
Learning interpretable cellular responses to complex perturbations in high-throughput screens

Mohammad Lotfollahi^{1,3,*}, Anna Klimovskaia Susmelj^{2,5,*}, Carlo De Donno^{1,7,**}, Yuge Ji^{1,**}, Ignacio L. Ibarra¹, F. Alexander Wolf^{1,○}, Nafissa Yakubova², Fabian J. Theis^{1,3,4,6†}, David Lopez-Paz²

1 Helmholtz Center Munich – German Research Center for Environmental Health, Institute of Computational Biology, Neuherberg, Munich, Germany.

2 Facebook AI, 6 Rue Ménars, Paris, 75002, France

3 School of Life Sciences Weihenstephan, Technical University of Munich, Munich, Germany.

4 Department Mathematics, Technical University of Munich, Munich, Munich, Germany.

5 Swiss Data Science Center, Zurich, Switzerland.

6 Wellcome Trust Sanger Institute, Wellcome Genome Campus, Hinxton, Cambridgeshire, UK.

7 Department of Stress Neurobiology and Neurogenetics, Max Planck Institute of Psychiatry, Munich, Bavaria, Germany.

* These authors contributed equally to the work.

** These authors contributed equally to the work.

○ Present address: Cellarity, Inc., Cambridge, MA.

† Correspondence to fabian.theis@helmholtz-muenchen.de

Abstract

1 Recent advances in multiplexed single-cell transcriptomics experiments are facilitating the high-
2 throughput study of drug and genetic perturbations. However, an exhaustive exploration of the
3 combinatorial perturbation space is experimentally unfeasible, so computational methods are needed
4 to predict, interpret, and prioritize perturbations. Here, we present the compositional perturbation
5 autoencoder (CPA), which combines the interpretability of linear models with the flexibility of
6 deep-learning approaches for single-cell response modeling. CPA encodes and learns transcriptional
7 drug responses across different cell type, dose, and drug combinations. The model produces easy-to-
8 interpret embeddings for drugs and cell types, which enables drug similarity analysis and predictions
9 for unseen dosage and drug combinations. We show that CPA accurately models single-cell pertur-
10 bations across compounds, doses, species, and time. We further demonstrate that CPA predicts
11 combinatorial genetic interactions of several types, implying that it captures features that distin-
12 guish different interaction programs. Finally, we demonstrate that CPA can generate *in-silico* 5,329
13 missing genetic combination perturbations (97.6% of all possibilities) with diverse genetic interac-
14 tions. We envision our model will facilitate efficient experimental design and hypothesis generation
15 by enabling *in-silico* response prediction at the single-cell level, and thus accelerate therapeutic
16 applications using single-cell technologies.

17 **Introduction**

18 Single-cell RNA-sequencing (scRNA-seq) profiles gene expression in millions of cells across tissues^{[1,}
19 ^{2]} and species^[3]. Recently, novel technologies have been developed that extend these measure-
20 ments to high-throughput screens (HTSs), which measure response to thousands of independent
21 perturbations^[4, 5]. These advances show promise for facilitating and thus accelerating drug development^[6].
22 HTSs applied at the single-cell level provide both comprehensive molecular phenotyping and capture
23 heterogeneous responses, which otherwise could not be identified using traditional HTSs^[4].

24 While the development of high-throughput approaches such as “cellular hashing” ^[4, 7, 8] facil-
25 itates scRNA-seq in multi-sample experiments at low cost, these strategies require expensive li-
26 brary preparation^[4], and do not easily scale to large numbers of perturbations. These shortcom-

27 ings become more apparent when exploring the effects of combination therapies[9–11] or genetic
28 perturbations[5, 12, 13], where experimental screening of all possible combinations becomes infeasible.
29 While projects such as the Human Cell Atlas[14] aim to comprehensively map cellular states
30 across tissues in a reproducible fashion, the construction of a similar atlas for the effects of pertur-
31 bations on gene expression is impossible, due to the vast number of possibilities. Since brute-force
32 exploration of the combinatorial search space is infeasible, it is necessary to develop computational
33 tools to guide the exploration of the combinatorial perturbation space to nominate promising candi-
34 date combination therapies in HTSs. A successful computational method for the navigation of the
35 combinatorial space must be able to predict the behaviour of cells when subject to novel combinations
36 of perturbations only measured separately in the original experiment. These data are referred to as
37 Out-Of-Distribution (OOD) data. OOD prediction would enable the study of perturbations in the
38 presence of different treatment doses [4, 15], combination therapies[8], multiple genetic knockouts[5],
39 and changes across time[15].

40 Recently, several computational approaches have been developed for predicting cellular responses
41 to perturbations[16–20]. The first approach leverages mechanistic modeling [18, 19] to predict cell
42 viability[19] or the abundance of a few selected proteins[18]. Although they are powerful at interpret-
43 ing interactions, mechanistic models usually require longitudinal data (which is often unavailable in
44 practice) and most do not scale to genome wide measurements to predict high-dimensional scRNA-
45 seq data. Linear models[12, 21] do not suffer from these scalability issues, but have limited predictive
46 power and are unable to capture nonlinear cell-type specific responses. In contrast, deep learning
47 (DL) models do not face these limitations. Recently, DL methods have been used to model gene
48 expression latent spaces from scRNA-seq data [22–25], and describe and predict single-cell responses
49 [16, 17, 20, 26]. However, current DL-based approaches also have limitations: they model only a
50 handful of perturbations; can be difficult to interpret; cannot handle combinatorial treatments; and
51 cannot incorporate continuous covariates such as dose and time, or discrete covariates such as cell
52 types, species, and patients while preserving interpretability. Therefore, while current DL methods
53 have modeled individual perturbations, none have been proposed for HTS.

54 Here, we propose the *compositional perturbation autoencoder (CPA)*, a novel, interpretable method
55 to analyze and predict scRNA-seq perturbation responses across combinations of conditions such
56 as dosage, time, drug, and genetic knock-out. The CPA borrows ideas from interpretable linear
57 models, and applies them in a flexible DL model to learn factorized latent representations of both
58 perturbations and covariates. Given a scRNA-seq dataset, the perturbations applied, and covariates
59 describing the experimental setting, CPA decomposes the data into a collection of embeddings
60 (representations) associated with the cell type, perturbation, and other external covariates. Since
61 these embeddings encode the transcriptomic effect of a drug or genetic perturbation, they can be used
62 by CPA users to study drug effects and similarities useful for drug repurposing applications. By virtue
63 of an adversarial loss, these embeddings are independent from each other, so they can be recombined
64 at prediction time to predict the effect of novel perturbation and covariate combinations. Therefore,
65 by exploring novel combinations, CPA can guide experimental design by directing hypotheses towards
66 expression patterns of interest to experimentalists. We demonstrate the usefulness of CPA on five
67 public datasets and multiple tasks, including the prediction and analysis of responses to compounds,
68 doses, time-series information, and genetic perturbations.

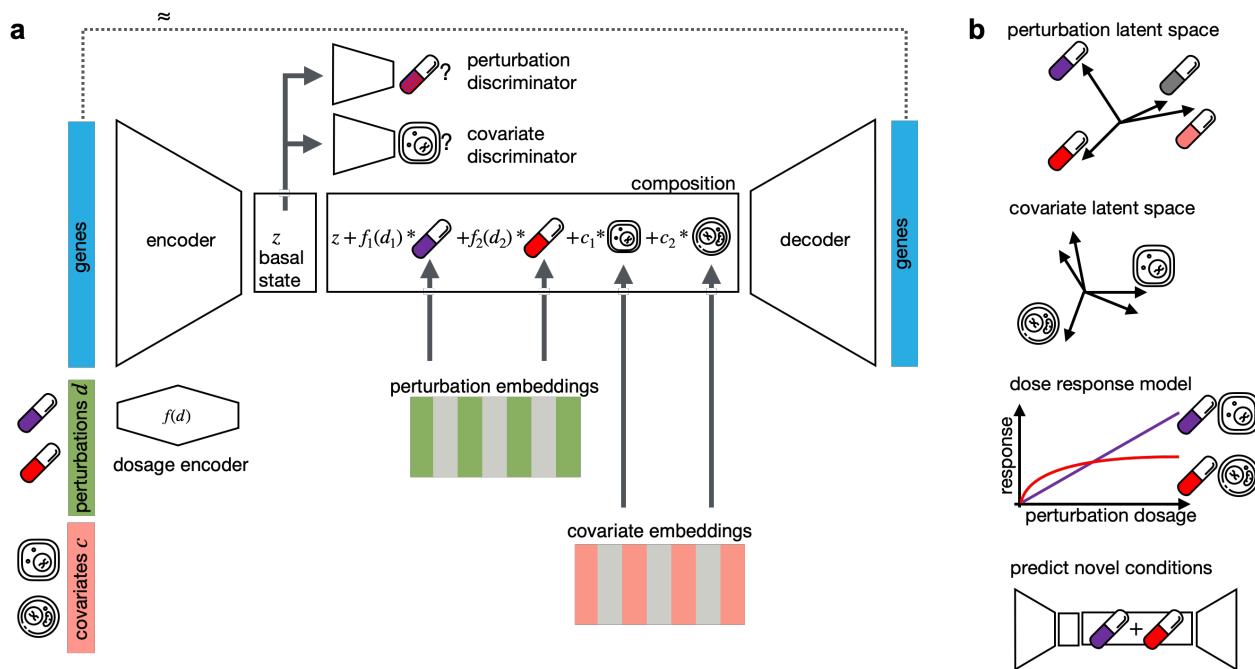


Figure 1: **Interpretable single-cell perturbation modeling using a compositional perturbation autoencoder (CPA).** (a) Given a matrix of gene expression per cell together with annotated potentially quantitative perturbations d and other covariates such as cell line, patient or species, CPA learns the combined perturbation response for a single-cell. It encodes gene expression using a neural network into a lower dimensional latent space that is eventually decoded back to an approximate gene expression matrix, as close as possible to the original one. To make the latent space interpretable in terms of perturbation and covariates, the encoded gene expression vector is first mapped to a “basal state” by feeding the signal to discriminators to remove any signal from perturbations and covariates. The basal state is then composed with perturbations and covariates, with potentially reweighted dosages, to reconstruct the gene expression. All encoder, decoder and discriminator weights as well as the perturbation and covariate dictionaries are learned during training. (b) Features of CPA are interpreted via plotting of the two learned dictionaries, interpolating covariate-specific dose response curves and predicting novel unseen drug combinations.

69

70 Results

71 Multiple perturbations as compositional processes in gene expression latent space

72 Prior work has modeled the effects of perturbations on gene expression as separate processes.
 73 While differential expression compares each condition separately with a control, modeling a joint
 74 latent space with a conditional variational autoencoder[17, 26, 27] is highly uninterpretable and not
 75 amenable to the prediction of the effects of combinations of conditions. Our goal here is to factorize
 76 the latent space of neural networks to turn them into interpretable, compositional models. If the
 77 latent space were linear, we could describe the observed gene expression as a factor model where
 78 each component is a single perturbation.

79 However, gene expression latent spaces, particularly in complex tissues, are nonlinear and best
 80 described by a graph or nonlinear embedding approximations[28, 29]. In scRNA-seq datasets, gene
 81 expression profiles of cell populations are often observed under multiple perturbations such as drugs,
 82 genetic knockouts, or disease states, in labeled covariates such as cell line, patient, or species. Each
 83 cell is labeled with its experimental condition and perturbation, where experimental covariates are

84 captured in categorical labels and perturbations are captured using a continuous value (e.g. a drug
85 applied with different doses). This assumes a sufficient number of cells per condition to permit the
86 estimation of the latent space in control and perturbation states using a large neural network.

87 Instead of assuming a factor model in gene expression space, we instead model the nonlinear super-
88 position of perturbation effects in the nonlinear latent space, in which we constrain the superposition
89 to be additive (see **Methods**). We decouple the effects of perturbations and covariates, and allow
90 for continuous effects such as drug dose by encoding this information in a nonlinearly transformed
91 scalar weight: a learned drug-response curve. The linear latent space factor model enables interpre-
92 tation of this space by disentangling latent space variance driven by covariates from those stemming
93 from each perturbation. At evaluation time, we are able to not only interpolate and interpret the
94 observed perturbation combinations, but also to predict other combinations, potentially in different
95 covariate settings.

96 Compositional perturbation autoencoder (CPA)

97 We introduce the CPA (see **Methods**), a method combining ideas from natural language processing
98 [30] and computer vision [31, 32] to predict the effects of combinations of perturbations on single-
99 cell gene expression. Given a single-cell dataset of multiple perturbations and covariates, the CPA
100 first uses an encoder neural network to decompose the cells' gene expression into three learnable,
101 additive embeddings, which correspond to its basal state, the observed perturbation, and the ob-
102 served covariates. Crucially, the embedding that the CPA encoder learns about a cell's basal state
103 is disentangled from (does not contain information about) the embeddings corresponding to the
104 perturbation and the covariates. This disentangling is achieved by training a discriminator classifier
105 [31] in a competition against the encoder network of the CPA. The goal of the encoder network in
106 the CPA is to learn an embedding representing a cell's basal state, from which the discriminator
107 network cannot predict the perturbation or covariate values. To perform well, the embedding of the
108 cell's basal state should contain all of the information about the cell's specifics. To account for con-
109 tinuous time or dose effects, the learned embeddings about each perturbation are scaled nonlinearly
110 via a neural network which receives the continuous covariate values for each cell, such as the time
111 or the dose. After integration of the learned embeddings about the cell's basal state, perturbations,
112 and covariate values into an unified embedding, the CPA uses a neural network decoder to recover
113 the cell's gene expression vector (**Figure 1**). Similar to many neural network models, the CPA is
114 trained using backpropagation [33] on the reconstruction and discriminator errors (see **Methods**),
115 to tune the parameters of the encoder network, the decoder network, the embeddings corresponding
116 to each perturbation and covariate value, and the dose/time nonlinear scalers. The learned embed-
117 dings allow the measurement of similarities between different perturbations and covariates, in terms
118 of their effects on gene expression. The main feature of the CPA is its flexibility of use at evaluation
119 time. After obtaining the disentangled embeddings corresponding to some observed gene expression,
120 perturbation, and covariate values, we can intervene and swap the perturbation embedding with any
121 other perturbation embedding of our choice. This manipulation is effectively a way of estimating
122 the answer to the counterfactual question: what would the gene expression of this cell have looked
123 like, had it been treated differently? This approach is of particular interest in the prediction of
124 unseen perturbation combinations and their effects on gene expression. The CPA can also visualize
125 the transcriptional similarity and uncertainty associated with perturbations and covariates, as later
126 demonstrated.

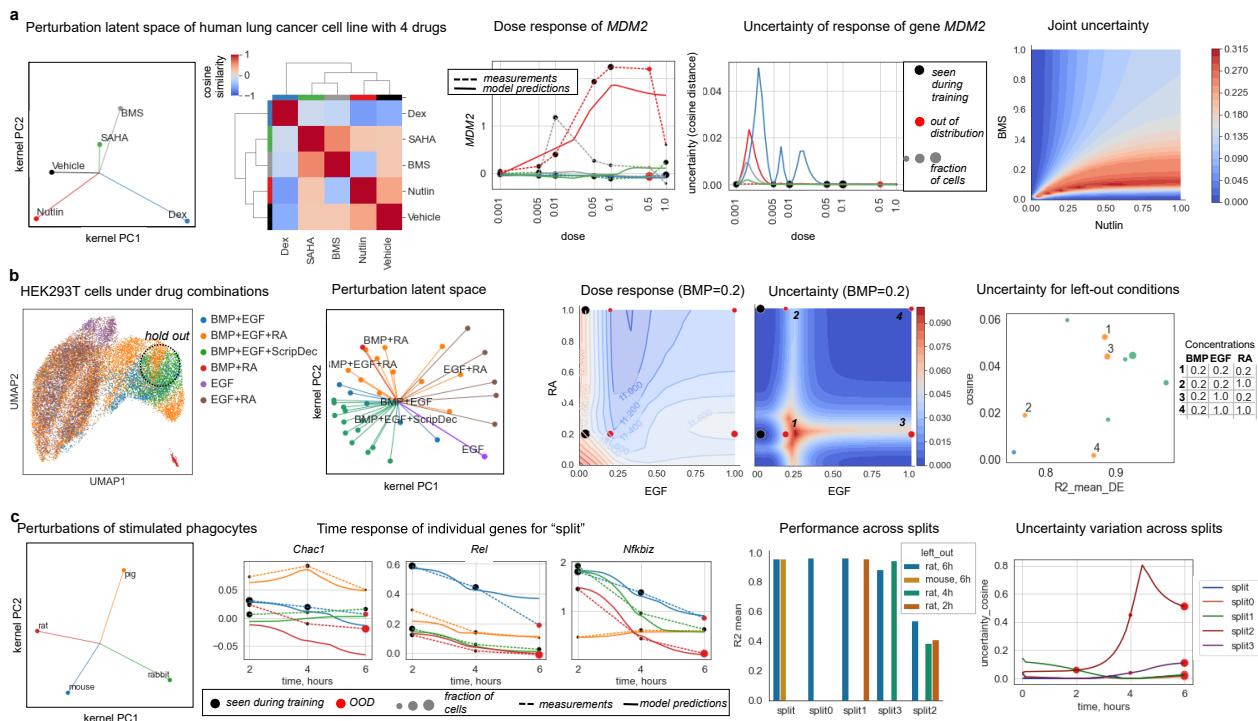


Figure 2: The CPA learns an interpretable latent space learning across drug dosages, drug combinations and experimental systems. (a) The sci-Plex 2 dataset from Srivatsan et al. [34]. Dose-response curves were generated using the CPA as a transfer from Vehicle cells to a given drug-dose combination. The *MDM2* gene, the top gene differentially expressed after treatment with Nutlin, was selected as an example. Black dots on the dose-response curve denote points seen at training time, red dots denote examples held out for OOD predictions. The sizes of the dots are proportional to the number of cells observed in the experiment. Solid lines correspond to the model predictions, dashed lines correspond to the linear interpolation between measured points. Nutlin and BMS are selected as examples of uncertainty in predictions for drug combinations. **(b)** 96-plex scRNA-seq experiment from Gehring et al. [8], with UMAP, showing variation of responses in gene expression space. The dashed circle on the UMAP represents the area on the UMAP where the majority of the cells from the left-out (OOD) condition lie. The experiment did not contain samples of individual drugs; therefore we represented the latent space of the drug combinations measured in the experiment. The dose-response surface was obtained via model predictions for a triplet of drugs: BMS at a fixed dose of 0.2, and EGF and RA changing on a grid. **(c)** Cross-species dataset from Hagai et al. [15], with samples of rat and mouse at time point 6 held out from training, and used as OOD. The latent space representation of individual species, and the individual average response of a species across time, demonstrates that the species are fairly different, with a small similarity between rat and mouse. The time response curves of individual genes demonstrate that the model is able to capture nonlinear behavior. The OOD splits benchmark demonstrates the way in which model performance on the distribution case changes when the model is trained on different subsets of the data. Split2 corresponds to the most difficult case, where all three time points for rat were held out from training. Red dots denote examples held out for OOD predictions; the size is proportional to the number of cells observed in the experiment.

127 CPA allows predictive and exploratory analyses of single-cell perturbation experiments.

128 We first demonstrated the performance and functionality of the CPA on three small single-cell
 129 datasets (**Figure 2**): a Sci-Plex2 dataset of human lung cancer cells perturbed by four drugs [35],
 130 a 96-plex-scRNA-seq experiment of HEK293T under different drug combinations [8], and a longi-
 131 toudnal cross-species dataset of lipopolysaccharide (LPS) treated phagocytes [15] (see **Methods**).

132 All three datasets represent different scenarios of the model application: (i) diverse doses; (ii) drug
133 combinations; and (iii) several Species and variation with respect to time instead of dose. We split
134 each dataset into three groups: train (used for model training), test (used for tuning of the model
135 parameters), and OOD (never seen during training or parameter setting, and intended to measure
136 the generalization properties of the model). **Supplementary Tables 1–5** shows the R^2 metrics (see
137 Methods) for the performance of the CPA on these datasets and various splits.

138 Sci-plex from Srivatsan et al. [35] contains measurements of a human lung adenocarcinoma cell
139 line treated with four drug perturbations at increasing doses. In this scenario, the model learns to
140 generalize to the unseen dosages of the drugs. To demonstrate the OOD properties, we withheld
141 cells exposed to the second to largest dose among all drugs. This choice was made because the vast
142 majority of cells are dead for most of the drugs at the highest dosage, and we would not have enough
143 cells to statistically test the generalizability of the CPA model. Since the latent space representation
144 learned by the CPA is still high-dimensional, we can use various dimensionality reduction methods
145 to visualize it, or simply depict it as a similarity matrix (**Figure 2a**). In **Supplementary Table 2**
146 we compare the performance of the CPA on the OOD example on two simple baselines: taking the
147 maximum dose as a proxy to the previous dose, and a linear interpolation between two measured
148 doses. These results demonstrate that the model consistently achieves high scores (a maximum
149 score of 1 yields perfect reconstruction) on all of the OOD cases, and on two of them significantly
150 outperform the baselines for Nutlin (0.92 vs 0.85) and BMS (0.94 vs 0.89). To demonstrate how well
151 the CPA captured the dose-response dynamics of individual genes, we looked at the top differentially
152 expressed genes upon Nutlin perturbation (**Figure 2a**). The dose-response curve agrees well with the
153 observed data. We additionally propose a simple heuristic to measure the model's uncertainty (see
154 **Methods**) with respect to unseen perturbation conditions. The model shows very low uncertainty
155 on the OOD split. This observation agrees well with the CPA's high R^2 scores on the OOD example.
156 However, when we tested the uncertainty of the model on a combination of two drugs (**Figure 2a**),
157 we saw that it produces much higher uncertainty compared to single drugs. This finding agrees with
158 the fact that the model never saw some drug combinations during training, and that such predictions
159 are more unreliable.

160 As the second working example, we took the 96-plex-scRNAse dataset from Gehring et al. [8].
161 This dataset contains 96 unique growth conditions using combinations of various doses of four drugs
162 applied to HEK293T cells. We hold out several combinations of these conditions as OOD cases, as
163 detailed in (**Supplementary Table 3**). We show that the CPA is able to reliably predict expression
164 patterns of unseen drug combinations (**Supplementary Table 3**) and produce a meaningful latent
165 perturbation latent space (**Figure 2b**). For this dataset, even simple baselines are not applicable
166 anymore, since the expression of cells exposed to the individual drugs were not measured. We also
167 confirmed that our heuristic for the measurement of uncertainty agreed with the model's performance
168 on OOD examples.

169 As our third example we studied the cross-species dataset from Hagai et al. [15]. Here we show that
170 the CPA can also be applied in the setting of multiple covariates, such as different species or cell
171 types, and the dynamics of the covariate can be a non-monotonic function, such as time instead
172 of the dose-response. In this example, bone marrow-derived mononuclear phagocytes from mouse,
173 rat, rabbit, and pig were challenged with LPS (**Figure 2c**). The learned CPA latent space agreed
174 with expected species similarities, with a relatively higher value found between rat and mouse. We
175 compared the generalization abilities of the model by withholding different parts of the data for OOD
176 cases: "split0" (rat at six hours), "split1" (rat at two and six hours), "split2" (rat at two, four,
177 and six hours), "split3" (rat at four and six hours), and "split" (rat and mouse at six hours. This
178 last split was used for the main analysis) (**Supplementary Table 4**). The model produced high
179 performance values compared to the performance on the test split (see **Supplementary Table 5**)
180 on the majority of the OOD splits, and showed a comparatively lower performance when the model
181 was not exposed to any LPS and rat examples with the exception of control cells. On this dataset,
182 we observed that the model with the lowest performance was the one with the highest number of

183 held-out examples, yet the model uncertainty also spiked for these OOD cases, suggesting that they
184 might be not reliable (**Figure 2c**). In contrast, for cases with high R^2 scores, models were more
185 certain about these predictions (**Supplementary Table 4**).

186 CPA finds interpretable latent spaces in large-scale single-cell high-throughput screens

187 The recently proposed sci-Plex assay [35] profiles thousands of independent perturbations in a single
188 experiment via nuclear hashing. With this high-throughput screen, 188 compounds were tested in 3
189 cancer cell lines. The panel was chosen to target a diverse range of targets and molecular pathways,
190 covering transcriptional and epigenetic regulators and diverse mechanisms of action. The screened
191 cell lines A549 (lung adenocarcinoma), K562 (chronic myelogenous leukemia), and MCF7 (mammary
192 adenocarcinoma) were exposed to each of these 188 compounds at four doses (10 nM, 100 nM, 1
193 μ M, 10 μ M), and scRNA-seq profiles were generated for altogether 290 thousand cells (**Figure 3a**).
194 As above, we split the dataset into 3 subsets: train, test, and OOD. For the OOD case, we held out
195 the highest dose (10 μ M) of the 36 drugs with the strongest effect in all three cell lines. Drug, dose,
196 and cell line combinations present in the OOD cases were removed from the train and test sets.

197 CPA is able to extrapolate to the unseen OOD conditions with unexpected accuracy, as it captures
198 the difference between control and treated conditions also for a compound where it did not see
199 examples with the highest dose. As one example, pracinostat has a strong differential response to
200 treatment compared to control, as can be seen from the distributions of the top 5 differentially
201 expressed genes (**Figure 3b**). Despite not seeing the effect of Pracinostat at the highest dose in any
202 of the three cell lines, CPA correctly infers the mean and distribution of these genes (**Figure 3b**).
203 CPA performs well in modeling unseen perturbations, as the correlation of real and predicted values
204 across OOD conditions is overall better than the correlation between real values (**Figure 3c**). When
205 looking at individual conditions (**Figure 3d**), CPA does well recapitulating genes with low and high
206 mean expression in the OOD condition.

207 CPA has lower performance when predicting experiments with more unseen covariates. To assess the
208 ability of the model to generalize to unseen conditions, we trained CPA on 28 splits with different
209 held-out conditions, with one of the doses held out in anywhere between 1-3 cell lines (**Figure**
210 **3e**). We see here that K562 is the hardest cell line to generalize, when considering training on two
211 cell lines to generalize to another. We also see that extrapolating to the highest dose is a harder
212 task than interpolating intermediate doses, which is consistent with the difficulty of anticipating the
213 experimental effect of a higher dose, versus fitting sigmoidal behavior to model intermediate doses.
214 When examining the shape of the sigmoid per compound learned by the model (**Figure 3f**), we
215 see that epigenetic compounds, which caused the greatest differential expression effects, have higher
216 latent response curves, indicating that CPA learns a general, cell-line agnostic response strength
217 measure for compounds. This learned sigmoid behavior can then be used in conjunction with the
218 latent vectors to reconstruct the gene expression of treated cells over interpolated doses (**Figure**
219 **3g**).

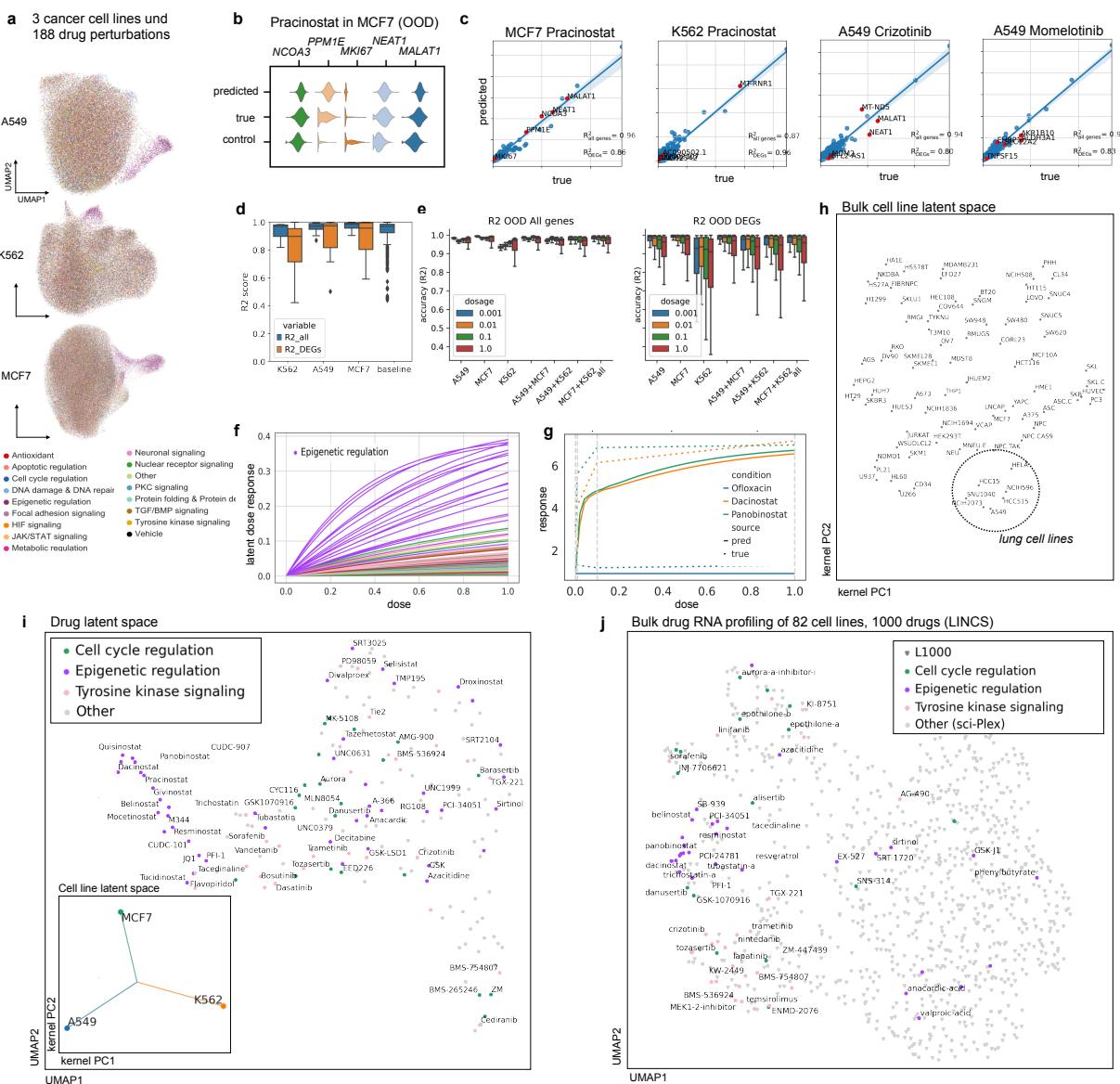


Figure 3: Learning drug and cell line latent representations from massive single-cell screens of 188 drugs across cancer cell lines. (a) UMAP representation of sci-Plex samples of A549, K562 and MCF7 cell-lines colored by pathway targeted by the compounds to which cells were exposed. (b) Distribution of top 5 differentially expressed genes in MCF7 cells after treatment with Pracinostat at the highest dose for real, control and CPA predicted cells. (c) Mean gene expression of 5,000 genes and top 50 DEGs between CPA predicted and real cells together with the top five DEGs highlighted in red for four compounds for which the model did not see any examples of the highest dose. (d) Box plots of R^2 scores for predicted and real cells for 36 compounds and 108 unique held out perturbations across different cell lines. Baseline indicates comparison of real compounds with each other. (e) R^2 scores box plot for all and top 50 DEGs. Each column represents a scenario where cells exposed with specific dose for all compounds on a cell line or combinations of cell lines were held from training and later predicted. (f) Latent dose response obtained from dose encoder for all compounds colored by pathways. (g) Real and predicted dose response curves based on gene expression data, for a single compound with differential dose response across three cell lines. (h) Latent representation of 80 cell lines from L1000 dataset. (i) Two dimensional representation of latent drug embeddings as learned by the CPA. Compounds associated with epigenetic regulation, tyrosine kinase signaling, and cell cycle regulation pathways are colored by their respective known pathways. The lower left panel shows latent covariate embedding for three cell lines in the data, indicating no specific similarity preference. (j) Latent drug embedding of CPA model trained on the bulk-RNA cell line profiles from the L1000 dataset, with focus on drugs shared with the sci-Plex experiment from (a).

220 After training, CPA learns a compressed representation of the 188 compounds, where each drug
221 is represented by a single 256 dimensional vector (**Figure 3i**). To test whether the learned drug
222 embeddings are meaningful, we asked if compounds with similar putative mechanisms of action are
223 similar in latent space. This holds for a large set of major mechanisms: we find that epigenetic,
224 tyrosine kinase signaling, and cell-cycle regulation compounds are clustered together by the model,
225 which suggests the effectiveness of drugs with these mechanisms on these three cancer cell-lines
226 which is in line with the findings in the original publication [4].

227 We additionally demonstrate that the model learns universal relationships between compounds which
228 remain true across datasets and modalities. Using the same set of compounds tested in the sci-Plex
229 dataset together with 853 other compounds (for a total of 1000 compounds), we trained CPA on
230 L1000 bulk perturbation measurement data across 82 cell lines [36]. We observed that CPA works
231 equally well on bulk RNA-seq data, and also that matched epigenetic and tyrosine kinase signaling
232 compounds present in sci-Plex were close to each other in the latent representation, suggesting that
233 the learned model similarities apply across datasets (**Figure 3j**). This holds also for the other learned
234 embeddings: Applying the same similarity metric to the covariate embedding - here the 82 cell lines
235 - we observed that the cell line embedding learned by the model also represents cell line similarity
236 in response to perturbation, as cell lines from lung tissue were clustered together (**Figure 3h**).

237 CPA allows modeling combinatorial genetic perturbation patterns

238 Combinatorial drug therapies are hypothesized to address the limited effectiveness of mono-therapies[37]
239 and prevent drug resistance in cancer therapies[37–39]. However, the combined expression of a small
240 number of genes often drives the complexity at the cellular level, leading to the emergence of new
241 properties, behaviors, and diverse cell types [5]. To study such genetic interactions (GIs), recent
242 perturbation scRNA-seq assays allow us to measure the gene expression response of a cell to the
243 perturbation of genes alone or in combination[12, 13]. While experimental approaches are necessary
244 to assess the effect of combination therapies, in practice, it becomes infeasible to experimentally
245 explore all possible combinations without computational predictions.

246 To pursue this aim, we applied our CPA model to scRNA-seq data collected from Perturb-seq (single-
247 cell RNA-sequencing pooled CRISPR screens) to assess how overexpression of single or combinatorial
248 interactions of 105 genes (i.e., single gene x, single gene y, and pair x+y) affected the growth of
249 K562 cells [5]. In total, this dataset contains 284 conditions measured across $\approx 108,000$ single-cells,
250 where 131 are unique combination pairs (i.e., x+y) and the rest are single gene perturbations or
251 control cells. We observed that the latent genetic interaction manifold placed GIs inducing known
252 and similar gene programs close to each other (**Figure 4a**). For example, consider *CBL* (orange
253 cluster in **Figure 4a**): the surrounding points, comprising its regulators (e.g., *UBASH3A/B*) and
254 multisubstrate tyrosine phosphatases (e.g., *PTPN9/12*), have all been previously reported to induce
255 erythroid markers [5]. Next, we sought to assess our ability to predict specific genetic interactions.
256 We examined a synergistic interaction between *CBL* and *CNN1* in driving erythroid differentiation
257 which has been previously validated [5]. We trained a CPA model with *CBL*+*CNN1* held out
258 from the training data. Overexpression of either gene leads to the progression of cells from control
259 to single perturbed and doubly perturbed cells (**Supplementary Fig.2a**) toward the erythroid
260 gene program. Overexpression of both *CBL* and *CNN1* up-regulate known gene markers[5] such as
261 hemoglobins (see *HBA1/2* and *HBG1/2* in **Figure 4b**). We observed that our model successfully
262 predicted this synergistic interaction, recapitulating patterns similar to real data and inline with the
263 original findings (**Figure 4c**). We further evaluated CPA to predict a previously reported[5] genetic
264 epistatic interaction between *DUSP9* and *ETS1*, leading to domination of the *DUSP9* phenotype in
265 doubly perturbed cells (**Figure 4c**).

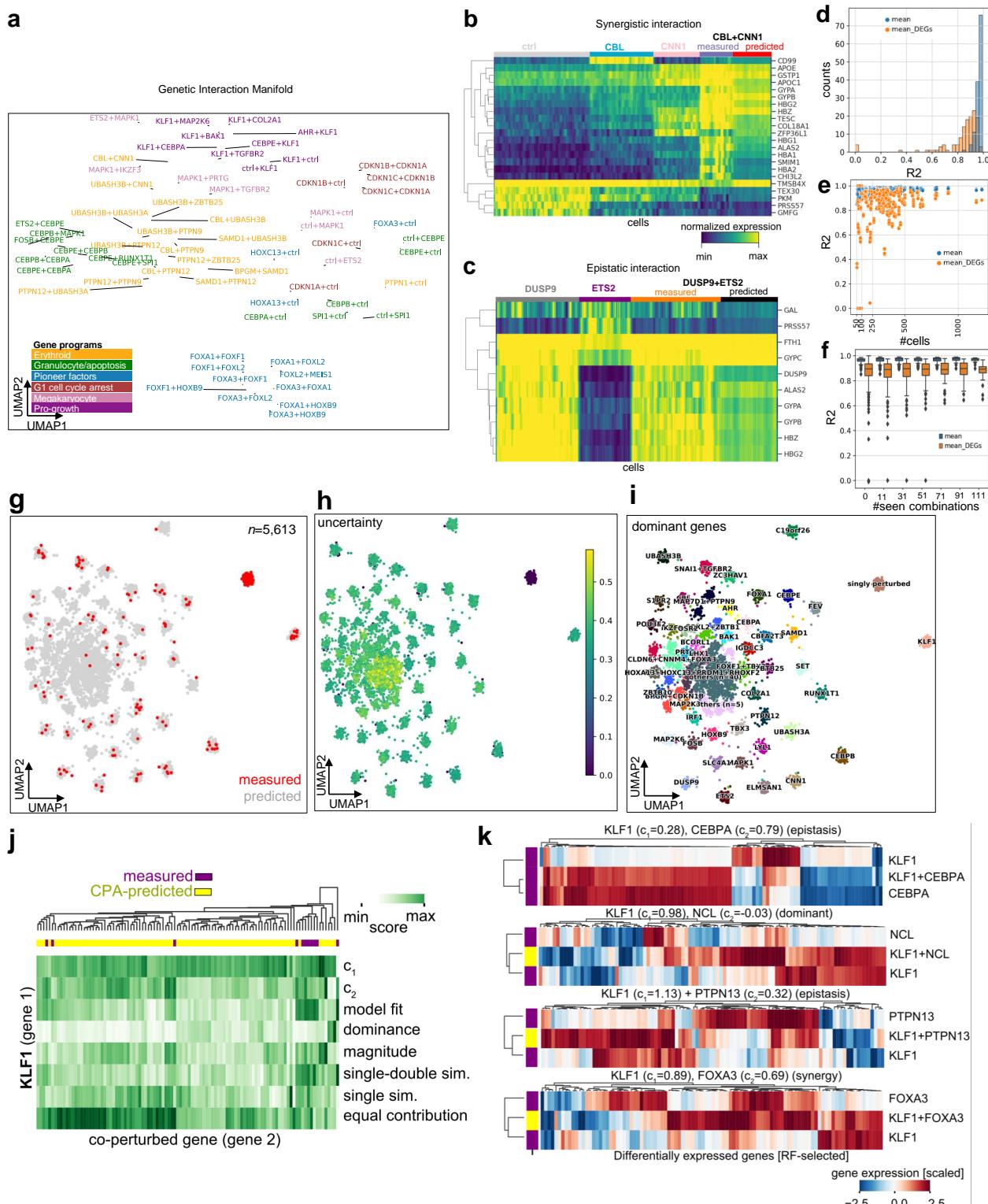


Figure 4: (Continued on the following page)

Figure 4: **Learning and predicting combinatorial genetic perturbations.** (a) UMAP inferred latent space using CPA for 281 single- and double-gene perturbations obtained from Perturb-seq[5]. Each dot represents a genetic perturbation. Coloring indicates gene programs associated to perturbed genes. (b) Measured and CPA-predicted gene expression for cells linked to a synergistic gene pair (*CBL*+*CNN1*). Gene names taken from the original publication. (c) As (b) for an epistatic (*DUSP9*+*ETS*) gene pair. Top 10 DEGs of *DUSP9*+*ETS* co-perturbed cells versus control cells are shown. (d) R² values of mean gene-expression of measured and predicted cells for all genes (blue) or top 100 DEGs for the prediction of all 131 combinations (13 trained models, with \approx 10 tested combinations each time) (orange). (e) R² values of predicted and real mean gene-expression versus number of cells in the real data (h) R² values for predicted and real cells versus number of combinations seen during training. (g) UMAP of measured (n=284, red dots) and CPA-predicted (n=5,329, gray dots) perturbation combinations. (h) As (g), showing measurement uncertainty (cosine similarity). (i) As (g), showing dominant genes in leiden clusters (25 or more observations). (j) Hierarchical clustering of linear regression associated metrics between *KLF1* with co-perturbed genes, in measured and predicted cells. (k) Scaled gene expression changes (versus control) of RF-selected genes (x-axis) in measured (purple) and predicted (yellow) perturbations (y-axis). Headers indicate gene-wise regression coefficients, and interaction mode suggestions[5].

266 To systemically evaluate the CPA's generalization behavior, we trained 13 different models while
267 leaving out all cells from \approx 10 unique combinations covering all 131 doubly perturbed conditions in
268 the dataset, which were predicted following training. The reported R^2 values showed robust prediction
269 for most of the perturbations: lower scores were seen for perturbations where the evaluation was
270 noisy due to sample scarcity ($n < 100$), or when one of the perturbations was only available as singly
271 perturbed cells in the data, leading the model to fail to predict the unseen combination (Figure 4d-e,
272 see **Supplementary Fig. 2**). To further understand when CPA performance deteriorated, we first
273 trained it on a subset with no combinations seen during training, and then gradually increased the
274 number of combinations seen during training. We found that overall prediction accuracy improved
275 when the model was trained with more combinations, and that it could fail to predict DEGs when
276 trained with fewer combinations (see $n < 71$ combinations in Figure 4f).

277 Hence, once trained with sufficiently large and diverse training data, CPA could robustly predict
278 unseen perturbations. We next asked whether our model could generalize beyond the measured
279 combinations and generate *in-silico* all 5,329 combinations, which were not measured in the real
280 dataset, but made up \approx 98% of all possibilities. To study the quality of these predictions, we
281 trained a model where all combinations were seen during training to achieve maximum training
282 data and sample diversity. We then predicted 50 single-cells for all missing combinations. We
283 found that, while the latent embeddings did not fully capture all the nuances in the similarity of
284 perturbations compared to gene space, it provided an abstract and easier to perform high-level
285 overview of potential perturbation combinations. Thus, we leveraged our latent space to co-embed
286 (Figure 4g) all measured and generated data while proving an uncertainty metric based on the
287 distance from the measured phenotypes (Figure 4h). We hypothesized that the closer the generated
288 embedding was to the measured data, the more likely it was to explore a similar space of the genetic
289 manifold around the measured data. Meanwhile, the distant points can potentially indicate novel
290 behaviors, although this would require additional consideration and validation steps. Equipped
291 with this information, we annotated the embedding clusters based on gene prevalence, finding that
292 single genes (i.e. gene x) paired with other genes (i.e., y) as combinations (i.e., x+y) are a main
293 driver of cluster separation (Figure 4i). Genes without measured double perturbations were less
294 likely to be separated as independent clusters using the newly predicted transcriptomic expression
295 (Supplementary Fig. 3a), suggesting that their interaction-specific effects were less variable than
296 cases with at least one double perturbation available in the training data.

297 To investigate the type of interaction between the newly predicted conditions, we compared the
298 differences between double and single perturbations versus control cells and thus annotated their

299 interaction modes (adapted from [5] for *in silico* predictions). In each gene-specific cluster, we ob-
300 served variability across these values, suggesting that our predictions contained granularity that went
301 beyond single gene perturbation effects, and could not be fully dissected by two dimensional embed-
302 dings. Upon curation of gene perturbations using these metrics and the levels of experimental data
303 available (**Supplementary Fig. 3b**), we decided to predict and annotate interaction modes based
304 on these values when double measurements were available for at least one gene. For example, we ob-
305 served clustering of *KLF1* and partner gene perturbation pairs solely from these metrics, suggesting
306 the existence of several interaction modes (**Figure 4j**). When we further examine the differenti-
307 ally expressed genes in each co-perturbation, our metrics validated previously reported epistatic
308 interactions (*CEBPA*), and proposed new ones with *KLF1*-dominant behavior (*NCL*), gene synergy
309 (*FOXA3*), and epistasis (*PTPN13*), among others (**Figure 4k**). Repeating this analysis across all
310 measured and predicted double perturbations, we found genes with potential interaction prevalences
311 (**Supplementary Fig. 3c**). Among genes which repeatedly respond to several perturbations, we
312 found common gene expression trends in both direction and magnitude (**Supplementary Fig. 3d**),
313 suggesting that variation is modulated by conserved gene regulatory principles that are potentially
314 captured in our learned model.

315 Altogether, our analysis indicated that double perturbation measurements can be generated by CPA
316 by leveraging genetic perturbation data, which when combined with an uncertainty metric allows us
317 to generate and interpret gene regulatory rules in the predicted gene-gene perturbations.

318

319 Discussion

320 *In-silico* prediction of cell behavior in response to a perturbation is critical for optimal experiment
321 design and the identification of effective drugs and treatments. With CPA, we have introduced a
322 versatile and interpretable approach to modeling cell behaviors at single-cell resolution. CPA is
323 implemented as a neural network trained using stochastic gradient descent, scaling up to millions of
324 cells and thousands of genes.

325 We applied CPA to a variety of datasets and tasks, from predicting single-cell responses to learning
326 embeddings, as well as reconstructing the expression response of compounds, with variable drug-
327 dose combinations. Specifically, we illustrated the modeling of perturbations across dosage levels
328 and time series, and have demonstrated applications in drug perturbation studies, as well as genetic
329 perturbation assays with multiple gene knockouts, revealing potential gene-gene interaction modes
330 inferred by our model predicted values. CPA combines the interpretability of linear decomposition
331 models with the flexibility of nonlinear embedding models.

332 While CPA performed well in our experiments, it is well known that in machine learning there is
333 no free lunch, and as with any other machine learning model, CPA will fail if the test data are very
334 different from the training data. To alert CPA users to these cases, it is crucial to quantify model
335 uncertainty. To do so, we implemented a distance-based uncertainty score to evaluate our predictions.
336 Additionally, scalable Bayesian uncertainty models are promising alternatives for future work[40].
337 Although we opted to implement a deterministic autoencoder scheme, extensions towards variational
338 models[17, 23], as well as cost functions other than mean squared error[22] are straightforward.

339 Aside from CPA, existing methods[17, 26] such as scGen[16] have also been shown capable of predict-
340 ing single-cell perturbation responses when the dataset contains no combinatorial treatment or dose-
341 dependent perturbations. Therefore, it may be beneficial to benchmark CPA against such methods
342 on less complicated scenarios with few perturbations. However, this approach might not be practical,
343 considering the current trend towards the generation of massive perturbation studies[4, 5, 12].

344 Currently, the model is based on gene expression alone, so it cannot directly capture other levels
345 of interactions or effects, such as those due to post-transcriptional modification, signaling, or cell
346 communication. However, due to the flexibility of neural network-based approaches, CPA could

347 be extended to include other modalities, for example via multimodal single-cell CRISPR[41, 42]
348 combined scRNA- and ATAC-seq[43, 44] and CUT&Tag[45, 46]. In particular, we expect spatial
349 transcriptomics[47, 48] to be a valuable source for extensions to CPA due to its high sample number
350 and the dominance of DL models in computer vision.

351 The CPA model is not limited to single-cell perturbations. While we chose the single-cell setting due
352 to the high sample numbers available, the CPA could readily be applied to large-scale bulk cohorts,
353 in which covariates might be patient ID or transcription factor perturbation. These and any other
354 available attributes could be controlled independently[31] to achieve compositional, interpretable
355 predictions. Any bulk compositional model may be combined with a smaller-scale single-cell model
356 to compose truly multi-scale models of observed variance. The flexibility of the DL setting will also
357 allow addition of constraints on perturbation or covariate latent spaces. These could, for example,
358 be the similarity of chemical compounds[49], or clinical-covariate induced differences of patient IDs.
359 The key feature of the CPA versus a normal autoencoder is its latent space disentanglement and the
360 induced interpretability of the perturbations in the context of cell states and covariates. Eventually,
361 any aim in biology is not only blind prediction, but mechanistic understanding. This objective is
362 exemplified by the direction that DL models are taking in sequence genomics, where the aim is not
363 only the prediction of new interactions, but also the interpretation of the learned gene regulation
364 code. We therefore believe that CPA can not only be used as a hypothesis generation tool for
365 *in-silico* screens but also as an overall data approximation model. Deviations from our assumed
366 data generation process (see **Methods**) would then tell us about missing information in the given
367 data set and/or missing aspects in the factor model. By including multiple layers of regulation,
368 the resulting model can grow in flexibility for prediction and for mechanistic understanding on for
369 example synergistic gene regulation or other interactions.

370 Finally, we expect CPA to facilitate new opportunities in expression-based perturbation screening,
371 not only to learn optimal drug combinations, but also in how to personalize experiments and
372 treatments by tailoring them based on individual cell response.

373

374 **Code availability**

375 Code to reproduce all of our results is available at <http://github.com/facebookresearch/CPA>.

376

377 **Data availability**

378 All datasets analyzed in this manuscript are public and have published in other papers. We have
379 referenced them in the manuscript and made available at <http://github.com/facebookresearch/CPA>.

381

382 **Author Contributions**

383 M.L., A.K.S, and D.L.P. conceived the project with contributions from F.J.T. D.L.P., M.L. and
384 A.K.S designed the algorithm and implemented the first version. Y.J., C.D. and A.K.S. performed
385 the first refactor. The final code is implemented by D.L.P. and A.K.S. with contributions from C.D.,
386 Y.J. and M.L. M.L. and C.D. curated all the datasets. F.A.W. helped interpret the model and
387 results. M.L., A.K.S. and F.J.T. designed analyses and use-cases. M.L., A.K.S., Y.J., I.L.I. and C.D
388 performed the analysis. F.J.T. and N.Y. supervised the research. All authors wrote the manuscript.

389

390 **Acknowledgments**

391 M.L. and F.J.T. are grateful for valuable feedback from Aviv Regev and Dana Pe'er. We appreciate
392 support from all members of Theis lab, specifically Malte D. Luecken and Fabiola Curion for their
393 feedback and proof-reading. M.L is thankful for early graphical designs by Monir Jazaeri (Jaz) which
394 did not make it to the final version of the paper. F.J.T. acknowledges support by the BMBF (grant
395 L031L0214A, grant 01IS18036A and grant 01IS18053A), by the Helmholtz Association (Incubator
396 grant sparse2big, grant ZT-I-0007) and by the Chan Zuckerberg Initiative DAF (advised fund
397 of Silicon Valley Community Foundation, 2018-182835 and 2019-207271). This work was further
398 supported by Helmholtz Association's Initiative and Networking Fund through Helmholtz AI [grant
399 ZT-I-PF-5-01]. I.L.I. has received funding from the European Union's Horizon 2020 research and
400 innovation programme under grant agreement No 874656.

401

402 **Competing interests**

403 F.J.T. reports receiving consulting fees from Roche Diagnostics GmbH and Cellarity Inc., and own-
404 ership interest in Cellarity, Inc. and Dermagnostix. F.A.W. is a full-time employee of Cellarity Inc.,
405 and has ownership interest in Cellarity, Inc.

406

407 **Methods**

408 **Data generating process**

409 We consider a dataset $\mathcal{D} = \{(x_i, d_i, c_i)\}_{i=1}^N$, where each $x_i \in \mathbb{R}^G$ describes the gene expression of G
410 genes from cell i . The perturbation vector $d_i = (d_{i,1}, \dots, d_{i,M})$ contains elements $d_{i,j} \geq 0$ describing
411 the dose of drug j applied to cell i . If $d_{i,j} = 0$, this means that perturbation j was not applied to
412 cell i . Unless stated otherwise, the sequel assumes column vectors. Similarly, the vector of vectors
413 $c_i = (c_{i,1}, \dots, c_{i,K})$ contains additional discrete covariates such as cell-types or species, where each
414 covariate is itself a vector. More specifically, $c_{i,j}$ is a K_j -dimensional one-hot vector.

415 We assume that an unknown generative model produced our dataset \mathcal{D} . The three initial components
416 of this generative process are a latent (unobserved) basal latent state z_i^{basal} for cell i , together with its
417 perturbation vector d_i and covariate vector c_i . We assume that the basal latent state is independent
418 from the perturbation vector d_i and covariate vector c_i . Next, we form the latent (also unobserved)
419 perturbed latent state z_i as:

$$z_i = z_i^{\text{basal}} + V^{\text{perturbation}} \cdot (f_1(d_{i,1}), \dots, f_M(d_{i,M})) + \sum_{j=1, \dots, K} V^{\text{cov}_j} \cdot c_{i,j} \quad (1)$$

420 In this equation, each column of the matrix $V^{\text{perturbation}} \in \mathbb{R}^{d \times M}$ represents a d -dimensional embed-
421 ding for one of the M possible perturbations represented in d_i . Similarly, each column of the matrix
422 $V^{\text{cov}_j} \in \mathbb{R}^{d \times K_j}$ represents a d -dimensional embedding for the j -th discrete covariate, represented as
423 a K_j -dimensional one-hot vector $c_{i,j}$. The functions $f_j : \mathbb{R} \rightarrow \mathbb{R}$ scale non-linearly each of the $d_{i,j}$ in
424 the perturbation vector d_i , therefore implementing M independent dose-response (or time-response)
425 curves. Finally, we assume that the generative process returns the observed gene expression x_i by
426 means of an unknown decoding distribution $p(x_i|z_i)$. This process builds the observation (x_i, d_i, c_i) ,
427 which is then included in our dataset \mathcal{D} .

428 **Compositional Perturbation Autoencoder (CPA)**

429 Assuming the generative process described above, our goal is to train a machine learning model
430 $x'_i = M((x_i, d_i, c_i), d')$ such that, given a dataset triplet (x_i, d_i, c_i) as well as a target perturbation d' ,
431 estimates the gene expression x'_i . The term x'_i represents what would the counterfactual distribution
432 of the gene expression x_i with covariates c_i look like, had it been perturbed with d' instead of d_i .

433 Given a dataset and a learning goal, we are now ready to describe our proposed model, the Com-
434 positional Perturbation Autoencoder (CPA). In the following, we describe separately how to train
435 and test CPA models.

436 **Training**

437 The training of a CPA model consists in auto-encoding dataset triplets (x_i, d_i, c_i) . That is, during
438 training, a CPA model does not attempt to answer counterfactual questions. Instead, the training
439 process consists in (1) encoding the gene expression x_i into an estimated basal state \hat{z}_i^{basal} that does
440 not contain any information about (d_i, c_i) , (2) combining \hat{z}_i^{basal} with learnable embeddings about
441 (d_i, c_i) to form an estimated perturbed state \hat{z}_i , and (3) decoding \hat{z}_i back into the observed gene
442 expression x_i .

More specifically, the CPA model first encodes the observed gene expression x_i into an estimated basal state:

$$\hat{z}_i^{\text{basal}} = \hat{f}^{\text{enc}}(x_i).$$

443 In turn, the estimated basal state is used to compute the estimated perturbed state \hat{z}_i :

$$\hat{z}_i := \hat{z}_i^{\text{basal}} + \hat{V}^{\text{perturbation}} \cdot (\hat{f}_1(d_i, 1), \dots, \hat{f}_M(d_i, M)) + \sum_{j=1, \dots, K} \hat{V}^{\text{cov}_j} \cdot c_{i,j} \quad (2)$$

444 Contrary to (1), this expression introduces three additional learnable components: the perturba-
445 tion embeddings $\hat{V}^{\text{perturbation}}$, the covariate embeddings \hat{V}^{cov} and the learnable dose-response curves
446 ($\hat{f}_1, \dots, \hat{f}_M$), here implemented as small neural networks constrained to satisfy $\hat{f}_j(0) = 0$.

447 As a final step, a decoder \hat{f}^{dec} accepts the estimated perturbed state \hat{z}_i and returns $\hat{f}_\mu^{\text{dec}}(\hat{z}_i)$ and
448 $\hat{f}_{\sigma^2}^{\text{dec}}(\hat{z}_i)$, that is, the estimated mean and variance of the counterfactual gene expression x'_i .

449 To train CPA models, we use three loss functions. First, the reconstruction loss function is the
450 Gaussian negative log-likelihood:

$$\ell_i := \frac{\log s(\hat{f}_{\sigma^2}^{\text{dec}}(\hat{z}_i))}{2} + \frac{(\hat{f}_\mu^{\text{dec}}(\hat{z}_i) - x'_i)^2}{2 \cdot s(\hat{f}_{\sigma^2}^{\text{dec}}(\hat{z}_i))}, \quad (3)$$

451 where $s(\sigma^2) = \log(\exp(\sigma^2 + 10^{-3}) + 1)$ enforces a positivity constraint on the variance and adds
452 numerical stability. This loss function rewards the end-to-end auto-encoding process if producing
453 the observed gene expression x_i .

Second, and according to our assumptions about the data generating process, we are interested in removing the information about (d_i, c_i) from \hat{z}_i^{basal} . To achieve this information removal, we follow an adversarial approach [31]. In particular, we set up the following auxiliary loss functions:

$$\begin{aligned} \ell_i^d &:= \text{CrossEntropy}(\hat{f}_d^{\text{adv}}(\hat{z}_i^{\text{basal}}), d_i), \\ \ell_{i,j}^c &:= \text{CrossEntropy}(\hat{f}_{c_{i,j}}^{\text{adv}}(\hat{z}_i^{\text{basal}}), c_{i,j}), \quad \forall j = 1, \dots, K. \end{aligned}$$

454 The functions $\hat{f}_d^{\text{adv}}, \hat{f}_{c_{i,j}}^{\text{adv}}$ are a collection of neural network classifiers trying to predict about (d_i, c_i)
455 given the estimated basal state \hat{z}_i^{basal} .

456 Given this collection of losses, the training process is an optimization problem that repeats the
457 following two steps:

458 1. sample $(x_i, d_i, c_i) \sim \mathcal{D}$, minimize $\ell_i^d + \sum_j \ell_{i,j}^c$ by updating the parameters of \hat{f}_d^{adv} and $\hat{f}_{c_i,j}^{\text{adv}}$, for
 459 all $j = 1, \dots, K$;

460 2. sample $(x_i, d_i, c_i) \sim \mathcal{D}$, minimize $\ell_i - \lambda \cdot (\ell_i^d + \sum_j \ell_{i,j}^c)$ by updating the parameters of the encoder
 461 \hat{f}^{enc} , the decoder \hat{f}^{dec} , the perturbation embeddings $\hat{V}^{\text{perturbation}}$, the covariate embeddings
 462 \hat{V}^{cov_j} for all $j = 1, \dots, K$, and the dose-response curve estimators $(\hat{f}_1, \dots, \hat{f}_M)$.

463 Testing

464 Given an observation (x_i, d_i, c_i) and a counterfactual treatment d' , we can use a trained CPA model
 465 to answer what would the counterfactual distribution of the gene expression x_i with covariates c_i
 466 look like, had it been perturbed with d' instead of d_i . To this end, we follow the following process:

467 1. Compute the estimated basal state $\hat{z}_i^{\text{basal}} = \hat{f}^{\text{enc}}(x_i)$;

468 2. Compute the counterfactual perturbed state \hat{z}'_i

$$\hat{z}'_i := \hat{z}_i^{\text{basal}} + \hat{V}^{\text{perturbation}} \cdot (\hat{f}_1(d'_{i,1}), \dots, \hat{f}_M(d'_{i,M})) + \sum_{j=1, \dots, K} \hat{V}^{\text{cov}_j} \cdot c_{i,j}.$$

468 Note that in the previous expression, we are using the counterfactual treatment d' instead of
 469 the observed treatment d_i .

469 3. Compute and return the counterfactual gene expression mean $x'_{i,\mu}$:

$$x'_{i,\mu} = \hat{f}_\mu^{\text{dec}}(\hat{z}'_i),$$

and variance x'_{i,σ^2} :

$$x'_{i,\sigma^2} = \hat{f}_{\sigma^2}^{\text{dec}}(\hat{z}'_i).$$

470 Hyper-parameters and training.

471 For each dataset, we perform a random hyper-parameter search of 100 trials. The table below
 472 outlines the distribution of values for each of the hyper-parameters involved in CPA training.

Group	Hyperparameter	Default value	Random search distribution
general	embedding dimension	256	RandomChoice([128, 256, 512])
	batch size	128	RandomChoice([64, 128, 256, 512])
	learning rate decay, in epochs	45	RandomChoice([15, 25, 45])
nonlinear scalers	hidden neurons, nonlinear scalers	64	RandomChoice([32, 64, 128])
	hidden layers	2	RandomChoice([1, 2, 3])
	learning rate	1e-3	10 ^{Uniform} (-4, -2)
	weight decay	1e-7	10 ^{Uniform} (-8, -5)
encoder and decoder	hidden neurons, encoder and decoder	512	RandomChoice([256, 512, 1024])
	hidden layers	4	RandomChoice([3, 4, 5])
	learning rate	1e-3	10 ^{Uniform} (-4, -2)
	weight decay	1e-6	10 ^{Uniform} (-8, -4)
discriminator	hidden neurons, discriminator	128	RandomChoice([64, 128, 256])
	hidden layers	3	RandomChoice([2, 3, 4])
	regularization strength	5	10 ^{Uniform} (-2, 2)
	gradient penalty	3	10 ^{Uniform} (-2, 1)
	learning rate	3e-4	10 ^{Uniform} (-5, -3)
	weight decay	1e-4	10 ^{Uniform} (-6, -3)
	number of learning steps	3	RandomChoice([1, 2, 3, 4, 5])

474 Model evaluation.

475 We use several metrics to evaluate the performance of our model: (1) quality of reconstruction for in
 476 and OOD cases and (2) quality of disentanglement of cell information from perturbation information.

477 We split each dataset into 3 subsets: train, test, and OOD. For OOD cases, we choose combinations
478 of perturbations that exhibit unseen behavior. This usually corresponds to the most extreme drug
479 dosages. We select one perturbation combination as "control". Usually these are Vehicle or DMSO
480 if real control samples are present in the dataset, otherwise we choose a drug perturbation at a
481 lower dosage as "control". For the evaluation, we use the mean squared error of the reconstruction
482 of an individual cell and average it over the cells for the perturbation of interest. As an additional
483 metric we use classification accuracy in order to check how well the information about the drugs was
484 separated from the information about the cells.

485 **Uncertainty estimation.**

486 To estimate the uncertainty of the predictions we use as a proxy the minimum distance between the
487 queried perturbation and the set of conditions (covariate + perturbation combinations) seen during
488 training (**Supplementary Fig.1**). Intuitively, we expect predictions on queried conditions that are
489 more distant from the set of seen conditions to be more uncertain. To estimate this distance we first
490 compute the set of embeddings of the training covariate and perturbation combinations:

$$\hat{z}^{comb} = \hat{V}^{\text{perturbation}} \cdot (\hat{f}_1(d'_1), \dots, \hat{f}_M(d'_M)) + \sum_{j=1, \dots, K} \hat{V}^{\text{cov}_j} \cdot c_j. \quad (4)$$

The latent vector for the queried condition is obtained in the same manner. The cosine and euclidean
distances from the training embedding set are computed and the minimum distance is used as a proxy
for uncertainty.

$$u_{\text{cosine}} = \min(1 - \mathbf{S}_C(\hat{z}^{query}, \hat{z}^{comb})) \quad (5)$$

$$u_{\text{eucl}} = \min(\mathbf{d}(\hat{z}^{query}, \hat{z}^{comb})) \quad (6)$$

491 Where $\mathbf{S}_C(\mathbf{x}, \mathbf{y})$ stands for the cosine similarity and $\mathbf{d}(x, y)$ for the euclidean distance between the
492 two vectors.

493 With this methodology, in the case of a drug screening experiment, if we query a combination of
494 cell type, drug, and dosage that was seen during training, we get an uncertainty of zero, since this
495 combination was present in the training set. It is important to note that with this method we obtain
496 a condition-level uncertainty, in that all cells predicted under the same query will have the same
497 uncertainty, thus not taking cell specific information into account.

498 **R2 score**

499 We used the `r2_score` function from *scikit-learn* which reports R2 (coefficient of determination)
500 regression score.

501 **Datasets**

502 Gehring et al.

503 This dataset[8] comprises of 21,191 neural stem cells (NSCs) cells perturbed with EGF/bFGF,
504 BMP4, decitabine, scriptaid, and retinoic acid. We obtained normalized data from the original
505 authors and after QC filtering 19,637 cells remained. We further selected 5,000 highly variable
506 genes (HVGs) using SCANPY's[50] `highly_variable_genes` function for training and evaluation of
507 the model.

508 Genetic CRISPR screening experiment

509 We obtained the raw count matrices from Norman *et al.*[5] from GEO (accession ID GSE133344).
510 According to authors guide, we excluded "NegCtrl1_NegCtrl0__NegCtrl1_NegCtrl0" control cells

511 and merged all unperturbed cells as one "ctrl" condition. We then normalized and log-transformed
512 the data using SCANPY and selected 5,000 HVGs for training. The processed dataset contained
513 108,497 cells.

514 Cross-species experiment

515 The data was generated by Hagai *et al.*[15] and downloaded from ArrayExpress (accession: E-MTAB-
516 6754). The data consists of 119,819 phagocytes obtained from four different species: mouse, rat, pig
517 and rabbit. Phagocytes were treated with lipopolysaccharide (LPS) and the samples were collected
518 at different time points: 0 (control), 2, 4, and 6 hours after the beginning of treatment. All genes
519 from non-mouse data were mapped to the respective orthologs in the mouse genome using Ensembl
520 ID annotations. We filtered out cells with a percentage of counts belonging to mitochondrial genes
521 higher than 20%, then proceeded to normalize and log-transform the count data. For training and
522 evaluation, we selected 5000 HVG using SCANPY. After filtering, the data consists of 113,400 cells.

523 sci-Plex 2

524 The data was generated by Srivatsan *et al.* [35] and downloaded from GEO (GSM4150377). The
525 dataset consists of A549 cells treated with one of the following four compounds: dexamethasone,
526 Nutlin-3a, BMS-345541, or vorinostat (SAHA). The treatment lasted 24 hours across seven different
527 doses. The count matrix obtained from GEO consists of 24,262 cells. During QC we filtered
528 cells with fewer than 500 counts and 720 detected genes. We discarded cells with a percentage of
529 mitochondrial gene counts higher than 10%, thus reducing the dataset to 14,811 cells. Genes present
530 in fewer than 100 cells were discarded. We normalized the data using the size factors provided by
531 the authors and log-transformed it. We selected 5000 HVGs for training and further evaluations.

532 sci-Plex 3

533 The data was generated by Srivatsan *et al.*[35] and downloaded from GEO (GSM4150378). The
534 dataset consists of three cancer cell lines (A549, MCF7, K562), which are treated with 188 different
535 compounds with different mechanisms of action. The cells are treated with 4 dosages (10, 100, 1000,
536 and 10000 nM) plus vehicle. The count matrix obtained from GEO consists of 581,777 cells. The data
537 was subset to half its size, reducing it to 290,888 cells. We then proceeded with log-transformation
538 and the the selection of 5000 HVGs using SCANPY.

539 **Interpretation of combinatorial genetic interactions by perturbation pairs and respon-
540 der genes**

541 In the case of genetic screening, previous work by [5] proposed a set of metrics to annotate and
542 classify gene-gene interactions based on responder genes. Based on this, here we used measured or
543 predicted gene expression differences with respect to control cells (δ), for gene perturbations **a** (δa),
544 **b** (δb) and double perturbations **ab** (δab), to calculate interaction types by similarity between these
545 three expression vectors.

546 More specifically, to calculate association coefficients, we use the linear regression coefficients c_1 and
547 c_2 obtained from the model

$$\delta ab = \delta ac_1 + \delta bc_2 \quad (7)$$

548 To describe interaction modes, we used the following metrics.

- 549 1. **similarity between predicted and observed values:** $dcor(\delta ac_1 + \delta bc_2, \delta ab)$.
- 550 2. **linear regression coefficients:** c_1 and c_2 .
- 551 3. **magnitude:** $(c_1^2 + c_2^2)^{1/2}$.
- 552 4. **dominance:** $|\log_{10}(c_1/c_2)|$.

553 5. **similarity of single transcriptomes:** $dcor(a, b)$

554 6. **similarity of single to double transcriptomes:** $dcor([a, b], ab)$.

555 7. **equal contributions:** $\frac{\min(dcor(a,b),dcor(b,ab)}{\max(dcor(a,b),dcor(a,ab))}$.

556 Following clustering and comparison of these metrics across measured and predicted cells, we decided
557 the following rules of thumb to define and annotate interaction modes:

558 1. **epistatic:** $\min(\text{abs}(c_1), \text{abs}(c_2)) > 0.2$ and either **(i)** ($\text{abs}(c_1) > 2\text{abs}(c_2)$) or **(ii)** ($\text{abs}(c_2) > 2\text{abs}(c_1)$)

560 2. **potentiation:** magnitude > 1 and $\text{abs}(dcor(a, b)) - 1 > 0.2$.

561 3. **strong synergy (similar phenotypes):** magnitude > 1 and $\text{abs}(dcor([a, b], ab)) - 1 > 0.2$

562 4. **strong synergy (different phenotypes):** magnitude > 1 and $\text{abs}(dcor(a, b)) - 1 > 0.5$.

563 5. **additive:** $\text{abs}(\text{magnitude}) - 1 < 0.1$.

564 6. **redundant:** $\text{abs}(dcor([a, b], ab)) - 1 < 0.2$ and $\text{abs}(dcor(a, b)) - 1 < 0.2$

565 More than one genetic interaction can be suggested from these rules. In those cases, we did not
566 assign any plausible interaction. For visualization purposes, we consider perturbed genes with 50 or
567 more interaction modes reported with other co-perturbed genes (**Supplementary Fig.3c**).

568 To visualize differentially expressed genes with similar response across perturbations (**Supplementary**
569 **Fig.3d**), we trained a random forest classifier using as prediction labels *control*, *a*, *b* and *ab* cells,
570 and gene expression as features. We retrieved the top 200 genes from this approach. Then, we
571 annotated the direction (positive or negative) and the magnitude of those changes versus control
572 cells, generating a code for clustering and visualization. To label genes with potential interaction
573 effects, we labeled them if the double perturbation predicted magnitude is 1.5x times or higher than
574 the best value observed in single perturbations.

575 References

576 [1] Pisco, A. O. *et al.* A single cell transcriptomic atlas characterizes aging tissues in the mouse.
577 *BioRxiv* 661728 (2019).

578 [2] Domcke, S. *et al.* A human cell atlas of fetal chromatin accessibility. *Science* **370** (2020).

579 [3] Han, X. *et al.* Construction of a human cell landscape at single-cell level. *Nature* 1–9 (2020).

580 [4] Srivatsan, S. R. *et al.* Massively multiplex chemical transcriptomics at single-cell resolution.
581 *Science* **367**, 45–51 (2020).

582 [5] Norman, T. M. *et al.* Exploring genetic interaction manifolds constructed from rich single-cell
583 phenotypes. *Science* **365**, 786–793 (2019). Publisher: American Association for the Advance-
584 ment of Science Section: Research Article.

585 [6] Yofe, I., Dahan, R. & Amit, I. Single-cell genomic approaches for developing the next generation
586 of immunotherapies. *Nature medicine* **26**, 171–177 (2020).

587 [7] McGinnis, C. S. *et al.* Multi-seq: sample multiplexing for single-cell rna sequencing using
588 lipid-tagged indices. *Nature methods* **16**, 619–626 (2019).

589 [8] Gehring, J., Park, J. H., Chen, S., Thomson, M. & Pachter, L. Highly multiplexed single-cell
590 rna-seq by dna oligonucleotide tagging of cellular proteins. *Nature Biotechnology* **38**, 35–38
591 (2020).

592 [9] Sachs, S. *et al.* Targeted pharmacological therapy restores β -cell function for diabetes remission.
593 *Nature Metabolism* **2**, 192–209 (2020).

594 [10] Kim, K.-T. *et al.* Application of single-cell rna sequencing in optimizing a combinatorial ther-
595 apeutic strategy in metastatic renal cell carcinoma. *Genome biology* **17**, 1–17 (2016).

596 [11] Al-Lazikani, B., Banerji, U. & Workman, P. Combinatorial drug therapy for cancer in the
597 post-genomic era. *Nature biotechnology* **30**, 679–692 (2012).

598 [12] Dixit, A. *et al.* Perturb-seq: Dissecting molecular circuits with scalable single-cell rna profiling
599 of pooled genetic screens. *Cell* **167**, 1853–1866.e17 (2016).

600 [13] Datlinger, P. *et al.* Pooled crispr screening with single-cell transcriptome readout.
601 *Nature methods* **14**, 297–301 (2017).

602 [14] Rozenblatt-Rosen, O., Stubbington, M. J., Regev, A. & Teichmann, S. A. The human cell atlas:
603 from vision to reality. *Nature News* **550**, 451 (2017).

604 [15] Hagai, T. *et al.* Gene expression variability across cells and species shapes innate immunity.
605 *Nature* **563**, 197–202 (2018).

606 [16] Lotfollahi, M., Wolf, F. A. & Theis, F. J. scgen predicts single-cell perturbation responses.
607 *Nature methods* **16**, 715–721 (2019).

608 [17] Lotfollahi, M., Naghipourfar, M., Theis, F. J. & Wolf, F. A. Conditional out-of-distribution
609 generation for unpaired data using transfer vae. *Bioinformatics* **36**, i610–i617 (2020).

610 [18] Yuan, B. *et al.* Cellbox: Interpretable machine learning for perturbation biology with application
611 to the design of cancer combination therapy. *Cell Systems* **12**, 128–140 (2021).

612 [19] Fröhlich, F. *et al.* Efficient parameter estimation enables the prediction of drug response using
613 a mechanistic pan-cancer pathway model. *Cell systems* **7**, 567–579 (2018).

614 [20] Rampášek, L., Hidru, D., Smirnov, P., Haibe-Kains, B. & Goldenberg, A. Dr.VAE: improving
615 drug response prediction via modeling of drug perturbation effects. *Bioinformatics* **35**, 3743–
616 3751 (2019).

617 [21] Kamimoto, K., Hoffmann, C. M. & Morris, S. A. Celloracle: Dissecting cell identity via network
618 inference and in silico gene perturbation. *bioRxiv* (2020).

619 [22] Eraslan, G., Simon, L. M., Mircea, M., Mueller, N. S. & Theis, F. J. Single-cell rna-seq denoising
620 using a deep count autoencoder. *Nature communications* **10**, 1–14 (2019).

621 [23] Lopez, R., Regier, J., Cole, M. B., Jordan, M. I. & Yosef, N. Deep generative modeling for
622 single-cell transcriptomics. *Nature methods* **15**, 1053–1058 (2018).

623 [24] Lotfollahi, M. *et al.* Query to reference single-cell integration with transfer learning. *bioRxiv*
624 (2020).

625 [25] Lopez, R., Gayoso, A. & Yosef, N. Enhancing scientific discoveries in molecular biology with
626 deep generative models. *Molecular Systems Biology* **16**, e9198 (2020).

627 [26] Russkikh, N. *et al.* Style transfer with variational autoencoders is a promising approach to
628 rna-seq data harmonization and analysis. *Bioinformatics* **36**, 5076–5085 (2020).

629 [27] Sohn, K., Lee, H. & Yan, X. Learning structured output representation using deep conditional
630 generative models. *Advances in neural information processing systems* **28**, 3483–3491 (2015).

631 [28] McInnes, L., Healy, J. & Melville, J. Umap: Uniform manifold approximation and projection
632 for dimension reduction. *arXiv preprint arXiv:1802.03426* (2018).

633 [29] Van der Maaten, L. & Hinton, G. Visualizing data using t-sne.
634 *Journal of machine learning research* **9** (2008).

635 [30] Mikolov, T., Sutskever, I., Chen, K., Corrado, G. & Dean, J. Distributed representations of
636 words and phrases and their compositionality. *arXiv preprint arXiv:1310.4546* (2013).

637 [31] Lample, G. *et al.* Fader networks: Manipulating images by sliding attributes. In
638 *Advances in neural information processing systems*, 5967–5976 (2017).

639 [32] Radford, A., Metz, L. & Chintala, S. Unsupervised representation learning with deep convolutional
640 generative adversarial networks. *arXiv preprint arXiv:1511.06434* (2015).

641 [33] Goodfellow, I., Bengio, Y., Courville, A. & Bengio, Y. *Deep learning*, vol. 1 (MIT press Cam-
642 bridge, 2016).

643 [34] Srivatsan, S. R. *et al.* Massively multiplex chemical transcriptomics at single-cell resolution.
644 *Science* **367**, 45–51 (2020). Tex.ids: srivatsanMassivelyMultiplexChemical2020a, srivatsanMas-
645 sivelyMultiplexChemical2020b publisher: American Association for the Advancement of Science
646 section: Research Article.

647 [35] Srivatsan, S. R. *et al.* Massively multiplex chemical transcriptomics at single-cell resolution.
648 *Science* **367**, 45–51 (2020).

649 [36] Musa, A. *et al.* Systems pharmacogenomic landscape of drug similarities from lincs data: drug
650 association networks. *Scientific reports* **9**, 1–16 (2019).

651 [37] Menden, M. P. *et al.* Community assessment to advance computational prediction of cancer
652 drug combinations in a pharmacogenomic screen. *Nature Communications* **10**, 2674 (2019).
653 Number: 1 Publisher: Nature Publishing Group.

654 [38] Adam, G. et al. Machine learning approaches to drug response prediction: challenges and recent
655 progress. *npj Precision Oncology* **4**, 19 (2020).

656 [39] Jia, J. et al. Mechanisms of drug combinations: interaction and network perspectives.
657 *Nature Reviews Drug Discovery* **8**, 111–128 (2009). Number: 2 Publisher: Nature Publishing
658 Group.

659 [40] Gal, Y. & Ghahramani, Z. Dropout as a bayesian approximation: Representing model uncer-
660 tainty in deep learning. In *international conference on machine learning*, 1050–1059 (PMLR,
661 2016).

662 [41] Frangieh, C. J. et al. Multimodal pooled perturb-cite-seq screens in patient models define
663 mechanisms of cancer immune evasion. *Nature genetics* 1–10 (2021).

664 [42] Papalexis, E. et al. Characterizing the molecular regulation of inhibitory immune checkpoints
665 with multimodal single-cell screens. *Nature Genetics* 1–10 (2021).

666 [43] Chen, S., Lake, B. B. & Zhang, K. High-throughput sequencing of the transcriptome and
667 chromatin accessibility in the same cell. *Nature biotechnology* **37**, 1452–1457 (2019).

668 [44] Clark, S. J. et al. scnmmt-seq enables joint profiling of chromatin accessibility dna methylation
669 and transcription in single cells. *Nature communications* **9**, 1–9 (2018).

670 [45] Kaya-Okur, H. S. et al. Cut&tag for efficient epigenomic profiling of small samples and single
671 cells. *Nature communications* **10**, 1–10 (2019).

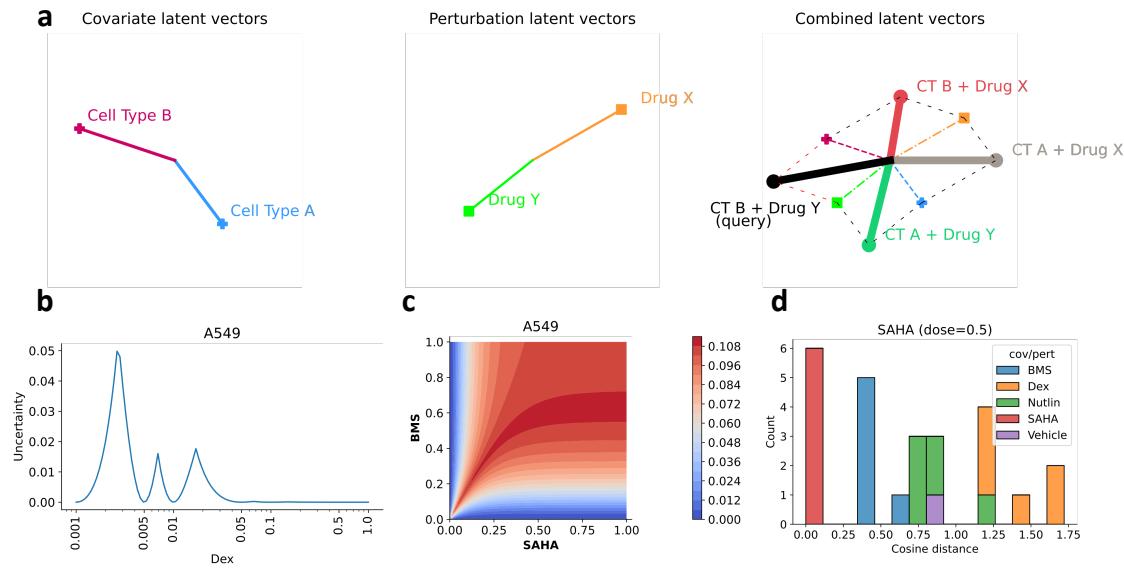
672 [46] Wu, S. J. et al. Single-cell CUT&Tag analysis of chromatin modifications in differentiation and
673 tumor progression. *Nature Biotechnology* 1–6 (2021). Publisher: Nature Publishing Group.

674 [47] van den Brink, S. C. et al. Single-cell and spatial transcriptomics reveal somitogenesis in
675 gastruloids. *Nature* **582**, 405–409 (2020).

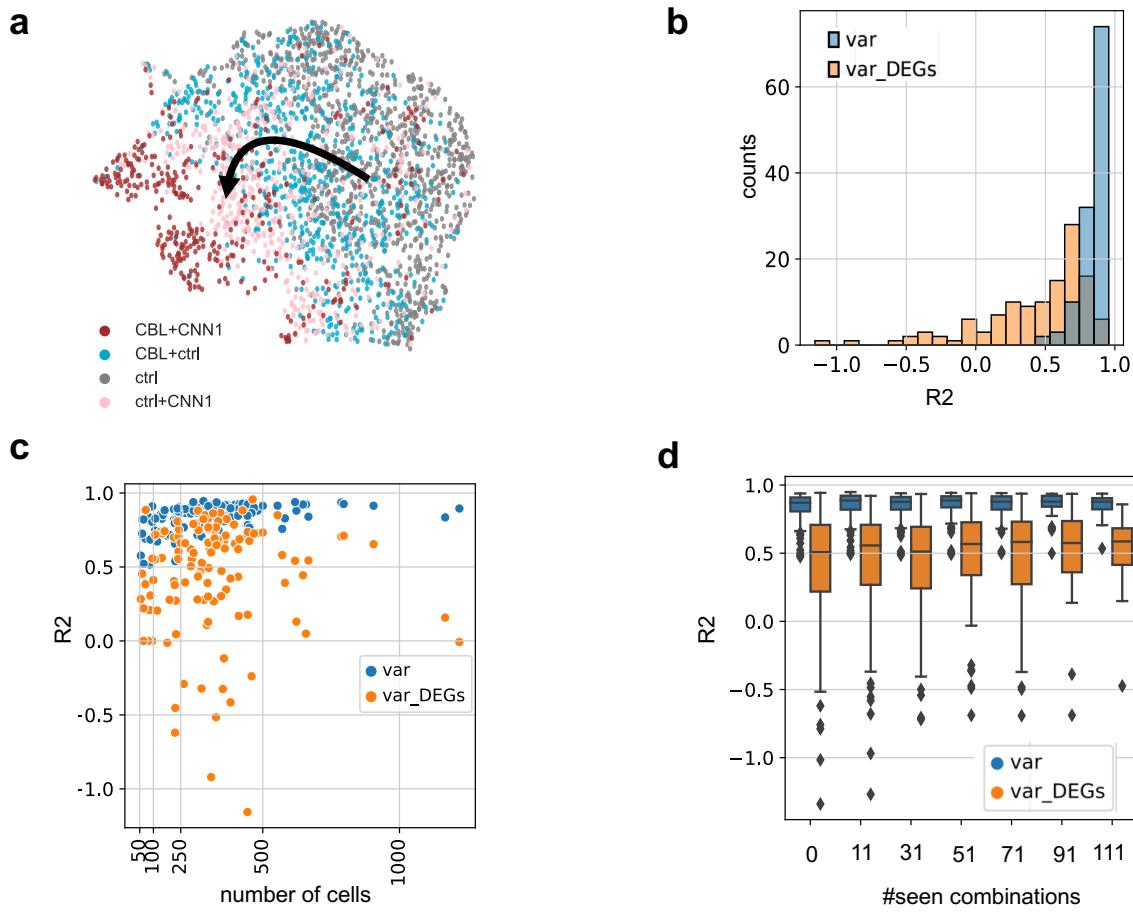
676 [48] Rodriques, S. G. et al. Slide-seq: A scalable technology for measuring genome-wide expression
677 at high spatial resolution. *Science* **363**, 1463–1467 (2019).

678 [49] Mater, A. C. & Coote, M. L. Deep learning in chemistry.
679 *Journal of chemical information and modeling* **59**, 2545–2559 (2019).

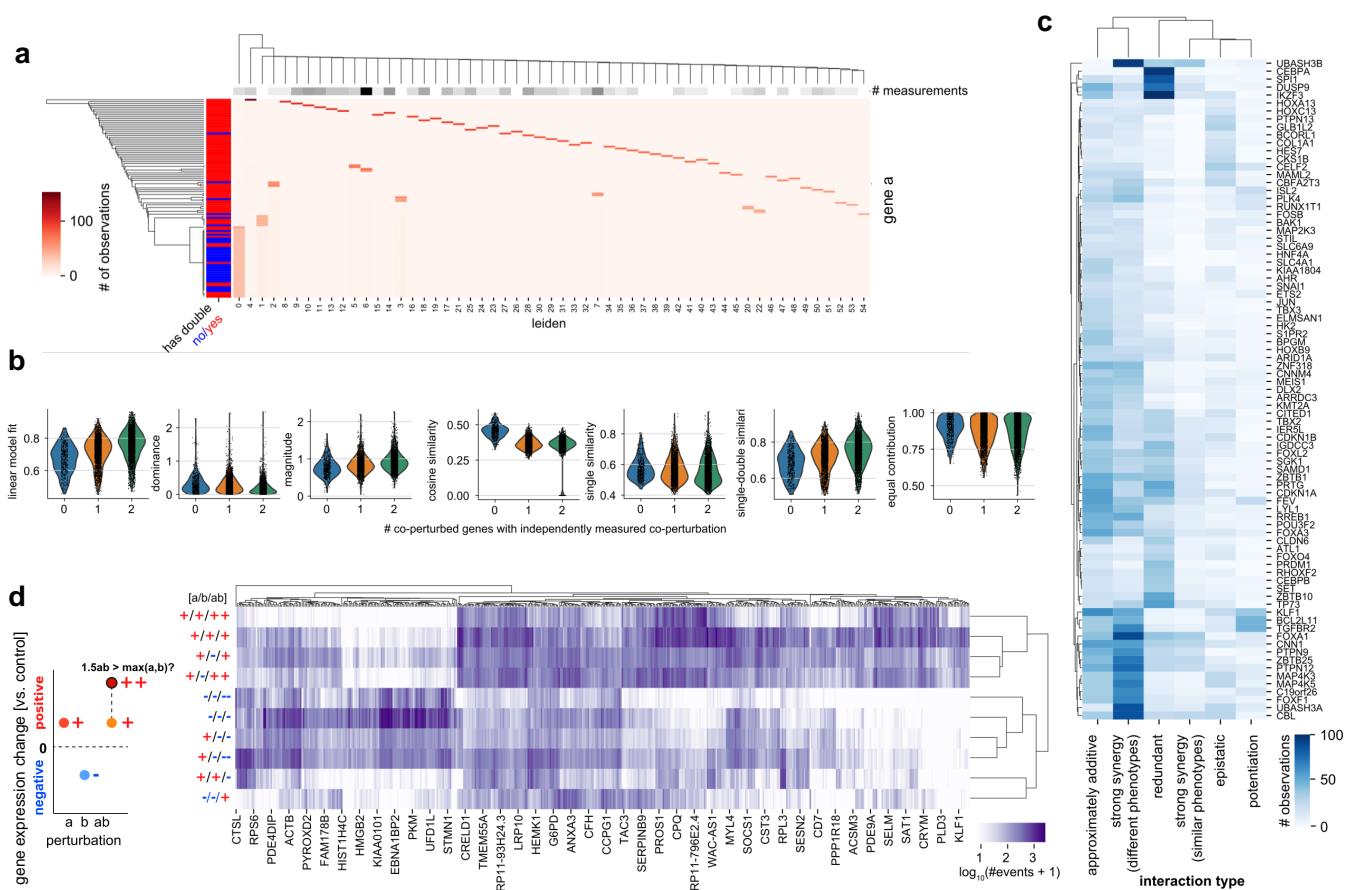
680 [50] Wolf, F. A., Angerer, P. & Theis, F. J. Scanpy: large-scale single-cell gene expression data
681 analysis. *Genome biology* **19**, 1–5 (2018).



Supplementary Figure 1: **CPA uncertainty estimation.** (a) Schematic representation of the steps involved in uncertainty estimation in the case of a dataset with two cell types and two drugs (single dosage per drug). The covariate and perturbation latent vectors are summed in order to generate the set of combinations in the training set. The distances from the query vector and all the vectors in the set are then computed. The closest distance is used as a proxy for uncertainty in the prediction of the model. (b) Example of uncertainty across dosages of Dexamethasone in the sci-Plex 2 dataset. The ticks on the x-axis (log-scaled) indicate dosages seen at training time for which the uncertainty is 0. The dosages were min-max normalized. (c) 2D plot of uncertainty across dosages (min-max normalized) of two different drugs and combinations thereof in the sci-Plex 2 dataset. (d) Example histogram of cosine distances between the SAHA (dose=0.5) and the vectors in the set of training perturbations. The distribution shows that training vectors belonging to the same perturbation but with different dosages have the lowest uncertainties, with other drugs being increasingly more distant.



Supplementary Figure 2: Performance evaluation for CPA combinatorial predictions. (a) UMAP representation of control (ctrl), singly perturbed (CBL+ctrl, ctrl+CNN1) and doubly perturbed (CBL+CNN1) cells. (b) R^2 scores for all genes (blue) or top 100 DEGs (orange) for the prediction of all 131 combinations in the data by training 13 different models and leaving out ≈ 10 combinations each time. (c) Scatter plots of number of samples in the real data for each combination (x-axis) versus R^2 values for the variance of predicted and real for that combination (d) Box-plots of R^2 values for variance for predicted and real cells while increasing the number of combinations seen during training.



Supplementary Figure 3: Gene-gene interaction insights revealed from genetic perturbation predictions using CPA. (a) Number of single gene observations in leiden clusters for generated measurements (from Figure 4i). Most leiden clusters contain a prevalence for one perturbed gene. The majority of genes without measured double perturbations share a limited number of clusters. (b) Quality control and interaction metrics to compare gene expression differences between single and double perturbations. Metrics vary based on number of genes with a measured double perturbation (See **Methods** for definitions). (c) Interaction mode counts predicted for all genes based on interaction metrics (based on [5]). (d) (left) Gene expression changes for double perturbations (ab) versus single perturbations (a, b), are compared by direction and magnitude. Positive (+) and negative (-) labels indicate increase/decrease versus control cells, and double positive/negative (++/- -) indicate values higher than 1.5 times the highest comparable value in single perturbations. (right) 500 genes with highest prevalence in differentially expressed genes across datasets, clustered by prevalent response types from single and double perturbations.

condition	dose_val	R2_mean	R2_mean_DE	R2_var	split	num_cells
SAHA	0.01	0.99	0.99	0.95	test	160
SAHA	0.005	0.98	0.98	0.93	test	143
SAHA	0.05	0.98	0.95	0.93	test	118
SAHA	1.0	0.98	0.95	0.91	test	137
SAHA	0.001	0.97	0.97	0.92	test	169
SAHA	0.1	0.96	0.86	0.94	test	129
SAHA	0.5	0.96	0.86	0.89	ood	604
Nutlin	0.001	0.98	0.98	0.94	test	135
Nutlin	0.05	0.98	0.98	0.94	test	136
Nutlin	0.005	0.98	0.97	0.94	test	107
Nutlin	0.1	0.98	0.97	0.94	test	200
Nutlin	0.01	0.98	0.97	0.93	test	180
Nutlin	0.5	0.92	0.86	0.84	ood	265
Nutlin	1.0	0.26	0.61	0.00	test	1
Dex	0.5	0.99	0.99	0.98	ood	864
Dex	1.0	0.99	0.98	0.95	test	222
Dex	0.1	0.99	0.94	0.96	test	218
Dex	0.05	0.98	0.90	0.93	test	210
Dex	0.001	0.95	0.87	0.89	test	123
Dex	0.01	0.95	0.61	0.89	test	238
Dex	0.005	0.94	0.53	0.88	test	108
BMS	0.001	0.98	0.97	0.92	test	212
BMS	0.005	0.97	0.99	0.92	test	151
BMS	0.5	0.95	0.89	0.78	ood	34
BMS	0.01	0.95	0.80	0.86	test	82
BMS	0.05	0.93	0.87	0.75	test	59
BMS	0.1	0.92	0.89	0.74	test	50
BMS	1.0	0.55	-0.87	0.21	test	6

Supplementary Table 1 | Performance scores for the sci-Plex 2 dataset. To improve readability the columns are sorted by: condition (first priority) and scores (second priority).

condition	R2_mean	R2_mean_DE	method
SAHA	0.98	0.94	linear
SAHA	0.96	0.86	CPA
Nutlin	0.92	0.86	CPA
Nutlin	0.85	0.80	linear
Dex	1.00	1.00	linear
Dex	0.99	0.99	CPA
BMS	0.95	0.89	CPA
BMS	0.89	0.85	linear

Supplementary Table 2 | A simple benchmark on OOD split for the sci-Plex 2 dataset.

condition	dose_val	R2_mean	R2_mean_DE	R2_var	split	num_cells
EGF+RA	0.2+1.0	0.98	0.95	0.84	test	553
EGF+RA	1.0+1.0	0.97	0.94	0.67	test	199
EGF+RA	0.2+0.2	0.96	0.98	0.89	test	87
EGF+RA	0.04+1.0	0.96	0.90	0.60	test	54
EGF+RA	1.0+0.2	0.95	0.91	0.75	test	30
EGF	0.2	0.98	0.96	0.86	test	90
EGF	1.0	0.94	0.71	0.60	test	73
BMP+RA	0.2+1.0	0.91	0.84	0.60	test	22
BMP+EGF+ScripDec	0.2+0.2+1.0	0.97	0.97	0.82	ood	166
BMP+EGF+ScripDec	0.04+0.04+1.0	0.97	0.96	0.61	test	28
BMP+EGF+ScripDec	0.04+0.2+1.0	0.97	0.94	0.50	test	39
BMP+EGF+ScripDec	0.2+0.2+0.2	0.97	0.92	0.65	ood	304
BMP+EGF+ScripDec	0.04+0.2+0.2	0.97	0.91	0.74	test	33
BMP+EGF+ScripDec	0.2+0.04+1.0	0.96	0.96	0.26	test	32
BMP+EGF+ScripDec	0.2+0.04+0.2	0.96	0.94	0.34	test	20
BMP+EGF+ScripDec	1.0+0.2+0.2	0.96	0.91	0.74	ood	113
BMP+EGF+ScripDec	0.2+1.0+1.0	0.95	0.89	0.51	ood	112
BMP+EGF+ScripDec	0.04+0.04+0.2	0.95	0.87	0.52	test	19
BMP+EGF+ScripDec	0.2+1.0+0.2	0.95	0.83	0.58	ood	105
BMP+EGF+ScripDec	0.04+1.0+1.0	0.94	0.96	0.57	test	17
BMP+EGF+ScripDec	1.0+0.04+0.2	0.91	0.86	0.51	test	15
BMP+EGF+RA	0.04+1.0+0.2	0.98	0.99	0.85	test	50
BMP+EGF+RA	0.2+0.04+0.2	0.98	0.98	0.74	test	63
BMP+EGF+RA	0.04+1.0+1.0	0.98	0.97	0.86	test	198
BMP+EGF+RA	0.04+0.2+1.0	0.97	0.96	0.88	test	552
BMP+EGF+RA	0.2+0.04+1.0	0.97	0.96	0.70	test	48
BMP+EGF+RA	0.04+0.2+0.2	0.97	0.92	0.80	test	73
BMP+EGF+RA	0.2+0.2+0.2	0.97	0.88	0.52	ood	206
BMP+EGF+RA	0.04+0.04+0.2	0.96	0.96	0.73	test	24
BMP+EGF+RA	0.2+1.0+0.2	0.96	0.89	0.66	ood	216
BMP+EGF+RA	0.2+1.0+1.0	0.96	0.87	0.66	ood	147
BMP+EGF+RA	0.2+0.2+1.0	0.95	0.77	0.59	ood	132
BMP+EGF	0.04+0.2	0.97	0.98	0.78	test	96
BMP+EGF	0.04+1.0	0.97	0.92	0.81	test	209
BMP+EGF	0.04+0.04	0.97	0.92	0.75	test	39
BMP+EGF	1.0+0.04	0.95	0.86	0.33	test	19
BMP+EGF	1.0+1.0	0.94	0.75	0.06	ood	113

Supplementary Table 3 | Performance scores for the 96-plex-scRNAseq dataset. For the readability the columns are sorted by: condition (first priority) and scores (second priority).

split	species	time	R2_mean	R2_mean_DE	R2_var	num_cells
split4	rat	6.0	0.97	0.91	0.85	7827
split4	rat	6.0	0.96	0.86	0.89	7827
split4	mouse	6.0	0.97	0.90	0.81	5280
split4	mouse	6.0	0.96	0.85	0.93	5280
split3	rat	6.0	0.89	0.70	0.72	7827
split3	rat	6.0	0.86	0.55	0.40	7827
split3	rat	4.0	0.95	0.80	0.80	5755
split3	rat	4.0	0.94	0.77	0.63	5755
split2	rat	6.0	0.55	-0.65	-2.18	7827
split2	rat	6.0	0.54	0.02	0.02	7827
split2	rat	4.0	0.74	-0.47	0.26	5755
split2	rat	4.0	0.39	-0.85	-0.47	5755
split2	rat	2.0	0.81	-0.41	0.49	7025
split2	rat	2.0	0.41	-0.96	-0.64	7025
split1	rat	6.0	0.97	0.91	0.92	7827
split1	rat	6.0	0.97	0.91	0.92	7827
split1	rat	2.0	0.96	0.90	0.90	7025
split1	rat	2.0	0.96	0.90	0.90	7025
split0	rat	6.0	0.97	0.89	0.90	7827
split0	rat	6.0	0.96	0.90	0.81	7827

Supplementary Table 4 | Performance scores for the cross-species dataset across different splits.

species	time	R2_mean	R2_mean_DE	R2_var	split	num_cells
rat	2.0	1.00	1.00	0.99	test	2138
rat	4.0	1.00	1.00	0.99	test	1715
rat	6.0	0.96	0.86	0.89	ood	7827
rabbit	6.0	0.99	0.99	0.99	test	2088
rabbit	2.0	0.99	0.99	0.98	test	2662
rabbit	4.0	0.99	0.96	0.98	test	1732
pig	6.0	0.99	0.99	0.99	test	1535
pig	4.0	0.99	0.98	0.99	test	1954
pig	2.0	0.99	0.98	0.98	test	1662
mouse	2.0	1.00	0.99	0.99	test	2904
mouse	4.0	1.00	0.99	0.99	test	2793
mouse	6.0	0.96	0.85	0.93	ood	5280

Supplementary Table 5 | Performance scores for the cross-species dataset.