by GSK.
$s^{\mathrm{base}}_{a,b,c}, s^{\mathrm{gs}}_{a,b,c}, s^{\mathrm{gsa}}_{a,b,c}$	Score contributions of the base bilinear decoder, GSK, and GSA.
$s_{a,b,c}, \widehat{y}_{a,b,c}$	Total decoder logit and predicted synergy probability.
$e_k \in \mathbb{R}^{d_a}$	Learnable embedding of gene k.
$h_k = e_k X_{c,k}$	Expression-modulated key of gene k.
$q = W_q(\widetilde{z}_i + \widetilde{z}_j)$	Drug-pair query used by GSA.
α_k	Attention weight assigned to gene k for the current drug pair.
W_{gsa}	Projection from the attended gene context to the drug space.
$T^{\mathrm{full}} \in \mathbb{R}^{d \times d \times G}$	Unrestricted interaction tensor in the raw-expression space.
$T^{\mathrm{GSK}} \in \mathbb{R}^{d \times d \times d_c}$	Structured interaction tensor in the latent cell-representation space.
$T_m \in \mathbb{R}^{d \times d}$	Matrix fiber associated with latent cell coordinate m.
δ_{rs}	Kronecker delta used to express the diagonal GSK structure.
$H \in \mathbb{R}^{G \times d_a}$	Gene-key matrix whose k-th row is $h_k^\top$.
$P_\perp$	Centering projection $I_G - G^{-1}\mathbf{1}\mathbf{1}^\top$.
J	Jacobian $\mathrm{diag}(\alpha) - \alpha\alpha^\top$ of the softmax function.
$R(w), \widehat{R}(w)$	Population risk and empirical risk for decoder coefficient vector w.
n, δ, ϵ	Sample size, failure probability, and target generalization gap.

Theoretical setup. This appendix analyzes the three decoder components from two complementary perspectives: expressiveness and coefficient-controlled capacity. The analysis isolates the decoder while treating the learned representations as fixed. This convention makes the comparison between alternative scoring functions explicit but does not characterize the data-dependent function class or generalization behavior of the end-to-end trained network. The dimensions and principal parameters are summarized above.

For theoretical comparison, consider a cell-independent symmetric bilinear reference decoder,

$$s_{ij} = z_i^\top W z_j. \tag{D.1}$$

GeneSyn-HGNN adds expression-derived gating, a structured drug-drug-latent cell interaction, and pair-dependent gene-attention weights. The following sections state the conditions under which these changes increase expressiveness and reduce the effective dimension of the scoring class relative to the specified reference families.

E Expressiveness of Cell-Conditioned Gating

CCG computes

$$g_c = \sigma(W_g x_c + b_g), \qquad \widetilde{z}_i = D_c z_i, \qquad D_c = \mathrm{diag}(g_c). \tag{E.1}$$

To isolate the incremental effect of CCG, this section temporarily omits the learned cell-indexed diagonal relation already present in the implemented base decoder. The resulting comparison is against a cell-independent symmetric bilinear reference and does not show that CCG is strictly more expressive than the complete base decoder containing w_c. The gated bilinear term is

$$s^c_{ij} = z_i^\top D_c W D_c z_j. \tag{E.2}$$

Thus, CCG converts a fixed interaction matrix into the cell-conditioned kernel $W_{\mathrm{gate},c} = D_c W D_c$.

Proposition E.1 (Strict expressiveness)**.** *Let symmetric W and the gate parameters be learnable. The gated family*

$$\mathcal{F}_{\mathrm{gate}} = \{(z_i, z_j, x_c) \mapsto z_i^\top D_c W D_c z_j\} \tag{E.3}$$

strictly contains the family of cell-independent bilinear functions

$$\mathcal{F}_{\text{fixed}} = \{(z_i, z_j, x_c) \mapsto z_i^\top W_{\text{ref}} z_j\}, \tag{E.4}$$

provided that the gate can take two distinct values on the set of cell states.

Proof. For inclusion, set $W_g = 0$. Then $g_c = \sigma(b_g)$ is independent of x_c, and every entry of g_c lies in $(0, 1)$. Consequently, $D_c \equiv D$ is invertible. Given an arbitrary symmetric W_{ref}, choose

$$W = D^{-1} W_{\text{ref}} D^{-1}. \tag{E.5}$$

It follows directly that $DWD = W_{\text{ref}}$, so the gated score equals $z_i^\top W_{\text{ref}} z_j$ for every (z_i, z_j, x_c). Therefore $\mathcal{F}_{\text{fixed}} \subseteq \mathcal{F}_{\text{gate}}$.

For strictness, let $g^{(1)} = g(x_{c_1})$ and $g^{(2)} = g(x_{c_2})$ be two distinct gate vectors. There is then a coordinate r for which $g_r^{(1)} \neq g_r^{(2)}$. Both values are positive, so $(g_r^{(1)})^2 \neq (g_r^{(2)})^2$. Choose W to have $W_{rr} = 1$ and all other entries zero, and take $z_i = z_j = \mathbf{e}_r$, the r-th standard basis vector. The resulting function satisfies

$$s(\mathbf{e}_r, \mathbf{e}_r, x_c) = g_r(x_c)^2, \tag{E.6}$$

and hence takes different values at x_{c_1} and x_{c_2}. Every function in $\mathcal{F}_{\text{fixed}}$ is independent of x_c, so no member of that family can equal this f on the common domain. The inclusion is therefore strict. □

Proposition E.2 (Rank preservation)**.** *If* $\operatorname{rank}(W) = r$*, then* $\operatorname{rank}(D_c W D_c) = r$ *for every finite gate input.*

Proof. For a finite gate input, every sigmoid output satisfies $0 < g_{c,k} < 1$. Hence

$$\det(D_c) = \prod_{k=1}^{d} g_{c,k} > 0, \tag{E.7}$$

so D_c is invertible. Right multiplication by an invertible matrix does not change the dimension of the column space, and left multiplication by one is a bijection on that space. Thus

$$\operatorname{rank}(D_c W D_c) = \operatorname{rank}(W D_c) = \operatorname{rank}(W) = r. \tag{E.8}$$

□

Corollary E.3 (Cell sensitivity)**.** *Assume that* Φ *is differentiable at* X_c*. Define*

$$u_c = z_i \odot W(g_c \odot z_j) + z_j \odot W^\top (g_c \odot z_i). \tag{E.9}$$

If

$$J_\Phi(X_c)^\top W_g^\top \operatorname{diag}\big(g_c \odot (1 - g_c)\big) u_c \neq 0, \tag{E.10}$$

then $\nabla_{X_c} s_{ij}^c \neq 0$*. Thus CCG is locally sensitive to the expression profile whenever the BioEncoder Jacobian, gate Jacobian, and drug-interaction direction do not annihilate one another.*

Proof. Writing $\widetilde{z}_i = g_c \odot z_i$ and $\widetilde{z}_j = g_c \odot z_j$ gives $s_{ij}^c = \widetilde{z}_i^\top W \widetilde{z}_j$. Differentiating the first and second gated factors separately yields

$$\nabla_{g_c} s_{ij}^c = z_i \odot W(g_c \odot z_j) + z_j \odot W^\top (g_c \odot z_i) = u_c. \tag{E.11}$$

Because $g_c = \sigma(W_g x_c + b_g)$ componentwise,

$$\frac{\partial g_c}{\partial x_c} = \operatorname{diag}\big(g_c \odot (1 - g_c)\big) W_g. \tag{E.12}$$

Applying the chain rule through $x_c = \Phi(X_c)$ therefore gives

$$\nabla_{X_c} s_{ij}^c = J_\Phi(X_c)^\top W_g^\top \operatorname{diag}\big(g_c \odot (1 - g_c)\big) u_c. \tag{E.13}$$

The stated condition is exactly that this gradient is nonzero. Proposition E.2 shows simultaneously that this cell dependence is introduced without changing the rank of W. □

F Coefficient Complexity and Conditional Capacity of the Gene-Synergy Kernel

A fully parameterized interaction between two d-dimensional drug vectors and a G-dimensional expression profile has the form

$$s^{\text{full}}_{i,j,c} = \sum_{r=1}^{d}\sum_{s=1}^{d}\sum_{k=1}^{G} T^{\text{full}}_{rsk}\widetilde{z}_{ir}\widetilde{z}_{js}X_{c,k}, \tag{F.1}$$

and requires $O(d^2G)$ decoder parameters. Enforcing exchange symmetry reduces the exact coefficient count of the unrestricted reference to $Gd(d+1)/2$ but leaves its $O(d^2G)$ order unchanged. GSK first compresses X_c to $x_c \in \mathbb{R}^{d_c}$ and then uses

$$v_c = x_c^\top W_{\text{gs}} \in \mathbb{R}^d, \qquad s^{\text{gs}}_{i,j,c} = (\widetilde{z}_i \odot \widetilde{z}_j)^\top v_c. \tag{F.2}$$

Equivalently,

$$s^{\text{gs}}_{i,j,c} = \sum_{r=1}^{d}\sum_{m=1}^{d_c} (W_{\text{gs}})_{mr}\widetilde{z}_{ir}\widetilde{z}_{jr}x_{c,m}. \tag{F.3}$$

This is a structured third-order tensor whose entries satisfy $T^{\text{GSK}}_{rsm} = \delta_{rs}(W_{\text{gs}})_{mr}$ in the learned cell-representation space. It retains interactions in shared latent drug dimensions, while the full matrix in the base bilinear term remains available to capture cross-dimension drug interactions.

Proposition F.1 (Parameter efficiency)**.** *Holding the BioEncoder, both RHGNN encoders, and CCG fixed, GSK reduces the number of decoder parameters for the three-way interaction from $O(d^2G)$ to $O(dd_c)$.*

Proof. The unrestricted tensor has one independent coefficient T^{full}_{rsk} for each $r, s \in \{1, \dots, d\}$ and $k \in \{1, \dots, G\}$, hence exactly d^2G coefficients. In GSK, the effective tensor in the cell-representation space satisfies

$$T^{\text{GSK}}_{rsm} = \delta_{rs}(W_{\text{gs}})_{mr}. \tag{F.4}$$

All entries with $r \neq s$ are therefore fixed at zero, and every remaining entry is determined by one of the $d_c d$ coefficients of $W_{\text{gs}} \in \mathbb{R}^{d_c \times d}$. With the BioEncoder, both RHGNN encoders, and CCG fixed, no other trainable decoder parameter is introduced by this three-way term. Thus the parameter count changes from d^2G to dd_c, which gives the stated orders. The two factors in the reduction are the diagonal restriction in the drug modes, from d^2 to d, and the replacement of the raw G-dimensional profile by the d_c-dimensional cell representation. □

Proposition F.2 (Decoder-level covering bound)**.** *Assume the BioEncoder, both RHGNN encoders, and CCG are learned from data independent of the n examples used in this bound and are then held fixed. Consequently, the learned representations define sample-independent feature maps. For either decoder, suppose its coefficient vector satisfies $\|w\|_2 \le B_w$, its feature map satisfies $\|\varphi\|_2 \le B_\varphi$ almost surely, and the loss takes values in $[0, 1]$ and is L_ℓ-Lipschitz in the score. Assume these norm and Lipschitz bounds are dimension-independent and that the $n \ge 2$ training examples are independent and identically distributed. Consider the feature maps*

$$\varphi_{\text{full}} = \text{vec}(\widetilde{z}_i \otimes \widetilde{z}_j \otimes X_c) \in \mathbb{R}^{d^2G}, \tag{F.5}$$

and

$$\varphi_{\text{diag}} = (\widetilde{z}_{ir}\widetilde{z}_{jr}x_{c,m})_{r,m} \in \mathbb{R}^{dd_c}. \tag{F.6}$$

Up to logarithmic factors and norm-dependent constants, the corresponding uniform generalization gaps are bounded above by

$$\text{gap}_{\text{full}} = O\left(\sqrt{\frac{d^2G\log n + \log(1/\delta)}{n}}\right), \tag{F.7}$$

and

$$\text{gap}_{\text{GSK}} = O\left(\sqrt{\frac{dd_c\log n + \log(1/\delta)}{n}}\right) \tag{F.8}$$

simultaneously with probability at least $1-\delta$ over the training examples, where the constants may depend on B_w, B_φ, and L_ℓ independently of n and the feature dimension.

Proof. We first derive the bound for a generic D-dimensional class

$$\mathcal{F}_D = \{(y, \varphi) \mapsto \ell(y, \langle w, \varphi \rangle) : \|w\|_2 \leq B_w\}. \tag{F.9}$$

Let $\mathcal{N}_\rho$ be a Euclidean ρ-net of the radius-B_w parameter ball. A volume comparison gives

$$|\mathcal{N}_\rho| \leq \left(1 + \frac{2B_w}{\rho}\right)^D. \tag{F.10}$$

For a fixed $w_0 \in \mathcal{N}_\rho$, Hoeffding's inequality applies because the loss lies in $[0, 1]$:

$$\Pr\Big(|R(w_0) - \widehat{R}(w_0)| > t\Big) \leq 2e^{-2nt^2}. \tag{F.11}$$

Taking a union bound over the net and solving for t shows that, with probability at least $1 - \delta$,

$$\max_{w_0 \in \mathcal{N}_\rho} |R(w_0) - \widehat{R}(w_0)| \leq \sqrt{\frac{D \log(1 + 2B_w/\rho) + \log(2/\delta)}{2n}}. \tag{F.12}$$

Now take any w in the parameter ball and choose $w_0 \in \mathcal{N}_\rho$ with $\|w - w_0\|_2 \leq \rho$. By the Lipschitz property and Cauchy-Schwarz,

$$|\ell(y, \langle w, \varphi \rangle) - \ell(y, \langle w_0, \varphi \rangle)| \leq L_\ell B_\varphi \rho. \tag{F.13}$$

This inequality holds pointwise, so it changes each of the population and empirical risks by at most $L_\ell B_\varphi \rho$. The triangle inequality therefore gives

$$\sup_w |R(w) - \widehat{R}(w)| \leq \sqrt{\frac{D \log(1 + 2B_w/\rho) + \log(2/\delta)}{2n}} + 2L_\ell B_\varphi \rho. \tag{F.14}$$

Choosing $\rho = (L_\ell B_\varphi \sqrt{n})^{-1}$ (with the zero-Lipschitz case being trivial) makes the last term $2/\sqrt{n}$ and changes the entropy term to $D \log(1 + 2B_w L_\ell B_\varphi \sqrt{n})$. Absorbing norm-dependent constants yields

$$O\left(\sqrt{\frac{D \log n + \log(1/\delta)}{n}}\right). \tag{F.15}$$

For the unrestricted tensor, vectorizing its coefficients and features gives $D = d^2G$. For GSK, vectorizing W_{gs} against φ_{diag} gives $D = dd_c$. Apply the generic result to each class with failure probability $\delta/2$. A final union bound makes both inequalities hold on the same event with probability at least $1 - \delta$. Replacing $\log(2/\delta)$ by $\log(4/\delta)$ affects only constants inside the displayed orders. Solving these particular upper bounds for a target gap ϵ gives sufficient rates of $\widetilde{O}(d^2G/\epsilon^2)$ samples for the full tensor and $\widetilde{O}(dd_c/\epsilon^2)$ for GSK, where $\widetilde{O}$ suppresses logarithmic factors. These rates need not be tight and do not order the true sample complexities. The comparison is conditional on sample-independent feature maps and is not a bound for the complete nonconvex end-to-end training procedure. □

Remark F.3 (Separation from the learning curve)**.** *On O'Neil, the full model reaches test AUC* 0.90 *at an interpolated training fraction of approximately* 0.21*, whereas the model without GSK reaches the same level at approximately* 0.48*. The vertical AUC gain is* 0.0311 *at fraction* 0.2 *and increases to approximately* 0.05 *at larger fractions. This horizontal displacement is an empirical observation, not a consequence of Proposition F.2: that proposition compares two fixed-feature decoder classes, whereas the ablation removes GSK from a jointly trained model. It supplies only a conditional upper-bound comparison and neither orders the true sample complexities nor predicts the size or direction of the empirical effect.*

F.1 Exact representability

Let $T_m \in \mathbb{R}^{d \times d}$ denote the interaction matrix associated with cell-representation coordinate m. Because drug combinations are unordered, the relevant interaction function depends only on the symmetric part of each fiber We therefore state the result for symmetric fibers. The following result characterizes when a diagonal GSK representation loses no information relative to them.

Proposition F.4 (Exact representability by a shared basis)**.** *Assume that every T_m is real and symmetric. There exists an orthogonal matrix U such that the GSK form exactly represents every fiber if and only if the matrices $\{T_m\}_m$ are simultaneously orthogonally diagonalizable. Equivalently, the fibers commute pairwise.*

Proof. If $D_m = U^\top T_m U$ is diagonal for every m, set the GSK coefficients for coordinate m to the diagonal of D_m and use $z' = U^\top z$. Then

$$z_i^\top T_m z_j = (z_i')^\top D_m z_j' = \sum_k (D_m)_{kk} z_{ik}' z_{jk}', \tag{F.16}$$

which is exactly the GSK form.

Conversely, suppose one shared orthogonal basis represents every fiber. Then for each m there is a diagonal D_m such that

$$z_i^\top T_m z_j = z_i^\top U D_m U^\top z_j \qquad \text{for all } z_i, z_j \in \mathbb{R}^d. \tag{F.17}$$

Taking $z_i = \mathbf{e}_r$ and $z_j = \mathbf{e}_s$ for every pair of standard basis vectors shows equality of every matrix entry, so $T_m = U D_m U^\top$. Hence, for any m and m',

$$T_m T_{m'} = U D_m D_{m'} U^\top = U D_{m'} D_m U^\top = T_{m'} T_m, \tag{F.18}$$

because diagonal matrices commute.

It remains to show that pairwise commutation is sufficient. Since T_1 is real and symmetric, the spectral theorem decomposes $\mathbb{R}^d$ into mutually orthogonal eigenspaces of T_1. If v belongs to the eigenspace with eigenvalue λ, then for every m,

$$T_1(T_m v) = T_m(T_1 v) = \lambda T_m v, \tag{F.19}$$

so each eigenspace of T_1 is invariant under every T_m. The restrictions of $T_2, \ldots, T_{d_c}$ to each such eigenspace remain real symmetric and pairwise commuting. Diagonalize the restriction of T_2 inside each invariant eigenspace. Its eigenspaces are again invariant under all remaining matrices by the same calculation. Repeating this finite procedure produces an orthonormal basis of simultaneous eigenvectors for all fibers. Letting U contain these vectors as columns makes every $U^\top T_m U$ diagonal, completing the equivalence. □

When the fibers do not commute, the best diagonal model in a learned basis can be viewed as a joint approximate diagonalization. The residual is the off-diagonal energy that remains after choosing U. Because the drug embeddings are learned end to end, the encoder can rotate their latent basis toward directions that reduce this residual.

G Pair-Specific Selectivity of Gene Attention

For gene k, define the expression-modulated key

$$h_k = e_k X_{c,k} \in \mathbb{R}^{d_a}. \tag{G.1}$$

The gated drug pair defines the query and attention weights

$$q^{ij} = W_q(\widetilde{z}_i + \widetilde{z}_j), \qquad \alpha_k^{ij} = \frac{\exp\big((q^{ij})^\top h_k / \sqrt{d_a}\big)}{\sum_{k'=1}^{G} \exp\big((q^{ij})^\top h_{k'} / \sqrt{d_a}\big)}. \tag{G.2}$$

The context and its contribution to the score are

$$u^{ij} = \sum_{k=1}^{G} \alpha_k^{ij} e_k, \qquad s_{i,j,c}^{\mathrm{gsa}} = (\widetilde{z}_i \odot \widetilde{z}_j)^\top \sigma(W_{\mathrm{gsa}} u^{ij}). \tag{G.3}$$

The attention logits separate gene relevance into expression and compatibility:

$$(q^{ij})^\top h_k = X_{c,k} (q^{ij})^\top e_k. \tag{G.4}$$

Let $H \in \mathbb{R}^{G \times d_a}$ have rows $h_k^\top$, and define the centering projection $P_\perp = I_G - G^{-1}\mathbf{1}\mathbf{1}^\top$.

Proposition G.1 (Drug-pair-specific selection)**.** *For two drug pairs with queries q and q', their attention vectors are equal if and only if*

$$H(q - q') \in \mathrm{span}\{\mathbf{1}\}. \tag{G.5}$$

In particular, if $P_\perp H$ has full column rank, then $q \neq q'$ implies $\alpha(q) \neq \alpha(q')$. Setting $W_q = 0$ gives a pair-independent uniform-attention special case, whereas attainable distinct queries yield distinct attention vectors under the rank condition.

Proof. Let $a = Hq/\sqrt{d_a}$ and $a' = Hq'/\sqrt{d_a}$. If $a - a' = \gamma\mathbf{1}$ for some scalar γ, multiplying all exponentials by e^γ leaves their normalized values unchanged, so $\mathrm{softmax}(a) = \mathrm{softmax}(a')$.

For the converse, suppose the two softmax vectors are equal. For any genes k and ℓ, equality of the ratios of their probabilities gives

$$e^{a_k - a_\ell} = \frac{\mathrm{softmax}(a)_k}{\mathrm{softmax}(a)_\ell} = \frac{\mathrm{softmax}(a')_k}{\mathrm{softmax}(a')_\ell} = e^{a'_k - a'_\ell}. \tag{G.6}$$

The exponential is injective on $\mathbb{R}$, hence $a_k - a'_k = a_\ell - a'_\ell$ for every k, ℓ. Thus $a - a'$ is a constant vector, which is equivalent to $H(q - q') \in \mathrm{span}\{\mathbf{1}\}$.

Applying $P_\perp$ removes precisely this constant component. Therefore equal attention implies $P_\perp H(q - q') = 0$. If $P_\perp H$ has full column rank, its null space is $\{0\}$, so $q - q' = 0$. Finally, $W_q = 0$ makes $q = 0$ for every drug pair and yields $\alpha_k = 1/G$ for all genes. If the query map instead attains two distinct values, the proved implication gives different attention vectors, which no pair-independent rule can produce. □

Proposition G.2 (Quantitative selectivity)**.** *Suppose every logit along the line segment between q and q' is bounded in magnitude by L. Then*

$$\|\alpha(q) - \alpha(q')\|_1 \geq \frac{e^{-4L}}{G\sqrt{d_a}} \|P_\perp H(q - q')\|_2. \tag{G.7}$$

If $\sigma_{\min}(P_\perp H) = \eta > 0$, this gives

$$\|\alpha(q) - \alpha(q')\|_1 \geq \frac{\eta e^{-4L}}{G\sqrt{d_a}} \|q - q'\|_2. \tag{G.8}$$

Moreover,

$$\|\alpha(q) - \alpha(q')\|_1 \leq \frac{\max_k \|h_k\|_2}{\sqrt{d_a}} \|q - q'\|_2. \tag{G.9}$$

Thus, away from saturated logits and under a nondegenerate key matrix, the change in gene selection is controlled above and below by the change in the drug-pair query.

Proof. Write $a(q) = Hq/\sqrt{d_a}$ and let

$$J(q) = \mathrm{diag}(\alpha(q)) - \alpha(q)\alpha(q)^\top \tag{G.10}$$

be the softmax Jacobian. Set $\Delta q = q - q'$ and parameterize the segment by $q(t) = q' + t\Delta q$. By the chain rule and the fundamental theorem of calculus,

$$\alpha(q) - \alpha(q') = \int_0^1 J(q(t)) \frac{H\Delta q}{\sqrt{d_a}}\, dt. \tag{G.11}$$

The Jacobian is symmetric and satisfies $J(q(t))\mathbf{1} = 0$, so it acts only on the centered component $u = P_\perp H\Delta q/\sqrt{d_a}$. Consequently,

$$\alpha(q) - \alpha(q') = \overline{J}u, \qquad \overline{J} = \int_0^1 J(q(t))\, dt. \tag{G.12}$$

For any $v \perp \mathbf{1}$, expanding the quadratic form gives

$$v^\top J(q(t))v = \sum_k \alpha_k v_k^2 - \Big(\sum_k \alpha_k v_k\Big)^2 = \frac{1}{2}\sum_{k,\ell} \alpha_k \alpha_\ell (v_k - v_\ell)^2. \tag{G.13}$$

Every logit along the path lies in $[-L, L]$, so each attention weight obeys

$$\alpha_k(q(t)) = \frac{e^{a_k(q(t))}}{\sum_\ell e^{a_\ell(q(t))}} \geq \frac{e^{-L}}{Ge^{L}} = \frac{e^{-2L}}{G} =: m. \tag{G.14}$$

It follows that

$$v^\top J(q(t))v \geq \frac{m^2}{2}\sum_{k,\ell}(v_k - v_\ell)^2 = Gm^2\|v\|_2^2 = \frac{e^{-4L}}{G}\|v\|_2^2, \tag{G.15}$$

where the middle equality uses $\sum_k v_k = 0$. Integrating over $t \in [0, 1]$ yields the same lower bound for $v^\top \overline{J}v$. In particular, Cauchy-Schwarz gives

$$\|\overline{J}u\|_2\|u\|_2 \geq u^\top \overline{J}u \geq \frac{e^{-4L}}{G}\|u\|_2^2. \tag{G.16}$$

The inequality also holds when $u = 0$. Otherwise, divide by $\|u\|_2$. Using $\|\overline{J}u\|_1 \geq \|\overline{J}u\|_2$ and substituting the definition of u proves the first lower bound. If $\sigma_{\min}(P_\perp H) = \eta > 0$, the definition of the smallest singular value gives $\|P_\perp H\Delta q\|_2 \geq \eta\|\Delta q\|_2$, proving the second.

For the upper bound, put $w = H\Delta q/\sqrt{d_a}$ and let $\mu = \sum_k \alpha_k(q(t))w_k$. The k-th component of $J(q(t))w$ is $\alpha_k(q(t))(w_k - \mu)$, whence

$$\begin{aligned}
\|J(q(t))w\|_1 &= \sum_k \alpha_k(q(t))|w_k - \mu| \\
&\leq \left(\sum_k \alpha_k(q(t))(w_k - \mu)^2\right)^{1/2} \\
&\leq \left(\sum_k \alpha_k(q(t))w_k^2\right)^{1/2} \\
&\leq \|w\|_\infty.
\end{aligned} \tag{G.17}$$

The first inequality is weighted Cauchy-Schwarz. The second follows because subtracting the weighted mean reduces the weighted second moment. Finally,

$$\|w\|_\infty = \frac{1}{\sqrt{d_a}}\max_k |h_k^\top \Delta q| \leq \frac{\max_k \|h_k\|_2}{\sqrt{d_a}}\|\Delta q\|_2. \tag{G.18}$$

Integrating the pointwise Jacobian bound over t and applying the triangle inequality to the path integral gives the stated upper bound. □

The factor e^{-4L}/G has a natural interpretation. At uniform attention, $J = G^{-1}I - G^{-2}\mathbf{1}\mathbf{1}^\top$ and its nonzero eigenvalues are exactly $1/G$. As attention approaches a one-hot vector, J approaches zero, so no positive global lower bound is possible without excluding saturated logits. These propositions establish selectivity of the attention weights under stated rank and logit conditions. They do not imply that two queries must yield different final GSA scores, that the rank condition holds for the learned keys, or that attention weights identify causal biological effects.

H Additional Details for the Generalization Bound

The proof of Proposition F.2 makes the covering argument explicit. Its bound is uniform over every decoder coefficient vector in the norm ball, not merely over a fixed coefficient vector. Consequently, on the event where the bound holds, the empirically selected decoder is covered as well. If $\widehat{w}$ minimizes empirical risk over that ball and w^* minimizes population risk over the same ball, the uniform bound $\sup_w |R(w) - \widehat{R}(w)| \leq \epsilon_n$ additionally implies

$$\begin{aligned}
R(\widehat{w}) &\leq \widehat{R}(\widehat{w}) + \epsilon_n \\
&\leq \widehat{R}(w^*) + \epsilon_n \\
&\leq R(w^*) + 2\epsilon_n.
\end{aligned} \tag{H.1}$$

The first and third inequalities use uniform convergence. The middle inequality uses empirical optimality. This excess-risk implication still applies only to the decoder class conditional on sample-independent feature maps. If the BioEncoder, either RHGNN encoder, or CCG is fitted on the same n examples, its data dependence is not covered by the stated argument and requires an additional capacity or stability analysis.

I Derivation of the Attention Path Identity

For completeness, this section expands the first step in the proof of Proposition G.2. Along the path

$$q(t) = q' + t(q - q'), \qquad t \in [0, 1], \tag{I.1}$$

the vector-valued fundamental theorem of calculus gives

$$\alpha(q) - \alpha(q') = \int_0^1 \frac{d\alpha(q(t))}{dt}\, dt. \tag{I.2}$$

Because $a(q) = Hq/\sqrt{d_a}$ is linear and $\partial\alpha/\partial a = J$, the chain rule yields

$$\frac{d\alpha(q(t))}{dt} = J(q(t))\frac{H}{\sqrt{d_a}}(q - q'). \tag{I.3}$$

The last two factors do not depend on t, and therefore

$$\alpha(q) - \alpha(q') = \left(\frac{1}{\sqrt{d_a}}\int_0^1 J(q(t))\, dt\right) H(q - q'). \tag{I.4}$$

This identity separates the local sensitivity of softmax, represented by the path-averaged Jacobian, from the change in the drug-pair query projected onto the gene keys.

J Cross-Cell-Line Gene Attribution

For transparency, we report the complete gene rankings used in the diagnostic case study in Appendix C. The analysis instantiates the fold-0 synergy hypergraph, loads the pretrained O'Neil checkpoint, and searches drug IDs 0-7 over the first ten cell lines. ABT-888 and DASATINIB (IDs 1 and 7) attain the largest range of predicted probabilities in this search subset, with A2780 scoring 0.9980 and COLO320DM scoring 0.0012 ($\Delta = 0.9968$). These values are model predictions rather than experimentally measured synergy probabilities.

For drug pair (a, b) in cell line c, the script backpropagates through the GSK contribution and defines the importance of raw expression feature k as

$$\mathrm{Imp}_{a,b,c,k} = \left|\frac{\partial s^{\mathrm{gs}}_{a,b,c}}{\partial X_{c,k}}\right|. \tag{J.1}$$

Table 4 reports the resulting top-20 lists for the evaluated checkpoint. Twelve genes occur in both lists: PYCR1, MEST, MEGF9, DNAJC15, KIAA0753, GNB5, TLE1, BCL7B, ATP1B1, IKBKE, ACAT2, and EIF5. The table is a checkpoint-level diagnostic. Assessing attribution stability across folds and random seeds remains future work.

Table 4: Complete top-20 GSK gradient attributions for ABT-888 and DASATINIB across A2780 and COLO320DM. Importance is $|\partial s^{\mathrm{gs}}_{a,b,c}/\partial X_{c,k}|$ for the evaluated pretrained O'Neil checkpoint.

Rank	A2780 gene	Importance	COLO320DM gene	Importance
1	PYCR1	0.170459	MEST	0.136277
2	PCM1	0.161561	PYCR1	0.128285
3	MEST	0.160677	EFCAB14	0.127072
4	MEGF9	0.159134	MEGF9	0.124363
5	SSBP2	0.156589	KIAA0753	0.123515
6	DMTF1	0.133312	DNAJC15	0.120057
7	KCTD5	0.131842	GNB5	0.115325
8	DNAJC15	0.130692	MBTPS1	0.114382
9	TRIB1	0.130034	TLE1	0.110331
10	KIAA0753	0.129807	BCL7B	0.108783
11	PCMT1	0.128809	ATP1B1	0.103821
12	EIF5	0.125982	IKBKE	0.101322
13	ATP1B1	0.125787	COQ3	0.099637
14	GNB5	0.125222	CD19	0.099143
15	TLE1	0.123856	ACAT2	0.098159
16	SERTAD3	0.123672	TCIRG1	0.097914
17	BCL7B	0.122624	RABGGTA	0.097894
18	ACAT2	0.120482	ABAT	0.097808
19	RAP1A	0.120107	CXCR4	0.097152
20	IKBKE	0.119904	EIF5	0.096603