RESIDUAL FUSION FOR COLD-START MATRIX COMPLETION AND REWARD-MODEL-GUIDED GENERATION: A CASE STUDY IN DRUG RESPONSE PREDICTION WITH SPECIFICATION GAMING

Anonymous authors
Paper under double-blind review

ABSTRACT

Content-conditioned sparse matrix completion—predicting an interaction between a query entity and a target entity from content alone, without an identity-indexed lookup table—is a recurring cold-start problem across recommendation, matching, and scientific prediction, since a lookup table is by construction undefined for an entity absent from training. We study this problem in drug response prediction, where both the compound and cell-line axes are extremely sparse (median one observed interaction per compound) and only the cell-line axis must generalize zero-shot to unseen entities. We present a Two-Tower Residual Late Fusion architecture in which a hard-memorization branch resolves fine-grained interaction structure while a gated residual branch, conditioned on cold-start content alone, supplies interaction corrections. Because the residual branch only *adds* to the memorization branch rather than replacing it, the architecture admits a structural non-degradation property that does not depend on the residual branch being useful at test time. Evaluated on 51 breast cancer cell lines (136,342 records), the architecture outperforms a hard-memorization-only baseline in 48/51 cell lines (94.1%) under leave-cell-line-out evaluation (mean $\Delta R^2 = +0.016$). It also generalizes zero-shot to 601 externally held-out cell lines across 27 tissue types with zero cell-line overlap with training (median $R^2 = 0.627$). Prediction error correlates significantly with embedding-space distance to the nearest training entity ($r = 0.159$, $p = 9.2 \times 10^{-5}$), a significant signal that the content encoder has learned a distance-aware, not merely memorized, representation. We then ask a distinct question: is a validated predictor of this kind safe to repurpose, frozen, as a reward model for a downstream generative policy? A six-run reinforcement learning (RL) study (v1–v6) found no instance, under reward-model-blind checks, of the generator gaming the frozen predictor itself. It did, however, uncover a specification-gaming failure in an auxiliary diversity objective, which we diagnosed mechanistically and resolved with a calibrated, similarity-clustered penalty. A final sample-efficiency variant then achieved a reproducible reward improvement, replicated across two independent seeds, from an identical predictor-call budget without eroding this fix. Together, these results indicate that gated residual late fusion offers a practical route to cold-start generalization without sacrificing memorized accuracy, and that reward models built on such architectures can be repurposed for RL fine-tuning with failure modes that are mechanistically diagnosable and correctable rather than silently exploited. Code, trained checkpoints, and the RL evaluation pipeline will be released upon acceptance.

1 INTRODUCTION

Predicting an interaction between two entities from content alone, rather than from a learned identity embedding, is a structural requirement whenever one axis of the interaction matrix must generalize to entities absent from training. This is the classic cold-start problem in recommendation (Volkovs et al., 2017): a lookup-table embedding indexed by entity identity has no entry for an unseen entity, so any content-agnostic model is provably incapable of zero-shot generalization on that axis,

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.

independent of how much training data is available for other entities. The same structural constraint recurs whenever the entity distribution at deployment differs from the entity distribution at training time — a setting we study here through drug response prediction, where the goal is to predict a compound's potency against a cancer cell line, and where practical use (screening a compound against a patient-derived cell line never seen during training) requires the cell-line axis to generalize zero-shot to genomically novel entities.

Drug response prediction additionally exhibits a second structural property relevant beyond this application: extreme interaction sparsity coexisting with a requirement for fine-grained resolution. Large pharmacogenomic screens (Garnett et al., 2012; Yang et al., 2013; Barretina et al., 2012) provide tens of thousands of compound–cell-line measurements, but the median compound is tested against only one cell line, so pairwise interaction effects are almost unobservable directly and must be inferred from a memorization baseline plus a learned correction. Two architectural families address complementary halves of this problem in the literature and in our own setting: fingerprint-based memorization (ECFP + MLP) resolves activity cliffs — abrupt outcome changes from small structural perturbations — by hashing structurally similar inputs to shared parameters (Rogers & Hahn, 2010; Stumpfe & Bajorath, 2012; van Tilborg et al., 2022), but is by construction unable to generalize beyond its hashed support; graph neural networks learn adaptive, generalizable representations (Kipf & Welling, 2017; Gilmer et al., 2017) but over-smooth with depth, provably erasing the high-frequency structural information that activity-cliff resolution requires (Dablain et al., 2024). Neither failure mode is specific to chemistry: memorization-without-generalization and generalization-without-fine-resolution recur in any sparse interaction-prediction setting with both common and rare local structure.

We address the cold-start axis and the memorization/generalization tension jointly with a **Two-Tower Residual Late Fusion** architecture, presented as a general content-conditioned matrix-completion method and validated on drug response prediction:

1. A **content tower** combines a hard-memorization head (fine-grained resolution, no generalization) with learned soft representations (generalization, coarser resolution) via a residual sum rather than a weighted average or gated choice between them — a structural distinction with a concrete consequence developed in Section 3.
2. A **cold-start tower** encodes the entity requiring zero-shot generalization as a content-derived representation with *no identity dimension*, so any entity with the same content type can be encoded at inference regardless of training membership.
3. **Residual fusion** composes the two towers as $f_{\text{hard}} + \lambda(\mathbf{z}) \cdot f_{\text{residual}}$, where a learned, input-dependent gate $\lambda \in [0, 1]$ controls how much interaction correction the residual branch supplies. Because f_{hard} appears unconditionally in the sum and the residual branch can only add to it, the worst case ($\lambda \to 0$ on an entity for which the residual branch has learned nothing useful) reduces *exactly* to the memorization baseline rather than requiring the model to learn a cancellation of a harmful residual. The architecture thus admits a structural, not merely empirical, non-degradation property (Proposition 1).

We validate the architecture on the full set of 51 breast cancer cell lines in DepMap (Tsherniak et al., 2017) and externally on 601 cell lines across 27 tissue types with zero training overlap, and separately probe the encoder's learned representation with gradient-based attribution.

A predictor, however, is not only a scoring function: a sufficiently accurate one is also a candidate reward model for a downstream generative policy. This raises a distinct question from predictive accuracy — *specification gaming*, in which a policy optimized against a fixed objective satisfies that objective's literal specification without achieving the intended outcome (Krakovna et al., 2020), is a documented general failure mode of reward-optimizing systems (Amodei et al., 2016), and a frozen predictor repurposed as a reward model is exactly the setting in which it can arise: if the predictor has exploitable blind spots, a policy trained against it will find them, regardless of how well the predictor scores held-out data drawn from its training distribution. Section 5 reports a six-run investigation of this question, using the cell-line predictor developed in Sections 3–4 as a frozen reward model for a conditional generative policy over compound structures. The central empirical finding is not classic reward hacking against the frozen predictor, but a mechanistically diagnosable specification-gaming failure in an auxiliary diversity term, and its resolution.

2 Problem Formulation and Related Work

Content-conditioned matrix completion. The task is sparse matrix completion over a query × target matrix $\mathbf{M} \in \mathbb{R}^{|\mathcal{Q}|\times|\mathcal{T}|}$ (compounds × cell lines in our validation domain), $M_{ij} = \ln(\mathrm{IC}_{50})$. The training matrix is 4.71% dense (136,342 of $56{,}786 \times 51$ entries); the median number of targets observed per query is 1, so pairwise interaction effects are almost unobservable from co-occurrence alone. Both axes are parameterized by content rather than identity, and a rigorous evaluation must exceed the additive baseline $\hat{y}_{ij}^{\mathrm{add}} = \bar{\mu}_j + \bar{\mu}_i - \bar{\mu}$ (target mean, query mean, global mean): a model that fails to beat it has learned no query–target specificity beyond marginal effects.

Relation to cold-start recommendation. Content-based cold-start conditioning is well studied in recommendation, where identity-indexed collaborative-filtering embeddings are undefined for a new user or item and content features must substitute (Volkovs et al., 2017). Our setting differs in degree rather than kind: only one axis (cell lines) needs zero-shot generalization at deployment, while the other (compounds) is itself so sparse that a content-based, generalizable representation is also required for compounds seen only once in training — both axes are simultaneously cold-start-adjacent, which is what motivates a residual decomposition rather than a single shared content encoder.

Reward-model specification gaming in RL-guided generation. A separate line of work documents that policies trained against a fixed, imperfect proxy objective reliably find behaviors that satisfy the proxy without satisfying the intended goal (Amodei et al., 2016; Krakovna et al., 2020) — a failure mode distinct from, and orthogonal to, the proxy's held-out predictive accuracy. Section 5 tests this directly for a trained content-conditioned predictor repurposed as a frozen reward model.

3 Method: Two-Tower Residual Late Fusion

Content tower. A hard-memorization head, $f_{\mathrm{hard}}(\mathbf{x}) = \mathbf{W}_2\sigma(\mathbf{W}_1\mathbf{x} + \mathbf{b}_1) + \mathbf{b}_2$, acts on a fixed structural fingerprint of the query entity (ECFP (Rogers & Hahn, 2010) in our validation domain) as a regularized lookup table that resolves fine-grained interaction structure (Stumpfe & Bajorath, 2012) but is structurally blind to novel inputs outside its hashed support. A learned soft representation supplies the complementary generalization signal: a D-MPNN (Yang et al., 2019; Heid et al., 2024) graph encoder (edge-level message passing, generalizes but over-smooths with depth (Dablain et al., 2024)) and an octave-convolutional image encoder (Chen et al., 2019; He et al., 2016) operating on a rendered 2D depiction of the same query entity (splits features into high- and low-frequency sub-bands, preserving local structure alongside global topology) jointly supply this branch. Two drop-in alternative encoders (a pretrained transformer in place of the D-MPNN, a hybrid convolution+transformer in place of the octave encoder; Appendix A) are evaluated as part of model selection.

Cold-start tower. The zero-shot-generalizing entity (cell line) is encoded as a $3 \times 139 \times 139$ content image: Channel R is gene expression (19,177 genes), Channel G is expression × mutation severity, Channel B is reserved for a copy-number-variation signal that is zero-filled and unused in this version. A CNN encoder maps this image to $\mathbf{z} = f_{\mathrm{enc}}(\mathbf{I}) \in \mathbb{R}^d$ with *no identity dimension* of any kind, so any entity with a content image — including one never seen during training — can be encoded at inference. GroupNorm (Wu & He, 2018), weight decay, and a shallow architecture regularize against overfitting given an effective sample size, for this tower specifically, of 51 (distinct training cell lines), not 136,342 (records).

Residual late fusion. The joint prediction is

$$f = \underbrace{f_{\mathrm{hard}}(\mathbf{x} \oplus \mathbf{z})}_{\text{main effects}} + \underbrace{\lambda(\mathbf{z}) \cdot f_{\mathrm{residual}}\big(\mathrm{CrossAttn}(\mathbf{z}, [\mathbf{z}_{\mathrm{img}}, \mathbf{z}_{\mathrm{graph}}])\big)}_{\text{gated interaction correction}}, \tag{1}$$

where $\lambda(\mathbf{z}) \in [0, 1]$ is a learned gate over cross-attention (Vaswani et al., 2017).

Proposition 1 (Structural non-degradation)**.** *Because equation 1 is a* sum*, not a weighted average or a learned choice between f_{hard} and f_{residual}, f_{hard} appears in the output unconditionally and the*

Table 1: Summary statistics, 51-cell-line leave-target-out benchmark.

Metric	Count	%
Cell lines evaluated	51	100%
Residual > memorization-only	48	94.1%
Substantial (ΔR^2>0.02)	26	51.0%
Residual < memorization-only	3	5.9%
Mean ΔR^2 (all 51)	+0.016	
Mean ΔR^2 (48 winners)	+0.018	

residual branch can only add to it. If the residual branch is unhelpful on some input — e.g. a cold-start entity on which $\mathbf{z}_{\text{img}}, \mathbf{z}_{\text{graph}}$ carry no useful signal — the model needs only to learn $\lambda \to 0$ at that input to fall back exactly *to f_{hard}; it does not need to learn to cancel out or override a harmful residual value, because there is no architectural path for the residual term to dominate or replace the memorization term. Because $\lambda \equiv 0$ recovers f_{hard} exactly, the memorization-only baseline is contained in the residual model's hypothesis class, so the residual model's optimal empirical risk is no worse than the baseline's; any degradation on an unseen cold-start entity is therefore a generalization gap rather than a representational one (Vapnik, 1998).*

This is a structural, input-independent property, not an empirical observation about this particular training run; Section 4 reports the observed degradation rate.

The full architecture diagram is given in Appendix A (Figure 3).

Training and evaluation protocol. We minimize a Huber loss (Huber, 1964) plus ℓ_2 regularization. We report three splits: random (80/10/10, upper bound), leave-query-out (GroupKFold by compound), and leave-target-out (GroupKFold by cell line) — the decisive test of cold-start generalization, since a lookup-table baseline is undefined for a held-out target. We verified the absence of data leakage at the group, row, and content-image level prior to reporting results (Appendix A).

4 Results

4.1 Cold-Start Generalization: Leave-Target-Out

On the held-out random-split test set, six representative cell lines achieve $R^2 = 0.61$–0.69 (illustrative subset, not the full 51-line distribution). On leave-target-out — the decisive cold-start test — Table 1 shows Residual Fusion outperforms the hard-memorization-only baseline in 48/51 cell lines (94.1%), mean $\Delta R^2 = +0.016$ across all 51, with only 3 cell lines showing negative ΔR^2 (sign test on the 48/51 win rate: $p \approx 1 \times 10^{-11}$) — consistent with the structural property of Proposition 1. Extended per-cell-line results and ablations are in Appendix B.

4.2 Leave-Compound-Out Generalization

On the leave-compound-out split — testing whether the compound tower predicts IC_{50} for molecules structurally absent from training, rather than testing cold-start generalization on the cell-line axis — the model achieves $R^2 = 0.48$–0.58 across the same six representative cell lines, with mean $\Delta R^2 \approx +0.08$ over the additive baseline, confirming that the D-MPNN and image-encoder representations transfer to structurally novel compounds and are not merely fit to the training compound library.

4.3 Zero-Shot Generalization to Unseen Entities

We evaluated the trained model on an external dataset (CellTiter-Glo): 601/603 cell lines spanning 27 tissue types, with 0 cell-line overlap with training. Of the 197 external compounds, 88.8% were already present in training, so this evaluation isolates the cold-start (cell-line) axis specifically and holds compound identity largely fixed; it is not a test of generalization to structurally novel compounds, which the leave-compound-out split above addresses separately. The model achieves

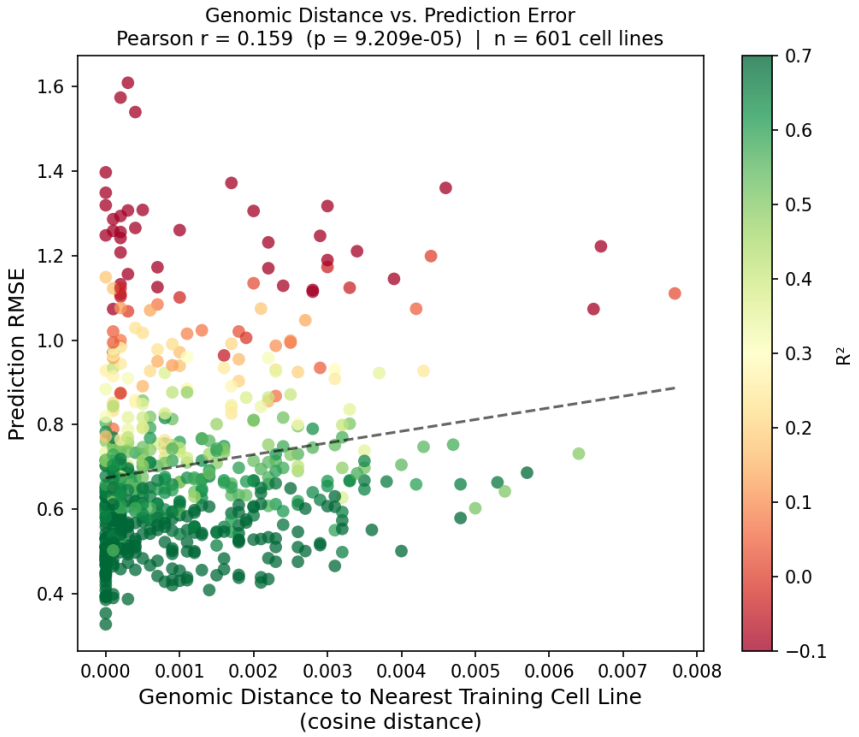

Figure 1: Embedding-space distance to nearest training entity vs. prediction RMSE, 601 externally held-out cell lines, zero training overlap. Pearson $r = 0.159$ ($p = 9.2 \times 10^{-5}$).

mean $R^2 = 0.515$, median $R^2 = 0.627$. Distance in the cold-start encoder's embedding space to the nearest training entity correlates positively and significantly with prediction error (Figure 1; Pearson $r = 0.159$, $p = 9.2 \times 10^{-5}$), and 23/23 tissue types with ≥ 2 cell lines achieve positive mean R^2 (breast highest, haematopoietic lowest) — evidence that the encoder's latent space carries directional, statistically significant reliability information about entities it has never seen, rather than treating all unseen entities as an undifferentiated, equally-trusted region.

4.4 What the Content Encoder Has Learned

Gradient-based attribution (GradCAM (Selvaraju et al., 2017)) on the cold-start tower recovers TP53 among the top-5 cross-cell-line genomic activators (5/51 cell lines) from interaction-prediction supervision alone, without any pathway annotation supplied at training time — partial evidence that the encoder's content representation carries biologically (not merely statistically) meaningful structure. However, TP53 does not clearly dominate: an uncharacterized gene, FSIP2, ranks marginally higher (6/51), and this driver-recovery signal is checkpoint-dependent (Appendix E); we report it as partial, not conclusive, evidence and discuss the caveat explicitly in Section 6.

5 Reward-Model-Guided Generation: A Specification-Gaming Investigation

We now ask the question motivated in Section 1: can the trained, frozen predictor from Sections 3–4 safely serve as a reward model for a downstream generative policy, or does the policy learn to exploit it? A conditional SMILES variational autoencoder (VAE) (Kingma & Welling, 2013; Gómez-Bombarelli et al., 2018), pretrained on the same compound corpus and conditioned on the target entity's content embedding $\mathbf{z}$, is fine-tuned with REINFORCE (Williams, 1992) against the frozen predictor as reward model, $R(\tau) = -\hat{y}(\tau)$ (Figure 2). Generation is autoregressive over SMILES

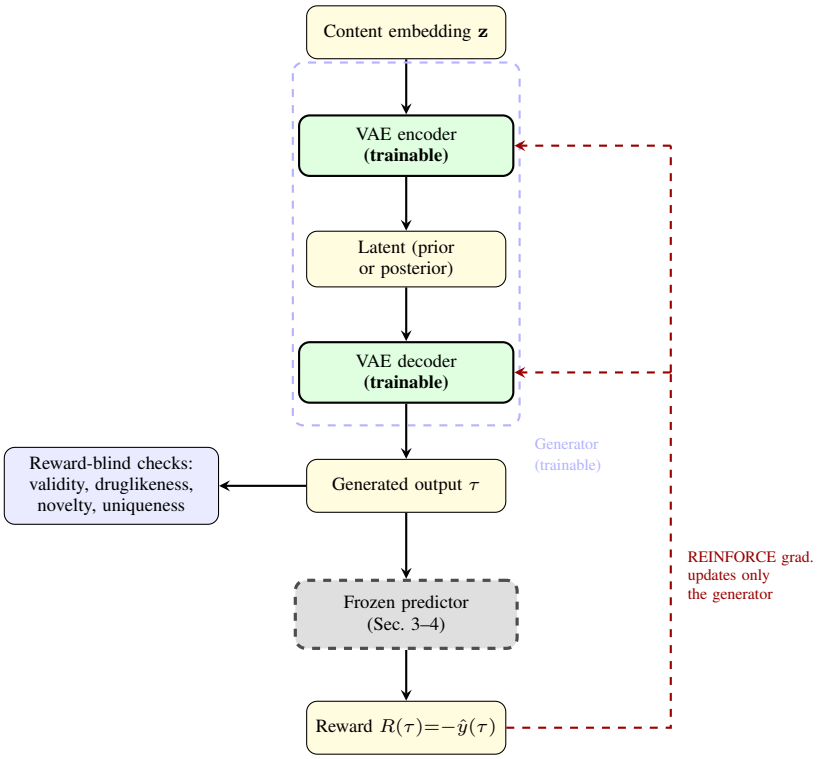

Figure 2: The reward-model-guided generation loop. The generator (encoder+decoder, green) is trained end-to-end across all six runs (v1–v5 single-step REINFORCE, v6 PPO-clipped batch reuse); the frozen predictor (grey, dashed) never receives a gradient update. Reward-blind checks are computed independently of the predictor and provide the reward-hacking test used throughout.

tokens, $\pi_\theta(\tau \mid \mathbf{z}) = \prod_{t=1}^{|\tau|} p_\theta(\tau_t \mid \tau_{<t}, \mathbf{z})$, with θ spanning *both* the VAE encoder (which maps $\mathbf{z}$ to the sampled prior/posterior latent) and the decoder (which autoregressively emits τ from that latent) — the two trainable blocks in Figure 2. Both are updated jointly by a single REINFORCE gradient with a batch-mean baseline b for variance reduction,

$$\nabla_\theta J(\theta) = \mathbb{E}_{\tau \sim \pi_\theta(\cdot|\mathbf{z})}\big[(R(\tau) - b)\, \nabla_\theta \log \pi_\theta(\tau \mid \mathbf{z})\big], \tag{2}$$

which is exactly the pair of dashed red arrows in Figure 2 routing the frozen predictor's reward back into the encoder and decoder; the frozen predictor itself receives no gradient. We score generated outputs both by the reward-model score — what the policy is directly optimized for — and by a battery of reward-model-*blind* checks the predictor cannot see or influence (validity, druglikeness (Lipinski et al., 2001), novelty, within-batch structural uniqueness); a widening gap between the two is the operational definition of reward hacking used throughout.

v1–v3: genuine reward improvement, but severe posterior-mode collapse. Across an initial three configurations, reward improved genuinely (validity and druglikeness rates remained at 97–100% throughout, so the predictor was not fooled by non-viable outputs), but posterior-mode structural uniqueness collapsed 7–8$\times$ after RL fine-tuning relative to a pre-RL baseline. We show this is not fixable by tuning the entropy bonus alone: for a per-step categorical policy $p = \mathrm{softmax}(z)$, the entropy-gradient $\partial H/\partial z_j = -p_j(\log p_j + H)$ provably vanishes as the policy saturates ($p_{j^*} \to 1$) — exactly the regime in which it is needed — and separately, the entropy bonus is capped at $\beta \ln |V|$ while the REINFORCE reward term is not (Appendix D). Empirically, even a $50\times$ larger entropy coefficient applied from the first epoch did not prevent collapse, falsifying the naïve prediction that early, large-coefficient entropy regularization alone preserves diversity.

v4: a repetition penalty, and the loophole exact-string matching leaves open. A batch-level penalty on exact-string duplicate count restores uniqueness 4–$5\times$, but a per-run "most-repeated output" diagnostic exposes a loophole invisible in the aggregate uniqueness number: the same two

Table 2: v4 vs. v5 head-to-head (largest similarity cluster's share of valid samples; lower is better).

Mode	v4	v5
prior	52.6–60.8%	14.8–24.3%
posterior	84.8–86.9%	12.1–20.3%

Table 3: v5 vs. v6 (two seeds), same evaluation protocol.

Mode	v5	v6 (seed 0)	v6 (seed 1)
prior reward	-3.55 to -3.38	-3.28 to -3.11	-3.29 to -3.12
posterior reward	-3.53 to -3.32	-3.28 to -3.14	-3.29 to -3.11
prior cluster share	$\approx 18\%$	$\approx 13.3\%$	$\approx 15.2\%$
posterior cluster share	$\approx 16\%$	$\approx 15.3\%$	$\approx 12.4\%$

outputs dominate the top-repeated slot across all 15 evaluated conditions, with close structural variants of one template filling most remaining top slots. Because the penalty matches only *exact* string equality, the policy satisfies the letter of the uniqueness reward by enumerating superficially distinct strings from one easy-to-decode template, rather than exploring genuinely different structures — specification gaming against the auxiliary diversity objective, distinct from and subtler than the mode collapse of v1–v3.

v5: similarity-clustered penalty, calibrated against the real failure. v5 replaces exact-string matching with fingerprint-similarity clustering, applying the same penalty form to cluster size rather than exact-match count. Rather than guessing a similarity threshold, we fingerprinted the actual dominant template family from v4 directly: its members' real pairwise similarities span 0.32–0.82, and a union-find merge in decreasing-similarity order shows the family fully connects into one cluster only after a 0.684-similarity merge. We set the clustering threshold at 0.65 — a small margin below this critical value, verified directly against the real family plus an unrelated control output — and separately introduce cold (least-drawn) sampling to correct for the highly skewed per-condition output pool. This reduces the largest cluster's share of valid samples from $\approx 86\%$ (v4, posterior mode) to $\approx 16\%$ (v5) (Table 2), reversing what a naive read of exact-string uniqueness alone would suggest, since exact-string uniqueness was measuring a property v4 was gaming rather than the property that actually matters.

v6: sample-efficient reuse via PPO-clipped importance sampling. v1–v5 share a structural inefficiency: single-step REINFORCE discards each collected batch after one gradient update, even though the expensive part of an outer step is collecting the batch (autoregressive decoding, one frozen-predictor call, structural checks), not the gradient step itself. v6 answers this with PPO's clipped, importance-sampled surrogate objective (Schulman et al., 2017; 2015), applied on top of v5's clustered penalty and cold sampling (both unchanged): for the same predictor-call budget, the decoder takes up to 4 clipped-surrogate gradient steps per collected batch instead of 1, so the comparison holds predictor calls fixed but not gradient steps. Across two independent training seeds, v6's reward-model score improves reproducibly over v5 in both sampling modes with non-overlapping ranges (Table 3), replicating within 0.006 across seeds, while largest-cluster share stays in the same $\approx$ 12–18% band as v5 — the efficiency gain is not bought at the cost of v5's diversity fix, nor of synthesizability, druglikeness, or novelty (Appendix D).

Summary. Across all six runs: (1) no instance of reward hacking against the frozen predictor itself was found — validity/druglikeness rates stayed $\geq 97\%$ throughout; (2) a real, controlled instance of specification gaming *was* found against the auxiliary diversity objective, diagnosed to a concrete mechanism at each stage (entropy-gradient vanishing under saturation; exact-string matching blind to structural similarity), and resolved with a calibrated fix; (3) the metric that initially suggested v5 was worse than v4 was itself an artifact of measuring the wrong property; and (4) reusing collected data more efficiently (v6) extracts a reproducible reward improvement from the identical predictor-call budget without eroding the diversity fix, replicated across seeds. Figure 12 (Appendix D) summarizes the full failure-mode taxonomy. These results establish that the RL fine-tuning process did not learn to exploit the frozen predictor mathematically or via degenerate outputs; they do *not*

establish that the predictor's accuracy is calibrated on the chemically novel region the policy is explicitly rewarded for exploring (mean nearest-neighbor similarity to the training compound library of ≈ 0.26–0.32) — a distinct, open limitation discussed in Section 6.

6 Discussion and Limitations

The residual decomposition matches the statistical structure of any sufficiently sparse interaction matrix: a memorization head captures main effects, and a gated residual captures interaction deviations, with $\lambda \to 0$ providing a structural, not merely empirical, fallback (Proposition 1) that explains why only 3/51 cold-start entities degrade at all. The leave-target-out and external zero-shot results together indicate that the content encoder learns transferable signal rather than memorizing identity — the decisive property for any content-conditioned cold-start method, independent of the specific validation domain used here.

On the generative side, the six-run investigation illustrates a general point about reward-model safety: the frozen predictor itself was not gamed, but an auxiliary term added on top of it (the diversity objective) was — twice, via two different mechanisms, at two different fix attempts. This suggests specification gaming risk in RL-guided generation is not a single property of "the reward model" to be checked once, but a property of every term in the effective objective, including terms added specifically to prevent an already-diagnosed failure mode.

Key limitations, detailed in Appendix E, include: (1) *training entity diversity* — the cold-start tower is trained on only 51 distinct entities of one type (breast cancer cell lines), and this pool's limited diversity may itself inflate apparent leave-target-out generalization independent of any data leakage, which we separately verified absent; (2) *content encoder architecture* — the CNN operates on raw content without exploiting known relational structure among content dimensions (e.g. gene regulatory network structure), motivating a structure-aware encoder as future work; (3) *reward-model extrapolation* — the RL-fine-tuned generator's reward-hacking checks are a coarse filter on outputs that satisfy them, and the frozen predictor's calibration on the chemically novel region the generator is rewarded for exploring has not been independently validated, unlike the cold-start axis, where distance-vs-error calibration was directly measured (Figure 1); and (4) *attribution checkpoint-dependence* — the content encoder's gradient-based attribution ranking is not stable across training checkpoints, so it should not be over-interpreted as a fixed architectural property.

7 Conclusion

We presented a Two-Tower Residual Late Fusion architecture for content-conditioned sparse matrix completion, with a structural non-degradation guarantee that follows from composing a memorization branch and a gated residual branch as a sum rather than a weighted choice. Validated on drug response prediction, it improves over a memorization-only baseline in 48/51 cold-start entities (94.1%) and generalizes zero-shot to 601 externally held-out entities across 27 tissue types (median $R^2 = 0.627$), with error correlating significantly with embedding-space distance to the nearest training entity. We then examined whether a frozen instance of this predictor can serve as a reward model for RL-guided generation: across six runs, no reward hacking against the frozen predictor itself was found, but a real specification-gaming failure in an auxiliary diversity objective was diagnosed to a concrete mechanism and fixed, and a sample-efficiency variant extracted a reproducible, cross-seed-validated reward improvement from an identical predictor-call budget without eroding that fix. The residual decomposition principle — memorize what is known, learn the deviation, and structurally bound the cost of the learned part being wrong — is a design pattern applicable to any sparse interaction-prediction setting where a cold-start axis coexists with a requirement for fine-grained resolution on the other axis.

AI use statement

Generative AI tools (e.g., Claude by Anthropic) were used to assist with LaTeX manuscript preparation, prose editing, figure/table formatting, and retrieval and discovery of related work for this submission, and with earlier interpretability-visualization development for the underlying project. All experimental results, model training, and quantitative findings reported in this paper were produced by the authors' own code and pipelines, not generated by an AI tool. All AI-assisted text

and formatting were reviewed and verified by the authors, including a dedicated fact-checking pass cross-referencing every numerical claim in this paper against the authors' own experiment logs and an independently maintained full working draft. The authors take full responsibility for the final content of this work, including all text, claims, and figures.

Ethics statement

This work uses only publicly available, de-identified cell-line pharmacogenomic data (DepMap/GDSC2, CellTiter-Glo) and involves no human subjects research, no patient-identifiable data, and no new data collection from human or animal subjects; no IRB approval was required. The generative extension (Section 5) produces novel small-molecule structures; we do not release, and the reward-model-blind checks in Section 5 do not screen for, synthesis-route feasibility or any assessment of dual-use potential, and we do not claim the generated compounds are safe, efficacious, or synthesizable without independent wet-lab validation.

Reproducibility statement

Section 3 specifies the full model architecture (Eq. 1) and training protocol; Appendix A specifies the data-leakage validation procedure (group, row, and content-image-level checks) applied prior to reporting any result. Section 5 and Appendix D specify the exact reward function, penalty forms, and hyperparameters (similarity threshold, cold-sampling power, PPO clip range) used in each of the six generative runs, including how the similarity threshold was calibrated against directly measured pairwise similarities of the real failure case rather than chosen a priori. Code, trained checkpoints, and the RL evaluation pipeline will be released at a public repository upon acceptance; the link is withheld during double-blind review.

References

Dario Amodei, Chris Olah, Jacob Steinhardt, Paul Christiano, John Schulman, and Dan Mané. Concrete problems in ai safety. *arXiv preprint arXiv:1606.06565*, 2016.

Jonathan B Baell and Georgina A Holloway. New substructure filters for removal of pan assay interference compounds (pains) from screening libraries and for their exclusion in bioassays. *Journal of Medicinal Chemistry*, 53(7):2719–2740, 2010.

Jordi Barretina, Giordano Caponigro, Nicolas Stransky, Kavitha Venkatesan, Adam A Margolin, Sungjoon Kim, Cameron J Wilson, Joseph Lehár, Gregory V Kryukov, Dmitriy Sonkin, et al. The cancer cell line encyclopedia enables predictive modelling of anticancer drug sensitivity. *Nature*, 483(7391):603–607, 2012. doi: 10.1038/nature11003.

Yunpeng Chen, Haoqi Fan, Bing Xu, Zhicheng Yan, Yannis Kalantidis, Marcus Rohrbach, Shuicheng Yan, and Jiashi Feng. Drop an octave: Reducing spatial redundancy in convolutional neural networks with octave convolution. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 3435–3444, 2019. doi: 10.1109/ICCV.2019.00353.

Seyone Chithrananda, Gabriel Grand, and Bharath Ramsundar. ChemBERTa: Large-scale self-supervised pretraining for molecular property prediction. *arXiv preprint arXiv:2010.09885*, 2020.

Haotian Cui, Chloe Wang, Hassaan Maan, Kuan Pang, Fengning Luo, Nan Duan, and Bo Wang. scGPT: Toward building a foundation model for single-cell multi-omics using generative AI. *Nature Methods*, 21:1470–1480, 2024. doi: 10.1038/s41592-024-02201-0.

Damien Dablain, Igor Krawczuk, Andras Benczur, and Gustavo Batista. Activity cliff prediction: Deep learning meets molecular graph representations. *Journal of Chemical Information and Modeling*, 64(5):1486–1499, 2024. doi: 10.1021/acs.jcim.3c01505.

Peter Ertl and Ansgar Schuffenhauer. Estimation of synthetic accessibility score of drug-like molecules based on molecular complexity and fragment contributions. *Journal of Cheminformatics*, 1(1):8, 2009.

Mathew J Garnett, Elena J Edelman, Sonja J Heidorn, Chris D Greenman, Anahita Dastur, King Wai Lau, Patricia Greninger, I Richard Thompson, Xi Luo, Jorge Soares, et al. Systematic identification of genomic markers of drug sensitivity in cancer cells. *Nature*, 483(7391):570–575, 2012. doi: 10.1038/nature11005.

Justin Gilmer, Samuel S. Schoenholz, Patrick F. Riley, Oriol Vinyals, and George E. Dahl. Neural message passing for quantum chemistry. In *Proceedings of the 34th International Conference on Machine Learning*, pp. 1263–1272, 2017.

Rafael Gómez-Bombarelli, Jennifer N Wei, David Duvenaud, José Miguel Hernández-Lobato, Benjamín Sánchez-Lengeling, Dennis Sheberla, Jorge Aguilera-Iparraguirre, Timothy D Hirzel, Ryan P Adams, and Alán Aspuru-Guzik. Automatic chemical design using a data-driven continuous representation of molecules. *ACS Central Science*, 4(2):268–276, 2018. doi: 10.1021/acscentsci.7b00572.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 770–778, 2016. doi: 10.1109/CVPR.2016.90.

Esther Heid, Kevin P Greenman, Yunsie Chung, Shih-Cheng Li, David E Graff, Florence H Vermeire, Haoyang Wu, William H Green, and Charles J McGill. Chemprop: A machine learning package for chemical property prediction. *Journal of Chemical Information and Modeling*, 64(1): 9–17, 2024. doi: 10.1021/acs.jcim.3c01250.

Peter J Huber. Robust estimation of a location parameter. *The Annals of Mathematical Statistics*, 35 (1):73–101, 1964. doi: 10.1214/aoms/1177703732.

Diederik P Kingma and Max Welling. Auto-encoding variational Bayes. *arXiv preprint arXiv:1312.6114*, 2013.

Thomas N Kipf and Max Welling. Semi-supervised classification with graph convolutional networks. *International Conference on Learning Representations*, 2017.

Victoria Krakovna, Jonathan Uesato, Vladimir Mikulik, Matthew Rahtz, Tom Everitt, Ramana Kumar, Zac Kenton, Jan Leike, and Shane Legg. Specification gaming: the flip side of ai ingenuity. DeepMind Blog, 2020.

Christopher A Lipinski, Franco Lombardo, Beryl W Dominy, and Paul J Feeney. Experimental and computational approaches to estimate solubility and permeability in drug discovery and development settings. *Advanced Drug Delivery Reviews*, 46(1–3):3–26, 2001. doi: 10.1016/S0169-409X(00)00129-0.

Francisco Martínez-Jiménez, Ferran Muiños, Inés Sentís, Jordi Deu-Pons, et al. A compendium of mutational cancer driver genes. *Nature Reviews Cancer*, 20(10):555–572, 2020. doi: 10.1038/s41568-020-0290-x.

Leland McInnes, John Healy, and James Melville. Umap: Uniform manifold approximation and projection for dimension reduction. *arXiv preprint arXiv:1802.03426*, 2018.

David Rogers and Mathew Hahn. Extended-connectivity fingerprints. *Journal of Chemical Information and Modeling*, 50(5):742–754, 2010. doi: 10.1021/ci100050t.

John Schulman, Sergey Levine, Pieter Abbeel, Michael Jordan, and Philipp Moritz. Trust region policy optimization. In *Proceedings of the 32nd International Conference on Machine Learning*, pp. 1889–1897, 2015.

John Schulman, Filip Wolski, Prafulla Dhariwal, Alec Radford, and Oleg Klimov. Proximal policy optimization algorithms. *arXiv preprint arXiv:1707.06347*, 2017.

Ramprasaad R Selvaraju, Michael Cogswell, Abhishek Das, Ramakrishna Vedantam, Devi Parikh, and Dhruv Batra. Grad-CAM: Visual explanations from deep networks via gradient-based localization. In *Proceedings of the IEEE International Conference on Computer Vision*, pp. 618–626, 2017. doi: 10.1109/ICCV.2017.74.

Dagmar Stumpfe and Jürgen Bajorath. Exploring activity cliffs in medicinal chemistry. *Journal of Medicinal Chemistry*, 55(7):2932–2942, 2012. doi: 10.1021/jm201706b.

The Cancer Genome Atlas Network. Comprehensive molecular portraits of human breast tumours. *Nature*, 490(7418):61–70, 2012. doi: 10.1038/nature11412.

Christina V. Theodoris, Ling Xiao, Anant Chopra, Mark D. Chaffin, Zeina R. Al Sayed, Matthew C. Hill, Helene Mantineo, Elizabeth M. Brydon, Zexian Zeng, X. Shirley Liu, and Patrick T. Ellinor. Transfer learning enables predictions in network biology. *Nature*, 618:616–624, 2023. doi: 10.1038/s41586-023-06139-9.

Aviad Tsherniak, Francisca Vazquez, Phil G Montgomery, Barbara A Weir, Gregory Kryukov, Glenn S Cowley, Stanley Gill, William F Harrington, Sandy Pantel, John M Krill-Burger, et al. Defining a cancer dependency map. *Cell*, 170(3):564–576, 2017. doi: 10.1016/j.cell.2017.06.010.

Derek van Tilborg, Alisa Alenicheva, and Francesca Grisoni. Exposing the limitations of molecular machine learning with activity cliffs. *Journal of Chemical Information and Modeling*, 62(23): 5938–5951, 2022. doi: 10.1021/acs.jcim.2c01073.

Vladimir N Vapnik. *Statistical Learning Theory*. Wiley-Interscience, New York, 1998.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in Neural Information Processing Systems*, 30, 2017.

Maksims Volkovs, Guangwei Yu, and Tomi Poutanen. Dropoutnet: Addressing cold start in recommender systems. In *Advances in Neural Information Processing Systems*, volume 30, 2017.

Ronald J Williams. Simple statistical gradient-following algorithms for connectionist reinforcement learning. *Machine Learning*, 8(3):229–256, 1992. doi: 10.1007/BF00992696.

Yuxin Wu and Kaiming He. Group normalization. *Proceedings of the European Conference on Computer Vision*, pp. 3–19, 2018. doi: 10.1007/978-3-030-01261-8_1.

Kevin Yang, Kyle Swanson, Wengong Jin, Connor Coley, Philipp Eiden, Hua Gao, Angel Guzman-Perez, Timothy Hopper, Brian Kelley, Miriam Mathea, et al. Analyzing learned molecular representations for property prediction. *Journal of Chemical Information and Modeling*, 59(8):3370–3388, 2019. doi: 10.1021/acs.jcim.9b00237.

Wanjuan Yang, Jorge Soares, Patricia Greninger, Elena J Edelman, Howard Lightfoot, Simon Forbes, Nidhi Bindal, Dave Beare, James A Smith, I Richard Thompson, et al. Genomics of drug sensitivity in cancer (GDSC): a resource for therapeutic biomarker discovery in cancer cells. *Nucleic Acids Research*, 41(D1):D955–D961, 2013. doi: 10.1093/nar/gks1111.

A Extended Method Details, Alternative Encoders, and Data-Leakage Validation

Alternative soft-branch encoders. The graph encoder and image encoder are the default choices for the two soft-branch slots in equation 1's cross-attention key/value set; each has a drop-in alternative evaluated as part of model selection. A pretrained chemical-language-model encoder (ChemBERTa (Chithrananda et al., 2020)) substitutes for the graph encoder: inputs are tokenized and passed through the pretrained transformer, fine-tuned end-to-end at a lower learning rate than the fusion head to preserve pretrained priors. A hybrid encoder substitutes for the octave image encoder: the octave convolutional branches are retained, and a parallel transformer stream patchifies the same rendered image and is trained from scratch at the same learning rate as the rest of the model. Formally, the two soft-branch representations feeding equation 1 are each drawn from a two-element set,

$$\mathbf{z}_{\text{graph}} \in \left\{ f_{\text{DMPNN}}(\mathbf{g}),\ \ f_{\text{ChemBERTa}}(\text{tok}(s)) \right\}, \tag{3}$$

$$\mathbf{z}_{\text{img}} \in \left\{ f_{\text{octave}}(\mathbf{x}_{\text{2D}}),\ \ f_{\text{octave}}(\mathbf{x}_{\text{2D}}) \oplus f_{\text{ViT}}(\text{patch}(\mathbf{x}_{\text{2D}})) \right\}, \tag{4}$$

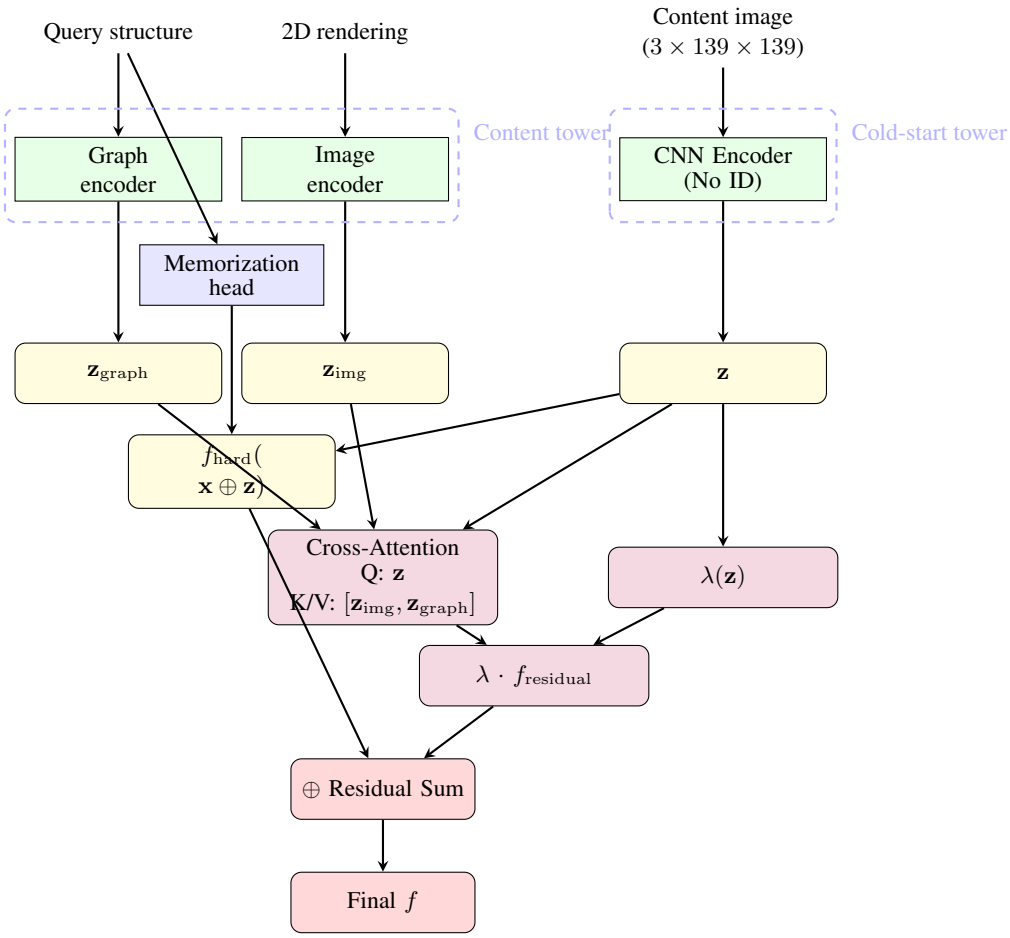

Figure 3: Two-Tower Residual Late Fusion architecture. Content tower (left) combines graph, image, and memorization-head representations. Cold-start tower (right) encodes content images without identity. Cross-attention produces a gated residual correction added to the hard-memorization baseline.

where $\mathbf{g}$ is the query compound's molecular graph, s its SMILES string, $\mathbf{x}_{2\text{D}}$ its rendered 2D depiction, and $\oplus$ concatenation. The 51-cell-line benchmark trains all four combinations per cell line and selects, independently per cell line c, the pair maximizing held-out validation R^2:

$$(k^*_{\text{graph}}, k^*_{\text{img}})(c) = \underset{(k_1,k_2) \in \{\text{DMPNN, ChemBERTa}\} \times \{\text{octave, hybrid}\}}{\arg\max} R^2_{\text{val}}(c;\, k_1, k_2). \tag{5}$$

Data leakage validation. We verified the absence of data leakage at three levels prior to reporting results. *Group exclusivity*: writing $\mathcal{G}_{\text{train}}, \mathcal{G}_{\text{val}}, \mathcal{G}_{\text{test}}$ for the group-identifier sets assigned to each partition (cell-line IDs for leave-target-out, compound IDs for leave-query-out),

$$\mathcal{G}_{\text{train}} \cap \mathcal{G}_{\text{val}} \;=\; \mathcal{G}_{\text{train}} \cap \mathcal{G}_{\text{test}} \;=\; \mathcal{G}_{\text{val}} \cap \mathcal{G}_{\text{test}} \;=\; \emptyset, \tag{6}$$

verified for both the leave-target-out split (17/16/18 cell lines, 0 overlap) and leave-query-out split (45,427/5,679/5,679 compounds, 0 overlap). *Row-level duplication*: writing each record as $r = (q, t) \in \mathcal{Q} \times \mathcal{T}$ (Section 2) and $\pi(r)$ for its assigned partition,

$$\forall\, r \in \mathcal{R} : \; \big|\{\, s \in \{\text{train}, \text{val}, \text{test}\} \;:\; \pi(r) = s \,\}\big| = 1, \tag{7}$$

i.e. no (compound, cell-line) pair appears in more than one partition. *Content-image uniqueness*: for the 51 cell-line content images $\{\mathbf{I}_c\}_{c=1}^{51}$,

$$\text{MD5}(\mathbf{I}_i) \neq \text{MD5}(\mathbf{I}_j) \quad \text{and} \quad \text{MSE}(\mathbf{I}_i, \mathbf{I}_j) = \tfrac{1}{3 \cdot 139^2} \|\mathbf{I}_i - \mathbf{I}_j\|_2^2 \;>\; 0 \qquad \forall\, i \neq j, \tag{8}$$

Table 4: Top-10 cell lines by Residual Fusion advantage (ΔR^2). All in the 100–500-record regime.

Cell Line	n	Memorization R^2	Best R^2	ΔR^2
JIMT1	312	0.017	0.065	+0.048
HCC70	402	0.185	0.234	+0.048
HCC1143	358	0.146	0.189	+0.043
BT20	487	0.243	0.285	+0.042
MDAMB415	402	0.290	0.331	+0.041
HCC1937	451	0.243	0.281	+0.038
HCC1395	360	0.323	0.360	+0.037
MDAMB157	392	0.222	0.258	+0.036
MDAMB436	364	0.039	0.073	+0.034
HDQP1	335	0.068	0.100	+0.032

verified across all $\binom{51}{2} = 1275$ pairs (min MSE 113.8, median 498.9), ruling out near-identical images as a trivial explanation for cross-entity generalization.

B Extended Results

For cell line c, the residual-fusion advantage tabulated throughout this paper is

$$\Delta R^2(c) = R^2_{\text{Residual}}(c) - R^2_{\text{memorization}}(c), \tag{9}$$

with R^2_{Residual} the best held-out validation R^2 over the model-selection set of equation 5 and $R^2_{\text{memorization}}$ the hard-memorization-only baseline (f_{hard} alone, $\lambda \equiv 0$).

The ChemBERTa-based soft-branch variant is the best configuration for all 10 of the top-10 cell lines by residual advantage, and together with the hybrid image-encoder+ChemBERTa combination accounts for 39/51 cell lines (76.5%) across the full benchmark, suggesting the pretrained-transformer branch supplies complementary signal to memorization particularly where residual fusion helps most.

Octave-ratio ablation. Following the octave-convolution channel split (Chen et al., 2019) used in the default image encoder (Section 3), input channels are partitioned into low- and high-frequency groups governed by ratio $\alpha \in [0, 1]$:

$$C_L = \lfloor \alpha C \rfloor, \qquad C_H = C - C_L, \tag{10}$$

with the low-frequency group $\mathbf{x}_L \in \mathbb{R}^{C_L \times H/2 \times W/2}$ processed at half spatial resolution and the high-frequency group $\mathbf{x}_H \in \mathbb{R}^{C_H \times H \times W}$ at full resolution ($\alpha = 0$ recovers a standard convolution). Ablations across $\alpha \in \{0.25, 0.50, 0.75, 0.90\}$ on two cell lines spanning distinct data regimes (HCC1937, BRCA1-mutant TNBC, sparser regime; SKBR3, HER2-amplified, denser regime) show residual models exceed the memorization-only baseline at most, but not all, settings tested: at HCC1937's largest ratio ($\alpha = 0.9$), both residual variants fall below the baseline, while $\alpha = 0.75$ yields the largest gain observed ($\Delta R^2 = +0.019$, D-MPNN variant). $\alpha = 0.75$ is the most consistently reasonable choice across the two cell lines tested, though its margin over $\alpha = 0.25$ at HCC1937 (0.001 ΔR^2 difference between the two configurations' best settings) is well within that cell line's own fold-level noise, $\sigma = \sqrt{\frac{1}{K-1}\sum_{k=1}^{K}(R^2_k - \bar{R}^2)^2} \approx 0.27$–$0.31$ across K cross-validation folds. Residual gains scale inversely with memorization-head saturation: largest in the sparser regime (HCC1937, up to $+0.019$), negligible in the denser regime (SKBR3, $\leq +0.002$, where ECFP memorization is already near-saturated).

Compound-level interpretability. Three query compounds spanning a >130-fold potency range were selected for dual-panel visualization: atom-importance via gradient backpropagation through the memorization head's input bits, and GradCAM on the image encoder's high-frequency branch. Across all three, the two independent attribution pathways identify consistent structural loci, and predicted potency tracks observed potency with errors of -9%, $+6\%$, and -24% respectively.

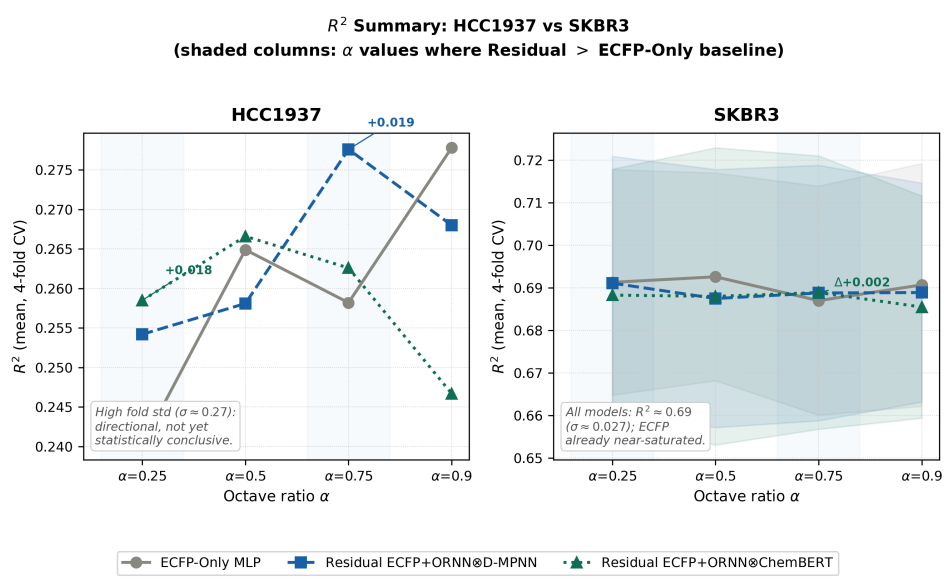

Figure 4: Alpha ablation profiles across $\alpha \in \{0.25, 0.50, 0.75, 0.90\}$: mean R^2 and RMSE ± 1 fold-std for three model configurations on two representative cell lines.

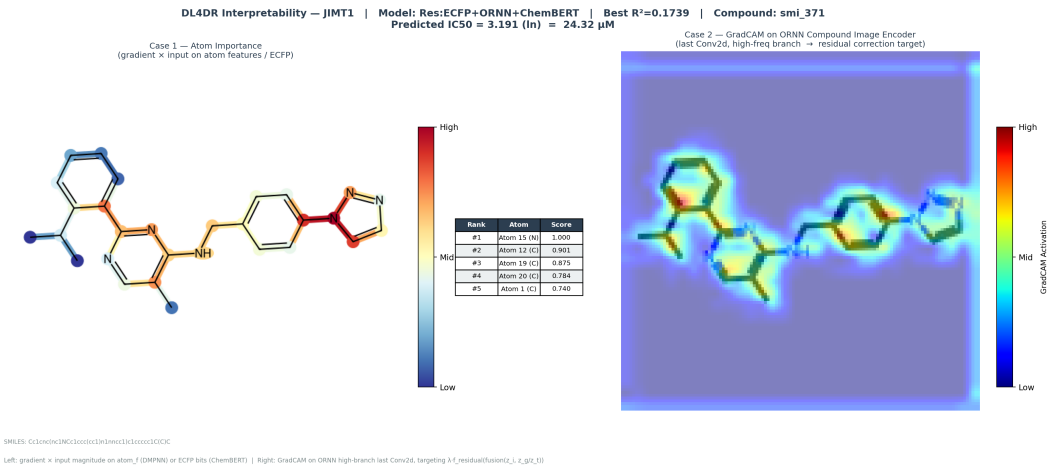

Figure 5: Dual-panel interpretability for a high-activity query compound (observed potency 26.6 μM, predicted 24.3 μM). Left: atom-importance heatmap. Right: GradCAM on the image encoder's high-frequency branch.

Cold-start-entity-level case studies. GradCAM case studies at two individual entities illustrate curated-driver recovery: TP53 ranks 8th of 15 top candidate genes at one cell line, and BRCA2 ranks 3rd of 15 at another, behind two novel candidates. These complement the aggregate driver-frequency analysis: each entity's content embedding foregrounds a distinct combination of curated and novel candidates.

C Extended Zero-Shot Validation

For entity c with true and predicted responses $\{y_i\}_{i=1}^{n_c}, \{\hat{y}_i\}_{i=1}^{n_c}$,

$$R^2(c) = 1 - \frac{\sum_i (y_i - \hat{y}_i)^2}{\sum_i (y_i - \bar{y})^2}, \tag{11}$$

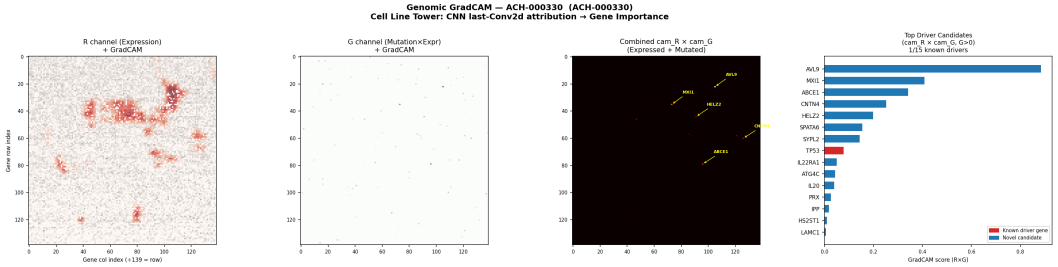

Figure 6: Content-encoder GradCAM case study. TP53 (red) ranks 8th of 15 candidate genes by combined attribution score.

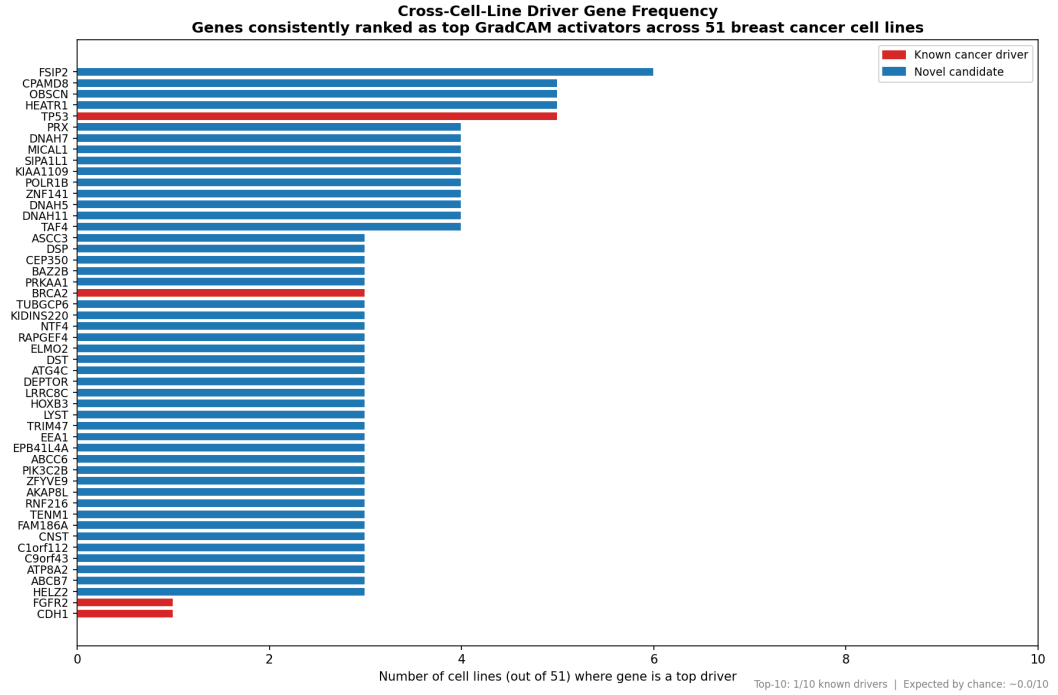

Figure 7: Cross-entity driver-gene frequency from cold-start-tower GradCAM (51 cell lines). Red: curated breast-cancer drivers (IntOGen/TCGA/COSMIC, (Martínez-Jiménez et al., 2020; The Cancer Genome Atlas Network, 2012)); blue: novel candidates. FSIP2 leads (6/51), followed by CPAMD8/OBSCN/HEATR1 (5 each) and TP53 (5/51).

the per-entity metric summarized in Figures 8–9. Writing $d(c) = \min_{c' \in \mathcal{C}_{\text{train}}} \|\mathbf{z}_c - \mathbf{z}_{c'}\|_2$ for embedding-space distance from held-out entity c to its nearest training entity, the correlation reported in Figure 1 is the Pearson coefficient between $d(c)$ and per-entity RMSE across the 601 held-out entities,

$$r = \frac{\sum_c (d(c) - \bar{d})\big(\mathrm{RMSE}(c) - \overline{\mathrm{RMSE}}\big)}{\sqrt{\sum_c (d(c) - \bar{d})^2}\,\sqrt{\sum_c \big(\mathrm{RMSE}(c) - \overline{\mathrm{RMSE}}\big)^2}} = 0.159 \quad (p = 9.2 \times 10^{-5}). \tag{12}$$

D Extended Generative-Extension Results

Why entropy regularization alone cannot fix collapse. For a per-step categorical policy $p = \mathrm{softmax}(z)$, differentiating the entropy bonus $H(p) = -\sum_a p_a \log p_a$ through the softmax Jacobian gives $\partial H / \partial z_j = -p_j(\log p_j + H)$. As the policy saturates ($p_{j^*} \to 1$), every term vanishes:

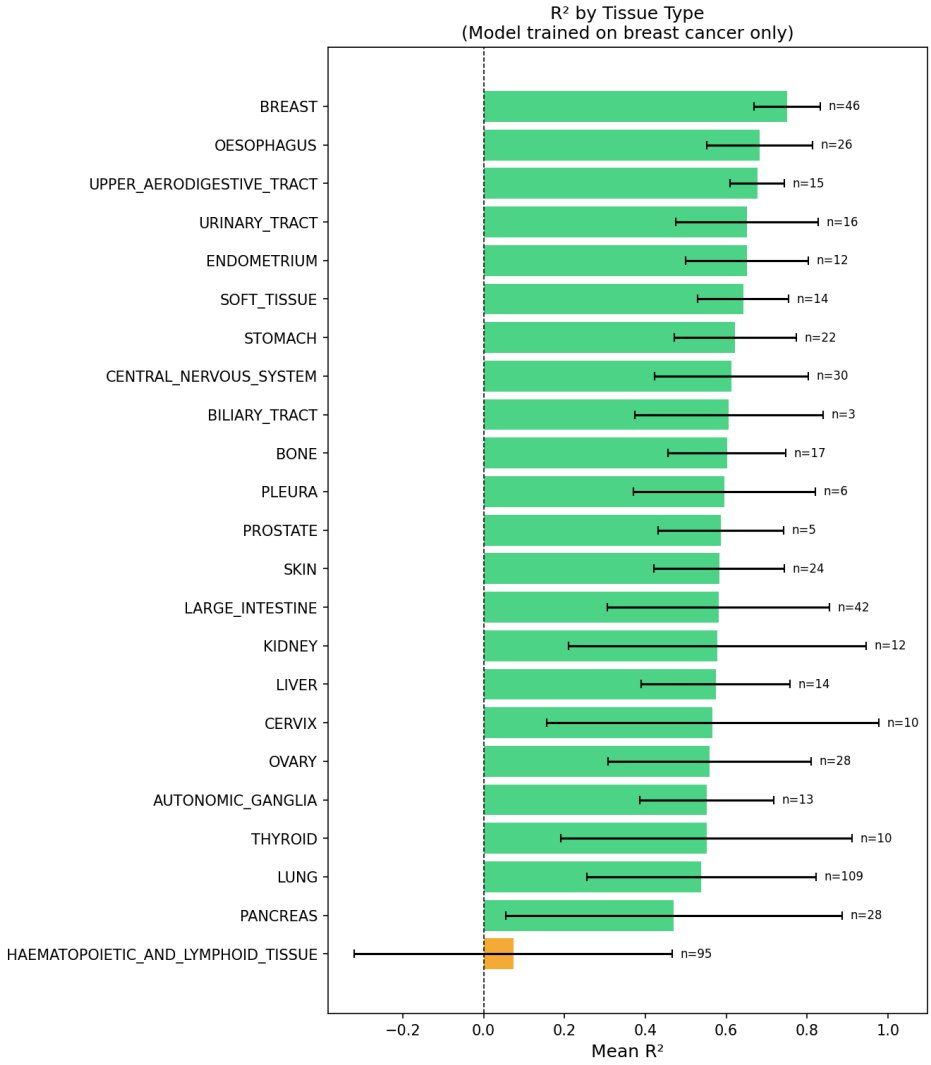

Figure 8: Mean R^2 by tissue type across 601 externally held-out cell lines. 23/23 tissue types with ≥ 2 cell lines achieve positive mean R^2; performance degrades with embedding-space distance from the breast-cancer training distribution, lowest for haematopoietic/lymphoid tissue (mean $R^2 = 0.073$).

Table 5: Pre-RL vs. post-RL posterior-mode diversity, 3 cell lines, 500 samples/line.

Metric	Pre-RL	v1	v2
Reward score	-4.51 to -3.54	-3.90 to -3.70	-3.24 to -3.14
Uniqueness	45.8–53.9%	6.3–7.8%	5.1–7.0%

the entropy gradient disappears exactly where it is most needed. Separately, the entropy bonus is capped at $\beta \ln |V|$ while the REINFORCE reward term is not; empirically, even a $50\times$ larger entropy coefficient applied from epoch 1 did not prevent collapse.

Formally, the v4 penalty subtracts a per-batch exact-match count from the frozen-predictor reward for output τ_i in batch $\{\tau_j\}$,

$$R_{\text{v4}}(\tau_i) = -\hat{y}(\tau_i) \; - \; \gamma \cdot \big|\{\, j \neq i : \tau_j = \tau_i \,\}\big|, \tag{13}$$

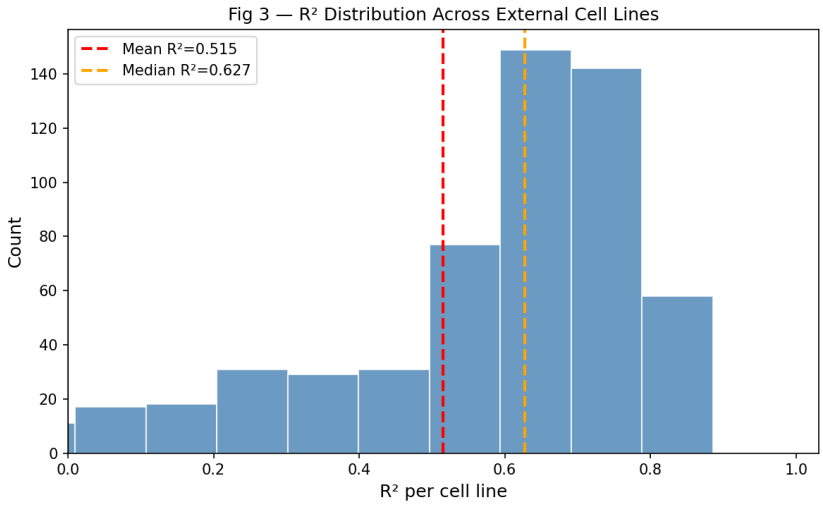

Figure 9: Distribution of per-entity R^2 across 601 externally held-out cell lines. Mean $R^2 = 0.515$, median $R^2 = 0.627$; right-skewed with a low/negative tail predominantly among haematopoietic cell lines.

Table 6: Synthesizability and reward-blind checks, v5 vs. v6 (both seeds).

Config	Mode	PAINS hit	Mean SA
v5	prior	0.07%	3.00
v5	posterior	0.00%	2.98
v6 seed 0	prior	0.07%	2.86
v6 seed 0	posterior	0.00%	2.87
v6 seed 1	prior	0.00%	2.88
v6 seed 1	posterior	0.00%	2.85

which the exact-string loophole (Section 5) satisfies by enumerating superficially distinct strings from one template. v5 replaces the exact-match count with the size of τ_i's similarity cluster,

$$R_{\text{v5}}(\tau_i) = -\hat{y}(\tau_i) \;-\; \gamma \cdot \big|\text{cluster}_\theta(\tau_i)\big|, \qquad \text{cluster}_\theta(\tau_i) = \{\, \tau_j : \tau_i \sim_\theta \tau_j \,\}, \tag{14}$$

where $\sim_\theta$ is the transitive closure, under union-find merging, of pairwise fingerprint similarity $\text{sim}(\tau_i, \tau_j) \geq \theta$ at the calibrated threshold $\theta = 0.65$ (Section 5). v4's higher raw exact-string uniqueness (28–32%) reflects only that one dominant template's members were distinct strings under equation 13; cluster_θ concentration under equation 14 reveals v4's posterior output was $\approx 86\%$ concentrated in a single similarity cluster, while v5's largest cluster never exceeds $\approx 24\%$ of any batch.

SA scores (Ertl & Schuffenhauer, 2009) cluster tightly between 2 and 4 across all six configurations, PAINS (Baell & Holloway, 2010) hit rate is effectively zero everywhere, and v5/v6 are statistically indistinguishable on novelty, druglikeness, PAINS, and mean SA — v6's reward improvement is not achieved by trading away synthesizability, druglikeness, or novelty.

Sample-efficient reuse via clipped importance sampling. v6 replaces the single-step REINFORCE update of equation 2 (used unchanged in v1–v5) with up to $m = 4$ clipped-surrogate gradient steps per collected batch (Schulman et al., 2017), reusing the same batch of $(\tau, \mathbf{z})$ pairs instead of discarding it after one update:

$$\mathcal{L}_{\text{PPO}}(\theta) = \mathbb{E}_\tau\Big[\min\big(\rho(\theta)\,A(\tau),\;\; \text{clip}(\rho(\theta),\,1-\epsilon,\,1+\epsilon)\,A(\tau)\big)\Big], \qquad \rho(\theta) = \frac{\pi_\theta(\tau \mid \mathbf{z})}{\pi_{\theta_{\text{old}}}(\tau \mid \mathbf{z})}, \tag{15}$$

with advantage $A(\tau) = R_{\text{v5}}(\tau) - b$ (equation 14, batch-mean baseline b) and π_θ the same autoregressive decoder likelihood as equation 2; θ_{old} denotes the parameters at batch-collection time, so

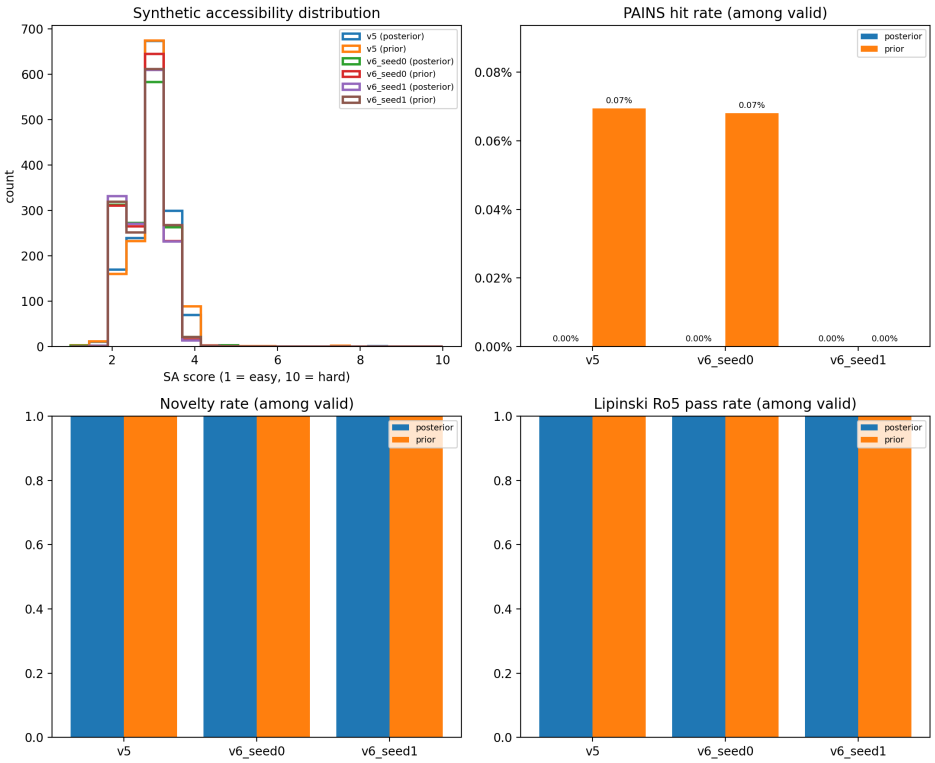

Figure 10: Reward-blind synthesizability checks across v5 and both v6 seeds, prior and posterior modes: SA-score distribution, PAINS hit rate, novelty rate, druglikeness pass rate.

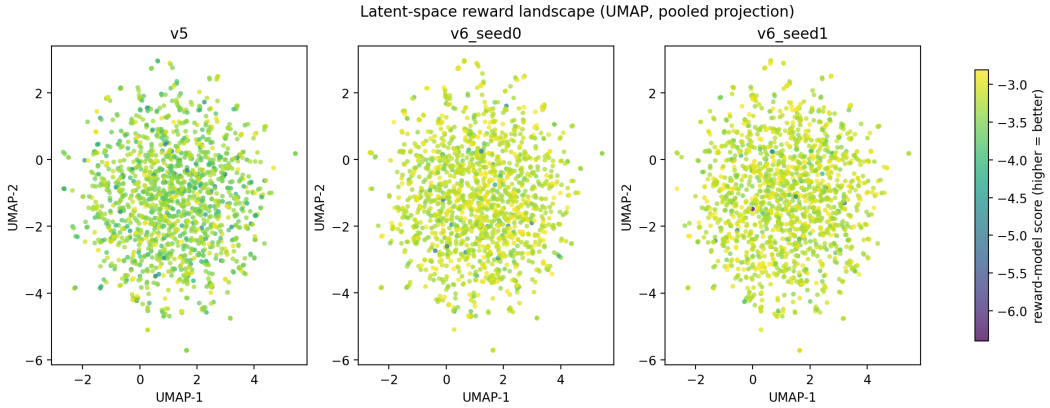

Figure 11: Latent-space reward landscape (UMAP (McInnes et al., 2018), prior-mode samples, pooled projection across all three checkpoints). Each point is one sampled output, colored by reward-model score. v6 seeds show a visibly higher concentration of high-reward points than v5, corroborating the aggregate improvement in Table 3.

$\rho(\theta) = 1$ at the first of the m steps and departs from 1 only as θ moves within the batch. The clustered penalty and cold sampling of v5 are otherwise unchanged.

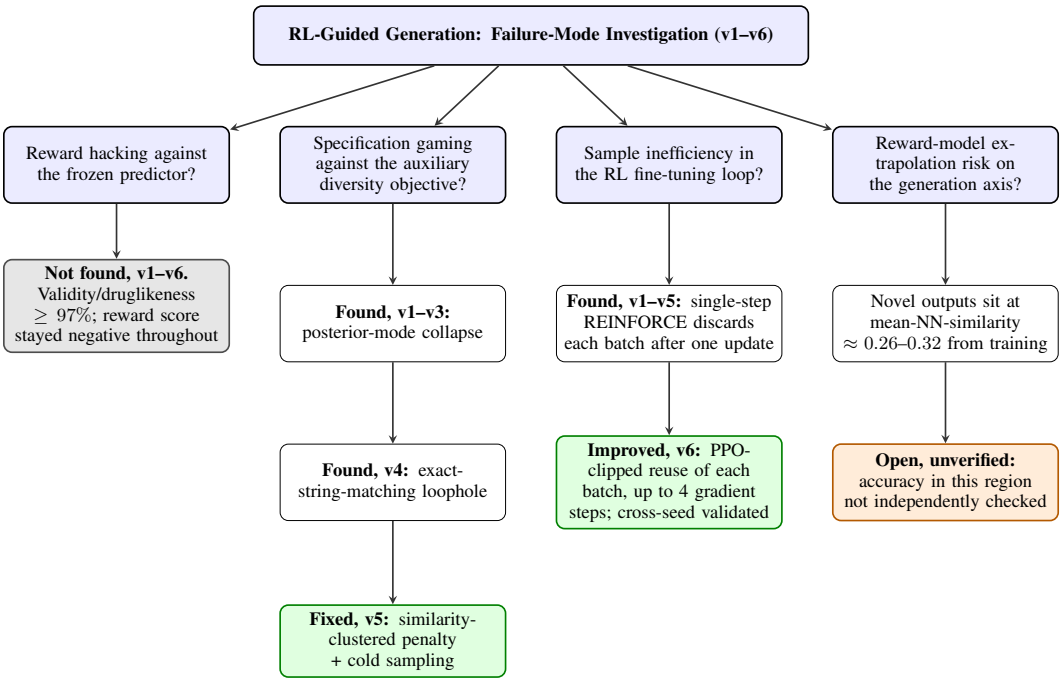

Figure 12: Failure-mode taxonomy for the six-run investigation. Gray = checked, not found (favorable). Green = a real problem found and fixed. Orange = an open, unverified risk carried into limitations.

E Extended Discussion and Limitations

Training entity diversity. The cold-start tower is trained exclusively on 51 breast cancer cell lines, and this pool's limited biological diversity may itself inflate apparent leave-target-out generalization independent of any leakage concern, which we separately verified absent (Appendix A).

Content encoder architecture. The current CNN operates on pixel-level content images without pathway or regulatory-network structure, motivating a structure-aware or pretrained foundation-model backbone as future work (Theodoris et al., 2023; Cui et al., 2024).

Compound coverage. 59% of compounds are tested in only one cell line, limiting what the model can learn about drug–cell-line interaction effects specifically, as opposed to marginal per-compound and per-cell-line effects. Focused experimental designs testing a panel of structurally diverse compounds across all cell lines would substantially improve the interaction signal available to the residual branch.

Reward-model extrapolation on the generation axis. The RL reward-hacking checks (validity, druglikeness, novelty, uniqueness) are a coarse filter: clearing them establishes only that an output looks superficially valid, not that the frozen predictor's specific score for it is trustworthy. Unlike the cold-start axis, where distance to training correlates significantly with prediction error (an explicit, testable reliability signal; Figure 1), no analogous generation-axis analysis was performed: writing $\mathcal{T}_{\text{gen}}$ for the set of generated outputs and sim for the same fingerprint-similarity metric underlying the v5/v6 clustering penalty (equation 14),

$$\overline{\text{NN-sim}} = \frac{1}{|\mathcal{T}_{\text{gen}}|} \sum_{\tau \in \mathcal{T}_{\text{gen}}} \max_{\tau' \in \mathcal{D}_{\text{train}}} \text{sim}(\tau, \tau') \;\approx\; 0.26\text{–}0.32, \tag{16}$$

a region where the predictor's accuracy is unverified. Closing this gap would require a per-output distance-vs-error analysis analogous to Figure 1, cross-validation against an independently trained oracle, and eventually independent experimental validation.

Checkpoint-dependence of attribution results. The cross-entity driver-gene-frequency ranking is not fixed across training: earlier checkpoints show TP53 as the dominant cross-entity activator, while the best-validation checkpoint reported throughout this work shows a weaker, non-dominant TP53 signal alongside the uncharacterized gene FSIP2. Predictive performance and attribution-signal clarity are not guaranteed to improve in lockstep; a single checkpoint's ranking should not be over-interpreted as a stable, architecture-level property.

Agentic retrieval-augmented interpretation. The GradCAM outputs reported in Section 4 and Appendix B — atom-level attributions for individual compounds and gene-level driver scores for individual cell lines — are numerical outputs that currently require manual literature lookup to contextualize biologically. Attaching an LLM agent downstream of these fixed, already-computed attributions — retrieving relevant pharmacological or pathway literature for a ranked list of attributed atoms or genes and composing a natural-language rationale linking the numerical attribution to known mechanism — is a natural extension. This role would be strictly interpretive and retrieval-based: the agent would not replace the D-MPNN, ECFP, or CNN encoders, nor generate or modify SMILES or genomic-image inputs, since natural-language descriptions of molecular or genomic structure are more lossy and hallucination-prone than the structured representations the model already uses.

Perturbation-aware modeling. Beyond static genomic backgrounds, drug activity is strongly modulated by biological perturbations — CRISPR knockouts, pathway activation states, epigenetic alterations, drug priming, and transcriptomic shifts induced by chemical or genetic interventions (e.g. L1000/CMap signatures) — which reshape target abundance, pathway wiring, compensatory mechanisms, and synthetic-lethal interactions, thereby altering drug sensitivity. Incorporating perturbations as a third modality, via a perturbation encoder or conditional modulation layers (e.g. FiLM, conditional attention), would enable context-specific response prediction under defined interventions and counterfactual queries such as how IC_{50} would change if a given gene were knocked out — moving the architecture toward a more mechanistic, intervention-capable framework, aligned with precision oncology.