Generating counterfactual explanations of tumor spatial proteomes to discover effective, combinatorial therapies that enhance cancer immunotherapy

Zitong Jerry Wang
Division of Biology and Biological Engineering
California Institute of Technology
Pasadena, CA 91125
zwang2@caltech.edu

Matt Thomson
Division of Biology and Biological Engineering
California Institute of Technology
Pasadena, CA 91125
mthomson@caltech.edu

Abstract

Recent advances in spatial omics methods enable the molecular composition of human tumors to be imaged at micron-scale resolution across hundreds of patients and ten to thousands of molecular imaging channels. Large-scale molecular imaging datasets offer a new opportunity to understand how the spatial organization of proteins and cell types within a tumor modulate the response of a patient to different therapeutic strategies and offer potential insights into the design of novel therapies to increase patient response. However, spatial omics datasets require computational analysis methods that can scale to incorporate hundreds to thousands of imaging channels (ie colors) while enabling the extraction of molecular patterns that correlate with treatment responses across large number of patients with potentially heterogeneous tumors presentations. Here, we have develop a machine learning strategy for the identification and design of signaling molecule combinations that predict the degree of immune system engagement with a specific patient tumors. We specifically train a classifier to predict T cell distribution in patient tumors using the images from 30-40 molecular imaging channels. Second, we apply a gradient descent based counterfactual reasoning strategy to the classifier and discover combinations of signaling molecules predicted to increase T cell infiltration. Applied to spatial proteomics data of melanoma tumor, our model predicts that increasing the level of CXCL9, CXCL10, CXCL12, CCL19 and decreasing the level of CCL8 in melanoma tumor will increase T cell infiltration by 10-fold across a cohort of 69 patients. The model predicts that the combination is many fold more effective than single target perturbations. Our work provides a paradigm for machine learning based prediction and design of cancer therapeutics based on classification of immune system activity in spatial omics data.

1 Introduction

Recent advances in spatial proteomic and spatial genomics methods enable the molecular composition of a human tissues to be captured across hundreds to thousands of imaging channles at micron-scale

NeurIPS 2022 AI for Science Workshop.

resolution [1–4]. Spatial genomics and proteomic methods use a variety of strategies to bar-code molecules including protein and mRNA so that many different types of proteins and mRNA can be detected in a single sample at high-resolution. Traditional tissue imaging methods like MRI, PET or H&E staining are either limited to gross tissue architecture, without providing molecular information, or are limited to a small number (1,2) of molecular channels. Therefore, conventional bioimaging data sets have one or two channels of information yielding one or two images per tissue region and are very similar to the conventional RGB images for which most ML methods from computer vision were developed [5]. Recent multiplexed imaging data contains orders-of-magnitude more channels per image due to new techniques such as imaging mass cytometry (IMC) [1] and sequential fluorescence in situ hybridization (seqFISH) [2], which enables simultaneous detection of proteins/mRNAs for 40/20,000+ genes in tissues at single micron resolution, respectively. These highly multiplexed data sets tends to yield thousands of images per tissue region where each image pixel provides information on the concentration of a specific molecular species (eg a protein abundance, an mRNA abundance) in a micron sized region of the tissue.

Spatial-omics and multiplexed imaging opens new opportunities for ML-based biomedical discovery. Classically, very little is known about the molecular composition of human tissues in health and disease as classic pathology methods focus primarily on cellular morphology and gross markers like nucleus size and tissue boundary shape. Spatial-omics methods provide detailed molecular information on the signaling molecules and cell-types within a tumor, allowing molecular scale investigation of how different cell-types, for example tumor cells and immune cells, interact within a tumor and what specific signaling and regulatory processes might distinguish patients that respond to a specific type of therapy from non-responders [6]. Importantly, generating highly-multiplexed spatial data across hundreds of patients/conditions opens up the possibility of creating ML methods that may be powerful enough to discover new disease treatment strategies directly from tissue images, going beyond computer vision-based disease detection [7].

As an important specific example, cancer immunotherapy treats cancer by enabling the body's own immune system to kill cancer cells, but many patients do not respond to immunotherapy and this is thought to be due to a lack of CD8+ T cell infiltration into the tumor core. T cell infiltration is influenced by numerous factors including stromal structures and the spatial distribution of different cell populations and signaling molecules in the tumor microenvironment (TME). Existing strategies for improving CD8+ T cell infiltration focus on single-target perturbations and have shown limited efficacy. In this work, we developed a machine learning strategy to discover molecular tumor perturbations predicted to drive T cell infiltration by identifying molecular determinants of T cell infiltration from large-scale spatial-omic datasets. Specifically, our strategy involves training a classifier network to predict the T cell marker channels (representing T cell distribution) using other (non-marker) channels in IMC images. Next, we apply a gradient descent-based counterfactual reasoning procedure to the classifier network to find combinations of molecular perturbations that increase the predicted level of T cell infiltration. We showed that convolutional neural networks (CNNs) can achieve over 70% precision and 60% recall in predicting T cell localization from IMC images of breast and melanoma tumors from human patients. Furthermore, our model discovered a combinatorial therapy that is predicted to increase T cell infiltration in the majority of patients (~ 30) with immune-excluded melanoma tumor from the dataset. This therapeutic strategy involves increasing the tumoral levels of chemokine CXCL9, CXCL10, CCL19, and CXCL12, and decreasing the level of CCL8. The combinatorial strategy motivates the development of engineered proteins that achieve combinatorial function in a small number of molecular vehicles while being retained in the TME after administration.

2 Methods

2.1 Framework overview

We developed a machine learning-based optimization framework to discover molecular perturbations of tumors that can drive T cell infiltration by learning molecular features of IMC images of TMEs that are predictive of immune cell engagement. Figure 2A shows an IMC image with d molecular channels, where we refer to a subset of the channels as marker channels and the rest as predictor channels. The general logic of our framework is to train a classifier network to predict a specific aspect of the marker channels in the IMC images given the predictor channels. In our example, we specifically use signaling molecules and chemokines (cell guidance molecules) to predict the

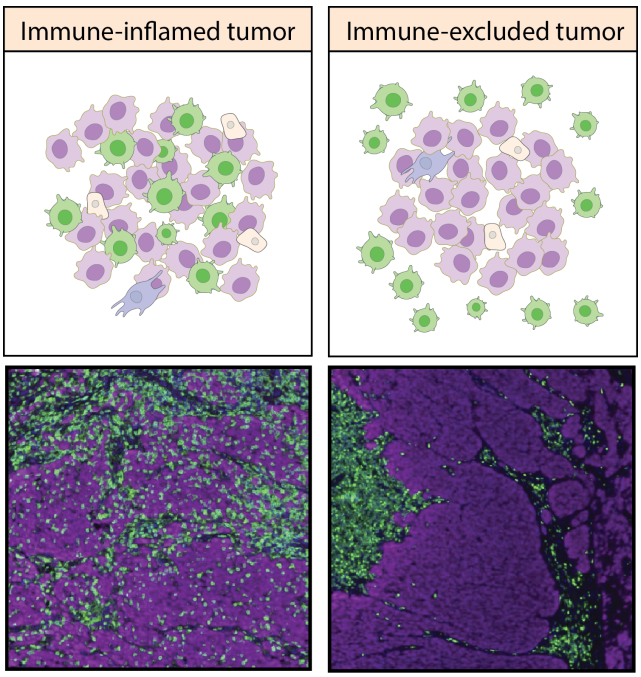

Figure 1: Images of tumors representing immune-inflamed and immune-excluded phenotypes visualized by a T cell marker (green) and a tumor marker (magenta), image adopted from [8].

abundance and state of T-cells within a local tumor region. Then we apply a formal 'counterfactual reasoning procedure' to the input data where we perform gradient descent on the predictor channels for patients with low T-cell infiltration. The gradient descent procedure adjusts molecular channels pixel by pixel to discover changes to the tumor signaling environment that the classifier network predicts will increase T-cell infiltration, allowing us to discover perturbations to the predictor channels that alter the classifier's prediction of the marker channels to a new, desired value.

Applying our framework to the problem of CD8+ T cell infiltration, we set marker channels to CD8+ T cell biomarkers. We train a classifier to predict the presence of T cells within an IMC image patch given all other channels (predictor). Then we use this classifier to search for changes to the predictor channels of tumor IMC images that can alter the classifier's prediction from T cells being absent to T cells being present. This altered image, known as a counterfactual instance in explainable AI [10], describes one perturbation of the TME that the model predicts will improve T cell infiltration.

2.2 Classifier training

We trained two classifiers, a multilayer perceptron (MLP) and a convolutional neural network (CNN), to predict the presence of non-exhausted CD8+ T cells from IMC images. Figure 2B shows that we first divided IMC images into $40 \times 40\,\mu\mathrm{m}$ (40×40 pixels) patches. To prepare each patch for network training, we removed four channels (CD3, CD8, Lag-3, PD-1) from which the label of each patch ("T cell" vs "no T cell") was determined, and we also removed signals from any channel that belong to the T cells whose distribution we are trying to predict (see [9] for detail). This second step of removing signals belonging to T cells can be carried out without cell segmentation, through an approximate procedure where all signals within a $5\,\mu\mathrm{m}$ radius of T cell marker signals are removed. We then trained a CNN to compute the probability an IMC patch contains T cells given all remaining channels. For the MLP model, instead of having individual pixel intensity as input features, we trained the model to perform the same prediction using only mean channel intensities $x_0 \in \mathbb{R}^d$ averaged over the whole patch (Figure 2B). Note the MLP is still taking advantage of spatial information in the data since each mean intensity vector characterizes a localized tissue region.

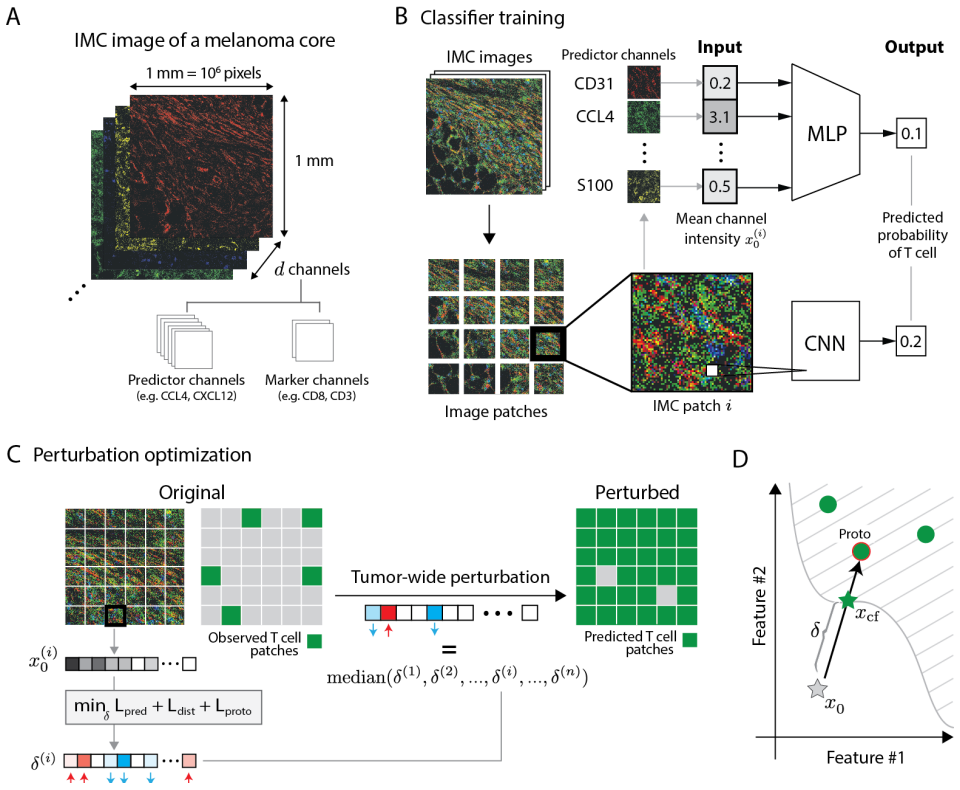

Figure 2: A ML-based optimization framework for discovering TME perturbations that drive T cell infiltration in immune-cold tumors. A) IMC image of a melanoma tumor with four channels shown [9]. B) Pipeline for MLP and CNN classifier training. C) Procedure for perturbation optimization which involves computing an optimal perturbation vector δ per IMC patch and taking the median across all such perturbations to obtain a tumor-wide perturbation vector. The positive values (shaded red) of δ correspond to increases in channel intensity, negative values (shaded blue) correspond to decreases in channel intensity. D) cartoon of the optimization algorithm in a 2D feature space, shaded region contains all instances that would be predicted by the classifier f as having T cells, green circles represent instances from training data, the one with red circle is proto from Equation 4.

2.3 Perturbation optimization via counterfactual reasoning

Given an IMC patch i without T cells, and a classifier f as described above, our goal is to find a perturbation $\delta^{(i)}$ for the patch such that f classifies the perturbed patch as having T cells (Figure 2C). For CNN models, $\delta^{(i)} \in \mathbb{R}^{w \times l \times d}$ is a 3D tensor that describes changes made for all d channel, at each pixel of the $w \times l$ pixel patch. For MLP models, $\delta^{(i)} \in \mathbb{R}^d$ is a vector where each element represents changes to the average intensity of a channel in patch i. This distinction between the two models highlights differences in the type of perturbations that each model can discover. The MLP model can find tissue-wide perturbation such as changing the overall level of an extracellular molecule, whereas the CNN model allows for cell-type specific perturbations (i.e. cell type-specific viral vectors) since perturbations are now being computed at the level of micron-scale pixels.

For simplicity, we focus on the MLP model for optimization. Given a MLP classifier f and a IMC patch i having mean channel intensity $x_0^{(i)}$ such that $f(x_0^{(i)}) = \mathbb{P}(\text{T cells present}) < p$, where $p > 0$ is the classification threshold below which the classifier predicts no T-cell, we aim to obtain a perturbation $\delta^{(i)}$ such that $f(x_0^{(i)} + \delta^{(i)}) > p$, by solving the following optimization problem adopted from [11],

$$\delta^{(i)} = \min_\delta L_{\text{pred}}(x_0^{(i)}, \delta) + L_{\text{dist}}(\delta) + L_{\text{proto}}(x_0^{(i)}, \delta), \tag{1}$$

such that

$$L_{\text{pred}}(x_0^{(i)}, \delta) = c \max(-f(x_0^{(i)} + \delta),\ -p), \tag{2}$$

$$L_{\text{dist}}(\delta) = \beta \|\delta\|_1 + \|\delta\|_2^2, \tag{3}$$

$$L_{\text{proto}}(x_0^{(i)}, \delta) = \theta \|x_0^{(i)} + \delta - \text{proto}\|_2^2, \tag{4}$$

where proto is an instance of the training set classified as having T cells, defined by first building a k-d tree of training instances classified as having T cells and setting the k-nearest item in the tree (in terms of euclidean distance to $x_0^{(i)}$) as proto. The first loss term in Equation 1 corresponds to an efficacy criteria whereas the second and third loss term represent key feasibility criteria that enables us to discover perturbations that are clinically feasible:

1. L_{pred} encourages $f(x_0^{(i)} + \delta)$ to become greater than p, meaning we want a δ that makes the model predict that the perturbed patch has T cells.
2. L_{dist} minimizes the distance between the original instance $x_0^{(i)}$ and the counterfactual instance $x_{\text{cf}}^{(i)} = x_0^{(i)} + \delta$ with elastic net regularization to generate sparse perturbations, in other words we favor perturbations that require making as few changes as possible since perturbations are often difficult to make in a real-world clinical setting.
3. L_{proto} explicitly guides the perturbation towards a counterfactual instance x_{cf} which falls in the distribution of the class of IMC patches containing T cells. This loss term helps to generate counterfactual instances that are similar to what has already been observed in IMC image of immune-inflamed tumors, making the predicted improvement in T cell infiltration more likely to hold in practice.

Figure 2D illustrates how the optimization algorithm selects the "best" perturbation through a simple schematic. Given the original instance x_0 which lacks T cells, the algorithm obtains x_{cf} by moving for the shortest distance possible, in the direction of the nearest training instance that contains T cells. Our strategy above may find different perturbations for different patches of a tumor, while in practice it is more feasible to apply the same perturbations to the whole tumor. We obtain a tumor-wide perturbation by taking the median across the set of patch-wise perturbations $\{\delta^{(i)}\}_i$.

3 Results

3.1 Datasets

We applied our optimization framework to two published IMC datasets of tumors from melanoma and breast cancer patients [9, 12]. The melanoma dataset contains IMC images of 159 tumor cores from 69 patients. Each core was approximately $1\,\text{mm}$ in diameter, imaged at $1\,\mu\text{m}$ resolution across 41 channels, corresponding to different proteins and mRNA measured [9]. The breast cancer dataset contains IMC images of 749 breast tumor cores from 693 patients [12]. Most tissue images (93%) were $0.6\,\text{mm}$ in diameter, has 1um resolution across 37 channels.

3.2 Classifier performance

Across IMC images of both melanoma and breast tumor, our simple MLP model was able to effectively predicted T cell distribution. Figure 3A shows some example predictions made by the trained MLP model, which was applied to every patch of different tumor cores. The model was able to capture the general distribution of T cells quite well. In addition, Figure 3B shows that our MLP model is able to accurately predict the proportion of patches in a tumor image that contains T cells, which is especially important given this metric is directly associated with the infiltration status of a tumor. Images with low proportion of tumor patches containing T cell correspond to immune-excluded tumors and images with high proportion of tumor patches containing T cells correspond to immune-inflamed tumors.

Across all classification metrics, Table 1 shows CNN models outperform MLP models, suggesting that even at $\sim$ 40um scale (size of patch), there may be meaningful spatial structures that influence T

cell localization. This performance difference between models is especially evident for breast tumors in comparison to melanoma tumors. Lastly, the precision of a classifier is especially important for our application because we want to be confident that when a perturbed patch is predicted to have T cells, that it is indeed true. By setting the classification threshold $p = 0.8$, we obtain high precision across all four models, so we set $p = 0.8$ for Equation 2. Confidence intervals in Table 1 obtained by retraining each model 8 times with different, randomized train-validation-test split and network initialization.

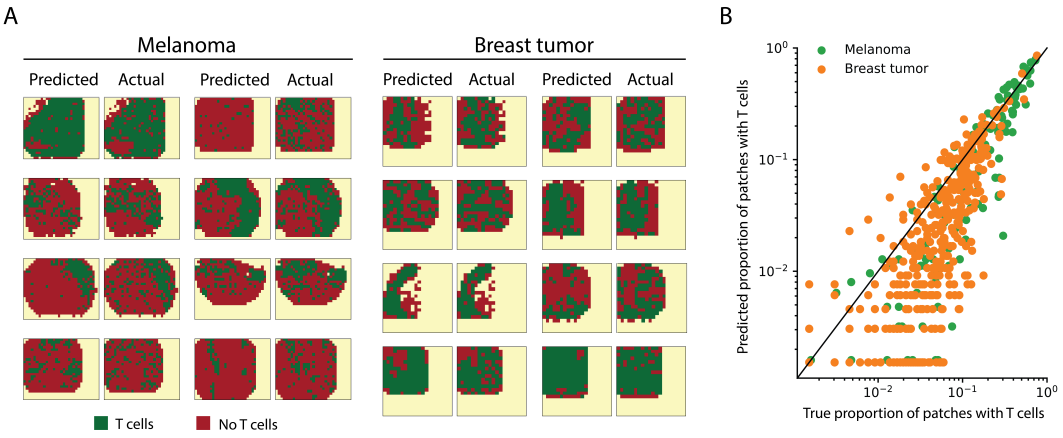

Figure 3: MLP prediction performance A) Predicted and actual T cell distribution from melanoma and breast tumor dataset (classification threshold $p = 0.5$). B) Predicted and true proportion of patches with T cells, each dot corresponds to a tumor core, diagonal line corresponds to perfect prediction.

Table 1: Performance of neural network models trained on different tumor IMC images to predict the presence of T cells, 95% confidence interval shown for different train-validation-test split and initialization, (see Table 3 for $p = 0.5$)

		$p = 0.8$				
Cancer type	Model	Precision	Recall	F1	AUC	MCC
Melanoma	MLP	$0.79 \pm 1\%$	$0.40 \pm 3\%$	$0.52 \pm 1\%$	$0.92 \pm .2\%$	$0.53 \pm 1\%$
Melanoma	CNN	$0.82 \pm 2\%$	$0.45 \pm 3\%$	$0.60 \pm 2\%$	$0.94 \pm .5\%$	$0.55 \pm 3\%$
Breast	MLP	$0.75 \pm 3\%$	$0.21 \pm 5\%$	$0.33 \pm 3\%$	$0.94 \pm .2\%$	$0.47 \pm 1\%$
Breast	CNN	$0.88 \pm 1\%$	$0.32 \pm 2\%$	$0.47 \pm 2\%$	$0.97 \pm .1\%$	$0.63 \pm 2\%$

3.3 Perturbation performance

Applying our counterfactual reasoning procedure using the MLP classifier trained on melanoma IMC images, we discovered a combinatorial therapy predicted to be highly effective in improving T cell infiltration in melanoma patients. For simplicity, we restricted the optimization procedure to only perturb the level of chemokines, which is a family of secreted proteins that are known for their ability to stimulate cell migration [13]. Figure 4A shows perturbations computed on IMC patches of melanoma tumors. Across all patches, the levels of CXCL9, CXCL10, CXCL12, and CCL19 are consistently raised and CCL8 levels are consistently reduced. Indeed, these five perturbations are all that remains after taking the median across all perturbations (bar graph). Figure 4B shows that after applying these five perturbations to the IMC image of a tumor, T cells infiltration level (roughly corresponds to the percent of tumor cells with T cells nearby) is predicted to increase by 10 fold. Furthermore, Figure 4C shows that this predicted improvement in T cell infiltration holds across nearly all 34 patients with immune-excluded tumors.

The combinatorial nature of our perturbation strategy is crucial to its effectiveness. Figure 4D shows that when each of the five perturbations are applied in isolation, the median T cell infiltration level

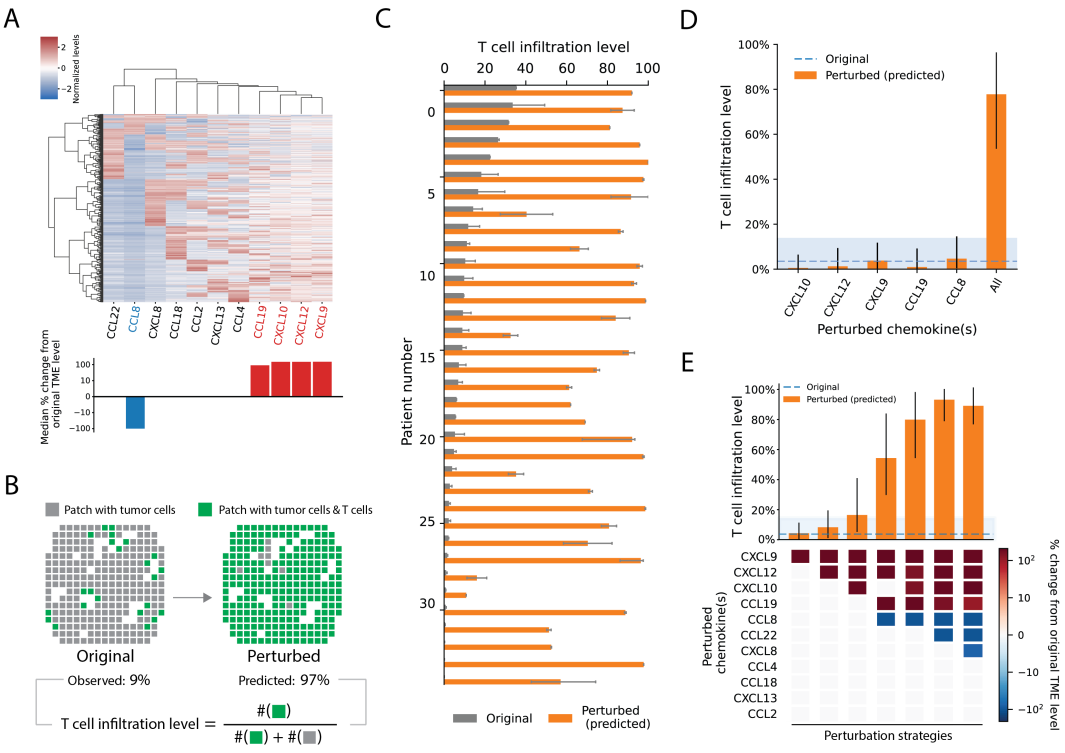

Figure 4: A) Normalized chemokine perturbations computed across IMC patches (row) lacking T cells; Bar graph shows the median of each column of the heatmap, in terms of relative change in intensity due to perturbations (only CCL8, CCL19, CXCL10, CXCL12, and CXCL9 are consistently perturbed). B) Matrix map computed from one patient's IMC image, showing the predicted distribution of T cells before and after applying the set of five major perturbations from panel A, overlaid on top of the observed tumor distribution. C) T cell infiltration level across all patients with immune-excluded tumors before and after (predicted) receiving the five perturbations, error bar represents the interquartile range computed from multiple IMC images from a patient. D) Predicted (median) T cell infiltration level across all IMC images, where each perturbation in the set of five is applied in isolation. E) Predicted (median) T cell infiltration across all IMC images for different optimized perturbation strategy of varying sparsity, error bar represents the first to third quartile. Each column corresponds to a therapeutic/perturbation strategy, where the colored boxes underneath each orange bar represents the set of chemokine perturbations involved in the strategy, while the orange bar is the predicted effect of the set of perturbation on T cell infiltration level

across all tumors does not undergo significant increase. Furthermore, we systematically explore the importance of combinatorial perturbation by changing parameters β of Equation 3 which adjusts the sparsity of the solution, where sparser solution means less molecules are perturbed. Figure 4E shows that although the model predicts the optimal single-target perturbation is not effective, perturbing four targets drastically increases T cell infiltration. Furthermore, perturbing more than five targets leads to diminishing returns in terms of the improvement in T cell infiltration. In conclusion, combinatorial perturbation of the TME appears necessary for improving T cell infiltration, at least within the scope of the chemokine targets considered.

4 Discussion

Our optimization framework combines deep neural networks with optimization methods from explainable AI to directly predict therapeutic interventions. Our predictions can be tested and the results fed back to update the model. One of the major strength of our model is that it can readily scale to deal with larger datasets and neural network models, which will be very important as more spatial

transcriptomics and proteomics datasets are quickly becoming available [14]. For future work, on the computational side, we plan to incorporate cell-type specific perturbations into our framework, which can be done by performing optimization on neural network architectures such as CNNs and vision transformers. On the experimental side, we plan to test our therapeutic predictions experimentally in murine models of melanoma.

References

[1] Charlotte Giesen, Hao AO Wang, Denis Schapiro, Nevena Zivanovic, Andrea Jacobs, Bodo Hattendorf, Peter J Schüffler, Daniel Grolimund, Joachim M Buhmann, Simone Brandt, et al. Highly multiplexed imaging of tumor tissues with subcellular resolution by mass cytometry. *Nature methods*, 11(4):417–422, 2014.

[2] Chee-Huat Linus Eng, Michael Lawson, Qian Zhu, Ruben Dries, Noushin Koulena, Yodai Takei, Jina Yun, Christopher Cronin, Christoph Karp, Guo-Cheng Yuan, et al. Transcriptome-scale super-resolved imaging in tissues by rna seqfish+. *Nature*, 568(7751):235–239, 2019.

[3] Vivien Marx. Method of the year: spatially resolved transcriptomics. *Nature methods*, 18(1):9–14, 2021.

[4] Sanja Vickovic, Gökcen Eraslan, Fredrik Salmén, Johanna Klughammer, Linnea Stenbeck, Denis Schapiro, Tarmo Äijö, Richard Bonneau, Ludvig Bergenstråhle, José Fernandéz Navarro, et al. High-definition spatial transcriptomics for in situ tissue profiling. *Nature methods*, 16(10):987–990, 2019.

[5] Sotiris Dimopoulos, Christian E Mayer, Fabian Rudolf, and Joerg Stelling. Accurate cell segmentation in microscopy images using membrane patterns. *Bioinformatics*, 30(18):2644–2651, 2014.

[6] George Adam, Ladislav Rampášek, Zhaleh Safikhani, Petr Smirnov, Benjamin Haibe-Kains, and Anna Goldenberg. Machine learning approaches to drug response prediction: challenges and recent progress. *NPJ precision oncology*, 4(1):1–10, 2020.

[7] Zilong Hu, Jinshan Tang, Ziming Wang, Kai Zhang, Ling Zhang, and Qingling Sun. Deep learning for image-based cancer detection and diagnosis- a survey. *Pattern Recognition*, 83:134–149, 2018.

[8] Lieke L van der Woude, Mark AJ Gorris, Altuna Halilovic, Carl G Figdor, and I Jolanda M de Vries. Migrating into the tumor: a roadmap for t cells. *Trends in cancer*, 3(11):797–808, 2017.

[9] Tobias Hoch, Daniel Schulz, Nils Eling, Julia Martínez Gómez, Mitchell P Levesque, and Bernd Bodenmiller. Multiplexed imaging mass cytometry of the chemokine milieus in melanoma characterizes features of the response to immunotherapy. *Science Immunology*, 7(70):eabk1692, 2022.

[10] Ruth MJ Byrne. Counterfactuals in explainable artificial intelligence (xai): Evidence from human reasoning. In *IJCAI*, pages 6276–6282, 2019.

[11] Arnaud Van Looveren and Janis Klaise. Interpretable counterfactual explanations guided by prototypes. In *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*, pages 650–665. Springer, 2021.

[12] Esther Danenberg, Helen Bardwell, Vito RT Zanotelli, Elena Provenzano, Suet-Feung Chin, Oscar M Rueda, Andrew Green, Emad Rakha, Samuel Aparicio, Ian O Ellis, et al. Breast tumor microenvironment structures are associated with genomic features and clinical outcome. *Nature genetics*, 54(5):660–669, 2022.

[13] Catherine E Hughes and Robert JB Nibbs. A guide to chemokines and their receptors. *The FEBS journal*, 285(16):2944–2971, 2018.

[14] Ao Chen, Sha Liao, Mengnan Cheng, Kailong Ma, Liang Wu, Yiwei Lai, Xiaojie Qiu, Jin Yang, Jiangshan Xu, Shijie Hao, et al. Spatiotemporal transcriptomic atlas of mouse organogenesis using dna nanoball-patterned arrays. *Cell*, 185(10):1777–1792, 2022.

[15] Martín Abadi, Paul Barham, Jianmin Chen, Zhifeng Chen, Andy Davis, Jeffrey Dean, Matthieu Devin, Sanjay Ghemawat, Geoffrey Irving, Michael Isard, et al. Tensorflow: A system for large-scale machine learning. In *12th {USENIX} Symposium on Operating Systems Design and Implementation ({OSDI} 16)*, pages 265–283, 2016.

A Appendix

A.1 Classifier training information

Our work is implemented using Keras with Tensorflow v2.10 [15] as backend. Classifier training was conducted on NVIDIA GeForce RTX 3090 Ti GPU with 24GB RAM workbench. The metric MCC (matthew correlation coefficient) was used to decide when to stop training. The same MLP architecture was trained on both the breast and melanoma dataset, consisting of two hidden layers of 30 and 10 nodes, respectively. The same CNN architecture was trained on both the breast and melanoma dataset, consisting of a convolutional layer, followed by max pooling, and another convolutional layer, a dropout layer, and finally two dense layers. The output of both the MLP and CNN represent the probability of T cells present.

A.2 Hyperparameters

For perturbation optimization, we used the following set of values for hyperparameters from Equation 1

Table 2: Optimization hyperparameters

Name	Value
p	0.8
c	100
β	0.1
θ	100
k	1

A.3 Classifier performance

Table 3: Performance of classifier trained on tumor IMC to predict the presence of T cells, $p = 0.5$.

Cancer type	Model	Precision	Recall	F1
Melanoma	MLP	$0.70 \pm 1\%$	$0.55 \pm 2\%$	$0.62 \pm 1\%$
Melanoma	CNN	$0.71 \pm 2\%$	$0.58 \pm 2\%$	$0.63 \pm 2\%$
Breast	MLP	$0.65 \pm 3\%$	$0.38 \pm 3\%$	$0.48 \pm .3\%$
Breast	CNN	$0.72 \pm 2\%$	$0.65 \pm 2\%$	$0.68 \pm 1\%$