SYSTEMMATCH: OPTIMIZING PRECLINICAL DRUG MODELS TO HUMAN CLINICAL OUTCOMES VIA GENERATIVE LATENT-SPACE MATCHING

Scott Gigante, Varsha G. Raghavan, Amanda M. Robinson,
Robert A. Barton, Adeeb H. Rahman, Drausin F. Wulsin,
Jacques Banchereau, Noam Solomon, & Luis F. Voloch
Immunai
New York, NY 10016, USA

Fabian J. Theis
Helmholtz Munich and Technische Universität München
Germany
fabian.theis@helmholtz-muenchen.de

ABSTRACT

Translating the relevance of preclinical models (*in vitro*, animal models, or organoids) to their relevance in humans presents an important challenge during drug development. The rising abundance of single-cell genomic data from human tumors and tissue offers a new opportunity to optimize model systems by their similarity to targeted human cell types in disease. In this work, we introduce SystemMatch to assess the fit of preclinical model systems to an *in sapiens* target population and to recommend experimental changes to further optimize these systems. We demonstrate this through an application to developing *in vitro* systems to model human tumor-derived suppressive macrophages. We show with held-out *in vivo* controls that our pipeline successfully ranks macrophage subpopulations by their biological similarity to the target population, and apply this analysis to rank a series of 18 *in vitro* macrophage systems perturbed with a variety of cytokine stimulations. We extend this analysis to predict the behavior of 66 *in silico* model systems generated using a perturbational autoencoder and apply a k-medoids approach to recommend a subset of these model systems for further experimental development in order to fully explore the space of possible perturbations. Through this use case, we demonstrate a novel approach to model system development to generate a system more similar to human biology.

1 INTRODUCTION

Among most therapeutic areas, failures in clinical trials are common and costly. Failure rates of drugs entering Phase I trials have hit 90% across most therapeutic areas (Mullard, 2016). Oncology in particular has one of the highest rates of clinical trial failures, with only 4% (Mullard, 2016) of therapies entering phase 1 FDA clinical trials ultimately being approved, despite having the most active clinical trials, with approximately 32% of Phase I FDA clinical trials being in oncology (Thomas et al., 2016). In part due to the desire to accelerate new medicines into the clinic to address unmet medical need, but also driven by competition in the industry, drug development organizations do not devote adequate time and resources to improve preclinical model systems that might be more predictive of clinical results (Honkala et al., 2021). Instead, the standard approach adopted by most of the industry relies on *in vitro* and *in vivo* tumor model systems that are poorly predictive of activity in patients due to the reductionist nature of the systems and inadequate attention devoted to understanding the molecular similarities and differences between preclinical and clinical data.

In this work, we describe an end-to-end machine learning pipeline, SystemMatch, that optimizes a preclinical model to best approximate the behavior of a target *in sapiens* population to enable drug

developers to quickly optimize preclinical models and identify those with the greatest predictive power for therapeutic priorities. SystemMatch uses single-cell genomic data from many preclinical models and evaluates them against a multi-study atlas of single-cell genomics data from the tumor or tissue of interest, helping decide which of these models is most likely to provide clinically meaningful predictions. Furthermore, SystemMatch utilizes a Compositional Perturbational Autoencoder (CPA) (Lotfollahi et al., 2021) to predict the behavior of single cells in previously untested combinations of experimental conditions. We use these predictions to recommend experimental changes to the preclinical models to enhance their similarity to the target population.

We demonstrate SystemMatch on a large multi-condition perturbational dataset of *in vitro* differentiating macrophages at the single-cell level, and we compare these model systems' proximity to a target population of human tumor suppressive macrophages obtained from a multi-study single-cell macrophage atlas we collated, integrated, and annotated for this purpose. We demonstrate that SystemMatch produces systems that contradict the standard dogma for generating *in vitro* suppressive macrophages, and we recommend further optimizations of our model system to generate systems with more predictive power for this purpose. This is, to our knowledge, the first computational pipeline for assessing and optimizing the fitness of preclinical models to an *in sapiens* target population.

2 BACKGROUND

2.1 SINGLE-CELL OMICS

Single cell omics refers to the quantitative characterization of biological phenotype at the cellular level. Early work in single-cell omics focused on the development of single-cell RNA sequencing technology (Kolodziejczyk et al., 2015), in which the quantity of gene transcripts (or mRNA) in each cell is counted through complex microfluidic assays. Further work has expanded this technology to the measurement of many modalities, including genes, proteins, metabolites, transcripts, lipids and more (Chen et al., 2020), as well as multimodal data such as CITE-seq, which measures both RNA expression and protein abundance in the same assay (Stoeckius et al., 2017). As the quality and quantity of single-cell omics data rise, it is increasingly straightforward to precisely define rare cell types and hitherto poorly understood cellular heterogeneity (see, e.g., Jaitin et al. (2014); Zeisel et al. (2015)). Bulk RNA sequencing, on the other hand, refers to the sequencing of transcripts present in a large number of cells at the resolution of a cell type or tissue. Most large bulk datasets are not cell type specific, which makes it impractical to understand the phenotypic profile of specific cell subpopulations. For example, with bulk RNAseq data obtained from solid tumors, it is typically not possible to accurately understand the transcriptomic profile of specific immune subpopulations, like CD8 memory T-cells or immunosuppressive macrophages, despite recent work to deconvolve bulk data to cell type resolution (Newman et al., 2015; Finotello et al., 2019).

2.2 PRECLINICAL MODEL SYSTEMS

Due to both the cost and ethical implications of testing novel drugs in humans, most or all drugs are first tested in preclinical models, which range from microorganisms, to cell- and tissue-based models, to animal models including mice and non-human primates. Preclinical data are typically required for FDA approval to test a drug in humans (McElvany, 2009) and are additionally used to prioritize selection of drugs to advance to clinical trial (Denayer et al., 2014). However, failures to translate success in preclinical models to humans have cast doubt on the predictive power of these models, prompting some to question their utility in drug prioritization (see, e.g., Schnabel (2008); Suckling (2008)), with some even going as far as to recommend forgoing animal models altogether and testing drugs directly in humans (Shanks et al., 2009). On the other hand, reverse translational approaches seek to inform the development of preclinical models through the study of clinical success, creating an iterative process between preclinical and clinical studies to optimize later generations of drugs, the success of which can be seen, for example, in the development of EGFR tyrosine kinase inhibitors for the treatment of non-small cell lung carcinoma (Honkala et al., 2021).

2.3 PERTURBATION PREDICTION

Predicting cellular responses to perturbations is an important goal in computational biology. Ji et al. (2021) detail several uses for modeling perturbational single-cell data, including perturbation response prediction, target and mechanism prediction, perturbation interaction prediction, and chemical property prediction. Here, we introduce a new use case for perturbational modeling, which is to predict perturbations that will generate an optimal model system. We generate *in silico* predictions for a wide variety of possible perturbations, then select those closest to our target model system. In this way, we can more rapidly converge on an ideal model system.

3 RELATED WORK

3.1 PRECLINICAL MODEL SYSTEM DEVELOPMENT

Classical protocols of preclinical model system design use direct measurement of the preclinical system's phenotype to evaluate the quality of the system. For example, Mia et al. (2014) select among a set of possible protocols to generate *in vitro* immunosuppressive macrophages by measuring a) secreted proteins known to be markers of immunosuppression and b) *in vitro* suppression of T cells, finding that *M-CSF + IL-4 + IL-10 + TGF*β generates the most suppressive macrophages. Fogg et al. (2020) take a different approach, culturing monocytes with ovarian cancer cell lines in order to understand the pathways activated by the cancer cells, finding an alternative pathway to macrophage polarization through *TGF*α. Reverse translational medicine, on the other hand, uses deep characterization of clinical response to existing drugs to understand the mechanisms of action of these drugs in order to design preclinical models that replicate the drug resistance mechanisms in humans (Honkala et al., 2021). However, to our knowledge, no prior work has been done to optimize a preclinical model's phenotype in the high-dimensional space made visible through single-cell genomics.

3.2 PERTURBATION RESPONSE PREDICTION

Machine learning models have been used in a number of different ways to predict cellular perturbation response Ji et al. (2021). A common setting is *causal imputation* Squires et al. (2020), where a model is required to predict a response to an intervention in a particular context after having been trained on related interventions and contexts. For example, Squires et al. (2020) predict genomic response data by training on perturbations in some cell types and predicting response in other cell types. Lotfollahi et al. (2019) introduces *scGen* for causal imputation on single-cell gene expression data, which uses a variational autoencoder to represent interventions in latent space and then adds the interventions in latent space to an unperturbed representation to obtain the perturbational response. In our setting, where we wish to predict perturbations that generate gene expression corresponding to an *in vivo* model system, combinations of treatments are also important to consider. Recently, the Compositional Perturbational Autoencoder Lotfollahi et al. (2021) was demonstrated to be able to predict combinations of gene knockouts from being trained on each knockout individually.

4 SYSTEMMATCH

Here, we introduce our iterative preclinical model optimization pipeline, SystemMatch. Fig. 1 shows a schematic representation of the SystemMatch process. First, in order to define the *target* population, we collect an atlas of single-cell data from relevant clinical cohorts giving a robust universal representation of the cell type of interest across multiple disease states. We apply a simple reference-based cell type annotation (see, e.g., Hao et al. (2021); Lopez et al. (2018)) to only retain the cells associated with the target population (e.g., all macrophages). Next, we integrate this single-cell atlas using a fixed gene list derived from expert domain knowledge to capture only the biology relevant to the system at hand and use this as input to a batch correction algorithm (see, e.g., Korsunsky et al. (2019); Hao et al. (2021); Haghverdi et al. (2018)). Within this integrated single-cell cell type–specific atlas, we annotate subtypes using a combination of graph-based clustering (Traag et al., 2019) and expert domain knowledge. We choose the target population from these subtypes as the subtype that most closely recapitulates the phenotype of the target population as described in the literature (e.g., macrophages that express known markers of immunosuppression).

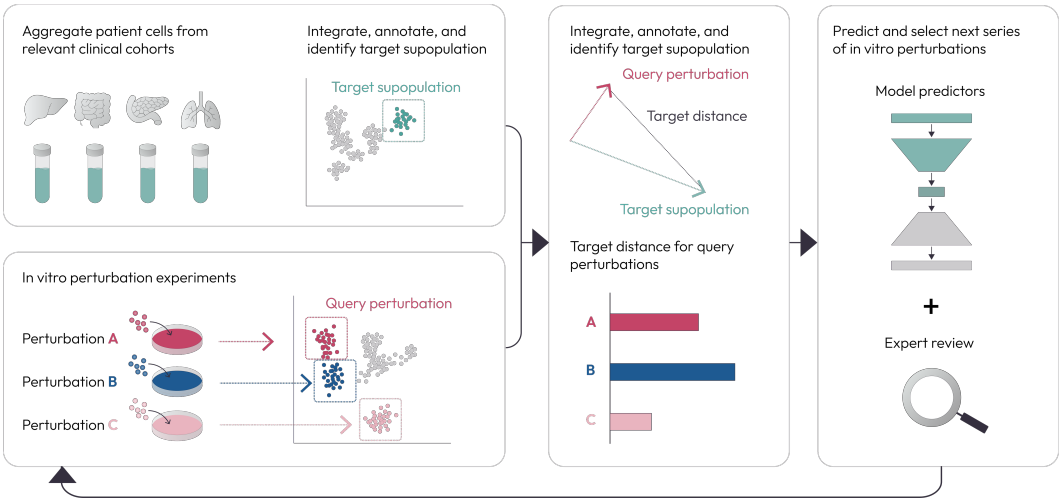


Figure 1: Schematic of the iterative SystemMatch pipeline.

Second, we collect the *query* model systems. We generate this data by developing many different model systems in parallel experiments (here, *in vitro* macrophage differentiation and polarization) with multiple experimental conditions (here, stimulation with different cytokines and combinations thereof) selected from prior domain knowledge to generate a heterogeneous set of model systems. We perform single-cell sequencing on the cells of interest from each model system and then perform quality control on these datasets to remove low quality and outlier cells in order to obtain a clean single-cell resolution representation of each model system, and project these datasets to a subspace of the full gene expression matrix by retaining only those genes known to be related to the biological function of interest.

Next, we compute the distances in this subspace between each query dataset and the target population using some distance metric m. In the simplest case, m is simply the L2 distance between the average (or "pseudobulked") gene expression of the query $\mathbb{X} := \{x_i \in \mathbb{R}^d\}$ and the target $\mathbb{Y} := \{y_i \in \mathbb{R}^d\}$

$$m_{L2}(\mathbb{X}, \mathbb{Y}) := \left\| \frac{\sum_i^{n_\mathbb{X}} x_i}{n_\mathbb{X}} - \frac{\sum_i^{n_\mathbb{Y}} y_i}{n_\mathbb{Y}} \right\|_2$$

where $n_\mathbb{X} := |X|$ and $n_\mathbb{Y} := |Y|$. However, other more complex distance metrics between single-cell datasets could also be used, e.g. the Earth Mover's Distance (EMD) or Wasserstein metric (Kantorovich, 1960)

$$m_{EMD}(\mathbb{X}, \mathbb{Y}) := \min_{\boldsymbol{F}} \sum_{i=1}^{n_\mathbb{Y}} \sum_{j=1}^{n_\mathbb{X}} F_{i,j} d(x_i, x_j)$$

where $\boldsymbol{F}$ is the $n_\mathbb{X} \times n_\mathbb{Y}$ flow matrix with $F_{i,j} \geq 0$ and d is a distance metric between cells, e.g. the Euclidean distance. Then, using our choice of m, we rank all query datasets to produce an ordered set of queries, with the query least distant from the target denoted the most representative model system of those tested.

However, this still leaves an open question: can we further improve the models beyond the original set of tested systems? To assist with this question, we employ generative deep neural networks to generate all possible combinations of conditions tested in the experimental queries to generate combinatorially many *in silico* model systems. Then, in order to avoid over-reliance on the accuracy of these predictions, we leverage the *in silico* queries to generate a search space of possible experimental conditions that are substantially different from the systems already tested. We use a modified form of the k-medoids algorithm (Kaufman & Rousseeuw, 2009) applied to the pseudo-bulked *in silico* queries. Briefly, k-medoids selects k equidistant model centroids as in k-means but requires

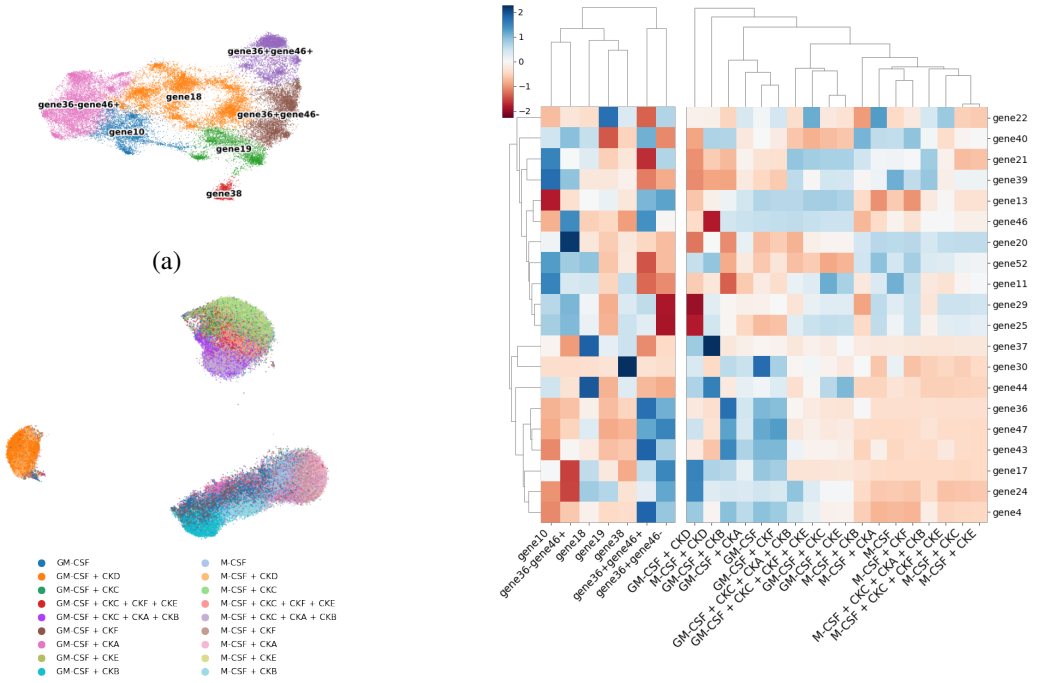


Figure 2: Characterization of human tumor macrophage atlas and *in vitro* stimulated macrophages. a) Human tumor macrophage atlas UMAP shows seven major clusters; b) *in vitro* macrophage UMAP shows *in vitro* conditions separate into three major groups; c) heatmap of major macrophage function genes in human macrophages (left) and *in vitro* macrophages (right) shows the *gene36-gene46+* cluster is most strongly associated with known suppressive response genes *gene20* and *gene46*, while this combination is absent in vitro.

that these centroids be selected from the existing data points. We extend this algorithm to enforce that all selected medoids are also equidistant from all existing experimentally tested queries, which leaves us with the smallest possible subset of experiments to run in the next iteration while ensuring that we do not leave any region of the search space untested. We can further combine these recommendations with expert knowledge by examining the genes driving the difference between our best-ranked query and the target population to propose additional experimental conditions.

Finally, we use the outputs of this pipeline to re-run the model system generation experiment, and we iterate upon this process until the generated model system is either sufficiently similar to the target population, or further iterations fail to improve upon the existing queries.

4.1 GENERATING *in silico* QUERIES

Since the space of possible model systems is extremely large and diverse, we are unable to test all possible experimental conditions *in vitro*. Hence, to better explore the space of untested systems that could best emulate the target population, we predict the gene expression of combinations of the cytokine stimulations in our *in vitro* experiment. We use CPA Lotfollahi et al. (2021) to generate *in silico* perturbed cells for 66 additional conditions. Briefly, CPA is an autoencoder trained to decompose the data into disentangled latent representations for three key attributes of each cell: the cell's "basal state", the perturbation effect, and the covariate effect. These latent representations are combined in the decoder and trained via a reconstruction loss. To enforce independence of the latent representations, CPA is trained using a discriminator network and adversarial loss such that no signal from the observed perturbation or covariates is captured in the cellular basal state embedding. As the latent representations are combined by summation, CPA can combine multiple perturbations, allowing us to generate *in silico* samples for combinations of cytokine stimulations. We further

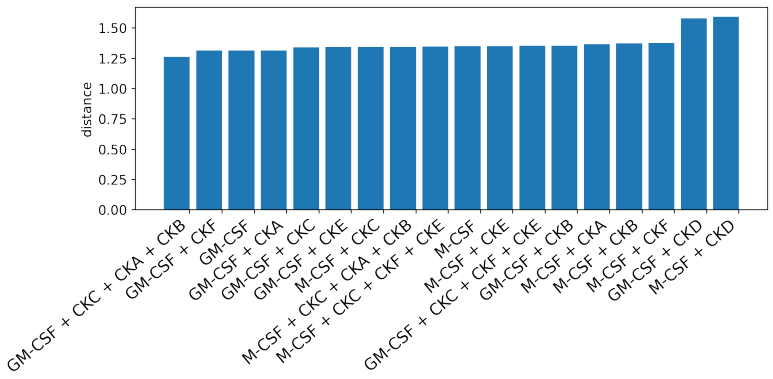


Figure 3: SystemMatch-produced ranking of *in vitro* conditions based on their computed similarity to the target condition.

modify CPA to allow the incorporation of multiple covariates in order to account for experimental batch covariates included in our *in vitro* dataset.

4.2 Selecting new queries for experimental iteration

To select a subset of the generated *in silico* perturbations for experimental validation, we devise a scheme based on k-medoids (Kaufman & Rousseeuw, 2009) which ensures that the conditions selected for experimental validation are a) sufficiently different from the conditions already tested; and b) sufficiently heterogeneous to cover the space of possible perturbations. Briefly, we run the regular k-medoids algorithm, but at the medoid-update step of the algorithm, we consider each of the existing *in vitro* perturbations as fixed medoids; this way, the k-medoids algorithm selects k new perturbations for testing which are both heterogeneous and far from any of the previously tested conditions.

5 Results

5.1 Characterizing *in vitro* and *in vivo* tumor associated macrophages

We collated public and proprietary data from tumour-infiltrating immune cells from one study we generated internally and four public studies (Zhang et al., 2021; Bassez et al., 2021; Yost et al., 2019; Qian et al., 2020), comprising of 125 patients across five tumor types. We filtered these data to just retain the macrophage populations using reference-based mapping (Hao et al., 2021). We then filtered low-quality cells (defined as having fewer than 1500 genes with non-zero counts) and run batch integration across all 125 patients using Harmony (Korsunsky et al., 2019) in a curated subspace of 319 macrophage function genes. We then ran Leiden clustering (Traag et al., 2019) and characterized each population independently according to expression of important marker genes. Finally, we selected the *gene36-gene46+* population as our target population, as it expressed known markers associated with suppressive macrophages and correlated with poor clinical response to common immunotherapies.

Fig. 2a shows the UMAP (McInnes et al., 2018) dimensionality reduction of the integrated macrophage atlas annotated by its marker genes. We identify inflammatory, suppressive, proliferating, and co-suppressive/inflammatory cell subsets, the last of which is not typically described in reviews of macrophage biology. For the purposes of this study, we target the suppressive population for *in vitro* system optimization.

To study the heterogeneity of macrophage development systems, we differentiated monocytes *in vitro* with either *M-CSF* or *GM-CSF* for 6 days and then stimulated these macrophages for an additional 4 days with one of nine combinations of cytokines typically understood to play a role in regulatory macrophage phenotype. Fig. 2b shows the UMAP dimensionality reduction of the *in*

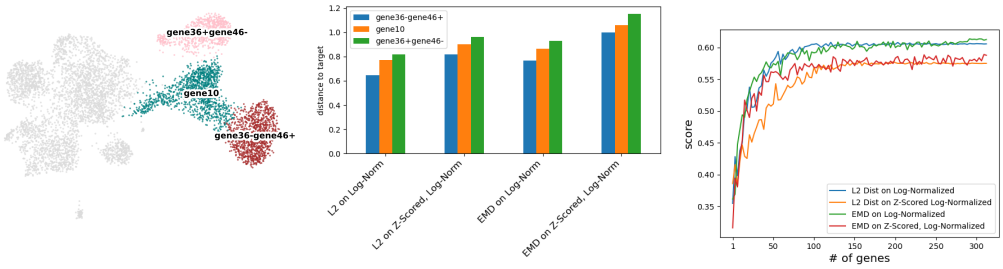


Figure 4: Quantitative evaluation of distance metrics for ranking data with known ground truth. UMAP (left) shows the annotated held-out, ground truth macrophage data; computed distances (middle) between the held-out macrophage populations and the target population for each method; evaluation scores (right) of each distance metric relative to the known ground truth.

vitro systems. We see that these systems primarily separate into three distinct groups, which can be described as pro-inflammatory, pro-suppressive, and basal.

Finally, to compare the *in vitro* cells to our human tumor macrophage atlas, we show a heatmap of selected marker genes for both datasets (Fig. 2c). We observe qualitatively that not one of the *in vitro* systems fully recapitulates the gene expression of the suppressive *gene36-gene46+* population. This is further recapitulated by the similarity ranking of the *in vitro* systems to the target population (Fig. 3), in which a) the best-ranked system using M-CSF, which serves as the basis for all systems most commonly used to generate suppressive macrophages (Mia et al., 2014; Fogg et al., 2020) is ranked 7th; and b) the closest system only reduces the distance to the target by 20% from the least common system, which is well known to produce a pro-inflammatory response.

5.2 Evaluating similarity ranking methods

To validate the effectiveness of SystemMatch, we compare held-out ground truth macrophage data from Tang-Huau et al. (2018) to our human tumor macrophage atlas. We preprocessed the held-out data by filtering out low-quality cells and annotated cells in the dataset using the same gene set and protocol as for the atlas, marking the macrophages as highly suppressive, moderately suppressive, or nonsuppressive. Fig.4 (left) shows the UMAP of the final annotated clusters of the ground truth macrophage dataset.

Because our target cells are suppressive macrophages, we expected the ranking of the ground truth cells regarding their proximity to the target suppressive cells to be from most suppressive (*gene36-gene46+*) to least suppressive (*gene36+gene46-*). The following distance metrics were considered (see Methods): 1) L2 distance on log-normalized pseudobulked gene expression; 2) L2 distance on z-scored, log-normalized pseudobulked gene expression; 3) Earth Mover's Distance on log-normalized single cell gene expression; 4) Earth Mover's Distance on z-scored, log-normalized single cell gene expression. On the full dataset, all four distance metrics correctly ranked the three datasets (Fig. 4 (middle)), so we define the following score to quantitatively evaluate a) how well it correctly ranked the held-out conditions and b) the separation of the most suppressive condition from the least suppressive condition.

For a given distance metric m, query subsets $\{\mathbb{X}_i \mid i \in \{$*gene36-gene46+*, *gene10*, *gene36+gene46-*$\}\}$ and target population $\mathbb{Y}$, we compute distances between the conditions in the held-out dataset to the target condition giving a distance vector

$$d^m = (m(\mathbb{X}_{\textit{gene36-gene46+}}, \mathbb{Y}), m(\mathbb{X}_{\textit{gene10}}, \mathbb{Y}), m(\mathbb{X}_{\textit{gene36+gene46-}}, \mathbb{Y}))$$

with corresponding rank vector r^m, where the expected rank vector $r^{\textit{true}} = (1, 2, 3)$. For a ranking r_m of length n, we define the score

$$\text{Score(m)} = \left(\frac{\sum_i^n I(r_i^m = r_i^{\textit{true}})}{n} + \frac{d^m_{gene36+gene46-} - d^m_{gene36-gene46+}}{d^m_{gene36+gene46-}} \right) /2$$

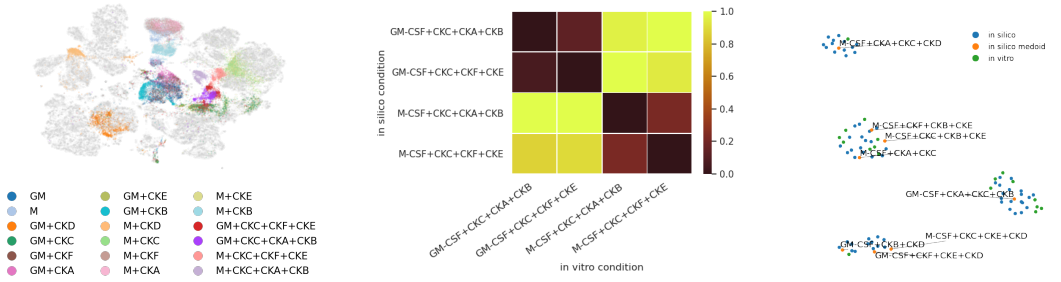


Figure 5: Sample output of the experiment recommender system. UMAP embedding of *in vitro* and *in silico* single cells (left), with *in silico* predictions shown in grey; heatmap of normalized distances from predicted triplets to held-out triplets (middle); UMAP embedding of average expression per condition (right), with k-medoids–recommended perturbations highlighted in orange.

We compute this score for different levels of dataset corruption, which we implement by subsampling genes to remove information. We repeat this subsampling up to 200 times, or until the score converges. Fig. 4 (right) shows the scores for the four metrics; EMD with log-normalized genes performs best with the full gene set, but pseudo-bulked L2 distance is most robust to corruption.

5.3 Recommending future experiments

In order to leverage the recommendations of SystemMatch to further improve the *in vitro* systems beyond those initially tested, we combine compositional perturbational autoencoders with a modified k-medoids algorithm to select a subset of possible combinations of the initially tested cytokines to test in the next experimental iterations. First, we use a modified form of CPA (Lotfollahi et al., 2021) to generate all possible double- and triple-combinations of cytokines (Fig. 5 (left)). We see that *M/GM-CSF* and *CKD* drive the most significant differences in the predicted cells, which is consistent with our understanding of these cytokines. Further, we see as expected that the double-combinations of cytokines are generally more similar to the *in vitro* single cytokine perturbations.

To validate the accuracy of our *in silico* predictions, we compared the similarity of the held-out *in vitro* triplet perturbations to the corresponding *in silico* perturbations. Fig. 5 (middle) shows the distance (using the L2 pseudobulk distance normalized to a range of $[0, 1]$) from each of the held-out triplets to the corresponding predicted triplets. All of the four are closest to the corresponding prediction. Equivalent heatmaps for the other distance metrics are shown in Fig. 7.

Next, we run our modified k-medoids algorithm on the average expression on each condition to select a subset of these double- and triple-combinations to test experimentally in future work to maximize the heterogeneity of all tested perturbations. Alternatively, we can use SystemMatch to test the proximity of the generated conditions; this will produce a less heterogeneous set of combinations for further testing but will optimize more directly towards the target condition. The result of this comparison is shown in Tab. 9; interestingly, the top predicted condition uses M-CSF, even though none of the top six *in vitro* conditions did. The *in silico* preference (however mild) for *M-CSF* over *GM-CSF* is consistent with commonly used models of suppressive macrophages (Mia et al., 2014; Fogg et al., 2020), giving confidence in our recommendations.

6 Discussion

In this work we provide an end-to-end ML pipeline for assessing the fitness of and optimizing preclinical models to maximize their predictive power of clinical results. We do this through an iterative process of comparing model systems to the target *in sapiens* population, applying perturbation prediction to suggest experimental changes, and iterating upon this process to produce preclinical models that are more representative of the relevant tumor or tissue data. We demonstrate our pipeline through a use case of developing an *in vitro* system optimized to produce macrophages most similar to human tumor-derived suppressive macrophages, and we recommend a set of further experiments

to fully explore the space of possible perturbations and ultimately generate a system more suited to developing drugs targeting these cells than those systems currently used in the literature.

We validate our pipeline's performance through a series of tests using ground truth data with a held-out *in sapiens* public dataset in which we manually annotate cells known to be close and far from the target population. We show that a) our pipeline is largely robust to the choice of distance metric; b) distance metrics computed on average expression are generally most robust to data corruption; and c) in the limit of low data corruption, single-cell optimal transport metrics may outperform these "pseudobulk" methods. We then validate our perturbation prediction using experimentally generated *in vitro* controls of cytokine combinations, through which we show that the *in silico* generated combinations are most similar to the corresponding matching *in vitro* populations.

We note that the robustness to the use of average expression (compared to single-cell) indicates that bulk RNA-seq could suffice in some cases for the query datasets; however, this would preclude the use of CPA, and the resulting lower-resolution data would provide less predictive power to generate *in silico* samples. Additionally, in the case of heterogeneous model systems (e.g. in systems with developmental trajectories or large numbers of cycling cells), single-cell metrics may score more highly. We also note that the use of *in silico* generated combinations of experimental conditions explores only a subset of possible model systems; systems generated with conditions not included in the initial experiment cannot be discovered in this way, and must instead be added to the iterative process by manual expert review.

In recent years, the utility of preclinical model data to evaluate the clinical relevance of a drug has been questioned. We show here, through a combination of single-cell genomic data and machine learning, a method by which these preclinical model systems can be optimized to enhance their potential predictive power. While ML methods have generated a great deal of impact in target discovery and validation, this is to our knowledge the first ML pipeline for actually establishing the preclinical model systems in which those targets can be evaluated.

REFERENCES

Ayse Bassez, Hanne Vos, Laurien Van Dyck, Giuseppe Floris, Ingrid Arijs, Christine Desmedt, Bram Boeckx, Marlies Vanden Bempt, Ines Nevelsteen, Kathleen Lambein, et al. A single-cell map of intratumoral changes during anti-pd1 treatment of patients with breast cancer. *Nature Medicine*, 27(5):820–832, 2021.

Wenyi Chen, Shumin Li, Anuja Shreeram Kulkarni, Lin Huang, Jing Cao, Kun Qian, and Jingjing Wan. Single cell omics: from assay design to biomedical application. *Biotechnology Journal*, 15 (1):1900262, 2020.

Tinneke Denayer, Thomas Stöhr, and Maarten Van Roy. Animal models in translational medicine: Validation and prediction. *New Horizons in Translational Medicine*, 2(1):5–11, 2014.

Francesca Finotello, Clemens Mayer, Christina Plattner, Gerhard Laschober, Dietmar Rieder, Hubert Hackl, Anne Krogsdam, Zuzana Loncova, Wilfried Posch, Doris Wilflingseder, et al. Molecular and pharmacological modulators of the tumor immune contexture revealed by deconvolution of rna-seq data. *Genome medicine*, 11(1):1–20, 2019.

Kaitlin C Fogg, Andrew E Miller, Ying Li, Will Flanigan, Alyssa Walker, Andrea O'Shea, Christina Kendziorski, and Pamela K Kreeger. Ovarian cancer cells direct monocyte differentiation through a non-canonical pathway. *BMC cancer*, 20(1):1–14, 2020.

Laleh Haghverdi, Aaron TL Lun, Michael D Morgan, and John C Marioni. Batch effects in single-cell rna-sequencing data are corrected by matching mutual nearest neighbors. *Nature biotechnology*, 36(5):421–427, 2018.

Yuhan Hao, Stephanie Hao, Erica Andersen-Nissen, William M Mauck III, Shiwei Zheng, Andrew Butler, Maddie J Lee, Aaron J Wilk, Charlotte Darby, Michael Zager, et al. Integrated analysis of multimodal single-cell data. *Cell*, 184(13):3573–3587, 2021.

Alexander Honkala, Sanjay V Malhotra, Shivaani Kummar, and Melissa R Junttila. Harnessing the predictive power of preclinical models for oncology drug development. *Nature Reviews Drug Discovery*, pp. 1–16, 2021.

Diego Adhemar Jaitin, Ephraim Kenigsberg, Hadas Keren-Shaul, Naama Elefant, Franziska Paul, Irina Zaretsky, Alexander Mildner, Nadav Cohen, Steffen Jung, Amos Tanay, et al. Massively parallel single-cell rna-seq for marker-free decomposition of tissues into cell types. *Science*, 343 (6172):776–779, 2014.

Yuge Ji, Mohammad Lotfollahi, F Alexander Wolf, and Fabian J Theis. Machine learning for perturbational single-cell omics. *Cell Systems*, 12(6):522–537, 2021.

Leonid V Kantorovich. Mathematical methods of organizing and planning production. *Management science*, 6(4):366–422, 1960.

Leonard Kaufman and Peter J Rousseeuw. *Finding groups in data: an introduction to cluster analysis*. John Wiley & Sons, 2009.

Aleksandra A Kolodziejczyk, Jong Kyoung Kim, Valentine Svensson, John C Marioni, and Sarah A Teichmann. The technology and biology of single-cell rna sequencing. *Molecular cell*, 58(4): 610–620, 2015.

Ilya Korsunsky, Nghia Millard, Jean Fan, Kamil Slowikowski, Fan Zhang, Kevin Wei, Yuriy Baglaenko, Michael Brenner, Po-ru Loh, and Soumya Raychaudhuri. Fast, sensitive and accurate integration of single-cell data with harmony. *Nature methods*, 16(12):1289–1296, 2019.

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, Yuge Ji, Ignacio L Ibarra, F Alexander Wolf, Nafissa Yakubova, Fabian J Theis, and David Lopez-Paz. Compositional perturbation autoencoder for single-cell response modeling. *bioRxiv*, 2021.

Karen D McElvany. Fda requirements for preclinical studies. In *Clinical Trials in the Neurosciences*, volume 25, pp. 46–49. Karger Publishers, 2009.

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

Sohel Mia, Andreas Warnecke, X-M Zhang, Vivianne Malmström, and Robert A Harris. An optimized protocol for human m2 macrophages using m-csf and il-4/il-10/tgf-β yields a dominant immunosuppressive phenotype. *Scandinavian journal of immunology*, 79(5):305–314, 2014.

Asher Mullard. Parsing clinical success rates. *Nature Reviews Drug Discovery*, 15(7):447–448, 2016.

Aaron M Newman, Chih Long Liu, Michael R Green, Andrew J Gentles, Weiguo Feng, Yue Xu, Chuong D Hoang, Maximilian Diehn, and Ash A Alizadeh. Robust enumeration of cell subsets from tissue expression profiles. *Nature methods*, 12(5):453–457, 2015.

Junbin Qian, Siel Olbrecht, Bram Boeckx, Hanne Vos, Damya Laoui, Emre Etlioglu, Els Wauters, Valentina Pomella, Sara Verbandt, Pieter Busschaert, et al. A pan-cancer blueprint of the heterogeneous tumor microenvironment revealed by single-cell profiling. *Cell research*, 30(9):745–762, 2020.

Jim Schnabel. Neuroscience: Standard model. *Nature*, 454(7205):682–685, Aug 2008. ISSN 1476-4687. doi: 10.1038/454682a. URL https://doi.org/10.1038/454682a.

Niall Shanks, Ray Greek, and Jean Greek. Are animal models predictive for humans? *Philosophy, ethics, and humanities in medicine*, 4(1):1–20, 2009.

Chandler Squires, Dennis Shen, Anish Agarwal, Devavrat Shah, and Caroline Uhler. Causal imputation via synthetic interventions. *arXiv preprint arXiv:2011.03127*, 2020.

Marlon Stoeckius, Christoph Hafemeister, William Stephenson, Brian Houck-Loomis, Pratip K Chattopadhyay, Harold Swerdlow, Rahul Satija, and Peter Smibert. Simultaneous epitope and transcriptome measurement in single cells. *Nature methods*, 14(9):865–868, 2017.

Keith Suckling. Animal research: too much faith in models clouds judgement. *Nature*, 455(7212): 460–460, Sep 2008. ISSN 1476-4687. doi: 10.1038/455460b. URL https://doi.org/10.1038/455460b.

Tsing-Lee Tang-Huau, Paul Gueguen, Christel Goudot, Mélanie Durand, Mylène Bohec, Sylvain Baulande, Benoit Pasquier, Sebastian Amigorena, and Elodie Segura. Human in vivo-generated monocyte-derived dendritic cells and macrophages cross-present antigens through a vacuolar pathway. *Nature communications*, 9(1):1–12, 2018.

David W Thomas, Justin Burns, John Audette, Adam Carroll, Corey Dow-Hygelund, and Michael Hay. Clinical development success rates 2006–2015. *BIO Industry Analysis*, 1(16):25, 2016.

Vincent A Traag, Ludo Waltman, and Nees Jan Van Eck. From louvain to leiden: guaranteeing well-connected communities. *Scientific reports*, 9(1):1–12, 2019.

Shun Yip, Pak Sham, and Junwen Wang. Evaluation of tools for highly variable gene discovery from single-cell rna-seq data. *Briefings in bioinformatics*, 20, 02 2018. doi: 10.1093/bib/bby011.

Kathryn E Yost, Ansuman T Satpathy, Daniel K Wells, Yanyan Qi, Chunlin Wang, Robin Kageyama, Katherine L McNamara, Jeffrey M Granja, Kavita Y Sarin, Ryanne A Brown, et al. Clonal replacement of tumor-specific t cells following pd-1 blockade. *Nature medicine*, 25(8):1251–1259, 2019.

Amit Zeisel, Ana B Muñoz-Manchado, Simone Codeluppi, Peter Lönnerberg, Gioele La Manno, Anna Juréus, Sueli Marques, Hermany Munguba, Liqun He, Christer Betsholtz, et al. Cell types in the mouse cortex and hippocampus revealed by single-cell rna-seq. *Science*, 347(6226):1138–1142, 2015.

Yuanyuan Zhang, Hongyan Chen, Hongnan Mo, Xueda Hu, Ranran Gao, Yahui Zhao, Baolin Liu, Lijuan Niu, Xiaoying Sun, Xiao Yu, et al. Single-cell analyses reveal key immune cell subsets associated with response to pd-l1 blockade in triple-negative breast cancer. *Cancer Cell*, 39(12): 1578–1593, 2021.

A Appendix

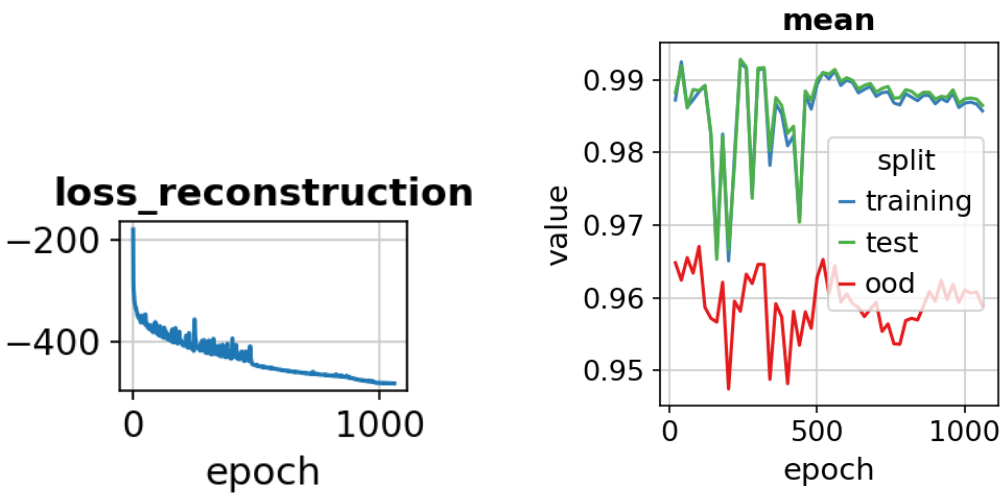
A.1 CPA Model Training Details

The CPA model was trained, tested, and evaluated on the *in vitro* dataset, in which cells were treated in eighteen separate conditions. Of these conditions, two were designated as control conditions (*GM-CSF* and *M-CSF*), four were considered triplet combinations of perturbations (e.g., *GM-CSF + CKC + CKA + CKB*), and the other twelve were single cytokine perturbations (e.g., *GM-CSF + CKF*). Since CPA is able to learn an independent transcriptional embedding per single perturbation and then add these embeddings to predict multiplet perturbations, we chose to train and validate the model on the control and singular cytokine perturbation conditions and evaluate its performance on the triplet perturbation conditions. The training dataset consisted of 417 genes deemed relevant to macrophage function by a combination of domain expertise and data-driven highly variable genes (Yip et al., 2018) from our human tumor macrophage atlas.

We ran CPA with an autoencoder width of 512, batch size of 128, and embedding size of 128, and we trained it for six hours over 1060 epochs on an NVIDIA Tesla A100 GPU. The reconstruction loss and model performance on the held-out triplet conditions (OOD) are shown in Fig. 6. We used this model to predict additional conditions, all of which were doublet or triplet cytokine conditions that were not present in the *in vitro* assay, in order to investigate which would best match our target model system.

A.2 GENERATION OF *in vitro* MACROPHAGE MODEL SYSTEMS

Human CD14+ Monocytes were isolated from peripheral mononuclear blood cells (PBMCs) using the EasySep™ Human Monocyte Isolation Kit (STEMCELL Technologies) following manufacturer's instructions. Monocytes were plated at $1x10^6$ cells per well in 6 well tissue culture treated plates (Corning). Cells were plated in 2.5mL RPMI media (Life Technologies) supplemented with 10% FBS (Life Technologies) and 1% Penicillin-Streptomycin (Thermo Fisher Scientific) and either *M-CSF* (PeproTech) or *GM-CSF* (PeproTech), respectively, at 20ng/mL. At day 3, media was aspirated and 2.5mL fresh *M-CSF/GM-CSF* media was added. At day 6, media was aspirated and 3mL *M-CSF/GM-CSF* media supplemented with additional differentiation cytokines was added to respective wells. Cells were isolated at day 10 with cell scrapers (Fisher Scientific) and counted with the Cellaca MX High-throughput Automated Cell Counter (Nexcelom) following manufacturer's instructions prior to single cell processing.



(a) (b)

Figure 6: Evaluation of model performance. (a) Reconstruction loss per training epoch of the CPA model. (b) R-squared between mean predicted gene expression versus actual gene expression over all 417 genes.

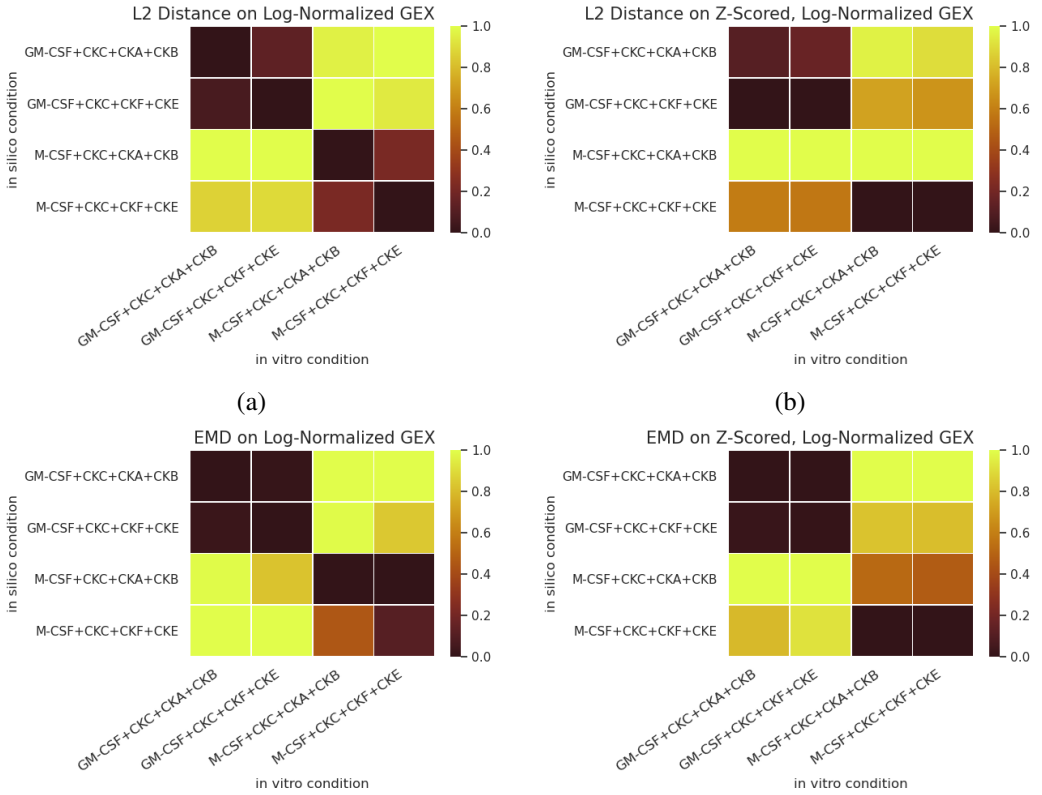


Figure 7: Evaluation of held-out *in vitro* triplet controls' proximity to predicted *in silico* triplets using all four tested distance metrics. Distances are normalized by column to the range $[0, 1]$. (a) L2 distance on log-normalized; (b) L2 distance on z-scored log-normalized; (c) EMD on log-normalized; (d) EMD on z-scored log-normalized.

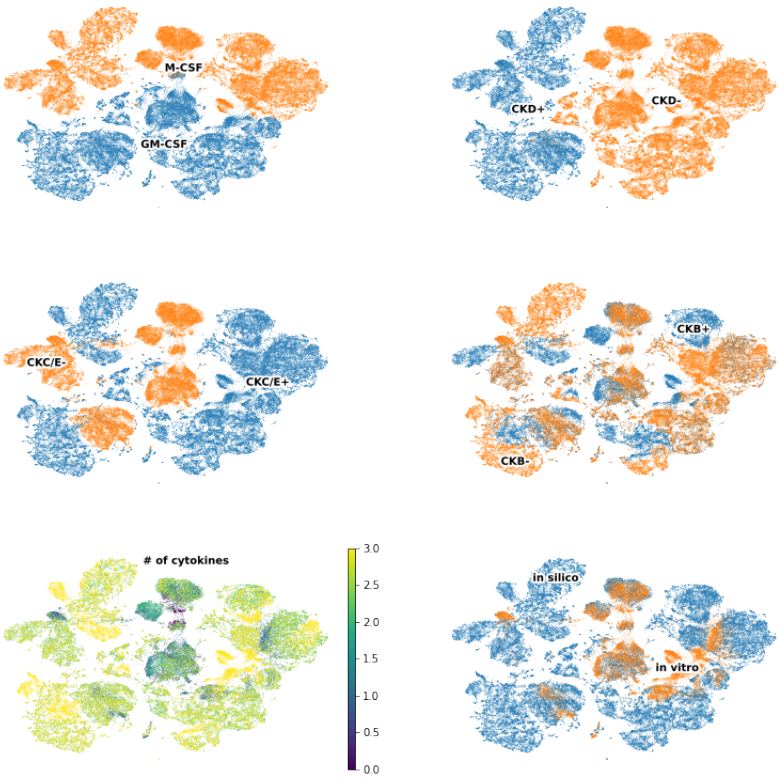


Figure 8: Detailed characterization of *in silico* and *in vitro* combined samples.

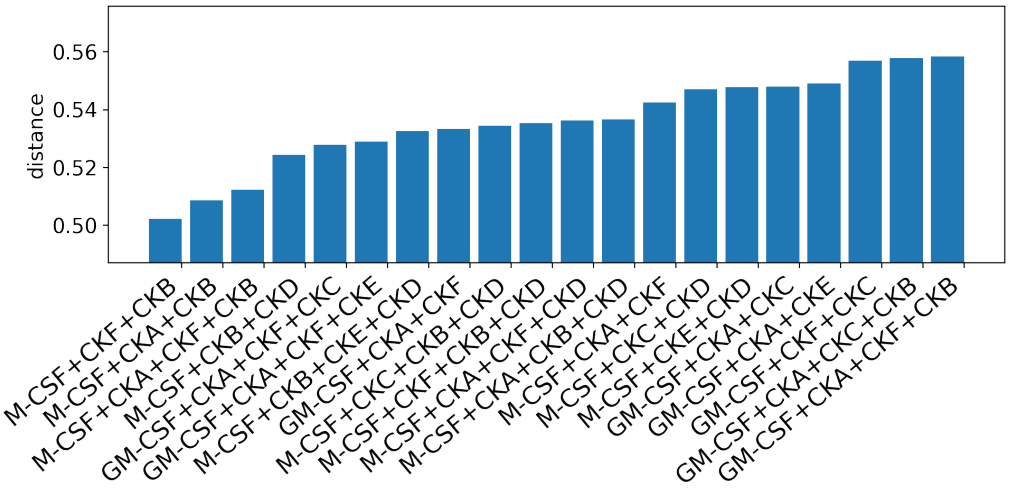


Figure 9: SystemMatch-produced ranking of top 20 *in silico* conditions based on their computed similarity to the target condition.