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

**Anonymous Author(s)**
Affiliation
Address
email

## Abstract

Recent advances in tissue imaging technologies have led to the generation of massive datasets of spatial profiles of human tissues, taken at micron scale resolution, spanning hundreds of patients, and across tens to thousands of molecular biomarkers. These high-dimensional imaging data necessitate the development of new, AI-based tools to uncover new biology and to support therapeutic developments against diseases. Currently, a major need in the treatment of solid tumor cancers are strategies that can drive the infiltration of T cells into the tumor. In this study, we developed an optimization strategy combining supervised ML and counterfactual explanations to discover clinically feasible tumor perturbations that drive T cell infiltration, and applied our framework to spatial proteomes of breast cancer and melanoma tissue. Our model predicts that altering the levels of four molecules (CCL4, CXCL12, CXCL13, CCL8) in immune-excluded melanoma tissues can increase T cell infiltration by 10 fold, across an entire cohort of 69 patients. Our work provides a paradigm for machine learning based prediction and design of cancer therapeutics based on classification of immune system activity in spatial proteomics data.

## 1 Introduction

Recent advances in bioimaging technology allow us to capture the spatial organization of cellular activity in tissue with unprecedented molecular detail [1–4]. Traditional bioimaging data were mostly in grayscale or contained only 1-2 color channels, corresponding to 1-2 different molecular biomarkers, similar to conventional RGB images for which most ML methods from computer vision were developed. With limited signals, bioimaging applications of ML focused mainly on object detection such as cell segmentation [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.

Spatial omics and multiplexed imaging opens new opportunities for ML-based biomedical discovery. The spatial organization of cell populations within tissues fundamentally shapes the health of our bodies by determining how cells interact [6], akin to how physical proximity between people shapes our society by structuring how we interact with one another. Although the cell interaction networks embedded in spatial omics data is likely a key determinant of health, we need new computational methods to interpret these high-dimensional spatial patterns in terms of health and disease [7]. Impor-

Submitted to NeurIPS 2022 AI for Science Workshop.

tantly, 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 [8].

As large amounts of spatials proteomics data on human tumors are becoming widely available [9, 10], ML-based methods may be applied to improve cancer immunotherapy . 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 T cell infiltration into the tumor core (Figure 1) [11]. Although many factors in the tumor microenvironment (TME) affect T cell infiltration, such as vasculature structure, anti-inflammatory cytokines and chemokines, existing strategies focus on single-target perturbations and have shown limited efficacy [12]. Perturbing multiple targets is likely more effective, but testing all possible combinations of molecules to perturb is experimentally infeasible. Existing IMC data contains spatial profiles (Figure 1) of both immune-inflamed (infiltrated by T cells) and immune-excluded tumors (T cells excluded from tumor core)[9, 10], which enables the training of ML models to predict T cell infiltration from IMC images and using the trained models as a substrate to search for perturbations that turn immune-excluded tumors immune-inflamed.

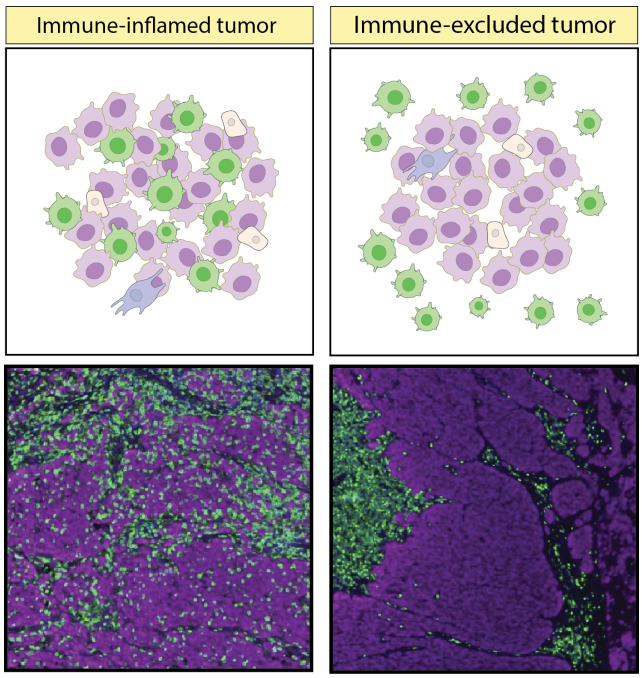


Figure 1: Images of tumors representing immune-inflamed and immune-excluded phenotypes visualized by a T cell marker (green) and a tumor marker (magenta), images from [12].

In this work, we apply supervised ML and counterfactual explanations to IMC images of breast and melanoma tumors to predict tumor perturbations that support T cell infiltration. Our strategy involves training a classifier to predict the presence of T cell markers using other (non-marker) channels in IMC images, then using this classifier to find tissue perturbations that increase the predicted probability of T cells. We showed that convolutional neural networks (CNNs) can achieve over 80% precision and 70% recall in predicting T cell localization from IMC images of breast and melanoma tumors. Furthermore, our model predicts altering the level of CCL4, CXCL12, CXCL13 and CCL8 in melanoma tumors can convert immune-excluded tumors in a cohort of 69 melanoma patients to be immune-inflamed. No single target perturbations were predicted to be effective. Our work provides a paradigm for machine learning based prediction and design of cancer therapeutics based on classification of immune system activity in spatial proteomics data.

# 2 Methods

## 2.1 Framework overview

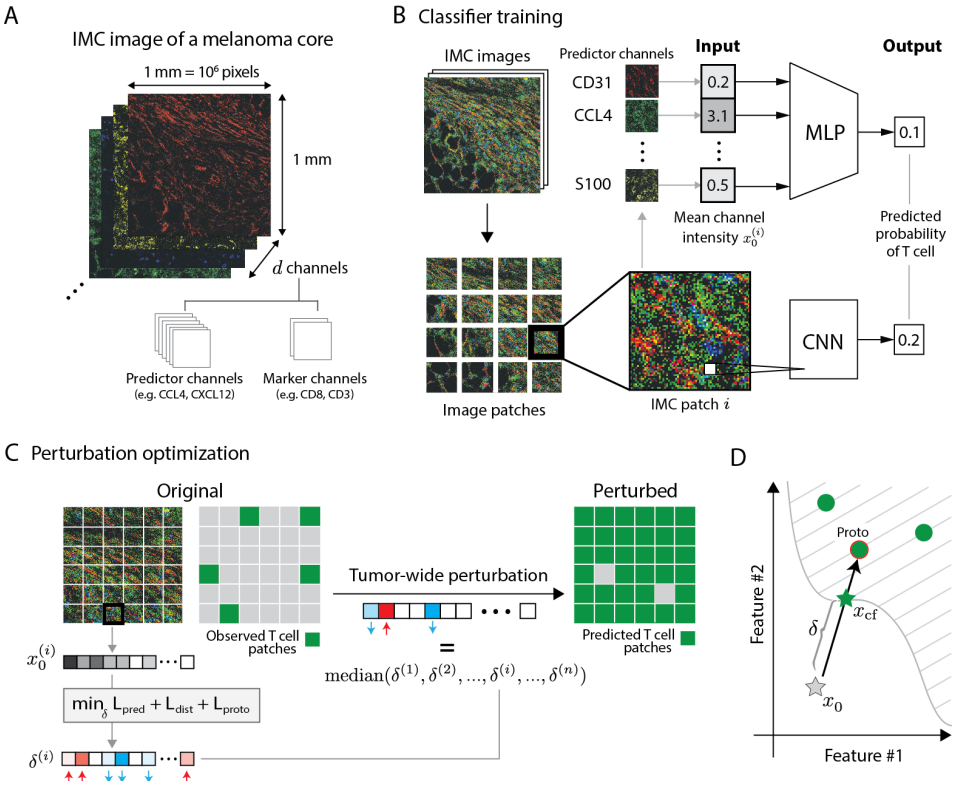


Figure 2: A ML-based optimization framework for discovering therapeutic strategies to support T cell infiltration in immune-cold melanoma. A) IMC image of a melanoma tumor with four channels shown [9]. B) Pipeline for classifier training. C) Procedure for perturbation optimization which involves computing an optimal perturbation vector $\delta$ per patch and taking the median across all such perturbations to obtain a tumor-wide perturbation vector. The positive values (shaded red) of $\delta$ 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 $\text{proto}$ from Equation 4.

We propose an optimization framework that uses IMC images of TMEs to generate tumor perturbation strategies that drive T cell infiltration. 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 to predict aspects of the marker channels in the IMC images given the predictor channels. Then we apply counterfactual reasoning to the data by performing gradient descent on the predictor channels, 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 T cell infiltration, we set marker channels to 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 [13], describes one perturbation of the TME that the model predicts will improve T cell infiltration.

### 2.2 Classifier training

We trained two classifiers, a MLP and a CNN, to predict the presence of T cells (non-exhausted, cytotoxic) from IMC images. Figure 2B shows that we first divided IMC images into $64 \times 64\,\mu\mathrm{m}$ (64 $\times$ 64 pixels) patches. For each patch, we then removed four channels (CD3, CD8, Lag-3, PD-1) from which the label of each patch ("T cell" vs "no T cell") was determined (see [9] for detail). We 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.

### 2.3 Perturbation optimization

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 every channel, at each pixel of the 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 region-wide perturbation such as changing the overall level of an extracellular molecule, whereas the CNN model allows for cell-type specific perturbations 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 [14],

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

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 $\text{proto}$ 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 $\text{proto}$. Each of the loss term in Equation 1 corresponds to a desirable property of the perturbation $\delta$:

1. $L_{\text{pred}}$ encourages $f(x_0^{(i)} + \delta)$ to become greater than $p$, meaning we want a $\delta$ that makes the model predict that the perturbed patch has T cells.
2. $L_{\text{dist}}$ 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
3. $L_{\text{proto}}$ explicitly guides the perturbation towards a counterfactual instance $x_{\text{cf}}$ 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.

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 $x_{\mathrm{cf}}$ by moving for the shortest distance possible, in the direction of the nearest training instance that is predicted to have 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 publicly available IMC datasets of tumors from melanoma and breast cancer patients [9, 10]. The melanoma dataset was obtained by IMC imaging of 159 tumor cores from 69 patients. Each tissue was approximately $1\,\mathrm{mm}$ in diameter, imaged at $1\,\mu\mathrm{m}$ resolution across 41 channels, corresponding to different proteins measured [9]. The breast cancer dataset was obtained by IMC imaging of 749 breast tumor cores from 693 patients [10]. Most tissues (93%) were $0.6\,\mathrm{mm}$ in diameter, image at 1um resolution across 37 channels.

## 3.2 Classifier performance

Across all classification metrics, Table 1 shows CNN models consistently outperform MLP models, suggesting that even at $\sim$ 60um 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. Err in Table 1 obtained by retraining each model 8 times with different, randomized train-validation-test split and network initialization.

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 | $0.86 \pm 1\%$ | $0.61 \pm 3\%$ | $0.71 \pm 1\%$ | $0.93 \pm .2\%$ | $0.65 \pm 1\%$ |
| Melanoma | CNN | $0.90 \pm 2\%$ | $0.67 \pm 3\%$ | $0.75 \pm 2\%$ | $0.95 \pm .5\%$ | $0.71 \pm 3\%$ |
| Breast | MLP | $0.83 \pm 3\%$ | $0.49 \pm 5\%$ | $0.66 \pm 3\%$ | $0.92 \pm .2\%$ | $0.57 \pm 1\%$ |
| Breast | CNN | $0.87 \pm 1\%$ | $0.56 \pm 2\%$ | $0.68 \pm 2\%$ | $0.95 \pm .1\%$ | $0.68 \pm 2\%$ |

## 3.3 Perturbation performance

Applying our optimization 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 algorithm to only perturb the level of chemokines, which is a family of secreted proteins that are known for their ability to stimulate cell migration [15]. Figure 3A shows perturbations computed on IMC patches of melanoma tumors. Across all patches, average levels of CCL4, CXCL12, and CXCL13 are consistently raised and CCL8 levels are consistently reduced. Indeed, these four perturbations are all that remains after taking the median across all perturbations (bar graph). Interestingly, one of these perturbations, injection of CCL4, have been tested in mouse melanoma [16]. Figure 3B shows that after applying these four perturbations to the IMC image of a tumor, T cells infiltration level (% of tumor cells with T

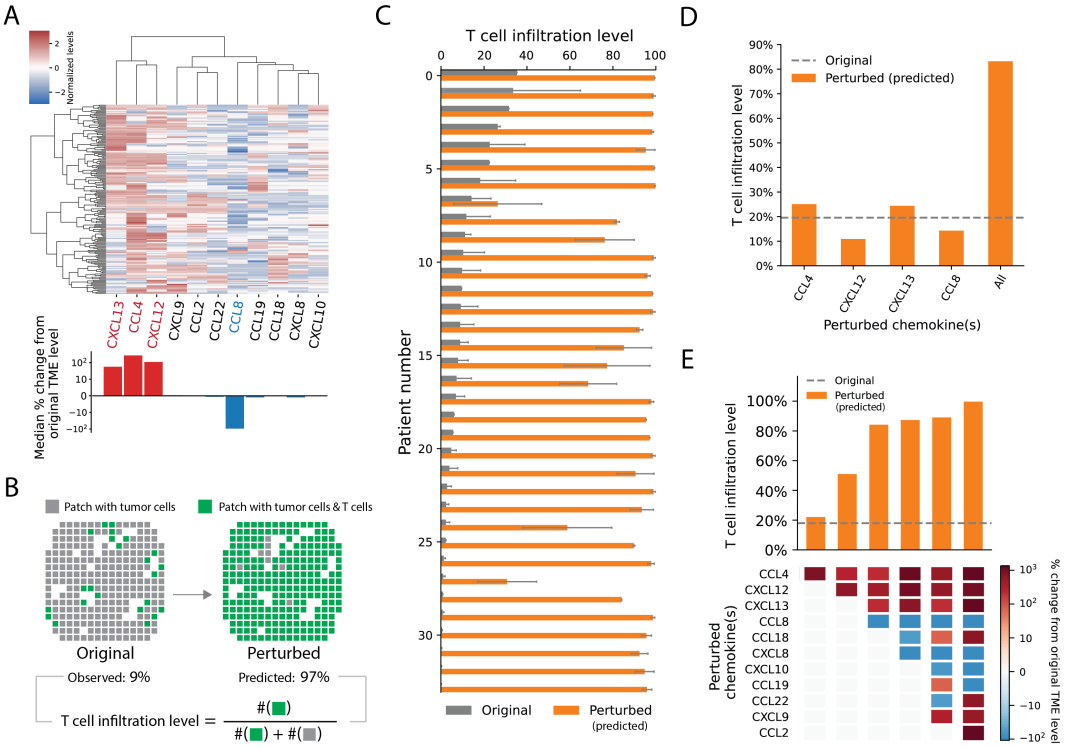


Figure 3: A) Normalized chemokine perturbations computed across IMC patches (row) lacking T cells; Bar graph shows the median of each column of the heatmap, in terms of relative change in intensity due to perturbations (only CCL4, CCL8, CXCL12, and CXCL13 are consistently perturbed). B) Matrix map computed from one patient's IMC image, showing the predicted distribution of T cells before and after applying the set of four 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 four perturbations, error bar computed from multiple IMC images from a patient. D) Predicted T cell infiltration level averaged across all IMC images, where each perturbation in the set of four is applied in isolation. E) Predicted T cell infiltration averaged across all IMC images for different optimized perturbations of varying sparsity.

cells nearby) is predicted to increase by 10 fold. Furthermore, Figure 3C 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 3D shows that when each of the four perturbations are applied in isolation, the T cell infiltration level averaged across all tumors does not undergo significant increase. Furthermore, we systematically explore the importance of combinatorial perturbation by changing parameters $\beta$ of Equation 3 which adjusts the sparsity of the solution, where sparser solution means less molecules are perturbed. Figure 3E shows that although the model predicts the optimal single-target perturbation is not effective, perturbing two targets drastically increases T cell infiltration. Furthermore, perturbing more than four 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 [17]. For future work, 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.

## 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] Leeat Keren, Marc Bosse, Diana Marquez, Roshan Angoshtari, Samir Jain, Sushama Varma, Soo-Ryum Yang, Allison Kurian, David Van Valen, Robert West, et al. A structured tumor-immune microenvironment in triple negative breast cancer revealed by multiplexed ion beam imaging. *Cell*, 174(6):1373–1387, 2018.

[7] 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.

[8] 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.

[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] 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.

[11] Paola Bonaventura, Tala Shekarian, Vincent Alcazer, Jenny Valladeau-Guilemond, Sandrine Valsesia-Wittmann, Sebastian Amigorena, Christophe Caux, and Stéphane Depil. Cold tumors: a therapeutic challenge for immunotherapy. *Frontiers in immunology*, 10:168, 2019.

[12] 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.

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

[14] 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.

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

[16] John-Michael Williford, Jun Ishihara, Ako Ishihara, Aslan Mansurov, Peyman Hosseinchi, Tiffany M Marchell, Lambert Potin, Melody A Swartz, and Jeffrey A Hubbell. Recruitment of cd103+ dendritic cells via tumor-targeted chemokine delivery enhances efficacy of checkpoint inhibitor immunotherapy. *Science advances*, 5(12):eaay1357, 2019.

[17] 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.

[18] Francois Chollet et al. Keras, 2015.

[19] 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 [18] with Tensorflow v2.10 [19] 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 |
| $\beta$ | 0.1 |
| $\theta$ | 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 | $0.78 \pm 1\%$ | $0.72 \pm 2\%$ | $0.75 \pm .8\%$ |
| Melanoma | CNN | $0.79 \pm 2\%$ | $0.81 \pm 2\%$ | $0.80 \pm 2\%$ |
| Breast | MLP | $0.75 \pm 3\%$ | $0.57 \pm 3\%$ | $0.65 \pm .3\%$ |
| Breast | CNN | $0.80 \pm 2\%$ | $0.70 \pm 2\%$ | $0.75 \pm 1\%$ |