DIAGNOSING REPRESENTATION AND PREDICTION ERRORS IN VIRTUAL-CELL MODELS UNDER DISTRIBUTION SHIFT

Anonymous authors
Paper under double-blind review

ABSTRACT

Predicting transcriptional responses to unseen drugs and cellular contexts is a central challenge in virtual-cell research. Overall prediction errors do not distinguish limited response-space coverage from inaccurate prediction within the covered space. We present a diagnostic framework that separates these components using response spaces learned from training data and characterizes within-space errors through prediction direction and magnitude. Analyses of LINCS and MoABox show that substantial response coverage can coexist with limited prediction accuracy. On MoABox, the response space captures 81.9% of target energy, yet most error from a direct predictor occurs within this space. Cross-cell experiments further show that fixed magnitude shrinkage improves transfer on LINCS and OP3 over unscaled predictions while preserving direction. The framework provides a clearer account of where virtual-cell predictions fail and how their errors change under specific model adjustments.

1 INTRODUCTION

Predicting transcriptional responses to perturbations is a central goal of virtual-cell research. Extending these predictions to unseen drugs, unseen cellular contexts, or both could broaden their applicability. Generalization beyond the training conditions is a central test of their usefulness. Prediction errors and correlations summarize performance, but interpreting that performance also requires understanding how predictions deviate from observed responses. A clearer account of these deviations can help identify which aspects of the prediction process need further investigation.

Existing approaches predict perturbation responses using compositional representations, optimal transport or mappings conditioned on molecular structure (Lotfollahi et al., 2023; Bunne et al., 2023; Hetzel et al., 2022; Yu et al., 2025). Comparisons based on aggregate error or correlation leave an important distinction unresolved: how much of a target response can be represented by a response space learned from training data, and how accurately a model predicts the corresponding component. A response may be well represented by this space while remaining difficult to predict from the available inputs. Separating these two aspects would make model comparisons more informative about where prediction succeeds or fails.

Therefore, we introduce a diagnostic framework that evaluates held-out responses relative to a response space learned from training data (Figure 1). The coverage metric measures the fraction of target-response energy represented by this space, while the within-space prediction accuracy metric measures how well a model recovers that component. We further examine prediction direction and magnitude to characterize errors within the space. Through this framework, we examine where prediction errors occur and how they change when model structure, directional supervision or prediction magnitude is adjusted.

Across LINCS, SciPlex3, MoABox, OP3 and Cellarity, we examine drug and cell OOD. LINCS and MoABox reveal broad response coverage alongside limited prediction accuracy; most error from a direct predictor on MoABox lies within the covered space. Direction-supervised factorization improves the original LINCS validation comparisons but underperforms direct prediction on MoABox and OP3. Fixed shrinkage improves cross-cell transfer over unscaled predictions on LINCS and OP3,

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.

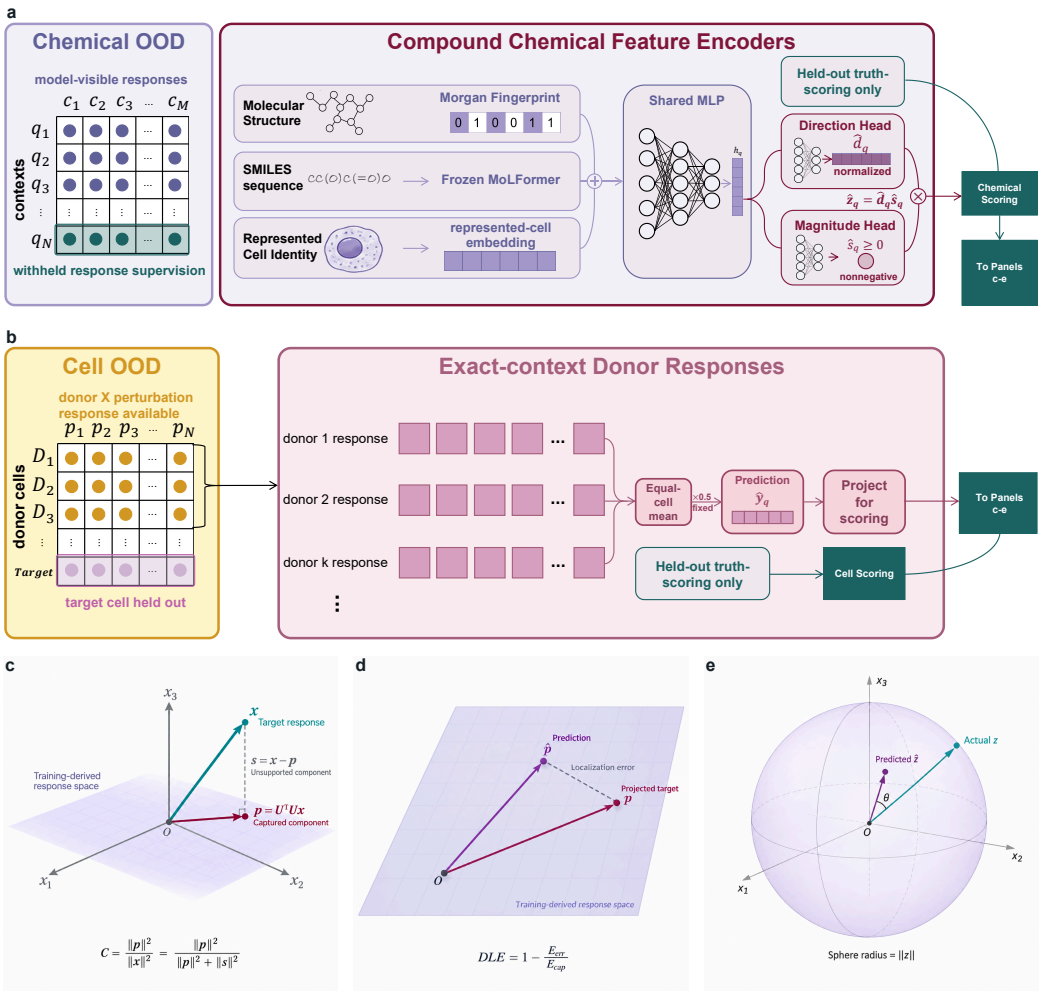


Figure 1: **Response-space diagnosis of virtual-cell predictions.** The framework separates target-response coverage from prediction accuracy within the covered space. **a**, Drug OOD predicts responses to unseen drugs using molecular and cell-context features. **b**, Cell OOD predicts responses in an unseen cell context by averaging responses to the same drug from other cells at matched dose and exposure time, then calibrating the prediction magnitude. **c**, Coverage measures the fraction of target-response energy represented by a space learned from training data. **d**, Within-space prediction accuracy measures how well the model recovers this represented component. **e**, Direction and magnitude further characterize prediction errors within the space. All panels are schematic.

with no clear pooled benefit on Cellarity. These results show how response-space diagnosis can help interpret both predictive gains and their variation across tasks.

2 Response-Space Evaluation

2.1 Tasks and Observable Information

A query specifies a compound, cellular context, dose, and exposure time. Drug OOD excludes all responses associated with the target standardized chemical parent; scaffold OOD additionally groups connected parent–scaffold components. The target cell may be represented by other drugs. Cell OOD excludes the target cellular context but permits the same drug at a matched dose and time in eligible

source cells. Table 1 summarizes these primary tasks. Supplementary analyses extend the evaluation to joint OOD, cross-accession validation, and dose transfer (Appendix J).

2.2 Coverage and Localization

For a training split, let $U \in \mathbb{R}^{r \times G}$ have orthonormal rows. For a target response x_q, oracle coordinates are $z_q = Ux_q$; they are used for scoring after the prediction $\widehat{z}_q$ is fixed. Orthogonality gives

$$\|x_q - U^\top \widehat{z}_q\|_2^2 = \|(I - U^\top U)x_q\|_2^2 + \|z_q - \widehat{z}_q\|_2^2. \tag{1}$$

For condition-mean responses, sum total, captured, and coordinate-error energies as $E_{\mathrm{tot}} = \sum_q \omega_q \|x_q\|^2$, $E_{\mathrm{cap}} = \sum_q \omega_q \|z_q\|^2$, and $E_{\mathrm{err}} = \sum_q \omega_q \|z_q - \widehat{z}_q\|^2$. Prespecified condition weights ω_q equal one unless stated otherwise. Define

$$C = \frac{E_{\mathrm{cap}}}{E_{\mathrm{tot}}}, \qquad \mathrm{DLE} = 1 - \frac{E_{\mathrm{err}}}{E_{\mathrm{cap}}}, \qquad S_{\mathrm{proj}} = C\,\mathrm{DLE}. \tag{2}$$

Coverage C measures support within the declared rank and gene axis. Localization efficiency, denoted DLE, measures recovery conditional on that support. Zero-coordinate predictions have DLE zero when captured target energy is positive; coordinate error exceeding captured energy gives negative DLE. Zero denominators remain undefined. Coverage is shared by predictors evaluated against the same targets and basis. The product S_{proj} is the full-response explained-energy skill of the reconstructed prediction relative to zero prediction.

The LINCS chemical analysis instead uses signed cross-replicate products to estimate repeatable energies (Appendix C). Those finite-sample ratios need not lie in $[0, 1]$ and cannot be interpreted as bounded Euclidean coverage. We distinguish these estimators throughout. A low rank or an incomplete gene axis also limits what "unsupported" means: it describes this representation, not biological impossibility.

2.3 Direction, Energy, and Reconstruction Error

For nonzero coordinates, define

$$\kappa_q = \frac{z_q^\top \widehat{z}_q}{\|z_q\|\|\widehat{z}_q\|}, \qquad R_{\mathrm{energy}} = \frac{\sum_q \omega_q \|\widehat{z}_q\|^2}{E_{\mathrm{cap}}}. \tag{3}$$

At each condition, localization error equals $\|z_q\|^2 + \|\widehat{z}_q\|^2 - 2\|z_q\|\|\widehat{z}_q\|\kappa_q$. Thus a higher mean cosine alone does not imply higher DLE. Energy is squared magnitude, and the energy minimizing error depends on directional agreement. Positive scalar shrinkage preserves direction while changing energy quadratically.

For a predictor emitting a full gene-space response $\widehat{x}_q$, raw-response MSE evaluates $\|x_q - \widehat{x}_q\|^2/G$, whereas projected-reconstruction MSE evaluates $\|x_q - U^\top U\widehat{x}_q\|^2/G$. The target remains the full response in both. Coordinate error alone excludes unsupported target energy. The two MSEs coincide for in-basis neural predictions but can differ for retrieved full-response baselines; we retain both where available.

3 Predictors and Experimental Design

3.1 Interventions and Controls

The neural drug models in the factorial comparisons use Morgan fingerprints, frozen MoLFormer embeddings, and the cell representation permitted by the task. DIRECT predicts coordinates with a single output head. COMPLETE uses a normalized direction v_q and nonnegative scale s_q from a shared representation h_q:

$$v_q = \frac{W_d h_q + b_d}{\|W_d h_q + b_d\|}, \quad s_q = \mathrm{softplus}(w_s^\top h_q + b_s), \quad \widehat{z}_q = s_q v_q. \tag{4}$$

Implementations guard zero norms. Crossing direct versus factorized outputs with absence versus presence of a direction loss produces a 2×2 comparison. For the original LINCS chemical-OOD

Table 1: **Datasets and primary evaluation tasks.** Counts describe the stated task axes; multiple tasks on one dataset are not independent datasets. Supplementary tasks are listed in Table 9.

Data source	Evaluation task	Analysis scope
LINCS GSE92742	Parent/scaffold; cell OOD	Original diagnostic and validation analyses; 9,694 held-out cell conditions
SciPlex3	Drug; cell OOD	1,812 drug conditions; 2,202 cell targets
Cellarity GSE306429	Cell OOD	659 targets in 10 backgrounds; 76 compound-label clusters
OP3 GSE279945	Drug; cell-type OOD	539 drug–cell-type conditions; 1,452 cell-type target rows in three fixed human donors
Novartis MoABox	Drug OOD	3,817 conditions, 3,816 parents; U-2 OS, 10 μM, 24 h

experiments, conditions with more consistent repeated measurements receive greater weight when learning response direction. We assess this consistency from responses relative to matched controls, before removing the average response for each cell context estimated from training data. These weights affect only directional supervision; the main prediction-error term treats all conditions equally. The weights guide model training and are distinct from the measures used to evaluate its predictions.

For MoABox and OP3 drug OOD, models learn the mean response for each condition, with conditions weighted equally in the main prediction loss. We scale this loss by the average energy of the training responses within the chosen response space. Models with directional supervision add a direction penalty weighted by 0.1. Conditions with negligible response in that space are excluded from this penalty but remain in the main loss. The scaling factor is estimated only from the data used to fit each model.

Losses, architectures, stopping rules, and task-specific settings are documented in Appendices E and J. Simple controls include zero prediction, a training-context mean, and molecular nearest-neighbor transfer where specified.

For cell OOD, eligible source cells D_q contribute one mean response each at the matched drug–dose–time context:

$$\widetilde{y}_q = |D_q|^{-1} \sum_{c' \in D_q} \bar{y}_{c',p_q,t_q,d_q}, \qquad \widehat{y}_q^{\mathrm{CRM}} = \alpha \widetilde{y}_q. \tag{5}$$

Calibrated Response Memory (CRM) uses $\alpha = 0.5$, selected in LINCS development and fixed in the later cell tasks. Comparisons with unscaled memory isolate shrinkage; comparisons with a global mean and a perturbation permutation probe query-specific information. Source-cell support differs by dataset. OP3 holds human-donor context fixed while transferring across cell types.

3.2 Response Construction and Split Boundaries

LINCS uses 978 landmark genes (Subramanian et al., 2017). Its original chemical analysis predicts residuals after subtracting training-fold cell-context means; its cell analysis retains complete treated-minus-control responses. The additional assays are SciPlex3 (Srivatsan et al., 2020), Cellarity (DeMeo et al., 2025), OP3 (Szałata et al., 2024), and MoABox (Hadjikyriacou et al., 2025; Ye et al., 2018). They use their specified condition-mean response constructions. SciPlex3 chemical validation also subtracts a training-only cell–dose context mean; MoABox, OP3, and Cellarity do not inherit the original LINCS chemical residual target. Outer held-out responses are excluded from fitted preprocessing and bases. The later nested protocols also refit these operations within the inner training role. The original LINCS mapper instead selects its checkpoint on an internal subset of the outer training role already used for context-mean and basis fitting; its validation evidence is interpreted under that recorded protocol. MoABox selects 5,000 genes using training treatments, matches same-plate DMSO, and reorders the verified plate-specific gene axes before response aggregation. Primary rank is 256 except for the two OP3 tasks, where it is 64.

Standardized-parent aliases stay within one role. For the MoABox amendment, one parent exceeding the frozen encoder's 202-token limit is excluded from the common support for all compared methods before expression modeling; remaining split assignments are inherited without backfilling. No SMILES truncation is used. Formal prediction locks precede target scoring. Raw-count integrity

inspection, previous target use in related tasks, and the retrospective status of GSE70138 are disclosed in Appendix J.

For MoABox and OP3 drug OOD, all models and seeds within each outer fold share a validation split that holds out chemical parents. Training duration is selected separately for each model and seed using MSE between reconstructed and observed responses across all selected genes. A newly initialized model is then fitted on the complete outer training set for the selected number of epochs. Test scoring begins only after predictions for every fold, model, and seed have been fixed.

3.3 Uncertainty and Comparison Scope

Within a task, additive energies are pooled before ratios are formed. For three-seed neural runs, metrics and paired contrasts are computed per seed and then averaged; this is not prediction ensembling. Paired bootstrap draws carry all aliases and contexts of the declared cluster, with 10,000 draws and common multiplicities across compared models and seeds. Cellarity uses compound-label clusters rather than verified chemical parents. OP3 cell-type inference conditions on the three observed human donors. Intervals are conditional on the observed assay, assignments, fitted bases, and fixed trained seeds. Secondary contrasts are pointwise and unadjusted. We compare effects within tasks and do not meta-average DLE across distinct targets, ranks, or estimators.

4 Results

4.1 Coverage and Prediction Accuracy

The original LINCS analysis estimates repeatable-energy coverage of 0.916670 under parent isolation and 0.915575 under parent–scaffold isolation at rank 256. Its mapper validation DLE is much smaller (0.087384 and 0.067242 for COMPLETE). These coverage and mapper values describe different held-out roles, so they support a qualitative comparison rather than a same-cohort product of metrics (Appendix H).

We use training-derived response spaces to distinguish representation limits from prediction error within those limits. On MoABox, the rank-256 basis captures 81.9% of target energy, whereas DIRECT attains a localization efficiency of 22.0%. Relative to total target energy, its reconstruction error comprises 18.1% outside the basis and 63.9% within it. Thus, most residual error occurs in directions already represented by the training basis. Coverage makes this gap to the best reconstruction within the fixed basis explicit, without implying that molecular features suffice to close it.

Coverage also quantifies how expanding the response space changes the geometric upper bound on explained-energy skill. In MoABox, increasing rank from 256 to 512 raises coverage from 81.92% to 83.46%, a gain of 1.53 percentage points. This increase is small relative to DIRECT's error within the rank-256 space, equivalent to 63.87% of total target energy. The comparison distinguishes additional representational coverage from the prediction gap within existing coverage. It does not establish whether retraining at higher rank would improve prediction.

SciPlex3 supplies a lower-support drug task: coverage is 0.577614 and Morgan kernel-ridge DLE is 0.041356 ($[0.015058, 0.069232]$). Its paired advantage over a Sattolo fingerprint control is 0.044707 ($[0.018415, 0.072330]$), indicating detectable but limited chemical prediction signal.

4.2 Direction and Within-Space Prediction Accuracy

COMPLETE improves over DIRECT in the original LINCS parent and scaffold validation contrasts, whereas the primary contrasts are negative in MoABox and OP3, with both intervals entirely below zero (Table 2). Coverage provides a shared reference for interpreting these model differences.

Table 2: **Drug-mapper effects depend on the evaluation task.** Paired DLE differences are COMPLETE minus DIRECT, with 95% grouped bootstrap intervals. Original LINCS rows are the saved specification-validation contrasts; later rows average three fixed seeds. They are separate task-specific estimands.

Source	Evaluation role	ΔDLE	95% CI
GSE92742	Parent validation	+.013960	[.004318,.023281]
GSE92742	Scaffold validation	+.012142	[.002984,.021620]
MoABox	Drug OOD	-.010469	[-.018972,-.002038]
OP3	Drug OOD	-.067458	[-.097148,-.035930]

On MoABox, DIRECT and COMPLETE use identical response bases, but COMPLETE increases mean coefficient cosine from 0.501 to 0.521 while reducing localization efficiency from 0.220 to 0.210 (paired difference -0.0105, 95% CI $[-0.0190, -0.0020]$). The performance difference therefore lies within the represented space. Higher mean directional agreement does not ensure lower coordinate error, which also depends on prediction magnitude and aggregation across conditions. OP3 shows the same pattern: cosine increases from 0.391 to 0.431 while DLE falls from 0.087 to 0.019 (Table 13). These comparisons show why directional agreement and within-space prediction accuracy need to be examined together. Complete factorial comparisons are reported in Appendices H.3 and J.

4.3 Magnitude Calibration in Cross-Cell Transfer

Table 3: **Cell-OOD shrinkage with fixed $\alpha = 0.5$.** The paired contrast is CRM minus the unscaled exact-context mean. All entries use task-specific condition-mean geometry; the uncertainty units and available source cells differ.

Source	Coverage	CRM DLE	ΔDLE	95% CI
GSE92742	.650061	.068098	+.069487	[.062935,.075719]
SciPlex3	.245375	.084442	+.921141	[.756911,1.152927]
Cellarity	.594753	.310898	+.003182	[-.032551,.057357]
OP3	.404935	.545974	+.045756	[.016844,.075698]

LINCS development selects shrinkage that preserves the retrieved direction. On its four jointly held-out evaluation cells, CRM improves DLE over raw memory by 0.069487 and over the global mean by 0.052304 ($[0.045289, 0.059710]$). Raw memory and CRM share cosine 0.149582; scaling reduces the energy ratio from 0.275169 to 0.068792. The gain therefore arises from radial rescaling at a fixed predicted direction.

On OP3, the same $\alpha = 0.5$ improves DLE by 0.045756 and reduces raw-response MSE by 0.001049 (95% CI for CRM minus unscaled: $[-0.001174, -0.000923]$). This transfers across cell types within three fixed human donors; it does not evaluate unseen human donors.

The benefit varies across datasets (Table 3). SciPlex3 shows a large gain over unscaled transfer, but CRM's own DLE interval and its contrast with the global mean include zero, and its K562 DLE is negative. Cellarity's shrinkage interval crosses zero, although exact-context memory clearly exceeds global and permuted controls. These comparisons distinguish the information provided by cross-cell responses from the additional gain of magnitude calibration.

5 Discussion

Across LINCS, SciPlex3, MoABox, OP3 and Cellarity, response-space diagnosis provides a clearer interpretation of virtual-cell prediction performance. It separates two aspects of generalization: how much of a target response is represented by the learned space, and how accurately the model predicts that component. The LINCS and MoABox analyses show that substantial coverage can coexist with limited prediction accuracy. On MoABox, most error from the direct predictor lies within the covered space. This finding identifies a concrete focus for model development: predicting the response variation already represented in the learned space more accurately.

Direction and magnitude further clarify how model changes affect prediction. On MoABox, the direction-supervised factorized model achieves higher directional agreement than the direct predictor, while its within-space prediction accuracy is lower. Fixed shrinkage provides a complementary example: it preserves direction while improving cross-cell transfer over unscaled predictions on LINCS, SciPlex3 and OP3. These gains demonstrate the contribution of magnitude adjustment. Cellarity shows that responses to the same drug in other cells contain predictive information beyond the global-mean baseline, with no clear additional pooled gain from fixed shrinkage. Together, these comparisons show how examining direction and magnitude helps interpret the different effects of a model adjustment across tasks.

Building on this work, our next goal is to develop virtual-cell models that generalize more reliably to unseen drugs and cellular contexts. We plan to use response-space diagnosis during model development to connect the learning of cellular response representations with the prediction of perturbation effects. By examining how changes to the representation, model structure and training objective affect both coverage and prediction accuracy, we aim to turn these diagnostics into practical guidance for model development. This provides a path from interpreting existing predictions toward constructing virtual-cell models informed by the response patterns observed across datasets.

6 Conclusion

We present a diagnostic framework for understanding generalization errors in virtual-cell response prediction. Across LINCS, SciPlex3, MoABox, OP3 and Cellarity, separating response-space coverage from within-space prediction accuracy helps explain where prediction errors occur. The results show that substantial coverage can coexist with limited prediction accuracy, while experiments that preserve prediction direction demonstrate how magnitude adjustment can improve cross-cell transfer. Together, coverage, prediction accuracy, direction and magnitude provide a clearer interpretation of model performance and a foundation for developing virtual-cell models that generalize to unseen drugs and cellular contexts.

Reproducibility Statement

The reproducibility ledger records inputs, split and donor assignments, response rank, hyperparameters, seeds, bootstrap units, and metric definitions. Versioned experiment archives preserve code, protocols, figure-source tables, input hashes, and source-data access records. The accompanying manuscript source archive is a separate writing artifact.

Ethics and Broader Impact

Transcriptomic response prediction can prioritize experiments. The framework supports transparent experimental prioritization by reporting model-visible information, negative controls, and uncertainty for each OOD analysis.

AI Assistance Statement

Generative AI tools assisted with literature searches, manuscript drafting, translation and revision, mathematical exposition, experiment-protocol development, data-processing and model code, figure preparation, and interpretation of saved results. The authors approved the research questions and protocols, reviewed the reported results, and remain responsible for the scientific claims and final manuscript.

References

Constantin Ahlmann-Eltze, Wolfgang Huber, and Simon Anders. Deep-learning-based gene perturbation effect prediction does not yet outperform simple linear baselines. *Nature Methods*, 22(8):1657–1661, 2025. URL https://www.nature.com/articles/s41592-025-02772-6. Published online 4 August 2025.

Charlotte Bunne, Stefan G. Stark, Gabriele Gut, Jacobo Sarabia del Castillo, Mitch Levesque, Kjong-Van Lehmann, Lucas Pelkmans, Andreas Krause, and Gunnar Rätsch. Learning single-cell perturbation responses using neural optimal transport. *Nature Methods*, 20(11):1759–1768, 2023.

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(8):1470–1480, 2024.

Benjamin DeMeo, Charlotte Nesbitt, Samuel A. Miller, Daniel B. Burkhardt, Inna Lipchina, Doris Fu, Peter Holderrieth, David Kim, Sergey Kolchenko, Artur Szalata, Ishan Gupta, Christine Kerr, Thomas Pfefer, Raziel Rojas-Rodriguez, Sunil Kuppassani, Laurens Kruidenier, Parul B. Doshi, Mahdi Zamanighomi, James J. Collins, Alex K. Shalek, Fabian J. Theis, and Mauricio Cortes. Active learning framework leveraging transcriptomics identifies modulators of disease phenotypes. *Science*, 390(6776):eadi8577, 2025.

Andrea Hadjikyriacou, Chian Yang, Martin Henault, Robin Ge, Leandra Mansur, Alicia Lindeman, Carsten Russ, Steffen Renner, Marc Hild, Jeremy Jenkins, Caroline Gubser-Keller, Jingyao Li, Daniel J. Ho, Marilisa Neri, Frederic D. Sigoillot, and Robert Ihry. Novartis/DRUG-seq U2OS MoABox Dataset. Zenodo, 2025. URL `https://zenodo.org/records/14291446`. Dataset, version 1.0.0.

Jens Hansen, Yuguang Xiong, Mustafa M. Siddiq, Priyanka Dhanan, Bin Hu, Bhavana Shewale, Arjun S. Yadaw, Gomathi Jayaraman, Rosa E. Tolentino, Yibang Chen, Pedro Martinez, Kristin G. Beaumont, Robert Sebra, Dusica Vidovic, Stephan C. Sch''urer, Joseph Goldfarb, James M. Gallo, Marc R. Birtwistle, Eric A. Sobie, Evren U. Azeloglu, Seth I. Berger, Angel Chan, Christoph Schaniel, Nicole C. Dubois, and Ravi Iyengar. Multiscale mapping of transcriptomic signatures for cardiotoxic drugs. *Nature Communications*, 15:7968, 2024. URL `https://www.nature.com/articles/s41467-024-52145-4`. Published online 11 September 2024; article 7968.

Leon Hetzel, Simon Böhm, Niki Kilbertus, Stephan Günnemann, Mohammad Lotfollahi, and Fabian J. Theis. Predicting cellular responses to novel drug perturbations at a single-cell resolution. In *Advances in Neural Information Processing Systems*, volume 35, pp. 26711–26722, 2022.

Mohammad Lotfollahi, F. Alexander Wolf, and Fabian J. Theis. scGen predicts single-cell perturbation responses. *Nature Methods*, 16(8):715–721, 2019.

Mohammad Lotfollahi, Anna Klimovskaia Susmelj, Carlo De Donno, Leon Hetzel, Yuge Ji, Ignacio L. Ibarra, Sanjay R. Srivatsan, Mohsen Naghipourfar, Riza M. Daza, Beth Martin, Jay Shendure, Jose L. McFaline-Figueroa, Pierre Boyeau, F. Alexander Wolf, Nafissa Yakubova, Stephan Günnemann, Cole Trapnell, David Lopez-Paz, and Fabian J. Theis. Predicting cellular responses to complex perturbations in high-throughput screens. *Molecular Systems Biology*, 19(6): e11517, 2023.

Pascal Passigan, Kevin Zhu, and Angelina Ning. Interpretable perturbation modeling through biomedical knowledge graphs. *arXiv preprint arXiv:2512.22251*, 2025.

Thai-Hoang Pham, Yue Qiu, Jucheng Zeng, Lei Xie, and Ping Zhang. A deep learning framework for high-throughput mechanism-driven phenotype compound screening and its application to COVID-19 drug repurposing. *Nature Machine Intelligence*, 3(3):247–257, 2021.

Xiaoning Qi, Lianhe Zhao, Chenyu Tian, Yueyue Li, Zhen-Lin Chen, Peipei Huo, Runsheng Chen, Xiaodong Liu, Baoping Wan, Shengyong Yang, and Yi Zhao. Predicting transcriptional responses to novel chemical perturbations using deep generative model for drug discovery. *Nature Communications*, 15(1):9256, 2024.

Yusuf Roohani, Kexin Huang, and Jure Leskovec. Predicting transcriptional outcomes of novel multigene perturbations with GEARS. *Nature Biotechnology*, 42(6):927–935, 2024.

Austin Spadaro, Alok Sharma, and Iman Dehzangi. ExPO: an exposure-conditioned neural operator for L1000 signature prediction. *Journal of Cheminformatics*, 18(1):94, 2026.

Sanjay R. Srivatsan, José L. McFaline-Figueroa, Vijay Ramani, Lauren Saunders, Junyue Cao, Jonathan Packer, Hannah A. Pliner, Dana L. Jackson, Riza M. Daza, Lena Christiansen, Fan Zhang, Frank Steemers, Jay Shendure, and Cole Trapnell. Massively multiplex chemical transcriptomics at single-cell resolution. *Science*, 367(6473):45–51, 2020.

Aravind Subramanian, Rajiv Narayan, Steven M. Corsello, David D. Peck, Ted E. Natoli, Xiaodong Lu, Joshua Gould, John F. Davis, Andrew A. Tubelli, Jacob K. Asiedu, et al. A next generation connectivity map: L1000 platform and the first 1,000,000 profiles. *Cell*, 171(6):1437–1452.e17, 2017.

Artur Szałata, Andrew Benz, Robrecht Cannoodt, Mauricio Cortes, Jason Fong, Sunil Kuppasani, Richard Lieberman, Tianyu Liu, Javier A. Mas-Rosario, Rico Meinl, Jalil Nourisa, Jared Tumiel, Tin M. Tunjic, Mengbo Wang, Noah Weber, Hongyu Zhao, Benedict Anchang, Fabian Theis, Malte D. Luecken, and Daniel B. Burkhardt. A benchmark for prediction of transcriptomic responses to chemical perturbations across cell types. In *Advances in Neural Information Processing Systems*, volume 37, pp. 20566–20616. Curran Associates, Inc., 2024. URL https://papers.nips.cc/paper_files/paper/2024/hash/24c4d51f3ef48dd2dbab78243ecb26a1-Abstract-Datasets_and_Benchmarks_Track.html.

Ramon Viñas Torné, Maciej Wiatrak, Zoe Piran, Shuyang Fan, Liangze Jiang, Sarah A. Teichmann, Mor Nitzan, and Maria Brbić. Systema: a framework for evaluating genetic perturbation response prediction beyond systematic variation. *Nature Biotechnology*, 44(6):1050–1059, 2026. Published online 25 August 2025.

Frederik Wenkel, Wilson Tu, Cassandra Masschelein, Hamed Shirzad, Liam Hodgson, Ihab Bendidi, Cian Eastwood, Shawn T. Whitfield, Craig Russell, Yassir El Mesbahi, Jiarui Ding, Marta M. Fay, Berton Earnshaw, Emmanuel Noutahi, and Alisandra K. Denton. TxPert: using multiple knowledge graphs for prediction of transcriptomic perturbation effects. *Nature Biotechnology*, 2026.

Zhenqin Wu, Bharath Ramsundar, Evan N. Feinberg, Joseph Gomes, Caleb Geniesse, Aneesh S. Pappu, Karl Leswing, and Vijay Pande. MoleculeNet: a benchmark for molecular machine learning. *Chemical Science*, 9(2):513–530, 2018.

Betty Xiong, Jan-Christian Huetter, Gabriele Scalia, Tommaso Biancalani, and Sepideh Maleki. LLM-guided retrieval for prediction of molecular perturbation responses. *arXiv preprint arXiv:2608.01734*, 2026. Workshop contribution, MLGenX at ICLR 2026.

Chaoyang Ye, Daniel J. Ho, Marilisa Neri, Chian Yang, Tripti Kulkarni, Ranjit Randhawa, Martin Henault, Nadezda Mostacci, Pierre Farmer, Steffen Renner, Robert Ihry, Leandra Mansur, Caroline Gubser Keller, Gregory McAllister, Marc Hild, Jeremy Jenkins, and Ajamete Kaykas. DRUG-seq for miniaturized high-throughput transcriptome profiling in drug discovery. *Nature Communications*, 9(1):4307, 2018.

Hengshi Yu, Weizhou Qian, Yuxuan Song, and Joshua D. Welch. PerturbNet predicts single-cell responses to unseen chemical and genetic perturbations. *Molecular Systems Biology*, 21(8): 960–982, 2025.

A Related Work

Virtual-cell models predict perturbational transcriptional responses through latent arithmetic, compositional representations, optimal transport, biological graphs, generative pretraining, and structure-conditioned mapping (Lotfollahi et al., 2019; 2023; Bunne et al., 2023; Roohani et al., 2024; Cui et al., 2024; Pham et al., 2021; Hetzel et al., 2022; Qi et al., 2024; Yu et al., 2025). Existing evaluations span unseen compounds, compound–cell pairs, molecular-scaffold splits, and transfer across cellular contexts (Wu et al., 2018; Passigan et al., 2025; Spadaro et al., 2026). They use squared error, correlation, gene-set scores, sign agreement, and distributional distances to summarize predictive quality.

Ahlmann-Eltze et al. (2025) evaluated genetic perturbation prediction using linear models with training-derived principal components, while Hansen et al. (2024) used SVD to identify drug-selective transcriptomic subspaces.

We use a response basis fitted on training data to account separately for held-out response energy outside the basis and prediction error within it. This diagnosis complements comparisons between predictive models.

Systema changes the reference used by response metrics and introduces centroid accuracy to assess perturbation-specific prediction beyond shared systematic variation (Viñas Torné et al., 2026). Our analysis quantifies target energy captured by a training-derived response subspace and prediction error within that subspace. These address complementary properties: a subspace can capture shared responses without a predictor learning perturbation-specific effects. Coverage therefore supplies a geometric reference for interpreting errors, while mean baselines and specificity controls remain necessary to assess whether predictions contain information about the queried perturbation.

TxPert uses split-half experimental reproducibility to contextualize prediction performance, while noting that this reference is not a strict upper bound (Wenkel et al., 2026). Our coverage analysis supplies a complementary reference: the fraction of observed target energy represented by a response basis fitted on training data. For predictions restricted to that basis, uncovered energy fixes a reconstruction error floor, and localization efficiency measures prediction skill within the covered space. This separates a representation constraint from coordinate prediction error; it neither estimates experimental reliability nor establishes how much biological signal is learnable.

We do not benchmark all of these model families. Chemical parent–scaffold isolation evaluates localization for unseen structures. Cell OOD evaluates transfer of an observed compound–time–dose response to an unseen cellular context. The cell case also connects to transcriptomic-signature retrieval and recent retrieve-and-aggregate prediction (Subramanian et al., 2017; Xiong et al., 2026) through radial calibration of exact-context donor responses.

B Derivations for the Response-Space Decomposition

This decomposition supplies the common hierarchy used across tasks. The following geometry is generic; the repeatability estimator and subsequent protocol sections describe the original LINCS analyses specifically. The additional datasets are documented in Appendix J.

Let $U \in \mathbb{R}^{K\times G}$ have orthonormal rows, let x_q be a target response, and let $z_q = Ux_q$. The model predicts coordinates $\widehat{z}_q$ and therefore a response $\widehat{x}_q = U^\top \widehat{z}_q$ lying in the declared response space. Orthogonal projection gives

$$x_q - \widehat{x}_q = (I - U^\top U)x_q + U^\top(z_q - \widehat{z}_q), \tag{6}$$

whose two terms are orthogonal. Hence

$$\|x_q - \widehat{x}_q\|_2^2 = \|(I - U^\top U)x_q\|_2^2 + \|z_q - \widehat{z}_q\|_2^2. \tag{7}$$

For a query set S, define

$$E_{\mathrm{tot}} = \sum_{q\in S}\|x_q\|_2^2, \quad E_{\mathrm{cap}} = \sum_{q\in S}\|z_q\|_2^2, \quad E_{\mathrm{err}} = \sum_{q\in S}\|z_q - \widehat{z}_q\|_2^2. \tag{8}$$

Summing Eq. (7) yields total prediction error $E_{\mathrm{tot}} - E_{\mathrm{cap}} + E_{\mathrm{err}}$. Normalizing the common explained-energy numerator gives

$$C = \frac{E_{\mathrm{cap}}}{E_{\mathrm{tot}}}, \qquad \mathrm{DLE} = 1 - \frac{E_{\mathrm{err}}}{E_{\mathrm{cap}}}, \qquad S_{\mathrm{proj}} = \frac{E_{\mathrm{cap}} - E_{\mathrm{err}}}{E_{\mathrm{tot}}} = C\,\mathrm{DLE}. \tag{9}$$

This formulation remains defined at $\mathrm{DLE} = 0$ and permits negative projected skill when coordinate error exceeds captured energy.

DLE combines angular and radial coordinate error. We therefore report the row-wise coefficient cosine

$$\cos_q = \frac{z_q^\top \widehat{z}_q}{\|z_q\|_2\|\widehat{z}_q\|_2}, \tag{10}$$

The original LINCS implementation sets this diagnostic to zero if either norm vanishes, then averages across evaluated conditions. Later tasks retain undefined directions and apply their recorded valid-direction masks; these conventions are not silently interchanged. Relative prediction energy is

$$R_{\mathrm{energy}} = \frac{\sum_q \|\widehat{z}_q\|_2^2}{E_{\mathrm{cap}}}. \tag{11}$$

Relative energy measures radial scale: 1 denotes matched oracle energy, values below 1 denote lower predicted energy, and values above 1 denote higher predicted energy. For any unscaled estimate w_q and global origin scale α,

$$\mathrm{DLE}(\alpha) = \frac{2\alpha\sum_q z_q^\top w_q - \alpha^2\sum_q \|w_q\|_2^2}{E_{\mathrm{cap}}}. \tag{12}$$

Positive scaling preserves Eq. (10) and changes DLE quadratically, allowing angular agreement and radial mismatch to be identified separately.

C LINCS Repeatability-Aware Chemical-OOD Estimation

Compound conditions contain unequal numbers of RNA-plate measurements. For a condition q with $R_q \geq 2$ residual responses x_{qj}^{chem}, we estimate repeatable squared energy from every ordered cross-replicate pair:

$$T_q = \frac{1}{R_q(R_q-1)} \sum_{j \neq k} (x_{qj}^{\mathrm{chem}})^\top x_{qk}^{\mathrm{chem}}, \tag{13}$$

$$A_q = \frac{1}{R_q(R_q-1)} \sum_{j \neq k} (U_{\mathrm{chem}} x_{qj}^{\mathrm{chem}})^\top (U_{\mathrm{chem}} x_{qk}^{\mathrm{chem}}). \tag{14}$$

Let $\bar{x}_q = R_q^{-1} \sum_j x_{qj}^{\mathrm{chem}}$ and $\bar{z}_q = U_{\mathrm{chem}} \bar{x}_q$. The prediction receives explained-energy credit

$$L_q(\widehat{z}_q) = 2\widehat{z}_q^\top \bar{z}_q - \|\widehat{z}_q\|_2^2. \tag{15}$$

The pooled estimators are

$$C_{\mathrm{rep}} = \frac{\sum_q A_q}{\sum_q T_q}, \qquad \mathrm{DLE}_{\mathrm{rep}} = \frac{\sum_q L_q}{\sum_q A_q}, \qquad S_{\mathrm{rep}} = \frac{\sum_q L_q}{\sum_q T_q} = C_{\mathrm{rep}} \mathrm{DLE}_{\mathrm{rep}}. \tag{16}$$

All conditions are pooled before ratios are formed, and every condition-level cross-replicate estimate, including negative values, enters the pooled ratios unchanged. This estimator removes independent measurement variance in expectation and defines a repeatability-adjusted finite-sample quantity. Chemical- and cell-OOD DLE share the same geometry and are interpreted within their respective task-specific estimands.

D Original LINCS Evidence Axes and Information Boundaries

D.1 Response construction

Both analyses use the 978 landmark-gene profiles from LINCS GSE92742 (Subramanian et al., 2017) at 24-h exposure and 10 μM. For each RNA plate, a response is the treated mean minus the matched same-plate DMSO mean. The cell-OOD target is the mean of all plate responses sharing cell, compound, time, and dose.

For compound OOD, the model predicts a context residual. A ContextMean is fit separately in each fold using training-role rows. In this fixed-dose, fixed-time cohort it is a training-fold cell-context mean, with declared cell and global fallbacks when a more specific context is absent. Each plate residual is $x_{qj}^{\mathrm{chem}} = y_{qj} - b_c^{\mathrm{tr}}$; the neural target and basis use the corresponding condition mean. Oracle coordinates are computed during evaluation after predictions are fixed.

D.2 Parent and scaffold isolation

Molecules are standardized to charge-neutral parent structures before roles are assigned. Parent OOD uses the standardized parent InChIKey as an indivisible group, preventing another record of an evaluated parent from providing response supervision.

The stricter scaffold cohort consists of cyclic compounds with a non-empty chiral Bemis–Murcko scaffold. We construct a bipartite graph whose nodes are standardized parent InChIKeys and scaffold identifiers and whose edges encode their observed association. Connected parent–scaffold components are the indivisible split groups. The transitive closure assigns every identical parent and every identical declared scaffold wholly to one role. Overlap assertions verify the assignments and exact-scaffold separation.

D.3 Cell block and donor eligibility

The development cells are SKB, ASC, HCC515, and HT29. A375, A549, HA1E, and PHH constitute the jointly held-out evaluation block. Response-basis fitting, response memory, calibration, and candidate selection use the development and admissible donor cells. Model-visible inputs consist of admissible donor responses.

Cell-OOD eligibility requires the exact compound–time–dose context in at least three distinct admissible donor cells after block exclusion. Each donor cell contributes one plate-aggregated condition mean, irrespective of its plate count. Equal-cell weighting gives each donor one contribution, and joint block exclusion removes every evaluation cell from all donor pools.

E Chemical Localization: Mapper Architecture and Training

The direction–magnitude mapper receives a 2,048-dimensional radius-2 Morgan fingerprint, a 768-dimensional embedding from a frozen MoLFormer encoder, and a learned 32-dimensional embedding of the represented cell

Table 4: **Original LINCS roles before molecular-feature filtering.** Entries list conditions/split groups/chemical parents, in that order. Split groups are standardized parents for parent OOD and connected parent–scaffold components for scaffold OOD. The scaffold cohort excludes 658 conditions from 119 parents in every fold. These counts precede the common feature support used for mapper evaluation.

Partition	Fold	Train	Validation	Test
Parent	0	24,036/3,790/3,790	8,013/1,264/1,264	8,014/1,264/1,264
Parent	1	24,038/3,791/3,791	8,012/1,263/1,263	8,013/1,264/1,264
Parent	2	24,039/3,792/3,792	8,012/1,263/1,263	8,012/1,263/1,263
Parent	3	24,039/3,791/3,791	8,012/1,264/1,264	8,012/1,263/1,263
Parent	4	24,037/3,790/3,790	8,014/1,264/1,264	8,012/1,264/1,264
Scaffold	0	23,646/2,234/3,686	7,881/743/1,235	7,878/700/1,278
Scaffold	1	23,643/2,189/3,731	7,881/745/1,233	7,881/743/1,235
Scaffold	2	23,643/2,189/3,746	7,881/743/1,220	7,881/745/1,233
Scaffold	3	23,640/2,188/3,746	7,884/746/1,233	7,881/743/1,220
Scaffold	4	23,643/2,231/3,688	7,878/700/1,278	7,884/746/1,233

line. The cohort fixes exposure at 24 h and 10 μM; molecular identity and represented cell are the varying model inputs. Morgan and MoLFormer inputs are separately mapped to 256 dimensions before concatenation with the cell embedding. The combined representation passes through layers of width 512 and 256. Each molecular branch uses a linear projection, LayerNorm, and GELU. The shared layers use GELU, dropout 0.1 after the first layer, and LayerNorm after the second. From shared representation h_q, the heads are

$$v_q = \frac{W_d h_q + b_d}{\max\{\|W_d h_q + b_d\|_2, 10^{-8}\}}, \qquad s_q = \mathrm{softplus}(w_s^\top h_q + b_s), \qquad \widehat{z}_q = s_q v_q. \tag{17}$$

The gene-space residual prediction is $U_{\mathrm{chem}}^\top \widehat{z}_q$.

For a training batch $B \subset F$, with fitting conditions F, let

$$\bar{E}_z = \frac{1}{|F|} \sum_{q \in F} \|\bar{z}_q\|_2^2, \qquad \nu_\epsilon(a) = \frac{a}{\max\{\|a\|_2, \epsilon\}}, \qquad \epsilon = 10^{-8}. \tag{18}$$

The loss is

$$\begin{aligned} \mathcal{L}_B = {} & \frac{1}{|B| \bar{E}_z} \sum_{q \in B} \left(\|\widehat{z}_q\|_2^2 - 2 \widehat{z}_q^\top \bar{z}_q \right) \\ & + \lambda_{\mathrm{dir}} \frac{\sum_{q \in B} w_q \left\{ 1 - \nu_\epsilon(\widehat{z}_q)^\top \nu_\epsilon(\bar{z}_q) \right\}}{\max\{\sum_{q \in B} w_q, \epsilon\}}, \qquad \lambda_{\mathrm{dir}} \in \{0, 0.1\}. \end{aligned} \tag{19}$$

Here $w_q = \mathrm{clip}(r_q, 0, 1)$. For condition q, r_q is the median Pearson correlation across distinct valid plate-level replicate pairs, using treated-minus-matched-control responses before context-mean subtraction. Replicate vectors with zero across-gene variance are excluded; fewer than two valid vectors leave r_q undefined. The directional coefficient is 0.1 for direction-supervised models and 0 otherwise.

The direct single-head comparator uses the same trunk and emits 256 coordinates. The direction-supervised factorized mapper adds the 257 parameters of its scalar magnitude head. The zero-directional-loss dual head has the same architecture and parameter count as the direction-supervised factorized mapper and sets $\lambda_{\mathrm{dir}} = 0$. The representation-permutation control randomly reassigns frozen MoLFormer embeddings among unique treatment identities across the response-free feature universe, keeping each treatment's incorrect vector consistent across its conditions. Morgan fingerprints remain correctly paired. The control isolates the contribution of correct drug–MoLFormer correspondence and retains Morgan information.

Architecture, rank 256, representations, λ_{dir}, optimizer, checkpointing rule, and comparator set are selected on parent-OOD fold 0. The development-selected specification uses AdamW with learning rate 10^{-3}, weight decay 10^{-4}, batch size 512, at most 100 epochs, early-stopping patience 12, and seed 20260731. A deterministic grouped holdout with a nominal 10% fraction selects the neural checkpoint: standardized parents for parent OOD and connected parent–scaffold components for scaffold OOD. ContextMean and the response basis use the declared training role; the deterministic 10% internal holdout governs neural gradient updates and checkpoint selection.

After development, parent folds 1–4 and scaffold folds 0–4 are each refit from their own permissible training role and evaluated on their specified held-out validation role. The parent-developed model specification is applied

Table 5: **Original LINCS mapper roles with both molecular features available.** Train and validation columns count conditions; inner fit and holdout entries list conditions/split groups. Group definitions match Table 4. All rows use seed 20260731: parent fold 0 is development, parent folds 1–4 are confirmatory, and scaffold folds 0–4 are transfer evaluations. The accompanying count summaries retain the factorial runs separately for each seed.

Partition	Fold	Train	Inner fit	Holdout	Validation
Parent	0	24,007	21,298/3,355	2,709/427	7,994
Parent	1	24,008	21,437/3,381	2,571/402	7,999
Parent	2	23,998	21,598/3,411	2,400/372	8,004
Parent	3	23,993	21,556/3,400	2,437/381	8,004
Parent	4	23,997	21,520/3,382	2,477/398	8,000
Scaffold	0	23,600	21,186/2,001	2,414/228	7,876
Scaffold	1	23,593	21,423/1,961	2,170/222	7,874
Scaffold	2	23,627	21,326/1,968	2,301/217	7,842
Scaffold	3	23,617	21,244/1,963	2,373/220	7,884
Scaffold	4	23,592	20,952/1,995	2,640/229	7,867

uniformly to every fold, and neural weights and response bases are re-estimated independently within each fold (Table 5).

Paired uncertainty uses 10,000 cluster-bootstrap draws. Parent OOD resamples standardized parent groups across folds; scaffold OOD resamples connected parent–scaffold components. Every condition belonging to a sampled group is retained. Primary contrasts compare the direction-supervised factorized mapper with the direct single head, the zero-directional-loss dual head, and the MoLFormer-permutation control.

F Cell-Case Calibration and Evaluation Protocol

For query $q = (c_q, p_q, t_q, d_q)$, define

$$D_q = \{c' \notin \mathcal{C}_{\text{eval}} : (c', p_q, t_q, d_q) \in \mathcal{Q}_{\text{adm}}^{\text{cell}}\}, \qquad |D_q| \geq 3. \tag{20}$$

The raw and calibrated memories are

$$\widetilde{y}_q = \frac{1}{|D_q|} \sum_{c' \in D_q} \bar{y}_{c', p_q, t_q, d_q}, \qquad \widehat{y}_q^{\text{CRM}} = \alpha \widetilde{y}_q, \qquad \alpha = 0.5. \tag{21}$$

The primary comparator is the perturbation-agnostic mean over all admissible donor conditions, projected through the same basis fitted from admissible donor responses.

During development, each of SKB, ASC, HCC515, and HT29 is treated in turn as the target. That target and all four evaluation cells are removed from its donor pool, and a separate uncentered rank-256 basis is fit to the remaining donor responses. The candidate set comprises the donor global mean, raw exact-context mean, exact-context coordinate median, CRM, and a perturbation-permuted control. The permutation reassigns predictions among perturbation identities. This preserves their energy distribution and destroys query–response correspondence.

The pooled B0 development sufficient statistics imply

$$\text{DLE}(\alpha) = 0.328534\,\alpha - 0.322822\,\alpha^2, \qquad \alpha^\star = \frac{0.328534}{2 \times 0.322822} \approx 0.509. \tag{22}$$

The pooled B0 development summary selects the coarse scale $\alpha = 0.5$, followed by a B1-A condition-level stability analysis with 10,000 paired bootstrap draws clustered by perturbation across development cells. The B1-A analysis yields the scale, candidate, rank, donor policy, and evaluation rule used for the held-out cells. The selected α is the predictive calibration parameter used in evaluation.

For held-out evaluation, the four evaluation cells are removed jointly and one uncentered rank-256 basis is fit to the remaining 26-cell donor pool. Rows are the leading eigenvectors of the raw donor-response Gram matrix with deterministic component signs. CRM, comparator, eligibility rule, rank, bootstrap unit, and decision gate use the development-selected configuration. Predictions are materialized separately from the scoring step. Development uses seed 20261731; the held-out evaluation and paired bootstrap use seed 20261831; prediction blocks contain 512 rows.

The evaluation gate requires: (i) positive pooled CRM-minus-global DLE with a 95% percentile interval excluding zero; (ii) one-sided empirical probability below 0.05 for a nonpositive DLE difference; (iii) coefficient-cosine noninferiority at margin 0.005 with one-sided probability below 0.05; and (iv) positive cell-wise DLE differences in at least three of four cells. Ten thousand draws resample perturbation identity jointly across cells. Empirical tails use $(b+1)/(B+1)$, so 0.0001 is the smallest attainable probability.

G Development Evidence for Transferable Direction and Radial Calibration

The development set contains 7,376 eligible conditions and 3,316 perturbation clusters; pooled oracle coverage is 0.673166. Table 6 reports all primary candidates (see also Figure 2). The raw exact-context mean has positive coefficient cosine, an energy ratio of 0.322822, and DLE 0.005712. The perturbation-permuted control preserves this energy and reduces cosine to 0.055299 and DLE to -0.249809. These results identify query-specific directional information in the unpermuted donor response. Useful prediction radius depends jointly on norm and alignment. Raw-memory energy is 0.322822 of oracle energy. Scaling by 0.5 preserves cosine, reduces relative energy to 0.080705, and raises DLE to 0.083561.

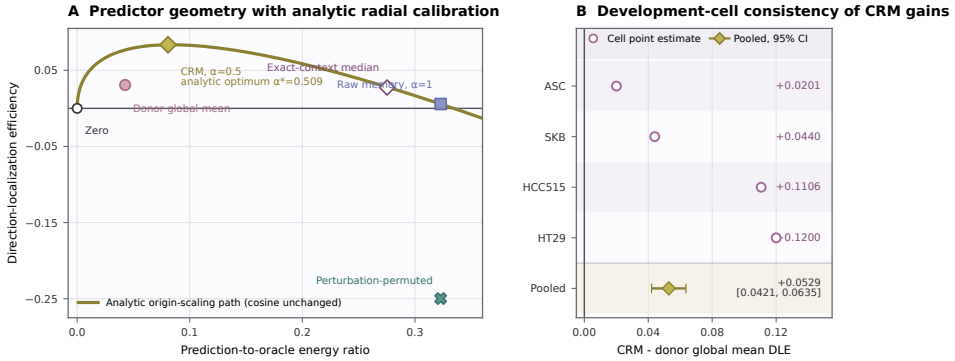


Figure 2: **Development evidence for transferable direction and radial calibration. A**, direction–localization efficiency versus prediction-to-oracle energy for all development candidates, overlaid with the analytic origin-scaling path of the raw exact-context donor mean. Positive scaling preserves coefficient cosine; the development-selected $\alpha = 0.5$ lies near the analytic optimum. **B**, descriptive development-cell CRM-minus-global DLE estimates and the pooled perturbation-clustered 95% interval.

Table 6: **Development evidence for query-specific direction and radial calibration.** These development diagnostics determine the global scale subsequently applied in the four-cell held-out evaluation.

Predictor	DLE	Projected skill	Mean cosine	Energy ratio
Donor global mean	.030685	.020656	.157845	.042593
Raw exact-context mean	.005712	.003845	.164452	.322822
Exact-context median	.027407	.018449	.162532	.275232
Perturbation-permuted	-.249809	-.168163	.055299	.322822
CRM ($0.5\times$ mean)	**.083561**	**.056251**	.164452	.080705

Relative to the donor global mean, CRM improves DLE by 0.052876 (95% interval $[0.042082, 0.063536]$, one-sided empirical $p = 0.0001$). The cosine difference is 0.006608 ($[-0.000252, 0.013441]$); the probability of falling below the -0.005 noninferiority margin is 0.0006. DLE differences are positive in all four development cells.

H Original LINCS Results and Sensitivity Analyses

H.1 Coverage–Localization Gap Across Reliability and Response-Energy Strata

Table 7 gives the complete chemical contrasts. Coverage measures target–basis support on the designated held-out role. Mapper effects measure localization under the shared model specification. Their side-by-side presentation shows the coverage–localization gap between these distinct estimands.

Table 7: **Persistence of the chemical coverage–localization gap.** Counts and mapper DLE refer to validation: parent folds 1–4 and scaffold folds 0–4. Groups are unique chemical parents and connected parent–scaffold components, respectively. Coverage uses the separate designated test roles. Intervals use 10,000 paired grouped bootstrap draws.

A. Coverage and direction-supervised-mapper performance

Partition	Conditions	Groups	Coverage ($r = 256$; 95% CI)	Factorized DLE
Standardized parent	32,007	5,043	.916670 [.912214,.921334]	.087384
Scaffold	39,343	3,668	.915575 [.910535,.920492]	.067242

B. Paired DLE: direction-supervised mapper minus comparator

Comparator	Parent OOD		Scaffold OOD	
	Δ	95% CI	Δ	95% CI
Direct single head	.013960	[.004318,.023281]	.012142	[.002984,.021620]
Factorized, zero dir. loss	.013544	[.000799,.027145]	.010346	[-.002195,.023551]
Permuted MoLFormer	.028451	[.017254,.039690]	.027823	[.016824,.039190]

Coverage computed directly from noisy replicate means is approximately 0.682, well below the cross-replicate estimate. Repeatable coverage remains about 0.895–0.897 in the highest replicate-consistency quartile and 0.908–0.909 in the strongest-response quartile. The coverage–localization gap therefore persists in both the high-reliability and high-energy strata.

Across all conditions, the direction-supervised factorized mapper has lower unweighted coefficient cosine. In high-response-energy and high-replicate-reliability strata it has higher cosine. Repeatable-energy-weighted DLE summarizes localization across response strengths. In scaffold OOD, the single-seed contrasts quantify the direction-supervised mapper relative to each comparator. The matched-seed factorial analysis estimates the conditional contribution of directional supervision across three optimization seeds.

H.2 Coverage and Localization Sensitivity to Response Rank

The compound analyses use a rank-256 response basis. To quantify how representation size affects the available signal, we evaluate repeatable coverage for $r \in \{32, 64, 128, 256, 512, 768\}$, using common group resamples across ranks. Coverage increases monotonically from 0.683010 to 0.988844 in parent OOD and from 0.681648 to 0.988479 in scaffold OOD. The rank-512 and rank-768 values quantify nested-basis target coverage. The two coverage curves differ by at most 0.16 percentage points across the audited ranks. Predictor sensitivity uses retained coordinates from the fixed rank-256 fits (Figure 3).

For localization sensitivity, the analysis uses the first $k \in \{32, 64, 128, 256\}$ coordinates from the saved validation targets and predictions of the fixed rank-256 fits. The $k = 256$ calculations exactly reproduce the saved condition metrics. The direction-supervised factorized mapper has the highest DLE point estimate at every audited k in both partitions and remains above the permuted-MoLFormer control.

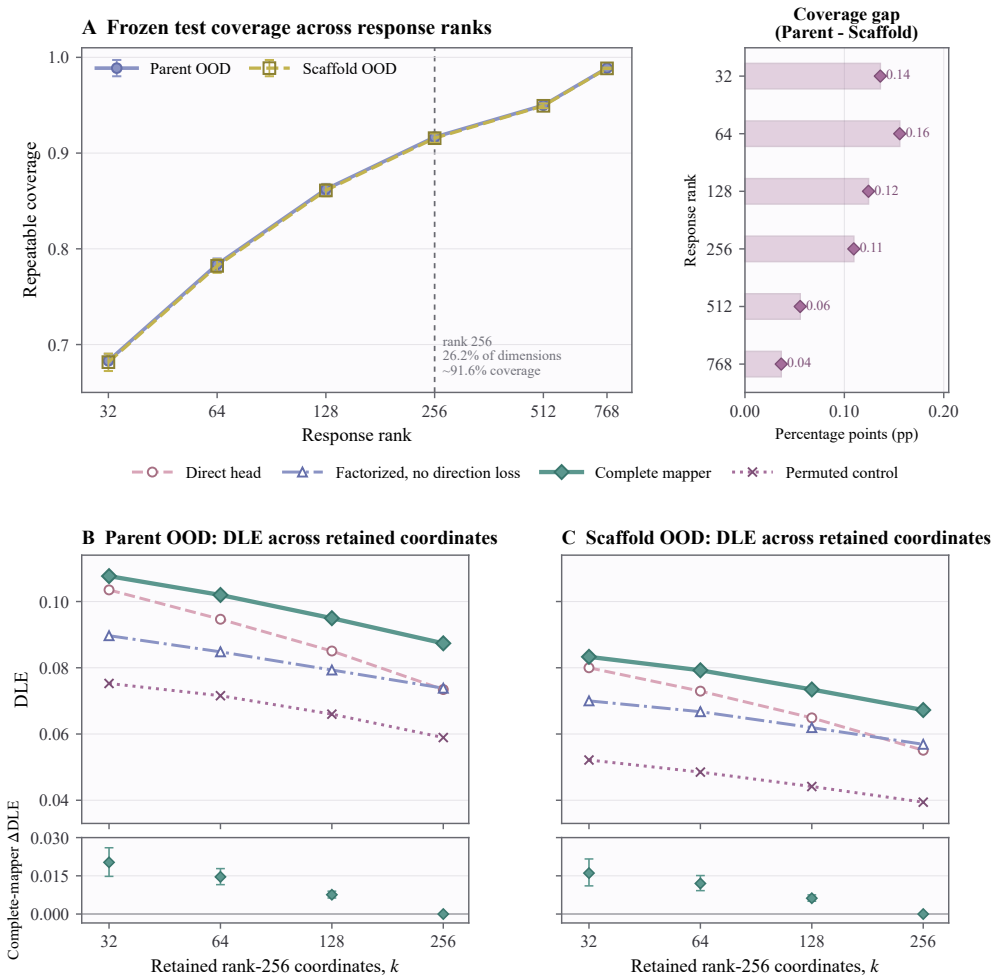


Figure 3: **Response-space coverage and retained-coordinate sensitivity of rank-256 models. A**, repeatable target coverage on held-out test conditions across nested response-basis ranks, with 95% cluster-bootstrap intervals from common group resamples; the vertical line marks rank 256. The right-hand panel reports the point-estimate coverage gap in percentage points (parent minus scaffold) on a zero-origin axis; the maximum gap is 0.16 percentage points. **B,C**, validation DLE obtained by truncating the saved target–prediction coordinate pairs from rank-256 fits to the first k coordinates. The lower panels report the paired difference from $k = 256$, with 95% intervals.

H.3 Parameterization-by-Supervision Interaction Across Seeds

We cross direct and direction–magnitude output heads with directional-loss weights 0 and 0.1, refitting all four design cells on every fold with matched seeds 20260731–20260733. The same folds, feature support, training rules, and checkpoint-selection procedure are used for all four cells. The aligned evaluation axes contain 32,007 conditions in 5,043 standardized-parent groups and 39,343 conditions in 3,668 scaffold-component groups.

Figure 4 reports the four design-cell means, all three seed-level values, and the five factorial contrasts. The interaction is $I = (\mathrm{FD} - \mathrm{FN}) - (\mathrm{DD} - \mathrm{DN})$. For each of 10,000 bootstrap draws, the same group multiplicities are reused across all four design cells and all three seeds before the seed-specific effects are averaged. The reported intervals are unadjusted 95% paired group-bootstrap intervals conditional on the three fixed optimization seeds. The parameterization–supervision interaction is 0.015726 ($[0.006075, 0.025510]$) for parent OOD and 0.010579 ($[0.000791, 0.021465]$) for scaffold OOD.

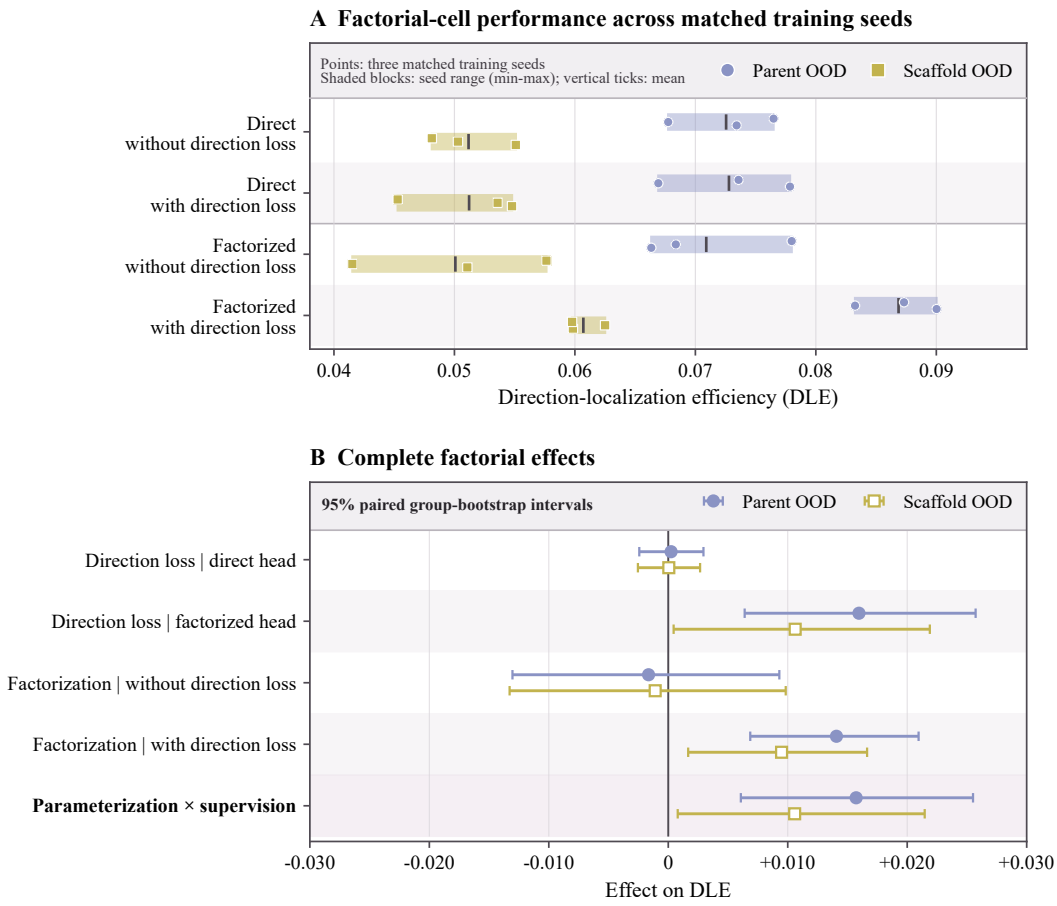


Figure 4: **Original LINCS chemical factorial analysis. A**, pooled DLE for the four factorial design cells in parent and scaffold OOD. Points show the three matched training seeds, shaded blocks span the seed minimum–maximum range, and vertical ticks mark seed means. **B**, five factorial effects in both partitions with 95% paired group-bootstrap intervals. The interaction $I = (\mathrm{FD} - \mathrm{FN}) - (\mathrm{DD} - \mathrm{DN})$ is estimated by averaging the three matched training seeds.

H.4 Radial Calibration Under Jointly Held-Out Cells

The held-out block contains 9,694 conditions and 4,790 perturbation clusters; coverage is 0.650061 for all predictors in the fixed basis. Table 8 reports complete pooled estimates and inference (see also Figure 5).

Table 8: **Cell-OOD performance on jointly held-out cells with a development-selected radial calibration.**

A. Absolute performance

Predictor	DLE	Projected skill	Mean cosine	Energy ratio
Donor global mean	.015794	.010267	.119562	.038369
Raw exact-context mean	-.001389	-.000903	**.149582**	.275169
CRM	**.068098**	**.044268**	**.149582**	.068792

B. Paired improvements

Outcome	Δ	95% interval	Inference
DLE, CRM minus global	.052304	[.045289,.059710]	$p_{\leq 0} = .0001$
DLE, CRM minus raw	.069487	[.062935,.075719]	—
Mean cosine, CRM minus global	.030020	[.024447,.035691]	$p_{<-0.005} = .0001$

Cell-wise DLE differences are 0.031693 in A375, 0.062761 in A549, 0.060840 in HA1E, and 0.029179 in PHH. The corresponding global-to-CRM DLE values are 0.072139 to 0.103832, −0.000206 to 0.062555, 0.055151 to 0.115991, and 0.000944 to 0.030124. DLE gains are positive in all four cells. Perturbation identity is the inferential unit for pooled estimates. The four cell-wise point estimates summarize breadth, and pooled perturbation-clustered intervals quantify uncertainty.

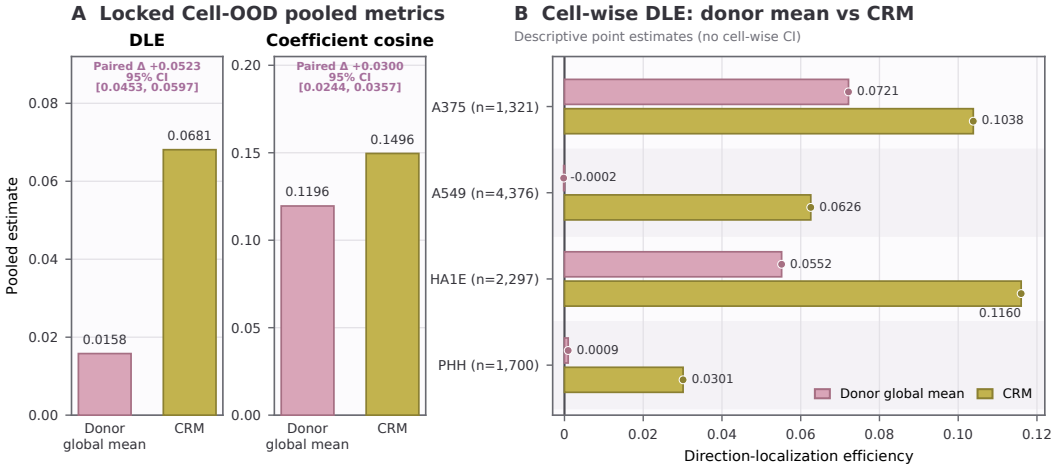


Figure 5: **LINCS cell-OOD retrieval and radial calibration on jointly held-out cells. A**, pooled DLE and coefficient cosine for the donor global mean and CRM; annotations give paired CRM-minus-global effects and 95% perturbation-clustered intervals. **B**, cell-wise DLE point estimates for the same predictors across four jointly held-out cells.

I Reproducibility and Evidence Scope

The experimental records specify input identities, response construction, parent and cell assignments, model-visible information, fitted bases, stopping rules, prediction locks, and bootstrap units. The manuscript source archive contains the text, references, and included figures; it is distinct from the versioned experiment code and result archives. The latter preserve the complete scientific provenance, including previously inspected data and protocol amendments. Dataset accessions and the task-specific procedures above identify the inputs and estimands needed to interpret each reported result.

J Additional datasets, task protocols, and primary-rank comparisons

These evaluations test different sources of distribution shift (Table 9). Reusing an assay for Drug OOD and Cell OOD defines two tasks, not two independent datasets. Ranks, response construction, and sampling units remain task-specific.

J.1 Estimands and uncertainty

For a row-orthonormal training basis $U \in \mathbb{R}^{r\times G}$, let $z_i = Uy_i$ and $a_i = U\widehat{y}_i$. We aggregate additive quantities before forming ratios:

$$E_{\mathrm{tot}} = \sum_i w_i \|y_i\|^2, \qquad E_{\mathrm{cap}} = \sum_i w_i \|z_i\|^2, \qquad E_{\mathrm{err}} = \sum_i w_i \|z_i - a_i\|^2, \tag{23}$$

$$C = E_{\mathrm{cap}}/E_{\mathrm{tot}}, \qquad \mathrm{DLE} = 1 - E_{\mathrm{err}}/E_{\mathrm{cap}}, \qquad R_P = \frac{\sum_i w_i \|a_i\|^2}{E_{\mathrm{cap}}}. \tag{24}$$

Projected skill is $C\,\mathrm{DLE}$. Raw-response MSE evaluates $\widehat{y}_i$ against y_i; projected-reconstruction MSE evaluates $U^\top a_i$ against the same original y_i, dividing by $G\sum_i w_i$. The two coincide for predictions confined to the scoring subspace, but can differ for full-response baselines and context-restored predictions. Mean cosine averages defined row-level coefficient cosines with weights w_i; R_P is a squared-norm ratio. Consequently, higher mean cosine with lower DLE does not isolate amplitude as the cause: DLE also depends on how alignment and amplitude vary with target energy.

Table 9: Evaluation populations and primary ranks. Parent counts denote standardized chemical groups except Cellarity's label clusters. OP3's two rows are tasks on the same assay. Dashes denote quantities not tabulated here.

Dataset and task	Evaluation units	Groups	Rank	Coverage C
SciPlex3, parent validation	1,812 conditions	150	256	0.577614
SciPlex3, Cell OOD	2,202 conditions	185	256	0.245375
GSE70138, external	4,549 conditions	690	256	–
SciPlex3, high dose	546 contexts	185	256	0.458541
Cellarity, Cell OOD	659 conditions	76 labels	256	0.594753
GSE92742, Joint OOD	7,827 conditions	3,892	256	0.647115
MoABox, Drug OOD	3,817 conditions	3,816	256	0.819246
OP3, Drug OOD	539 conditions	140	64	0.627320
OP3, Cell-type OOD	1,452 rows	132	64	0.404935

Intervals use 10,000 paired block-bootstrap draws, carrying all aliases and contexts of each chemical parent; Cellarity instead uses compound-label blocks. Bases, predictions, and assay backgrounds are fixed, without refitting. Three-seed results average seed-specific pooled ratios, not predictions or fold ratios. Secondary intervals are pointwise and unadjusted for multiplicity. Mean-response coverage is bounded; the separate SciPlex3 signed cross-replicate diagnostic can exceed one and is not interpreted as this coverage estimand.

J.2 SciPlex3 chemical, cellular, and dose shifts

The 24-h SciPlex3 chemical evaluation used four parent-validation folds, with 5,000 genes and context-residual bases fitted on training data. Morgan–Tanimoto kernel ridge regression selected regularization by five-fold inner parent validation; every inner context mean and basis was refitted. The negative control permuted unique training fingerprint classes using a fixed-point-free Sattolo permutation. Bounded geometry was subsequently computed for the mean of two replicate residual responses with predictions unchanged. This is validation evidence, not a previously unused final test.

The cellular task held out A549, K562, and MCF7 in turn. Each of 734 shared drug–dose contexts used the other two cells as exact-context response donors. Responses averaged two plate-matched treated-minus-DMSO replicates, requiring at least 20 assigned cells per replicate. Donor-only responses defined the basis. CRM applied the LINCS-fixed multiplier 0.5 to the donor mean; the global and permuted controls used the same multiplier.

Dose extrapolation used 10, 100, and 1,000 nM to predict 10,000 nM. Library-normalized log-CP10K responses subtracted matched DMSO controls and averaged two replicates. The 5,000-gene axis was selected using only 10/100-nM treated samples and their controls. Low-dose development predicted 1,000 nM from 10/100 nM and fixed a nonnegative least-squares multiplier $\alpha = 0.0827983115539$ in full gene space. Final extrapolation was $\alpha(2y_{1000} - y_{100})$; predictions were locked before highest-dose scoring. Those highest-dose responses had already appeared in the chemical and cellular tasks, so this lock does not establish historically untouched data (Table 10). Paired intervals support improvement over uncalibrated local extrapolation and nearest-dose transfer. The higher point estimate alone does not establish superiority over the low-dose mean.

J.3 Cross-accession validation and Cellarity

GSE70138 evaluation retained 978 landmark genes and 10-μM, 24-h conditions with at least two controlled RNA plates. Targets subtracted the GSE92742 training context. Frozen reference models were evaluated on 2,477 exactly overlapping conditions and 2,072 new drug–cell combinations, without external tuning. Prior external scoring and the earlier two-model comparison make the three-seed factorial analysis retrospective (Table 11). COMPLETE has lower point-estimate DLE than DIRECT, with a primary interval crossing zero. Direction loss improves the factorized model, partially offsetting the negative factorization effect without that loss (Table 12).

Cellarity GSE306429 used the published count layer and 13,290 genes. Each well was transformed by $\log(1 + 10^4 x / \sum_g x_g)$ and matched to same-library DMSO controls. Aggregation gave equal weight successively to wells, plates, batches, and two replicate labels. Each held-out cell required at least three visible cells with the same context. Ten cellular backgrounds were evaluated; uncertainty resampled 76 compound labels, whose standardized chemical-parent equivalence was not established.

Table 10: SciPlex3 chemical validation and high-dose extrapolation at rank 256. Predictor DLE values are point estimates. Dose-contrast intervals use 10,000 paired bootstrap draws over 185 chemical parents; all draws are defined.

Task	Predictor or contrast	DLE or ΔDLE
Chemical validation	Morgan kernel ridge	0.041356
	95% interval	[0.015058, 0.069232]
	Morgan minus Sattolo	0.044707
	95% interval	[0.018415, 0.072330]
High-dose extrapolation	Same-cell low-dose mean	0.050125
	Nearest-dose transfer	−0.251445
	Local log-dose extrapolation	−3.033155
	Calibrated local extrapolation	0.073774
	Calibrated minus local extrapolation	3.106929
	95% interval	[2.413803, 4.107577]
	Calibrated minus nearest-dose transfer	0.325219
	95% interval	[0.152084, 0.551794]

J.4 Simultaneously unseen drugs and cells

GSE92742 Joint OOD used four outer folds, holding out chemical parents and all four target cells (A375, A549, HA1E, PHH). Inner selection likewise held out parents and cells; single-axis visible regions could not supply held-out responses. All 978 landmarks were retained. Cell-context means, residual mapper bases, and full-response scoring bases used only the current fit partition. Joint-OOD implementations use training-side replicate-reliability weights. For each unseen drug, model-predicted residual responses in visible cells were restored to their fitted contexts and averaged equally. These are predicted donor responses, not observed responses to the unseen drug. Inner selection minimized full-response MSE for the unscaled prediction; fresh outer models used the selected epoch count. The prespecified primary contrast applied $\alpha = 0.5$ to both models. Its positive difference must be distinguished from the negative unscaled contrast (Tables 15–16).

J.5 MoABox and OP3 Drug OOD

MoABox used U-2 OS at 10 μM and 24 h. Verified gene permutations aligned 59,594 raw genes. Wells required at least 10,000 counts and 1,000 detected genes, with all three treated plate replicates passing and 16 matched DMSO wells per plate. Individually normalized log-CP10K responses were averaged across replicates, then equally across batches. Fifteen candidate labels failed QC; one additional chemical parent exceeded the fixed molecular token limit and was explicitly excluded before training. No sequences were truncated and retained split assignments were unchanged. Its coverage interval was [0.811332, 0.826789]. Increasing the descriptive basis rank from 32 to 512 raises coverage from 0.726746 to 0.834583; those extra ranks were not new neural training experiments.

OP3 GSE279945 used 1-μM, 24-h perturbations in B, Myeloid, NK, and T cells from three fixed human donors. Responses used per-cell log-CP10K, equal-cell well means, and DMSO means matched by donor, plate, and cell type. Eligibility required 20 treated cells, 20 matched control cells, two nonempty control wells, and at least two eligible donors per drug–type condition. Reference compounds were excluded. Equal-donor aggregation yielded 539 conditions, with complete Morgan/MoLFormer support for all 140 parents.

Both assays used the architecture in Appendix E at the task-specific output rank, five outer parent folds, three seeds, and the four factorial models, with 2,048-bit Morgan fingerprints and 768-dimensional frozen MoLFormer embeddings. Directional supervision used condition-level weights $w_q = \mathbf{1}\{\|z_q\|_2 > 10^{-8}\}$, evaluated in float64 before conversion to float32, without LINCS replicate-reliability weights. Inner fit data alone selected 5,000 genes by detection, count, and gene-ID tie-break, and fitted feature transformations and uncentered bases. Early stopping used inner full-gene MSE; fresh outer refits used the chosen epoch count. AdamW used learning rate 10^{-3}, weight decay 10^{-4}, batch size 512, at most 100 epochs, patience 12, and gradient clipping at 5. Main ranks were 256 and 64, respectively. The negative COMPLETE–DIRECT comparisons were retained (Table 12).

J.6 OP3 Cell-type OOD and limits of transfer

Each held-out OP3 cell type was predicted from the other three types for the same drug and human donor. Complete four-type contexts yielded 363 drug–donor contexts and 1,452 target rows. A row received weight equal to the reciprocal of its drug's eligible donor count, giving total weight 528. Source-only gene selection preceded a rank-64 basis fit to equal-donor source drug–type means. Predictions averaged the three source types

Table 11: Three-seed mean DLE for chemical prediction. External GSE70138 uses rank 256, MoABox rank 256, and OP3 rank 64. External baseline values are not tabulated.

Predictor	GSE70138	MoABox	OP3
DIRECT	0.026664	0.220340	0.086924
DIRECT + direction loss	0.025534	0.221115	0.084239
Factorized without direction loss	−0.004098	0.209242	0.067302
COMPLETE	0.018814	0.209871	0.019466
Training cell-context mean	–	0.177071	0.187054
Same-cell/type Morgan nearest neighbor	–	−0.218740	−0.416137
Zero	–	0.000000	0.000000

Table 12: Reported paired chemical-prediction contrasts. All entries are ΔDLE [pointwise 95% interval]. COMPLETE minus DIRECT is the primary comparison for each dataset; factorial and baseline contrasts are secondary. Untabulated contrasts are not inferred from rounded model means.

Dataset	Contrast	Estimate [95% interval]
GSE70138	COMPLETE − DIRECT	−0.007851 [−0.017336, 0.001075]
	Direction loss, factorized head	0.022912 [0.010285, 0.036190]
	Factorization, no direction loss	−0.030763 [−0.045660, −0.016923]
MoABox	COMPLETE − DIRECT	−0.010469 [−0.018972, −0.002038]
	Direction loss, direct head	0.000775 [−0.003987, 0.005642]
	Direction loss, factorized head	0.000630 [−0.002045, 0.003360]
	Factorization, no direction loss	−0.011098 [−0.019220, −0.003190]
	Factorization, direction loss	−0.011243 [−0.017042, −0.005510]
	Factorial interaction	−0.000145 [−0.004926, 0.004723]
	COMPLETE − zero	0.209871 [0.199759, 0.219853]
	COMPLETE − training mean	0.032800 [0.023835, 0.041714]
	COMPLETE − Morgan neighbor	0.428612 [0.373274, 0.491778]
OP3	COMPLETE − DIRECT	−0.067458 [−0.097148, −0.035930]
	Direction loss, direct head	−0.002684 [−0.005405, 0.000064]
	Direction loss, factorized head	−0.047836 [−0.076942, −0.017141]
	Factorization, no direction loss	−0.019622 [−0.043951, 0.004671]
	Factorial interaction	−0.045152 [−0.073609, −0.014610]
	COMPLETE − zero	0.019466 [−0.036184, 0.066935]
	COMPLETE − training mean	−0.167589 [−0.204430, −0.135330]

and applied fixed $\alpha = 0.5$. Parent blocks retained all donor/type rows and weights. The permutation control shuffled drug contexts within each human donor, requiring different parent identities and sharing the permutation across target types. This tests transfer to cell types conditional on the same three human donors; it does not test unseen human donors.

Shrinkage improved over the unshrunk mean in SciPlex3 and OP3, while the Cellarity primary interval included zero (Table 14). SciPlex3 CRM itself had interval [−0.048187, 0.178815], and K562 DLE was −0.035416; superiority over zero or global means is therefore not universal. These diagnostics describe predictive geometry, not biological causes. Intervals condition on fixed donors, plates, splits, and fitted bases; OP3's two tasks and the SciPlex3 tasks reuse responses. Assay-specific gene sets, normalization, rank, and measurement noise prevent treating their MSEs as a common-scale benchmark.

Table 13: Direction, projected energy ratio R_P, and errors at the primary rank. M_{raw} and M_{rec} denote raw-response and projected-reconstruction MSE. Dashes indicate untabulated quantities; no values are reconstructed from rounded ratios. Neural predictions in these two assays lie in the scoring subspace.

Dataset	Predictor	Mean cosine	R_P	M_{raw}	M_{rec}
MoABox	DIRECT	0.500860	0.274702	0.011353	0.011353
	DIRECT + direction	0.504455	0.259794	0.011344	0.011344
	Factorized, no direction	0.517528	0.247258	0.011479	0.011479
	COMPLETE	0.520777	0.262847	0.011472	0.011472
	Training mean	0.505832	0.177822	0.011844	0.011844
	Morgan neighbor	0.328530	0.986070	0.018136	0.016336
	Zero	–	0.000000	0.013854	0.013854
OP3	DIRECT	0.391431	0.297524	0.006878	0.006878
	DIRECT + direction	0.395044	–	0.006891	0.006891
	Factorized, no direction	0.427015	–	0.006968	0.006968
	COMPLETE	0.430989	0.205502	0.007186	0.007186
	Training type mean	0.497415	–	0.006418	–
	Morgan neighbor	0.317302	–	0.010371	–
	Zero	–	0.000000	0.007275	0.007275

Table 14: Cell and cell-type transfer. CRM is the exact-context response mean multiplied by the fixed value 0.5. Global and permuted controls receive the same multiplier. SciPlex3 coverage has interval [0.207363, 0.281627]. All effect intervals below are pointwise 95%.

Dataset	Predictor or contrast	DLE or ΔDLE [interval]
SciPlex3	Unshrunk mean	-0.836699
	CRM	0.084442
	Global mean	0.039772
	Permuted response memory	-0.427836
	CRM $-$ unshrunk	0.921141 [0.756911, 1.152927]
	CRM $-$ global	0.044670 [-0.080183, 0.134836]
	CRM $-$ permuted	0.512278 [0.474062, 0.557491]
Cellarity	Unshrunk mean	0.307715 [0.159673, 0.405051]
	CRM	0.310898 [0.216337, 0.373214]
	Global mean	0.023926 [0.014324, 0.031578]
	Permuted response memory	-0.120074 [-0.224861, -0.065503]
	CRM $-$ unshrunk	0.003182 [-0.032551, 0.057357]
	CRM $-$ global	0.286972 [0.198977, 0.345633]
	CRM $-$ permuted	0.430971 [0.370428, 0.497755]
OP3	Unshrunk mean	0.500219
	CRM	0.545974
	CRM $-$ unshrunk	0.045756 [0.016844, 0.075698]

OP3 raw-response MSE was 0.008857 for CRM and 0.009906 for the unshrunk mean; the paired difference was -0.001049 [-0.001174, -0.000923]. Cellarity raw MSE was 0.001136 (CRM), 0.001141 (unshrunk), 0.001373 (global), 0.001494 (permuted), and 0.001393 (zero).

Table 15: GSE92742 Joint OOD at rank 256. D and F denote DIRECT and COMPLETE predicted-donor means; subscript s denotes the prespecified 0.5 multiplier. Both mean cosine and R_P concern coordinates in the separate scoring basis. Context restoration can leave energy outside that basis.

Predictor	DLE	Projected skill	Mean cosine	R_P	M_{raw}	M_{rec}
Zero	0.000000	0.000000	–	0.000000	0.416575	0.416575
Global mean	0.017120	0.011078	0.122526	0.038084	0.411972	0.411960
D	0.008136	0.005265	0.110136	0.083941	0.414643	0.414382
F	0.003321	0.002149	0.115667	0.109716	0.415923	0.415680
D_s	0.025053	0.016212	0.110136	0.020985	0.409855	0.409821
F_s	0.029090	0.018824	0.115667	0.027429	0.408763	0.408733

Table 16: GSE92742 Joint OOD paired contrasts. The first DLE row is the prespecified primary comparison; remaining contrasts and raw-MSE intervals are descriptive. All intervals use 10,000 valid parent-block draws and condition on fixed cells, folds, bases, and three trained seeds. GSE92742 had prior response-analysis history.

Contrast	ΔDLE [95% interval]	ΔM_{raw} [95% interval]
$F_s - D_s$	0.004036 [0.002617, 0.005541]	-0.001092 [-0.001514, -0.000703]
$F - D$	-0.004815 [-0.007509, -0.002194]	0.001280 [0.000580, 0.001987]
$D_s - D$	0.016917 [0.013622, 0.020319]	-0.004788 [-0.005602, -0.003968]
$F_s - F$	0.025768 [0.021495, 0.030126]	-0.007160 [-0.008199, -0.006118]
$D_s - \mathrm{global}$	0.007934 [0.005477, 0.010518]	-0.002118 [-0.002813, -0.001457]
$F_s - \mathrm{global}$	0.011970 [0.008839, 0.015352]	-0.003210 [-0.004144, -0.002345]
$D_s - \mathrm{zero}$	0.025053 [0.021822, 0.028368]	-0.006720 [-0.007742, -0.005746]
$F_s - \mathrm{zero}$	0.029090 [0.024809, 0.033477]	-0.007812 [-0.009147, -0.006563]