SEASON COMBINATORIAL INTERVENTION PREDICTIONS WITH SALT & PEPER

Thomas Gaudelet, Alice Del Vecchio, Eli M Carrami, Juliana Cudini,
Chantriolnt-Andreas Kapourani,
Relation Therapeutics
London, UK
{thomas, alice, eli, juliana, andreas}@relationrx.com

Caroline Uhler,
Laboratory for Information and Decision Systems, MIT
Eric and Wendy Schmidt Center, Broad Institute of MIT and Harvard
Boston, USA
cuhler@mit.edu

Lindsay Edwards
Relation Therapeutics
London, UK
lindsay@relationrx.com

ABSTRACT

Interventions play a pivotal role in the study of complex biological systems. In drug discovery, genetic interventions (such as CRISPR base editing) have become central to both identifying potential therapeutic targets and understanding a drug's mechanism of action. With the advancement of CRISPR and the proliferation of genome-scale analyses such as transcriptomics, a new challenge is to navigate the vast combinatorial space of concurrent genetic interventions. Addressing this, our work concentrates on estimating the effects of pairwise genetic combinations on the cellular transcriptome. We introduce two novel contributions: SALT, a biologically-inspired baseline that posits the mostly additive nature of combination effects, and PEPER, a deep learning model that extends SALT's additive assumption to achieve unprecedented accuracy. Our comprehensive comparison against existing state-of-the-art methods, grounded in diverse metrics, and our out-of-distribution analysis highlight the limitations of current models in realistic settings. This analysis underscores the necessity for improved modelling techniques and data acquisition strategies, paving the way for more effective exploration of genetic intervention effects.

1 INTRODUCTION

Interventions are an essential tool when striving to decode the functioning of complex systems and in particular, to identify causal dependencies (as exemplified by causal learning theory (Eberhardt & Scheines, 2007)). In the field of biology, researchers routinely use chemical or genetic perturbations to probe and manipulate cellular systems. Notably, interventions are of fundamental importance to drug discovery, where the goal is to design an intervention that can suppress or even reverse disease-related biological processes in targeted and specific ways. Genetic interventions are of special interest as they allow more precise manipulation of individual genes, as opposed to (for example) small molecules that often have multiple cellular targets and wide-ranging effects that are poorly understood (Hopkins, 2008; Lin et al., 2017). Recent advancements in CRISPR technologies (that allow precise editing of genetic sequences) have empowered scientists to unravel the intricate workings of biological systems (Dixit et al., 2016; Frangieh et al., 2021; Norman et al., 2019; Replogle et al., 2022). From a drug discovery standpoint, manipulating specific elements and observing outcomes enables scientists to pinpoint promising hypotheses for the development of new therapeutics (Liu et al., 2020; Chavez et al., 2023; Kim & Lee, 2023) or understand the mode of action of existing ones (Jost & Weissman, 2018).

The application of deep learning to the task of estimating intervention effects in biological systems is a relatively recent development (Ma et al., 2018; Lotfollahi et al., 2019; Zhang et al., 2023). With the emergence of large, single-cell resolution datasets, there has been a rapid increase in research activity aiming to model gene expression patterns (Lopez et al., 2018) and how they are impacted by interventions (Roohani et al., 2023; Lotfollahi et al., 2023). However, as elsewhere, uncertainty remains about the best task framing and most appropriate metrics. For example, straightforward performance metrics (such as the RMSE or correlation coefficient between predicted and actual gene expression) often fail to appropriately weight signals that are likely to be of importance to biologists. By contrast, selected researchers have moved to using custom performance metrics (e.g. the error on only the tails of the predicted distributions) that capture more relevant detail. As a modality of interest, the transcriptome has a number of benefits. First, there are established and widely used technologies, such as scRNA-seq, that accurately measure gene expression levels at single cell resolution. Second, transcriptome variation captures low-level cellular states that often proxy higher-level effects, acting as a generic foundation for predictive models designed for specific phenotypic endpoints (this concept – that a phenotype can be effectively modelled by a gene expression pattern, sometimes referred to as a 'gene signature' – is widely used in drug discovery and biology). Because of this, we can motivate research and model development around transcriptomics using the same logic as foundation models in machine learning: addressing a low-level, generic task that can be built upon for specific applications.

Progress in CRISPR technologies that measure interventional impacts on a cell's transcriptome has opened the door to genome-scale analyses in which most / many genes can be intervened upon individually (Replogle et al., 2022). Although conducting a genome-wide perturb-seq (the name used for an experiment where all genes are intervened upon, and the transcriptomic effect measured at the single-cell level) screen remains prohibitively expensive at present, it is anticipated that costs will drop considerably in the coming years. The primary challenge moving forward is expected to revolve around understanding the complex effects of *combinatorial interventions* given the vast number of potential combinations (approximately $\binom{2\times10^4}{n}$), where n is the number of simultaneous perturbations). Further, studying combinatorial interventions becomes critical due to the limitations of single-target approaches in tackling the complexity of biological systems (Hopkins, 2008). Combinatorial strategies perturbing multiple genes may help unravel the organisation of complex networks, overcoming redundancies and disrupting adaptive responses. This is perhaps best exemplified by the combinatorial therapeutics leveraged to treat some forms of cancer (Shen et al., 2017; Kim & Lee, 2023).

In this work, we focus only on pairwise genetic combinations due to existing constraints in available public datasets. Additionally, as we expect the cost of genome-scale screens to decrease, we assume that we have already measured the outcomes of targeting each individual gene. Our key contributions are

(1) a careful comparison across current state-of-the-art methods using different metrics and grounded with SALT, a biologically-motivated, non-parametric baseline making the assumption that combination effects are mostly additive,

(2) PEPER, a deep learning model built from the additive inductive bias from SALT and achieving state-of-the-art performances,

(3) an out-of-distribution analysis indicating that current solutions are ill-suited to address realistic scenarios and calling for better modelling approaches and data acquisition strategies.

2 RELATED WORK

Context transfer. The task at hand can be related to the context transfer problem, whereby the effect of an intervention has been measured in some context, for instance a cell line, and we aim to predict the effect it would have in a different context (e.g. a different cell line). This task has been garnering interest from the deep learning community for some years, particularly for transcriptomics response prediction (Lotfollahi et al., 2019; Wu et al., 2022; Lotfollahi et al., 2023). Notably, both context transfer and combinatorial interventions tasks can be addressed in the same framework, as

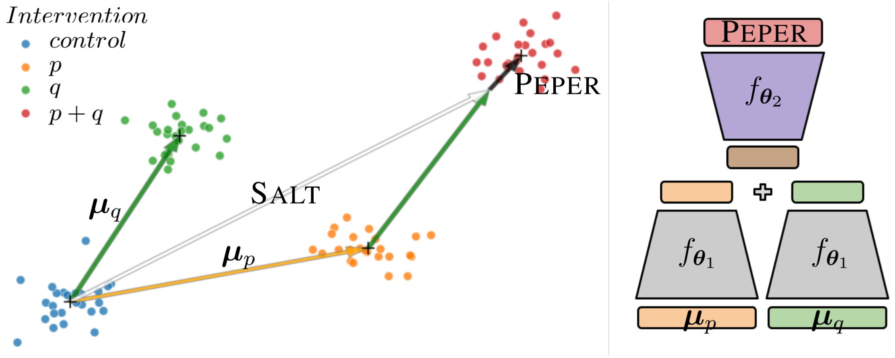

Figure 1: Illustration of SALT and PEPER predictions for the combinations of interventions on genes p and q.

exemplified by CPA (Lotfollahi et al., 2023). Conceptually, the subtle difference between the two tasks arise from the *sequential* nature of context transfer, the context precedes the intervention, as opposed to the *parallel* nature of the task at hand, i.e. all targets are affected simultaneously.

Potential outcomes and counterfactual estimation. A large majority of approaches (Lotfollahi et al., 2019; Wu et al., 2022; Lotfollahi et al., 2023) tackling either the context transfer or combinatorial intervention tasks aim to predict what state an input cell, typically a control cell, would have been in had it been treated or intervened upon. These can naturally be viewed under the lens of counterfactual estimation from the potential outcomes causal framework (Rubin, 2005), which is concerned with resolving similar statements in order to estimate causal effects.

Modelling cellular dynamics. Generating counterfactual predictions would collapse to a simpler task given a complete model of cellular dynamics that can be sampled at will. Despite a long history of this class of models in metabolism (e.g. (Edwards et al., 2011)), dynamic models of the complexity of transcription are rarer (Erbe et al., 2023; Bhaskar et al., 2023; Ishikawa et al., 2023); the lack of large-scale, time-resolved, interventional datasets has limited the development of methods targeting our task of interest through the modelling of cellular dynamics.

3 METHODS

3.1 NOTATION

In the following, we define $\mathbf{X}_p$ as the matrix representing cells measured after gene p intervention and $\tilde{\mathbf{X}}_p^m$ as the matrix of estimated cells for the same intervention predicted by method m, with both matrices structured such that rows correspond to individual cells and columns to measured genes. We use the letter c to refer to control cells which have only been treated with non-targeting guides and are typically used as controls experimentally. We use lower case $\mathbf{x}$ to denote a single cell, i.e. a row of a matrix as defined previously. If the intervention correspond to a pairwise combination, we write $\mathbf{X}_{p+q}$ where p and q indicate the two genes targeted.

We introduce the *average intervention effect vector* as the average difference between perturbed cells and control cells, we denote it by $\boldsymbol{\mu}_p = \overline{\mathbf{X}}_p - \overline{\mathbf{X}}_c$, where $\overline{\mathbf{X}}$ denotes the mean over the columns of $\mathbf{X}$, also known as *pseudo-bulk*, and gives the average expression of each gene across cells.

3.2 EXISTING ART

We use two published baselines to ground our work: GEARS (Roohani et al., 2023), the current state-of-the-art, and CPA (Lotfollahi et al., 2023).

CPA follows from a long lineage of autoencoder architectures applied to single-cell (sc)RNA-seq data (Lopez et al., 2018; Lotfollahi et al., 2019; Wu et al., 2022). In CPA, the authors model combinatorial effects as an additive process in a learnt (potentially nonlinear) latent space.

By contrast, GEARS takes its inspiration from the causal and graph neural network literature, relying on graph mutilation of biological graph priors and the message passing paradigm. In practice, GEARS applies a fixed, intervention-specific translation to an input cell. Formally, given a control cell input (i.e. a cell treated with non-targeting guides), an intervention targeting gene p and q jointly, GEARS's counterfactual prediction can be written as

$$\tilde{\mathbf{x}}_{p+q}^{GEARS} = \mathbf{x}_c + g_{\boldsymbol{\phi}}(p, q|\mathcal{G}),$$

where $g_{\boldsymbol{\phi}}$ is a graph neural network architecture with parameters $\boldsymbol{\phi}$ and $\mathcal{G}$ denotes the biological graph priors.

3.3 SALT

The aim of the study of combinations is often the identification of interactions such as synergy or antagonism (Norman et al., 2019; Bertin et al., 2023). However, these interactions are fairly rare (Martin, 2023) and notably genetic variations tend to be largely additive (Hill et al., 2008). As such, we expect that a simple additive heuristic would provide a strong non-parametric baseline for the task. We refer to it as SALT (Simply Assume Linear combinations of Transcriptomes) and an illustration can be seen on Figure 1. Following our notations, we can write SALT as

$$\tilde{\mathbf{x}}_{p+q}^{\text{SALT}} = \mathbf{x}_c + \boldsymbol{\mu}_p + \boldsymbol{\mu}_q.$$

It is worth noting that similar vector arithmetic baselines in transcriptome space have been used for the context transfer task (Lotfollahi et al., 2019).

3.4 PEPER

We spice up SALT by introducing a learnable non-linear correction term. This is reminiscent of statistical models that add a non-interactive term with interactive terms to model and study synergistic effects (Norman et al., 2019; Rønneberg et al., 2021). The resulting method – Perturbation Effect Prediction by Error Reduction (PEPER) – can be written as

$$\tilde{\mathbf{x}}_{p+q}^{\text{PEPER}} = \tilde{\mathbf{x}}_{p+q}^{\text{SALT}} + f_{\boldsymbol{\theta}}(\boldsymbol{\mu}_p, \boldsymbol{\mu}_q),$$

where $f_{\boldsymbol{\theta}}$ is a neural network with parameters $\boldsymbol{\theta}$. In practice, we decompose $f_{\boldsymbol{\theta}}$ as follows

$$f_{\boldsymbol{\theta}}(\boldsymbol{\mu}_p, \boldsymbol{\mu}_q) = f_{\boldsymbol{\theta}_2}(f_{\boldsymbol{\theta}_1}(\boldsymbol{\mu}_p) + f_{\boldsymbol{\theta}_1}(\boldsymbol{\mu}_q)),$$

with both $f_{\boldsymbol{\theta}_{1,2}}$ corresponding to multi-layer perceptrons. See Figure 1 for illustration. From the definitions, it is important to remark that GEARS, SALT and PEPER would at best capture the average intervention effect vector, $\boldsymbol{\mu}_{p+q}$.

To train PEPER we use the Central Moment Discrepancy (CMD) which is a distributional loss that compares moments (Zellinger et al., 2016). Because PEPER only models the average effect of interventions, we only use the first moment to compute the loss. Note that with this loss, we batch the data by intervention, i.e. each batch contains data from a single intervention.

4 EXPERIMENTS

4.1 EVALUATION

Datasets. We use two distinct CRISPR interventional datasets to evaluate prediction of the outcome of genetic combinations for our experiments (Norman et al., 2019; Wessels et al., 2023). Details can be found in Appendix A.

Splits. For each datasets, we define two in-distribution splits, *in-distribution-25* and *in-distribution-75*, and one out-of-distribution split based on similarities between intervention effects in transcriptomics space, reasoning that the task is rendered more challenging if a model is forced to estimate out-of-distribution transcriptomic responses. We detail the split creation in Appendix A.

Table 1: RMSE results on the in-distribution-25 splits. The numbers correspond to average score and standard deviation (in parenthesis) across 5 different seeds.

	NORMAN-DATASET		WESSELS-DATASET	
Model	**All**	**Top 20 ES**	**All**	**Top 20 ES**
Peper	0.030 (0.002)	**0.104 (0.006)**	**0.034 (0.001)**	**0.094 (0.002)**
GEARS	**0.029 (0.001)**	0.117 (0.0099)	**0.034 (0.001)**	0.101 (0.004)
CPA	0.032 (0.002)	0.119 (0.008)	0.036 (0.000)	0.131 (0.007)
Salt	0.034	0.124	0.061	0.143

Table 2: RMSE results on the in-distribution-75 splits. The numbers correspond to average score and standard deviation (in parenthesis) across 5 different seeds.

	NORMAN-DATASET		WESSELS-DATASET	
Model	**All**	**Top 20 ES**	**All**	**Top 20 ES**
Peper	**0.030 (0.000)**	**0.118 (0.000)**	0.043 (0.001)	**0.112 (0.005)**
GEARS	0.038 (0.001)	0.145 (0.011)	**0.038 (0.002)**	0.113 (0.005)
CPA	0.032 (0.000)	0.139 (0.004)	0.071 (0.004)	0.164 (0.018)
Salt	0.033	0.134	0.065	0.160

Hyperparameters selection. We tune each model on each split, using the validation sets to identify the best configurations. For CPA and GEARS, we use the tuning configurations provided by the authors. For Peper, we provide the hyperparameters and associated ranges considered in Appendix D. We retrain each model post-tuning with the identified best configurations using 5 different seeds.

Metrics. Following recommendations from Ji et al. (2023), we use root mean squared error (RMSE) as our primary metric to evaluate models' predictions (see Appendix B for details). We also include energy distance scores, a distributional metric, in the Appendix. Following previous work (Roohani et al., 2023), we calculate metrics in two ways: for all genes (labelled All in tables) and for the top 20 genes ranked by absolute effect sizes (labeled Top 20 ES in tables). These sets of 20 genes are naturally intervention-specific. As noted in the Introduction, metrics computed on a smaller, meaningful set of genes are generally preferred by the community as many lowly expressed genes are noisy and the effect of genetic interventions tend to be observed on a small set of genes (Peidli et al., 2024). We include both for completeness and report metrics computed on the held-out test sets using the 5 models obtained from the different seeds.

4.2 Results

Peper achieves state-of-the-art for in-distribution splits, as shown in Tables 1 and 2, indicating that our recipe relying on a strong inductive bias motivated by biological observations is an effective approach to the problem. The relatively good performances of the Salt heuristics also speak to the quality of the inductive bias. In particular, it is notable that Salt beats both CPA and GEARS in the hardest in-distribution-75 split of the norman-dataset. Our results demonstrate that the inductive bias provided by Salt's prior is more useful to the task at hand than the priors that GEARS relies on.

We investigate further model performances on the in-distribution-75 split for the norman-dataset by grouping interventions based on the types of interaction identified by Norman et al. (2019).The definition for the different interaction types can be found in Appendix E. Overall, we observe similar performances across the different groups, with Peper outperforming all other methods (see Figure 2). However, the *potentiation* group stands out with significantly worse performances for all models. Note that potentiation occurs when one of the two interventions combined has no effect on its own but enhances the effect of specific interventions when combined. Given this definition, the results could be explained by the fact that the average intervention effect vector is uninformative and does not help predicting the enhancing effect of the intervention. This result exemplifies the limitation of the inductive bias provided by Salt.

It is worth remarking that the only method that actually models single-cell variation, CPA, does not benefit from it. One could argue that this is an artifact of the metric choice as RMSE only compares

Table 3: RMSE results on held-out clusters splits. The numbers correspond to average score and standard deviation (in parenthesis) across 5 different seeds.

	Norman-dataset		**Wessels-dataset**	
Model	**All**	**Top 20 ES**	**All**	**Top 20 ES**
Peper	0.032 (0.000)	0.162 (0.000)	**0.049** (0.003)	0.189 (0.005)
GEARS	0.041 (0.001)	0.274 (0.01)	0.053 (0.003)	0.3 (0.068)
CPA	0.051 (0.024)	0.330 (0.142)	0.070 (0.014)	0.566 (0.096)
Salt	**0.031**	**0.157**	0.065	**0.181**

the first moment of the predicted and actual cell distributions. However, the gap with all three other approaches grows even larger when considering the energy distributional distance as reported in Appendix Tables 4 and 5. This might suggests that the noise-to-signal ratio in single-cell data is too high to model higher moments accurately for the task and evaluation we chose. In other terms, architectures that only model average effects and otherwise conserve the higher moments of the distribution of control cells, such as Peper and GEARS, may be advantaged due to the noisiness of the single-cell data.

When out-of-distribution, all models have significant performance drops and Salt becomes the best performing model for all but one metric (see Table 3). As Peper simply augments Salt, the performance drop is less striking when compared to GEARS. However, the results are symptomatic of an overfit to the seen data and a lack of generalisability beyond the support from the training set.

5 Discussion & Conclusion

In this paper, we investigated the estimation of the effect of genetic interventions targeting pairs of genes in diverse data scenarios. We proposed to use a biologically-motivated, non-parametric baseline, that we call Salt, to help put results into perspective. We introduced Peper, a deep learning architecture leveraging Salt's prior and demonstrated that it achieves state-of-the-art performances on our key metric.

However, our out-of-distribution splits uncovered significant weaknesses in all current modeling approaches when predicting previously unseen regions of the transcriptional response manifold. Unfortunately, this is precisely the setting in which these models are likely to be the most useful to biologists - predicting previously unseen phenotypic responses to combinatorial perturbations. This indicates a need for further research, accompanied by refined approaches to generating data. Creating situations where we consistently find ourselves within an in-distribution scenario for a given biological question is a challenging endeavour. One promising avenue is through efficient data acquisition strategies, such as active learning (Settles, 2012; Bertin et al., 2023; Scherer et al., 2022), which provides a level of control over data generation and experimental design.

Concerning model development, the incorporation of inductive biases derived from prior knowledge, typified by Bayesian learning theory (Fortuin, 2022) and here exemplified by Peper, has proven to be a robust approach in many applications. However, the observed performance drops in out-of-distribution splits highlight the limitations with current priors, which may not always be of high quality or suitable for generalization. The insufficiency of the inductive bias from Salt becomes clear from the performance drop over combinations that interact through potentiation. Moreover, the use of priors derived from measurements taken at a single timepoint has inherent limitations. These *static* priors lack insight into any underlying dynamics, and interactions that could arise were sequential measures available. To address these issues and enhance performance on this task, we need both better priors, better data and improved architectures capable of leveraging or rejecting inappropriate prior knowledge.

These challenges can also be indicative of a broader issue related to the scarcity of data, notably around combinatorial interventions. We do not yet know whether gene interactions tend to be conserved across different contexts, facilitating the use of transfer learning approaches to fill in the gap, or whether significant novel data would need to be generated for each new context. In all cases, we expect the situation to improve as the volume of publicly available data increases to a scale of

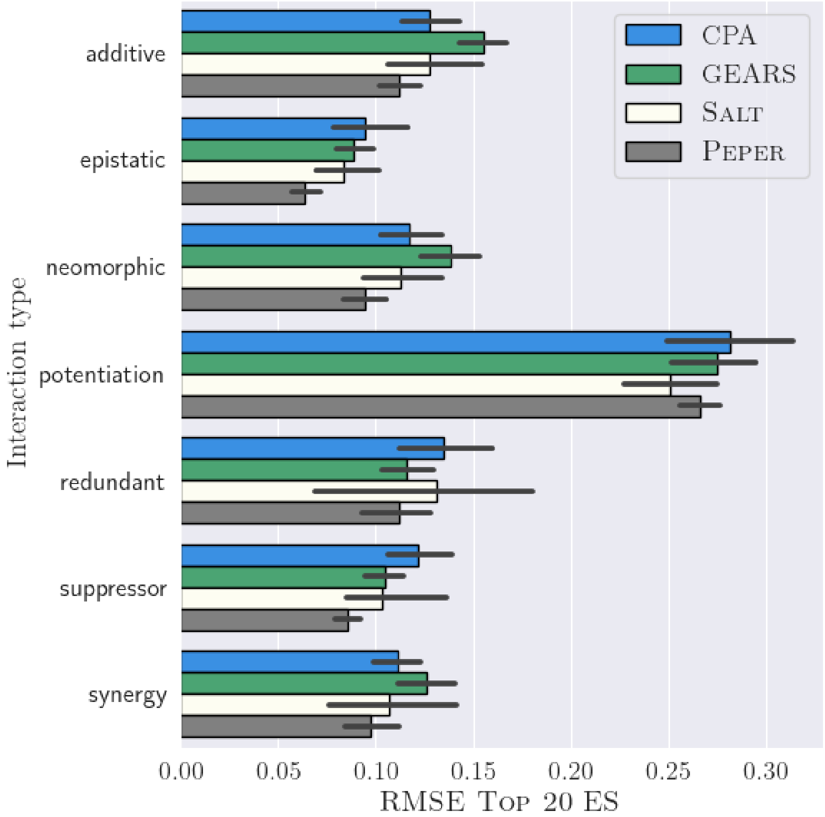

Figure 2: Score breakdown of GEARS, CPA, SALT, and PEPER (lower is better) based on intervention subgroups identified by Norman et al. (2019). The definition for each label can be found in Appendix E. As we can see, there are some variability in performances based on interaction type, but the potentiation interaction type stands out particularly.

dataset sufficient to train large modern architectures. In the interim, the community should employ clever inductive biases and priors while efficiently acquiring data to address specific questions.

On a conceptual note, our results suggest that modelling higher moments of single-cell distributions may be detrimental to overall performances on our metrics of interest. This should be caveated by the fact that our observation rests only on a single model, namely CPA. However, it can be further compounded by the fact that Ji et al. (2023) identified mean-squared error between pseudo-bulks as the best metric to compare single-cell transcriptomics distributions in multiple applications. Put together, these two results might imply limited utility gain from considering higher moments for selected applications; for these, a more productive focus might be centered on (pseudo-)bulk data. However, this statement needs to be tested in specific cases, and it is important to recognise that it might not hold for many possible downstream applications. Lastly, and perhaps most importantly of all, the choice of metrics should be carefully thought through such that improvement on a chosen metric captures meaningful improvements in practice.

ACKNOWLEDGMENTS

This work would not have been possible without the support of the entire Relation Therapeutics team, with special mentions for Edith M. Hessel, Marco P. Licciardello, Mohammad Lotfollahi, Rebecca E. McIntyre, Charles E.S. Roberts, Jake P. Taylor-King, and John C. Whittaker, for the many insightful discussions.

References

Paul Bertin, Jarrid Rector-Brooks, Deepak Sharma, Thomas Gaudelet, Andrew Anighoro, Torsten Gross, Francisco Martínez-Peña, Eileen L Tang, MS Suraj, Cristian Regep, et al. Recover identifies synergistic drug combinations in vitro through sequential model optimization. *Cell Reports Methods*, 3(10), 2023.

Dhananjay Bhaskar, Sumner Magruder, Edward De Brouwer, Aarthi Venkat, Frederik Wenkel, Guy Wolf, and Smita Krishnaswamy. Inferring dynamic regulatory interaction graphs from time series data with perturbations. *arXiv preprint arXiv:2306.07803*, 2023.

Michael Chavez, Xinyi Chen, Paul B Finn, and Lei S Qi. Advances in crispr therapeutics. *Nature Reviews Nephrology*, 19(1):9–22, 2023.

Atray Dixit, Oren Parnas, Biyu Li, Jenny Chen, Charles P Fulco, Livnat Jerby-Arnon, Nemanja D Marjanovic, Danielle Dionne, Tyler Burks, Raktima Raychowdhury, et al. Perturb-seq: dissecting molecular circuits with scalable single-cell rna profiling of pooled genetic screens. *Cell*, 167(7): 1853–1866, 2016.

Frederick Eberhardt and Richard Scheines. Interventions and causal inference. *Philosophy of Science*, 74(5):981–995, 2007.

Lindsay M. Edwards, Houman Ashrafian, and Bernard Korzeniewski. In silico studies on the sensitivity of myocardial pcr/atp to changes in mitochondrial enzyme activity and oxygen concentration. *Mol. BioSyst.*, 7:3335–3342, 2011.

Rossin Erbe, Genevieve Stein-O'Brien, and Elana J Fertig. Transcriptomic forecasting with neural ordinary differential equations. *Patterns*, 4(8), 2023.

Vincent Fortuin. Priors in bayesian deep learning: A review. *International Statistical Review*, 90(3): 563–591, 2022.

Chris J Frangieh, Johannes C Melms, Pratiksha I Thakore, Kathryn R Geiger-Schuller, Patricia Ho, Adrienne M Luoma, Brian Cleary, Livnat Jerby-Arnon, Shruti Malu, Michael S Cuoco, et al. Multimodal pooled perturb-cite-seq screens in patient models define mechanisms of cancer immune evasion. *Nature Genetics*, 53(3):332–341, 2021.

William G Hill, Michael E Goddard, and Peter M Visscher. Data and theory point to mainly additive genetic variance for complex traits. *PLoS Genetics*, 4(2):e1000008, 2008.

Andrew L Hopkins. Network pharmacology: the next paradigm in drug discovery. *Nature Chemical Biology*, 4(11):682–690, 2008.

Masato Ishikawa, Seiichi Sugino, Yoshie Masuda, Yusuke Tarumoto, Yusuke Seto, Nobuko Taniyama, Fumi Wagai, Yuhei Yamauchi, Yasuhiro Kojima, Hisanori Kiryu, et al. Renge infers gene regulatory networks using time-series single-cell rna-seq data with crispr perturbations. *Communications Biology*, 6(1):1290, 2023.

Yuge Ji, Tessa Green, Stefan Peidli, Mojtaba Bahrami, Meiqi Liu, Luke Zappia, Karin Hrovatin, Chris Sander, and Fabian Theis. Optimal distance metrics for single-cell rna-seq populations. *bioRxiv*, pp. 2023–12, 2023.

Marco Jost and Jonathan S Weissman. Crispr approaches to small molecule target identification. *ACS Chemical Biology*, 13(2):366–375, 2018.

Yuna Kim and Hyeong-Min Lee. Crispr-cas system is an effective tool for identifying drug combinations that provide synergistic therapeutic potential in cancers. *Cells*, 12(22):2593, 2023.

Ann Lin, Christopher J Giuliano, Nicole M Sayles, and Jason M Sheltzer. Crispr/cas9 mutagenesis invalidates a putative cancer dependency targeted in on-going clinical trials. *Elife*, 6:e24179, 2017.

Dan Liu, Xuan Zhao, Anqun Tang, Xiyue Xu, Shuci Liu, Li Zha, Wen Ma, Junnian Zheng, and Ming Shi. Crispr screen in mechanism and target discovery for cancer immunotherapy. *Biochimica et Biophysica Acta (BBA)-Reviews on Cancer*, 1874(1):188378, 2020.

Romain Lopez, Jeffrey Regier, Michael B Cole, Michael I Jordan, and Nir Yosef. Deep generative modeling for single-cell transcriptomics. *Nature Methods*, 15(12):1053–1058, 2018.

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, et al. Predicting cellular responses to complex perturbations in high-throughput screens. *Molecular Systems Biology*, pp. e11517, 2023.

Jianzhu Ma, Michael Ku Yu, Samson Fong, Keiichiro Ono, Eric Sage, Barry Demchak, Roded Sharan, and Trey Ideker. Using deep learning to model the hierarchical structure and function of a cell. *Nature Methods*, 15(4):290–298, 2018.

Olwenn V Martin. Synergistic effects of chemical mixtures: how frequent is rare? *Current Opinion in Toxicology*, pp. 100424, 2023.

Thomas M Norman, Max A Horlbeck, Joseph M Replogle, Alex Y Ge, Albert Xu, Marco Jost, Luke A Gilbert, and Jonathan S Weissman. Exploring genetic interaction manifolds constructed from rich single-cell phenotypes. *Science*, 365(6455):786–793, 2019.

Stefan Peidli, Tessa D Green, Ciyue Shen, Torsten Gross, Joseph Min, Samuele Garda, Bo Yuan, Linus J Schumacher, Jake P Taylor-King, Debora S Marks, et al. scperturb: harmonized single-cell perturbation data. *Nature Methods*, pp. 1–10, 2024.

Joseph M Replogle, Reuben A Saunders, Angela N Pogson, Jeffrey A Hussmann, Alexander Lenail, Alina Guna, Lauren Mascibroda, Eric J Wagner, Karen Adelman, Gila Lithwick-Yanai, et al. Mapping information-rich genotype-phenotype landscapes with genome-scale perturb-seq. *Cell*, 185(14):2559–2575, 2022.

Leiv Rønneberg, Andrea Cremaschi, Robert Hanes, Jorrit M Enserink, and Manuela Zucknick. bayesynergy: flexible bayesian modelling of synergistic interaction effects in in vitro drug combination experiments. *Briefings in Bioinformatics*, 22(6):bbab251, 2021.

Yusuf Roohani, Kexin Huang, and Jure Leskovec. Predicting transcriptional outcomes of novel multigene perturbations with gears. *Nature Biotechnology*, pp. 1–9, 2023.

Donald B Rubin. Causal inference using potential outcomes: Design, modeling, decisions. *Journal of the American Statistical Association*, 100(469):322–331, 2005.

Paul Scherer, Thomas Gaudelet, Alison Pouplin, Jyothish Soman, Lindsay Edwards, Jake P Taylor-King, et al. Pyrelational: A library for active learning research and development. *arXiv preprint arXiv:2205.11117*, 2022.

Burr Settles. Uncertainty sampling. In *Active Learning*, Synthesis Lectures on Artificial Intelligence and Machine Learning, pp. 11–21. Morgan & Claypool Publishers, 2012.

John Paul Shen, Dongxin Zhao, Roman Sasik, Jens Luebeck, Amanda Birmingham, Ana Bojorquez-Gomez, Katherine Licon, Kristin Klepper, Daniel Pekin, Alex N Beckett, et al. Combinatorial crispr–cas9 screens for de novo mapping of genetic interactions. *Nature Methods*, 14(6):573–576, 2017.

Gordon K Smyth. Limma: linear models for microarray data. In *Bioinformatics and Computational Biology Solutions using R and Bioconductor*, pp. 397–420. Springer, 2005.

Jing Tang, Krister Wennerberg, and Tero Aittokallio. What is synergy? the saariselkä agreement revisited. *Frontiers in Pharmacology*, 6:181, 2015.

Vincent A Traag, Ludo Waltman, and Nees Jan Van Eck. From louvain to leiden: guaranteeing well-connected communities. *Scientific Reports*, 9(1):5233, 2019.

Hans-Hermann Wessels, Alejandro Méndez-Mancilla, Yuhan Hao, Efthymia Papalexi, William M Mauck III, Lu Lu, John A Morris, Eleni P Mimitou, Peter Smibert, Neville E Sanjana, et al. Efficient combinatorial targeting of rna transcripts in single cells with cas13 rna perturb-seq. *Nature Methods*, 20(1):86–94, 2023.

Yulun Wu, Layne C Price, Zichen Wang, Vassilis N Ioannidis, Rob Barton, and George Karypis. Variational causal inference. In *NeurIPS 2022 Workshop on Causality for Real-world Impact*, 2022.

Werner Zellinger, Thomas Grubinger, Edwin Lughofer, Thomas Natschläger, and Susanne Saminger-Platz. Central moment discrepancy (cmd) for domain-invariant representation learning. In *International Conference on Learning Representations*, 2016.

Jiaqi Zhang, Chandler Squires, Kristjan Greenewald, Akash Srivastava, Karthikeyan Shanmugam, and Caroline Uhler. Identifiability guarantees for causal disentanglement from soft interventions. *arXiv preprint arXiv:2307.06250*, 2023.

A Data

Datasets. The first dataset was generated with CRISPR activation interventions in K562 cells (Norman et al., 2019), and comprise 124 combinations derived from a pool of 106 individual potential targets. The second dataset utilises CRISPR Cas13 technology in THP1 cells and encompasses 142 combinations within a set of 27 genes (Wessels et al., 2023). In the following, we refer to each dataset as norman-dataset and wessels-dataset. We detail our data processing in Appendix A.

Pre-processing. We follow standard pre-processing steps for both datasets. We first remove cells that do not pass the usual quality control thresholds in terms of total number of counts, number of genes expressed, and percentage of mitochondrial counts. We also remove the subset of cells associated to a target if the subset is too small. Then, we filter out measured genes if they are not expressed consistently for at least one of the cell subsets associated to targets. Finally, we normalise cell counts to a 10^4 total and log-transform the data.

Splits. To create our splits, we first cluster interventions based on pseudo-bulk profiles using the Leiden algorithm (Traag et al., 2019). From this we obtain clusters that group interventions that have relatively similar average responses in terms of transcriptomics. We use the clustering to define two *in-distribution* splits and one *out-of-distribution* split. For the *in-distribution* splits, we define train, validation, and test sets from the set of combinatorial interventions within each cluster. The two splits are defined with different amounts of held combinations, we hold-out either 25% of combinations (i.e. 12.5% in val and test sets) or 75% (i.e. 37.5% in val and test sets). Henceforth, we will refer to these splits as *in-distribution-25* and *in-distribution-75*. We define *out-of-distribution* splits by holding out entire clusters of combinations for validation and test, choosing the farthest clusters away from the cluster containing control cells. Note that all cells associated to single interventions are added to the training sets for all splits. Visualisations based on UMAP of the various splits can be found in the Appendix (Figure 3 and Figure 4).

B Log-fold change and RMSE metric

As discussed in the main document, we compute the root mean-square error between predicted and actual log-fold change vectors with respect to the control distribution. Specifically, we use Limma's definition of log-fold change (Smyth, 2005), which defines the log-fold change associated to perturbation targeting gene p with respect to the control distribution as

$$\boldsymbol{\delta}_p = \frac{\boldsymbol{\mu}_p}{\log(2)},$$

assuming the data is log-transformed in base 10. This formulation of log-fold change from Limma is motivated by the assumption that gene expression follows a normal distribution in log space, rather than in count space. We then denote for each perturbation p the set of k genes that have the greatest absolute log-fold change value by $\mathcal{S}_{p,k}$. With these definitions, we can simply write our metrics over a set of perturbations $\mathcal{P}$ as

$$\mathrm{RMSE}_{\mathrm{ALL}}(\mathcal{P}) = \frac{1}{|\mathcal{P}|}\sum_{p\in\mathcal{P}}\sqrt{\frac{\sum_{i=1}^{N}\left(\boldsymbol{\delta}_p[i] - \tilde{\boldsymbol{\delta}}_p[i]\right)^2}{N}},$$

$$\mathrm{RMSE}_{\mathrm{TOP\ 20\ ES}}(\mathcal{P}) = \frac{1}{|\mathcal{P}|}\sum_{p\in\mathcal{P}}\sqrt{\frac{\sum_{i\in\mathcal{S}_{p,20}}\left(\boldsymbol{\delta}_p[i] - \tilde{\boldsymbol{\delta}}_p[i]\right)^2}{|\mathcal{S}_{p,20}|}},$$

where N is the total number of genes post-QC, $|\cdot|$ denotes the cardinal of a set, and $\tilde{\delta}$ denotes predicted log-fold changes.

C Additional results.

To get a view of the prediction quality in distributional terms, we compute the energy distance between predicted and actual distribution of cells on the test set for both in-distribution splits. The results are reported in Tables 4 and 5. The results are mostly consistent with the RMSE scores, apart from CPA which performs significantly worse on this metric. As discussed in the main document, this suggest that modelling single-cell variations is detrimental when compared with methods which conserve the higher moments from the control cell distribution.

D Peper hyperparameters tuning configuration.

We list below the name and range of the hyperparameters tuned for Peper on each split.

- batch size: $[256, 6136]$,
- batch accumulation: $[1, 100]$,
- learning rate: $\left[10^{-5}, 10^{-3}\right]$,
- weight decay: $\left[0, 10^{-4}\right]$,
- number of encoder layers $(f_{\boldsymbol{\theta}_1})$: $[1, 4]$,
- number of decoder layers $(f_{\boldsymbol{\theta}_2})$: $[1, 4]$,
- latent dimension: $[1000, 6500]$,
- encoder and decoder hidden dimension: $[1000, 6500]$.

E Glossary.

Epistasis A combination is epistatic if the effect of targeting one of the genes mask the effect of targeting the other.

Neomorphism A combination is neomorphic if the resulting effect is atypically new or unexpected given the effects observed from intervening on each gene individually.

Potentiation Potentiation is a sub-type of synergy in that the effect of the combination is greater than expected. The defining aspect is that one of the target does not have any effect on its own and only enhances the effect of targeting the other.

Redundancy A combination is redundant when the two targets have similar effect when targeted individually and the combination leads also to a similar output.

Suppression There is suppression when the combined intervention on two genes attenuates the effect of each gene when individually targeted.

Synergy A combination is synergistic when the effect of targeting both genes together is greater than expected when compared to the effects of targeting each gene individually. There exist multiple mathematical definition of synergy, each making different assumption about what should be expected, e.g. Bliss independence and Loewe additity (Tang et al., 2015).

Table 4: Energy distance results on in-distribution-25 split. The numbers correspond to average score and standard deviation (in parenthesis) across 5 different seeds.

	NORMAN-DATASET		**WESSELS-DATASET**	
MODEL	**ALL**	**TOP 20 ES**	**ALL**	**TOP 20 ES**
PEPER	0.151 (0.013)	**0.056 (0.003)**	0.151 (0.002)	**0.041 (0.001)**
GEARS	**0.144 (0.009)**	0.062 (0.007)	**0.126 (0.003)**	**0.041 (0.002)**
CPA	1.883 (0.048)	0.129 (0.012)	2.365 (0.059)	0.181 (0.010)
SALT	0.145	0.065	0.292	0.059

Table 5: Energy distance results on in-distribution-75 split. The numbers correspond to average score and standard deviation (in parenthesis) across 5 different seeds.

	NORMAN-DATASET		**WESSELS-DATASET**	
MODEL	**ALL**	**TOP 20 ES**	**ALL**	**TOP 20 ES**
PEPER	0.151 (0.001)	**0.067 (0.000)**	0.211 (0.006)	0.050 (0.003)
GEARS	0.178 (0.005)	0.085 (0.007)	**0.152 (0.009)**	**0.048 (0.003)**
CPA	1.873 (0.028)	0.132 (0.005)	2.467 (0.133)	0.147 (0.020)
SALT	**0.146**	0.075	0.331	0.069

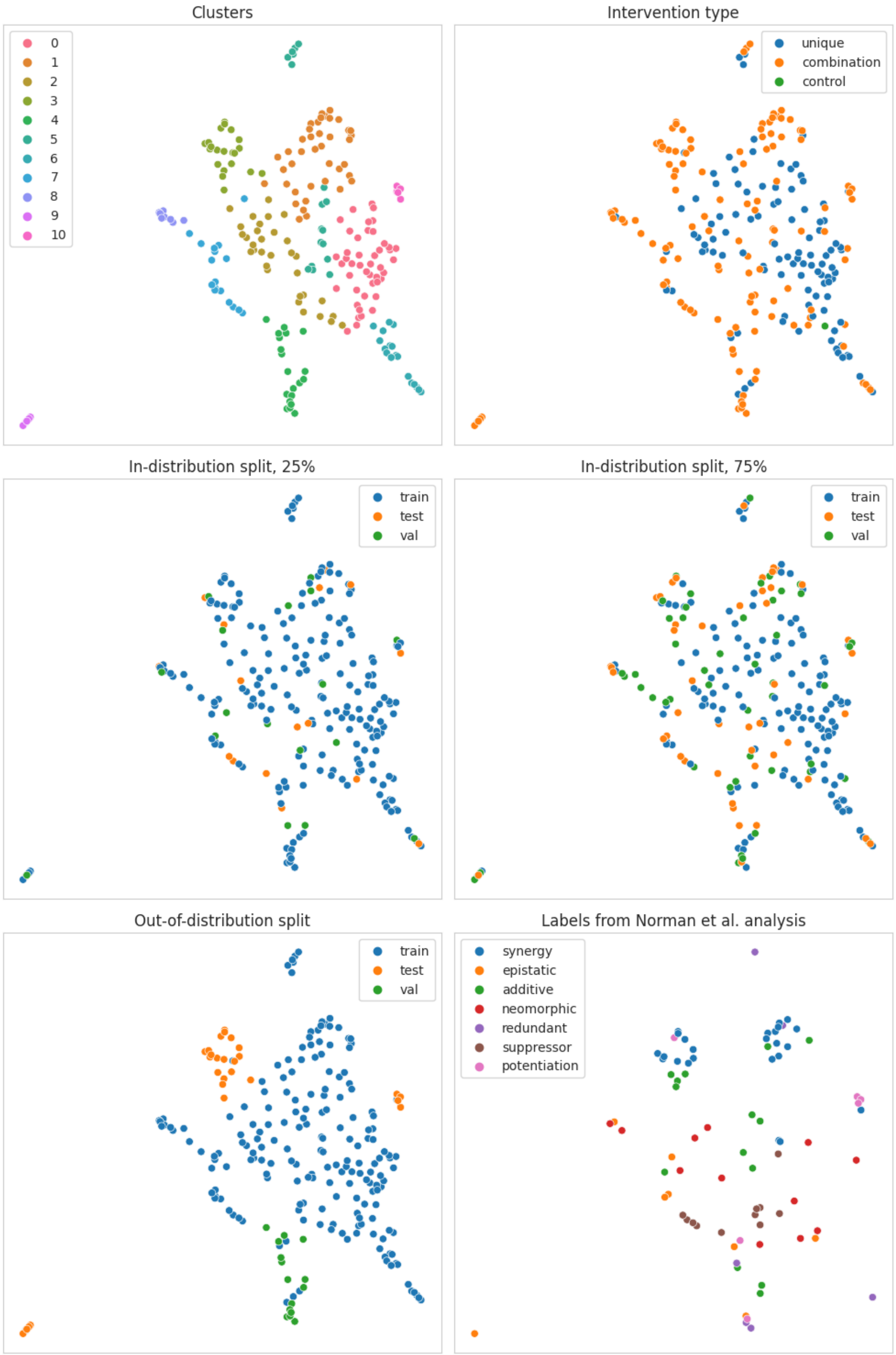

Figure 3: UMAP visualisation of pseudo-bulked data from the norman-dataset, showing the clustering of interventions, their types, and whether they are in train, validation, or test sets in the different splits we use for the analysis. The last panel show the interaction labels associated to combinations by Norman et al. (2019).

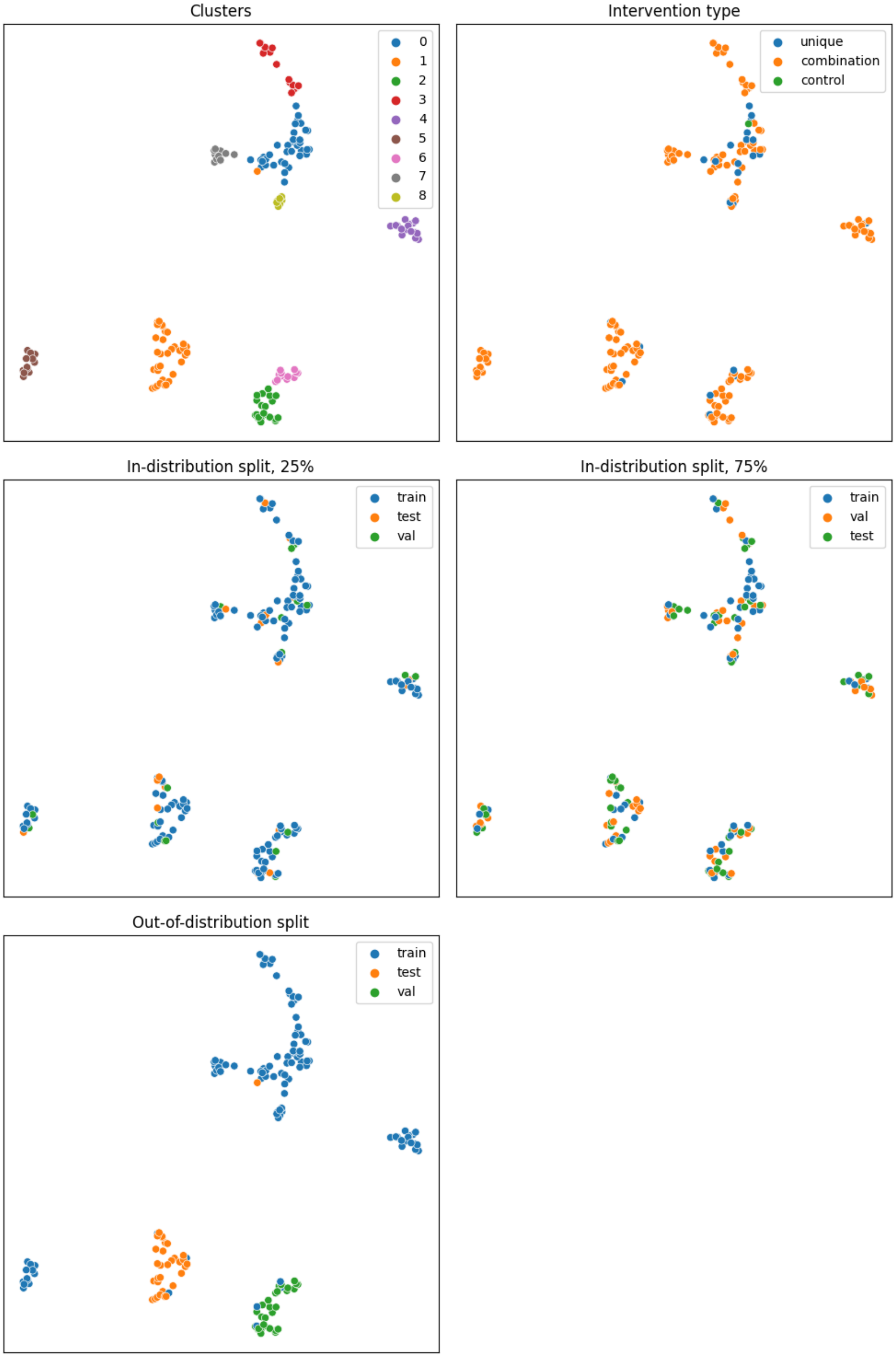

Figure 4: UMAP visualisation of pseudo-bulked data from the wessels-dataset, showing the clustering of interventions, their types, and whether they are in train, validation, or test sets in the different splits we use for the analysis.