Parameter Inference In Tumor Environments

Alber David[1]

[1]Stockholm University: alber.p.david@gmail.com

September 9, 2020

Keywords: Tumor Biology, Cellular Potts Model, Parameter Inference, Maximum Pseudo Likelihood, Monte Carlo Optimisation

The tumor environment is a conglomeration of many different cell types. Modern laboratory techniques allow us to study tissue samples of cancer patients, revealing the heterogeneous nature of this tissue type. It is believed that understanding the microscopic rules of interaction between each cell type to every other cell type will reveal the fundamentals of the macroscopic organisation patterns. This work provides a data driven approach to learn the key parameters that are guiding the spatial organisation of the tumor environment. A method to learn a parameter set of manageable size from pathological images is presented. This set of connectivity parameters is used as input for a generative model. The goal is to generate a pattern configuration from a simulation, which represents the macroscopic key features of the original pathological image, to identify key principles of the cellular organisation and to quantify the interaction strength that is responsible for the resulting organisation. The quest to be able to simulate a pattern configuration that is indistinguishably similar to a tissue sample of a cancer patient and opens insights in the complex mechanics of cancer and provides perspectives for effective treatment methods.

Contents

1 Introduction

Research of the last decades revealed that tumors follow common objectives. Six hallmarks have been identified [1]. Recent effort in the field of bioinformatics showed that it is possible to classify jobs and expertise of different cancer types [2], [3] by analysing DNA samples of cancer patients. Still there exists no fundamental theory describing how the interplay between different cell types shapes the heterogeneous tumor environment. What are the key parameters guiding the spatial organisation of a tumor, is the question this project tries to answer. Can we find a theory that is capable of describing the vast diversity in the organisation we are able to observe in pathological tissue samples of cancer patients? To get a deeper insight in the rules that drive the formation of such a system are provided by being able to simulate some of its key features. A theory based on pairwise interactions between different cell types is formulated. This provides the framework for a generative tumor model. Also, a data driven approach is used to infer quantitative principles and learn the interaction strength between cell types. The combination of these two lines of thought will provide the fundament for a descriptive model of the spatial organisation within the tumor environment and give insights about how macroscopic tissue configurations emerge from microscopic cell to cell interactions.

The hope is to find a deterministic formulation of the spatial organisation of tumor tissue by analysing the seemingly chaotic tissue samples of triple negative breast cancer (TNBC) patients. This will provide necessary fundamentals for effective and customized treatment. More over, a successful theoretical description could provide a cost and time efficient way to make predictions that go beyond the possibilities in a laboratory environment.

2 Methods

This research project aims to get a better understanding on the spatial organisation of the tumor environment. More precisely, our goal is to develop a generative model that is able to capture key features of pathological images. Further we want to develop a framework that is able to learn the set of parameters which drive the tumor organisation in the first place. The hope is to objectively quantify the interaction strength on a cell to cell level and uncover connections that might remain unnoticed by looking at a composite of thousands of cells with the bare eye.

To analyze the large scale properties of the tissue samples under consideration, the radial distribution function (RDF) provides answers about the average length scale of direct interactions from one cell to another one, see sec. 2.1.
Defining a description of the systems energy allows us to formulate a generative model:

We define a Hamiltonian for the system under consideration and optimize the systems energy with Markov Chain Monte Carlo (MCMC) optimisation, see sec. 2.2.
For a well chosen set of input parameters, the MCMC framework provides a tool that is capable of generating cell configurations which capture key features, observed in the pathological images [4]. These are compartmentalized, cold and mixed configurations. Assuming the energy we defined provides a representative description of the system, we aim to learn the parameters that are most likely to drive a model configuration, similar to the cellular configuration we observe in the pathological images. The framework of maximum pseudo likelihood (MPL) inference provides a computationally feasible formalism to tackle this challenge, see sec. 2.3 .
No mathematical model is capable of portraying the full underlying truth. For this work the reader has to consider the results with respect to the following approximations: The method to infer inter cellular connectivities relies on 2D pathological tissue samples of the size of $780\,\mu m^2$. The author assumes that the given data is representative for the patients tumor environment as a whole, i.e: the tumor environment is assumed to be isotropic. Also the concentration of cell types found in a tissue sample is maintained for the presented models. This means that the framework of the generative model considers an initial number of cells from different cell types which are only allowed to switch positions in order to generate an optimal configuration w.r.t. the systems energy description. This is justified by a separation of time scales: The process of cell division is slow compared to the movement of cancer and immune cells. Finally, the modeling approach is inspired by discrete element simulations. The formation of a configuration is described by pairwise interaction between cells. An underlying distribution of molecules or temperature as mentioned in [5] is not yet considered. The interaction parameters the presented approach delivers implicitly infer the forces, which might originate from an underlying field implicitly.

2.1 Typical interaction scale

The starting point of the analysis are tissue samples of cancer patients from [4]. They contain information about the cell location in 2D and its corresponding cell type.
Inspecting these images the question arises if large scale structures can be detected in the tissue?
This will provide knowledge about the typical direct interaction scale between cells.

The radial distribution function $g(r)$ describes how, in a system of N cells, density varies as a function of distance from a reference cell. This is: How much more likely is it to find a cell within a distance r, compared to random organisation. This pair correlation function can be measured as:

$$g(r) = \frac{f(r)}{\eta} \tag{1}$$

In eq. (1) the function $f(r)$ describes how often a distance r between two cells is found for a given set of cell positions on a domain of size L^2. This means that the occurrence of cellular distances in bins of size Δr is measured. The denominator in eq. (1) normalizes the distribution $f(r)$. It is defined as:

$$\eta = \frac{N(2\pi r \Delta r \rho)}{2}, \text{ with } \rho = \frac{N}{L^2} \tag{2}$$

We can associate a potential and force to the radial distribution through a logarithmic relation [6]. This provides a physics intuition about forbidden contact regions and how fast the direct impact from one cell to another one decays with distance. The potential associated to the radial distribution function is defined as:

$$u(r) = -\log\left(g(r)\right) \tag{3}$$

To obtain an analytical expression, the data for the potential $u(r)$ is fitted to a Lennard Jones like potential V_{LJ} minimizing the χ^2 function. This means that the potential $u(r)$, obtained from the positional data of the cells is fitted to:

$$V_{LJ}(r, \sigma, \epsilon) = 4\epsilon \left[\left(\frac{\sigma}{r}\right)^8 - \left(\frac{\sigma}{r}\right)^4 \right] \tag{4}$$

The free parameters ϵ and σ are computationally optimized s.t. $\chi^2 = \left(u(r) - V_{LJ}(r, \sigma, \epsilon)\right)^2$ is minimal. As shown in the results section 3.1 the obtained analytical expression delivers an expression to do back on the envelope estimations or describe the length scale of direct contact interactions in a sophisticated model.

2.2 A Generative Model for Tumor Organisation

Having analysed the typical interaction radius between cells, we seek a generative model description which considers the heterogeneous tumor environment.
Can we find a generative model description, that is capable of producing the pathological images we see? We seek interaction laws on an inter cellular level, that result in large scale patterns. Since different tumor environments are found in the same type of triple negative breast cancer (TNBC) the aspiration is that a single model is capable of producing compartmentalized, hot and cold tumors, by tweaking as little parameters as possible.
The presented formalism teaches us about key principles of the driving mechanisms in the spatial organisation of tumors. It lets us explore parameter regimes that are not possible to examine physically.

Inspiration comes from statistical mechanics. The well understood Ising mode describes the magnetisation of spin sights $s_i = \{+1/2, -1/2\}$ on a lattice. Every sight interacts with other sights within its local neighborhood. Local interaction rules on a sight to

sight level, result in a large scale pattern formation. Moreover, continuously changing the system temperature T leads to a phase transition from a randomly organised system to system in which all spins are aligned in the same direction.
We extend the idea of the Ising model to continuous space and multiple cell types. The q state extension of the two state Ising model is called a Cellular Potts model. It is already explored in the field of cellular molecular biology (CMB) [7] to model inter and intra cellular processes. Note that for the biological interpretation, there is no physical meaning to the temperature T. Therefor our model treats this parameter as a constant set to $T = 1$.

The Hamiltonian describes a rule for interaction between cells. The key parameters are, the symmetric connectivity matrix $\mathbf{J}$ and the length scale $\mu \in \mathbb{R}$ at which cells interact. The length scale of interactions is controlled with a step function Θ. The systems energy description reads as follows:

$$\mathcal{H} = \sum_{i=1}^{N-1} \sum_{j=i+1}^{N} J_{s_i s_j} \Theta(||\vec{x}_i - \vec{x}_j|| \leq \mu) \tag{5}$$

In expression (5) the first sum goes over cells i and the second one sums over all other cells j up to the total number of cells in the configuration N. The connectivity strength between two cells of type s_i, s_j for $l = \{1, ..., q\}$ possible different states is governed by the symmetric square matrix $\mathbf{J} \in \mathbb{R}^{q \times q}$ s.t the entry $J_{s_i s_j} \in \mathbb{R}$. The spatial dependency for the interaction range between cells at positions $\vec{x}_i$, $\vec{x}_j$ is defined by the function Θ. In this implementation we use a step function which is 1, whenever the distance between two cells is smaller or equal to a limiting radius μ and 0 else. This implies that the function $\Theta(||\vec{x}_i - \vec{x}_j|| \leq \mu)$ controls the size of the of the local neighborhood of every cell, i.e: number of surrounding cells j every cell i interacts with. On a regular square lattice cells are separated by a distance r. The k^{th} order neighborhood n_i^k of cell i is defined with $\mu = r\sqrt{2^k}$. On pathological images the cells are not aligned on a regular square lattice. The limiting interaction length is chosen to be as small as possible but such that every cell still has at least one neighbor. Cell sorting has been shown for a similar energy description $\mathcal{H}$ by [8]. Opposed to the standard formulation of Potts model Hamiltonians, our model assigns every cell type a connectivity strength to every other cell type. However, an energy description with only self interactions, as in [8], without a surface friction term, can be reached by setting all diagonal elements of $\mathbf{J}$ to the same value and the off diagonal elements to 0. Assuming symmetric interactions our model consists of $\frac{q(q+1)}{2}$ connectivities. These are the models free parameters. Negative nodal connections $J_{s_i s_j}$ indicate affinity from cell type s_i to cell type s_j and positive nodal connections indicate repulsion. If $J_{s_i s_j} = 0$, this indicates that the interaction between cell type s_i to cell type s_j is passive.
Given the connectivity strength from every cell type to every other cell type ($\mathbf{J}$), Markov Chain Monte Carlo (MCMC) optimisation is used to generate a configuration with the lowest system energy. The algorithm is set up as follows:

- evaluate the systems energy $\mathcal{H}$
- choose a random sight $i \in [1, N]$
- choose a random target sight $j \in [1, N] \setminus N_i^k$, excluding the local neighborhood of i
- switch the cell types s_i, s_j of the sights i and j
- evaluate the systems energy $\hat{\mathcal{H}}$ considering the type switch
- accept the trail move if: $unirand[0, 1] \leq e^{\mathcal{H}-\hat{\mathcal{H}}}$

This agent based model reduces the inter cellular interaction of multiple cell types to pair interactions. As we will see in sec. 3.3.4 the model is able to produce large scale pattern formations, similar to the ones observed in the pathological images in [4].

2.3 Parameter Inference

We found a model description that is capable of generating compartmentalized, sorted, random cell configurations. These are the classes in which [4] subdivides the pathological tissue samples. The configurations are generated by minimizing the systems energy, which is described by our definition of a Hamiltonian, a function of connectivity strength $J_{s_i s_j}$ between cell type s_i and s_j and μ, the distance at which cells are able to interact with each other. The connectivity strength however is still not explored.
Can we learn the Potts model descriptive parameters? We want to infer the connectivity strength $J_{s_i s_j}$ between every cell type, to every other cell type, only by using image data, this is, the location in 2D and the type of every sight s_i with $i \in 1, ..., N$ and $s_i \in \{1...q\}$. Further, is it possible to infer the connections that are above the obvious structures of random, sorted and compartmentalized by analyzing a pathological image?
A computer guided method to learn the connectivity strength between different cell types is an objective analytical tool that could reveal organisation patterns beyond what we can identify by eye sight.

A Bayesian approach on likelihood inference provides framework to learn the symmetric connectivity matrix $\mathbf{J}$ for a given configuration of sights $\vec{s}$, that most likely generated the image. This is done by evaluating the probability of finding an image in a configuration $\vec{s}$ for given connectivities $\mathbf{J}$:

$$P(\mathbf{J}|\vec{s}) = \frac{P(\vec{s}|\mathbf{J})P(\mathbf{J})}{P(\vec{s})} \tag{6}$$

Since the normalisation $P(\vec{s})$ is constant for a given configuration $\vec{s}$, we can simplify $P(\mathbf{J}|\vec{s}) \propto P(\vec{s}|\mathbf{J})P(\mathbf{J})$, assuming no prior knowledge on the parameters $P(\mathbf{J}) = const.$ This means, maximizing the posterior probability $P(\mathbf{J}|\vec{s})$ is equivalent to maximizing the likelihood function $P(\vec{s}|\mathbf{J})$ provided by the observations and the prior knowledge about the parameters $P(\mathbf{J})$ [9], [10]. According to the Hammersley-Clifford theorem, the joint Gibbs distribution for the Hamiltonian $\mathcal{H}$ under consideration, is defined as:

$$P(\vec{s}|\mathbf{J}) = \frac{e^{-\beta\mathcal{H}}}{Z} \tag{7}$$

Where the partition function Z is defined as:

$$\sum_{s_0=1}^{q} \dots \sum_{s_N=1}^{q} e^{-\beta\mathcal{H}} \tag{8}$$

Evaluating the partition function is forbiddingly hard as its complexity scales with q^N. Making use of the conditional independent assumption $P(A,B) = P(A|B)P(B) \xrightarrow{\text{cond. independent}} P(A)P(B)$ as proposed by Besag (1974), it is possible to evaluate the maximum pseudo likelihood (MPL). This provides a computationally feasible estimation of the maximum likelihood. The local conditional density function for a single cell i reads as:

$$P(s_i|\mathbf{J}) = \frac{e^{-\beta\mathcal{H}_{s_i}}}{\sum_{l=1}^{q} e^{-\beta\mathcal{H}_l}} \tag{9}$$

Similarly to [9], now only energy contributions from the k^{th} order neighborhood n_i^k of cell i are considered in eq.(9), s.t the energy contribution from cell i reads as:

$$\mathcal{H}_{s_i} = \sum_{j=[0,\dots i-1,i+1,\dots N]} J_{s_i s_j} \Theta(||\vec{x}_i - \vec{x}_j|| \leq \mu) \tag{10}$$

$$= \sum_{j\in[0,N]:j\in n_i^k} J_{s_i s_j} \tag{11}$$

The definition of (9) allows the evaluation of the maximum pseudo likelihood for the Potts Markov Random Field as:

$$\log(\mathcal{L}) = -\sum_{i=1}^{N} \left[\beta\mathcal{H}_{s_i} + \log\left(\sum_{l=0}^{q} e^{-\beta\mathcal{H}_l} \right) \right] \tag{12}$$

Exploring the expression of the pseudo likelihood (PL) (12) we can show that it is translational symmetric w.r.t. the Hamiltonian $\mathcal{H}$: $\log(\mathcal{L}(\mathcal{H})) = \log(\mathcal{L}(\mathcal{H}+K))$

$$-\log(\mathcal{L}(\mathcal{H}+K)) = \sum_{i=1}^{N}\left[\beta \sum_{j\in[0,N]:j\in n_i^k}\left(J_{s_i s_j}+K\right) + \log\left(\sum_{l=0}^{q}\exp(-\beta \sum_{j\in[0,N]:j\in n_i^k} J_{l s_j}+K)\right)\right] \tag{13}$$

$$= \sum_{i=1}^{N}\left[\beta\Theta K + \sum_{j\in[0,N]:j\in n_i^k}\beta J_{s_i s_j} + \log\left(\exp(-\beta\Theta K)\cdot\sum_{l=0}^{q}\exp(\sum_{j\in[0,N]:j\in n_i^k} -\beta J_{l s_j})\right)\right] \tag{14}$$

$$= \sum_{i=1}^{N}\left[\beta\mathcal{H}_{s_i} + \log\left(\sum_{l=0}^{q} e^{-\beta\mathcal{H}_l}\right)\right] = -\log(\mathcal{L}(\mathcal{H})) \tag{15}$$

In eq. (13) the integer constant Θ indicates the number of neighboring cells of the local neighborhood of cell i, i.e. $\sum_{j\in[0,N]:j\in n_i^k} 1$.
The impact of the translational symmetry in the PL, see eq. (13), is, that from originally $\frac{q(q+1)}{2}$ free parameters one will remain undetermined. This reduced the range of parameters we can possibly infer with the method under consideration to $\frac{q(q+1)}{2}-1$.

3 Results

The data which we use to drive the before discussed models and analysis stems from tissue samples of 41 TNBC patients. Above information about the up regulation of several biochemical markers, each cell is classified as one out of 17 different cell types. The procedure of immunostaining the samples and the machine learning models behind the classification of the cells, among other analysis, was done by [4]. Our models process information about the cells location, in 2D, and its type. As a simplification we approximate the cells expansion to a dot, at the location of its center of mass.
The strategy to test the success and the accuracy of the MPL is to first generate a Markov random field with set inter cellular connectivities $\mathbf{J}$. We refer to this process as *encoding*. Then the MPL framework is supposed to infer these connectivities $\mathbf{J}$. This process is referred to as *decoding*. The goal is that the Markov random field, generated with the decoded connectivities matches qualitatively with the initially encoded configuration and the decoded parameter connectivities match quantitatively with the set of parameters that encode the initial configuration, see fig. 1.

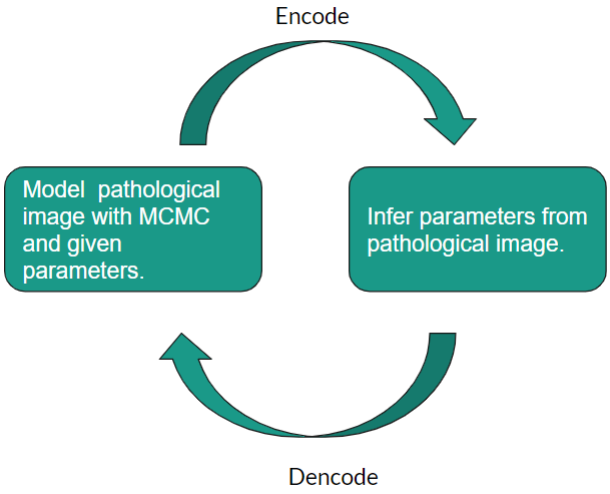

Figure 1: Visualisation of the workflow strategy: A Markov random field is generated (encoded) with a custom chosen set of parameters. The resulting pattern configuration is decoded, i.e: the parameters are inferred with the MPL framework. The inferred parameters are used to again generate a pattern configuration. If the configuration generated with the initially chosen parameters and the inferred parameters agree, the process was successful.

First the findings of the radial distribution analysis are discussed in sec. 3.1. This provides an overview about the general structure of the TNBC tissues under consideration. In the following chapters, 3.2 and 3.3.4, the success and limitations of the two pillars decoding (referred as MPL framework), encoding (referred as MCMC framework), see fig. 1 are discussed.

3.1 The radial distribution

We analyze the radial distribution function of the cells that form the TNBC tissue samples under consideration. The samples contain $\mathcal{O}(10^4)$ cells. Given the information of the cells location, the distance from every cell i to every other cell j is binned and normalized according to (1). This analysis reveals the RDF of the configuration and hence provides information about how much more likely it is to find a cell within a neighborhood of distance r, compared to random organisation.

As proposed in (3), we associate a potential $u(r)$ to the analyzed data of the RDF. To obtain an analytical expression, the data is fitted to a Lennard Jones potential V_{LJ}, as formulated in (4). Note that from an expression of energy we can find also an expres-

sion of force f. All of these findings are depicted in figure 2.

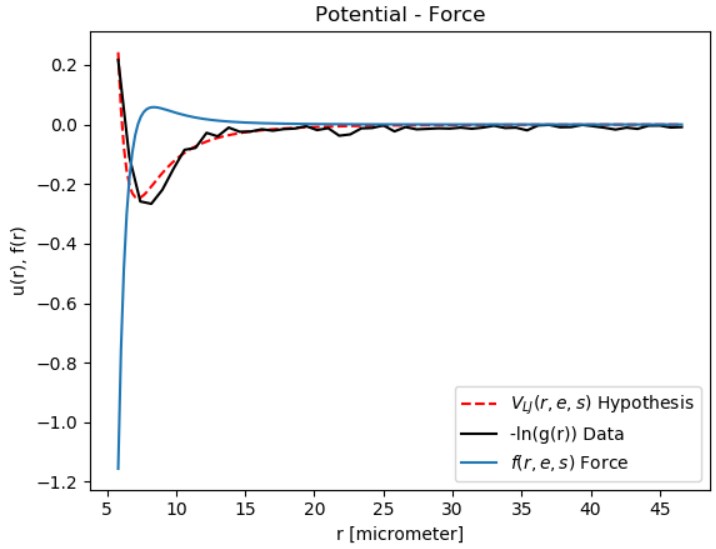

Figure 2: The potential potential $u(r) = \ln(g(r))$ associated to the RDF $g(r)$ measured in a tissue sample. The red dotted line is the χ^2 fit of the measured potential to a Lennard Jones like potential. The function $f(r)$ depicts the force, which is derived from V_{LF}.

The analysis in fig. 2 shows that for the tissue under consideration no long range order can be measured. The radial distribution resembles the one of a gas like particle configuration. Cells interact most likely at 8 µm, this is, nearest neighbor interaction. Note that this analysis is done, treating the cells as indistinguishable sights. The Lennard Jones potential V_{LJ} and force that can be associated to the radial distribution $g(r)$ through a logarithmic relation gives us more detailed information at which distance the interaction is strongest and how fast the direct influence from one cell to another one decays. The function that minimize the χ^2 fit is:

$$V_{LJ}(r)\text{-Hypothesis } = 0.988 \left[\left(\frac{-6.076}{r} \right)^8 - \left(\frac{-6.076}{r} \right)^4 \right] \tag{16}$$

3.2 Parameter learning via MPL

For two cell types aligned on one axis, every configuration can be generated by six universal configuration of three sights. The configuration as a whole is made from

combinations of these basic three cell building blocks. For the six universal 3 sight configurations the maximum of the pseudo likelihood can be derived analytically as $\nabla_J \log(\mathcal{L}(\mathbf{J})) = 0$, see eq. (12). Denoting 'x' to one, 'o' to another cell type and bold letters (**x**,**o**) to the reference cell, table 1 provides a detailed analytical insight on the six universal 3 sight combinations individually. The energy contribution $\mathcal{H}$, the local partition $\mathcal{Z}$, the multi directional derivative $\nabla_J = \begin{bmatrix} J_{11}, & J_{12}, & J_{22} \end{bmatrix}^T$ of the logarithmic pseudo likelihood $\nabla_J \left(\beta\mathcal{H} + \log(\mathcal{Z})\right)$ and the tendency to its maximum (**max PL**) are inspected. Note that since the MC temperature $T = 1$ is constant, this also holds for the inverse MC temperature $\beta = 1$.

Table 1: Analytical analysis of different state configurations w.r.t the reference cell (bold).

State ψ	$\mathcal{H}$	$\mathcal{Z}$	$\nabla_J \left(\beta\mathcal{H} + \log(\mathcal{Z})\right)$	**max PL**
$I1$: o-**x**-x	$J_{12} + J_{22}$	$e^{-\beta J_{12}}(e^{-\beta J_{22}} + e^{-\beta J_{11}})$	$\begin{bmatrix} \frac{\beta I_1 e^{\beta J_{11}}}{e^{\beta J_{11}} + e^{\beta J_{22}}} \\ 0 \\ -\frac{\beta I_1 e^{\beta J_{11}}}{e^{\beta J_{11}} + e^{\beta J_{22}}} \end{bmatrix}$	$J_{11} \to -\infty$
$I2$: o-**o**-x	$J_{12} + J_{11}$	$e^{-\beta J_{12}}(e^{-\beta J_{22}} + e^{-\beta J_{11}})$	$\begin{bmatrix} -\frac{\beta I_2 e^{\beta J_{22}}}{e^{\beta J_{11}} + e^{\beta J_{22}}} \\ 0 \\ \frac{\beta I_2 e^{\beta J_{22}}}{e^{\beta J_{11}} + e^{\beta J_{22}}} \end{bmatrix}$	$J_{22} \to -\infty$
$M1$: o-**o**-o	$2J_{11}$	$e^{-2\beta J_{11}} + e^{-2\beta J_{12}}$	$\begin{bmatrix} \frac{2\beta M_1 e^{2\beta J_{11}}}{e^{2\beta J_{11}} + e^{2\beta J_{12}}} \\ -\frac{2\beta M_1 e^{2\beta J_{11}}}{e^{2\beta J_{11}} + e^{2\beta J_{12}}} \\ 0 \end{bmatrix}$	$J_{11} \to -\infty$
$M2$: x-**x**-x	$2\mathrm{J}_{11}$	$e^{-2\beta J_{22}} + e^{-2\beta J_{12}}$	$\begin{bmatrix} 0 \\ -\frac{2\beta M_2 e^{2\beta J_{22}}}{e^{2\beta J_{12}} + e^{2\beta J_{22}}} \\ \frac{2\beta M_2 e^{2\beta J_{22}}}{e^{2\beta J_{12}} + e^{2\beta J_{22}}} \end{bmatrix}$	$J_{22} \to -\infty$
$A1$: o-**x**-o	$2J_{12}$	$e^{-2\beta J_{11}} + e^{-2\beta J_{12}}$	$\begin{bmatrix} 0 \\ \frac{2A_1 \beta e^{2\beta J_{12}}}{e^{2\beta J_{12}} + e^{2\beta J_{22}}} \\ -\frac{2A_1 \beta e^{2\beta J_{12}}}{e^{2\beta J_{12}} + e^{2\beta J_{22}}} \end{bmatrix}$	$J_{12} \to -\infty$
$A2$: x-**o**-x	$2J_{12}$	$e^{-2\beta J_{22}} + e^{-2\beta J_{12}}$	$\begin{bmatrix} -\frac{2A_2 \beta e^{2\beta J_{12}}}{e^{2\beta J_{11}} + e^{2\beta J_{12}}} \\ \frac{2A_2 \beta e^{2\beta J_{12}}}{e^{2\beta J_{11}} + e^{2\beta J_{12}}} \\ 0 \end{bmatrix}$	$J_{12} \to -\infty$

The minimum of every universal 3 sight configuration, analyzed at its own, is found for one of the three interaction parameters $J_{s_i s_j}, s_i, s_j \in \{x, o\}$ tending towards $-\infty$, leaving the other two interaction parameters undetermined.
How often each of these basic building blocks appears in the configuration as a whole shapes the PL landscape. Even if every universal 3 sight configuration at its own is minimal for one connectivity $J_{s_i s_j}$ tending towards negative infinity, the overall configu-

ration, i.e: the sum of universal 3 sight combinations, has a minimum $m \in \mathbb{R}$. This idea is supported by the following contour plots 3, 4, 5, 6 which show the negative maximum pseudo likelihood as a function of the free parameters J_{11}, J_{22}. Remember that from the translation symmetry shown in (13) one of the originally three free parameters will be undetermined, which is why it is chosen to be $J_{12} = 0$.

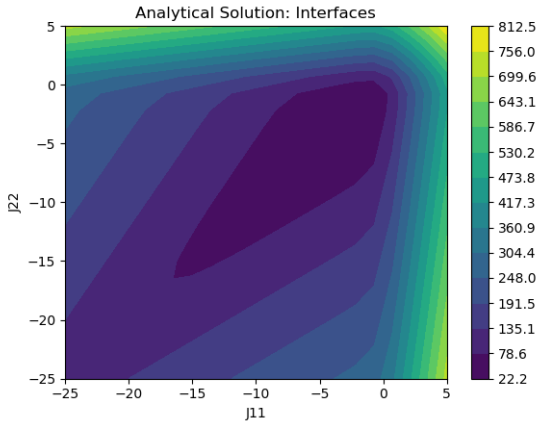

Figure 3: Analytical result of the negative pseudo likelihood terms in tab.1. The linear combination of $I1$, $I2$, $M1$, $M2$, $A1$, $A2$ terms matches the number of set configurations from fig. 4: $9\,(\log(\mathcal{L}_{I1}) + \log(\mathcal{L}_{I2})) + 40\,(\log(\mathcal{L}_{M1}) + \log(\mathcal{L}_{M2})) + (\log(\mathcal{L}_{A1}) + \log(\mathcal{L}_{A2}))$.

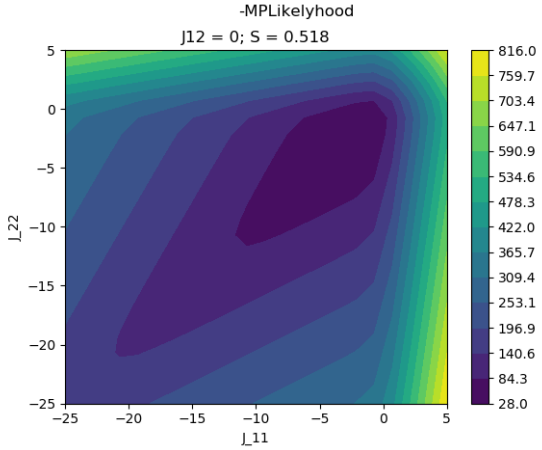

Figure 4: Negative pseudo likelihood landscape evaluated on a set configuration that is dominated by middle $(M1, M2)$ terms: cells surrounded by its own kind: $I1 = I2 = 9$, $M1 = M2 = 40$ and $A1 = A2 = 1$ terms (see tab. 1).

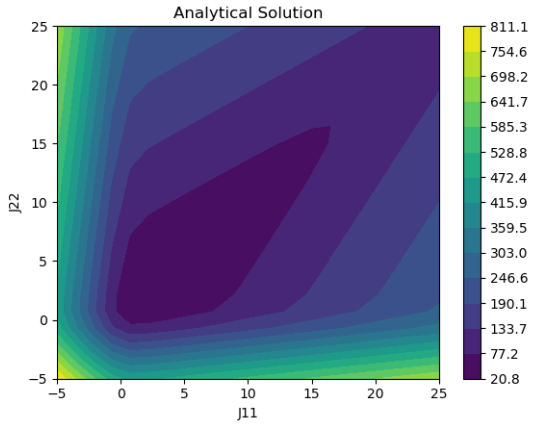

Figure 5: Analytical result of the negative pseudo likelihood terms in tab.1. The linear combination of $I1$, $I2$, $M1$, $M2$, $A1$, $A2$ terms matches the number of set configurations from fig. 6: $9\,(\log(\mathcal{L}_{I1}) + \log(\mathcal{L}_{I2})) + (\log(\mathcal{L}_{M1}) + \log(\mathcal{L}_{M2})) + 40\,(\log(\mathcal{L}_{A1}) + \log(\mathcal{L}_{A2}))$.

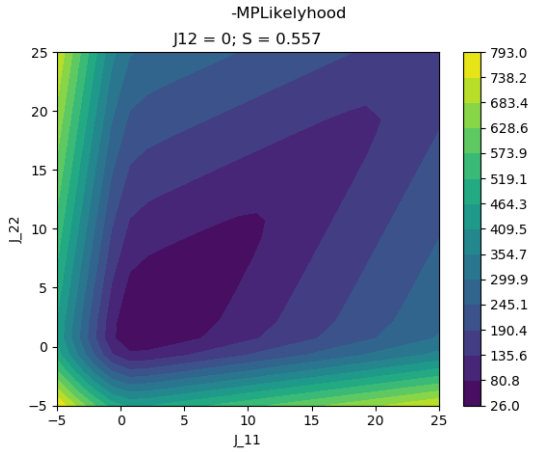

Figure 6: Negative pseudo likelihood landscape evaluated on a set configuration that is dominated by asorted $(A1, A2)$ terms: cells surrounded by a different type: $I1 = I2 = 9$, $M1 = M2 = 1$ and $A1 = A2 = 40$ terms (see tab. 1).

These four images compare the computer evaluation of the PL for a set combination of 3 sight configurations to the analytical evaluation the PL. The analytical PL is set up from a linear combination of the terms from tab. 1. How often a particular term appears is resembled in the size of the coefficient in front of each $\log(\mathcal{L})$ term. For the scenario in which the configuration is determined by middle terms (3, 4), as well as for the scenario in which the configuration is determined by asorted terms (5, 6) we find that the analytical contour landscape is in accordance with the computer evaluation.

More over we can see that the contour landscape is shaped by how often every term appears: The contour landscape of fig. 4 is evaluated from a configuration that is dominated by cells of its own type being next to each other. Correctly the negative PL is minimal for negative connectivity parameters J_{11}, J_{22} of equal size. This suggests that the configuration results from affinity from cells to other cells of its own kind. On the contrary fig. 6 suggests positive connectivity parameters of equal size. The reason why in both scenarios the negative PL landscape is minimal for finite connectivities is due to the fact that the middle term dominated scenario also contains asorted terms and vice versa. Both scenarios have in common that connectivities of equal size are preferred. This can by explained by the appearance of interface terms: The special setting of configurations is that the number of interfaces terms from one type appears equally often than the interface terms from the second type: $I1 = I2$. If the proportion of the two cell types is equally big, this condition is implied if an interface exists. The PL is minimal for the choice $J_{11} = J_{22}$, leaving no further constraint on the magnitude of the two parameters. This can be observed in fig. 7, where the PL of a configuration of only interface terms (... x-x-o-o-x-x-o-o...) $I1 = I2 = 50$, $M1 = M2 = A1 = A2 = 0$ is evaluated. This is, every reference cell is surrounded by one cell of its own type and another one of a different type.

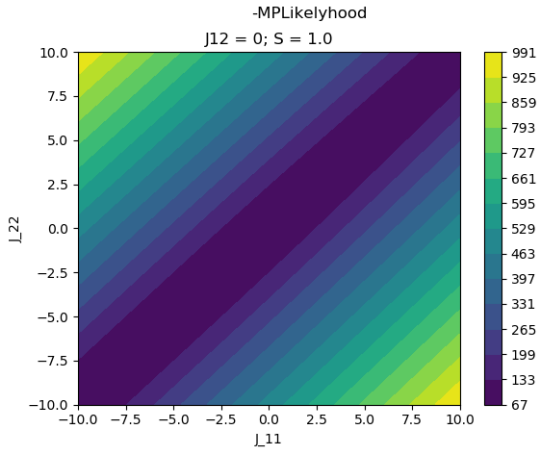

Figure 7: Negative pseudo likelihood evaluated on a configuration that is composed of $I1 = I2 = 50$, $M1 = M2 = A1 = A2 = 0$. The MPL framework leaves the size of the connectivities undetermined but suggests that they are equally big $J_{11} = J_{22}$.

For this case the size of the parameters is undetermined, as long as they are equally big. Most importantly one should point out that this result is not in accordance with what we find when driving the model with a set of given connectivity parameters. From the way the MCMC encoding framework to create a configuration from a given set of connectivity parameters is set up, it is not possible to find a set of parameters that drives

a pattern that is dominated by interfaces. Choosing positive parameters results in self repulsion, a configuration dominated by $A1$, $A2$ terms and negative parameters results in self affinity. The pattern is then dominated by $M1$, $M2$ terms. If the size of these parameters is, above a certain threshold, equally big, this explanation still holds. The fact that the Hamiltonian we constructed is not capable of driving a pattern, dominated by interface contributions, suggests that some modifications have to be done to be able to encode any configuration possible.
Within the MPL framework, this finding suggests that, if equally many interface terms occur in some configuration, from originally $\frac{q\cdot(q+1)}{2}$ inter-cellular connectivities, we can only infer $\frac{q\cdot(q+1)}{2}-2$, because of the transnational symmetry of the PL and the $J_{11}=J_{22}$ symmetry enforced for equal amount of interface configurations from the two cell types.

3.3 Modeling spatial organisation via MCMC

3.3.1 Equivalence patterns

Transnational symmetry in pseudo likelihood analysis, see eq. (13), suggests that this indeterminacy of one parameter is also reflected in the choice of the parameter set of connectives $\mathbf{J}$, when generating a configuration. This symmetry implies that from generally $\frac{q(q+1)}{2}$ degrees of freedom only $\frac{q(q+1)}{2}-1$ can be inferred. Indeed, similar [1] configurations can be generated for different sets of parameters, which are transnational symmetric. Consider a prototypical parameter sets that drives a sorted configuration between two cell types. All of the following parameter sets $\mathbf{J}$ generate similar configurations of a sorted pattern.

- $\mathbf{J}=\begin{bmatrix}-a & b\\ b & -a\end{bmatrix}$: Equally high self affinity among the two types and repulsion to other types.
- $\mathbf{J}-b=\begin{bmatrix}-a-b & 0\\ 0 & -a-b\end{bmatrix}$: Self affinity among the two types.
- $\mathbf{J}+a=\begin{bmatrix}0 & a+b\\ a+b & 0\end{bmatrix}$: Repulsion to the other types.

In the considered model the number of parameters to drive a configuration can even be reduced to $\frac{q(q+1)}{2}-2$. Again a configuration of two cell types can be used as an instructional example: If the connection between the two types is passive ($J_{12}=J_{21}=0$), and

[1] Since the MCMC process is stochastic, the final pattern configurations are never identical. Still, the large scale structure of two cell types being sorted, random or maximally far away from its own type is clearly identifiable.

one self interaction is also passive ($J_{22} = 0$), random, sorted and asorted configurations can be generated by defining the self interaction of the other cell type ($J_{11} = a$). Note that this parameter reduction applies only to the process of generating a configuration. The most likely hypothesis, when infering the parameters, is still the one with $\frac{q(q+1)}{2} - 1$ parameters. Reducing the number of free parameters in the MPL to $\frac{q(q+1)}{2} - 2$ results in a lower maximum likelihood opposed to the one that can be reached with $\frac{q(q+1)}{2} - 1$. This key finding suggests that with the presented framework it is not possible to infer unique inter cellular interactions from the observation of pathological images. The transnational symmetry in the parameter inference MPL model and in the generative MCMC model suggest that several hypothesis are equally likely to have generated a configuration.

3.3.2 Pattern configurations

With the definition of a systems energy, see (5), we can use the MCMC optimisation strategy to generate pattern configurations of cells. For two cell types $l = \{1, 2\}$ our model can generate three configurations: random, sorted, asorted (which we call the state of maximal repulsion from every cell to every other cell of its own type). The magnitude and sign of the value in the connectivity matrix $J_{s_i s_j}$ defines a pairwise affinity or repulsion between cell type s_i and s_j. Consider the following figures 8, 9, 10 as instructional example for the possible pattern formations of two cell types on a regular square lattice:

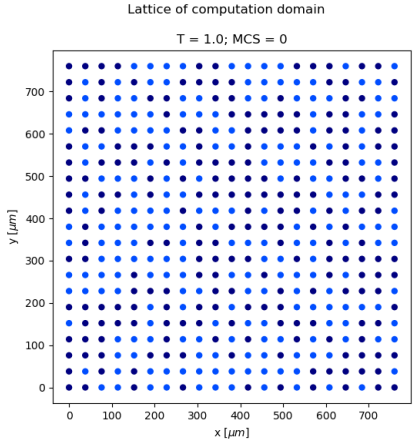

Figure 8: Initially the configuration of cell types are randomly distributed.

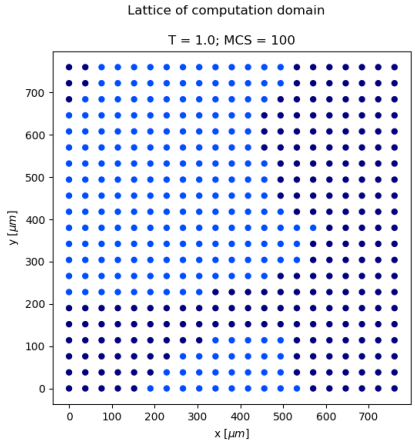

Figure 9: A sorted configuration results from a MCMC optimisation after 100 Monte Carlo epochs (MCS) with a parameter set of $\mathbf{J} = \begin{bmatrix} -5 & 0 \\ 0 & -5 \end{bmatrix}$ or its translations.

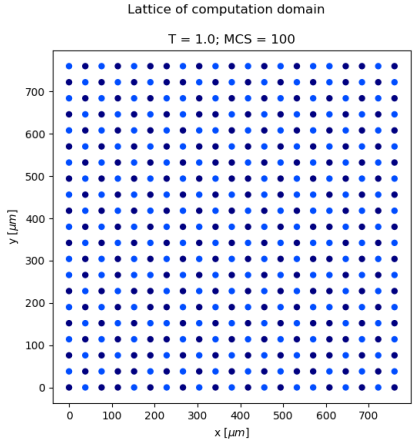

Figure 10: A asorted configuration results from a MCMC optimisation after 100 Monte Carlo epochs (MCS) with a parameter set of $\mathbf{J} = \begin{bmatrix} 5 & 0 \\ 0 & 5 \end{bmatrix}$ or its translations.

The model is extendable to multiple cell types, i.e: $l = \{0, ..., q\}$. The final pattern configuration is still determined by pairwise interactions. Again the following figure 11 is an instructive example on how a chosen parameter set results in a large scale pattern.

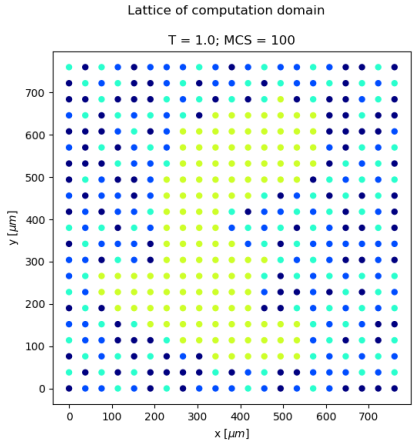

Figure 11: The final pattern configuration is determined by the pairwise interactions of the four cell types. The configuration results after 100 Monte Carlo epochs (MCS) from MCMC optimising the parameter set $\mathbf{J} = \begin{bmatrix} 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 5 & 0 \\ 0 & 0 & 0 & -5 \end{bmatrix}$.

In fig. 11 it is possible to identify a cluster among cells of cell type four (green). This matches with the parameter $J_{44} = -5$, which indicates self affinity towards cells of its own type. Cell type three (turquoise) is chosen to be self repulsive $J_{33} = 5$. Indeed within the first order neighborhood of cells from type three, there is no other cell of this type. Finally, cell type one (dark blue) and cell type two (light blue) are passive towards its own kind $J_{11} = J_{22} = 0$ and all other cell types. As a result we can observe random formation between these two types. To find quantitative evidence for these observations we measure the probability of finding a cell from type s_j in the local (first order) neighborhood of a cell from type s_i. The entry $P_{s_i s_j}$ in the square matrix $\mathbf{P} \in \mathbb{R}^{q \times q}$ contains these probabilities. Note that the probability of having a cell of its own type within the local neighborhood can be read off from the diagonal elements of $\mathbf{P}$. Let us consider the configuration of figure 11 as an example:

$$\mathbf{P} = \begin{bmatrix} 0.30 & 0.27 & 0.37 & 0.05 \\ 0.32 & 0.20 & 0.41 & 0.07 \\ 0.47 & 0.45 & 0.0 & 0.08 \\ 0.05 & 0.06 & 0.065 & 0.85 \end{bmatrix} \tag{17}$$

The self affinity of cell type four is set to be high $J_{44} = -5$ which is reflected by a high probability of finding its own type within a local neighbor hood ($P_{44} = 0.85$).

On the other hand cell type three is given the attribute of self repulsion $J_{33} = 5$. This is reflected in a vanishing probability of finding its own type within a first order neighborhood ($P_{33} = 0$).

3.3.3 Phase transitions

We introduce entropy as the amount of randomness one image configuration contains. The entropy is defined as:

$$S = -\sum_{i=1}^{2} p_i \log(p_i) \tag{18}$$

In eq. (18) p_i is the probability of finding a cell of its own type or of a different type in the first order neighborhood. These are for example the entries along a row of the local probability matrix $\mathbf{P}$ (17).
The entropy is maximal if the order of the configuration is low. In our case this is the case if the probability of finding cells of the same type as is equally likely then finding a cell of any other type within the first order neighborhood. With this definition of entropy we can measure phase transitions from maximally sorted, random and maximally asorted configurations. These are primarily depending on the magnitude of the ordering parameter βJ, see fig. 12. Further, the the phase transition is influenced by the number of states included in a configuration. The model follows the theoretically predicted logarithmic relation ship w.r.t. the increase in the phase parameter (see eq. (19)), as more states are added to the configuration [11]. This is demonstrated in fig. 13.

$$l = \{1, 2\} : (\beta J)_c = \frac{1}{2} \log(1 + \sqrt{2}) \tag{19}$$

$$l = \{1, ..., q\} : (\beta J)_c = \log(1 + \sqrt{q}) \tag{20}$$

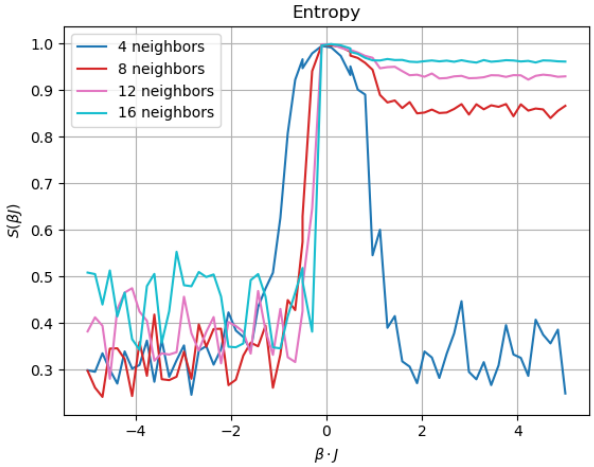

Figure 12: Entropy of a 2 cell type configuration aligned on a regular 2D lattice. The entropy in the asorted regime (positive phase parameters βJ) increases with the order of neighborhood k. This is because a global configuration optimum of maximally self repulsive cell types involves having cells of the same type within a local neighborhood for $k > 1$, see appendix 4.2.

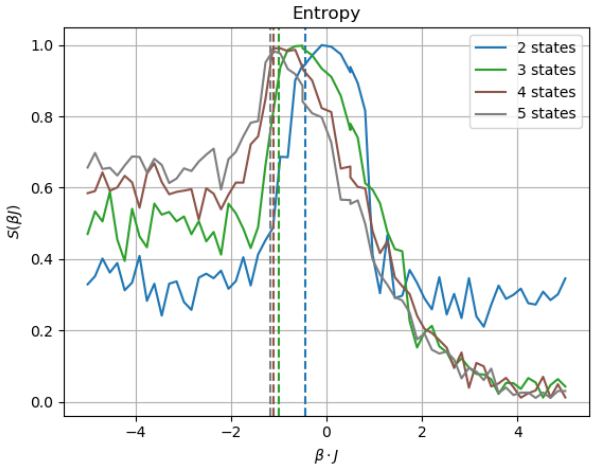

Figure 13: A sharp drop of entropy can be observed when decreasing the phase parameter βJ. This is the result of cell types of its own kind being clustered together. As a consequence of adding more states to the configuration, the interface region between the clusters increases, which is why the entropy reaches lower minima as the number of states increases. Considering positive connectivity parameters $\beta J > 0$, the interface region increases naturally as more states are added, which is why the entropy in the asorted regime decreases when more states are added.

3.3.4 An approach on real data

So far it could be shown that the MPL framework, sec. 3.2 is capable of correctly suggesting the tendency of a parameter set **J**, which could have been used to generate a pattern configuration similar to the one which was analyzed, see fig. 3, 4, 5, 6.
Also it has been shown that due to a translational symmetry eq. (13) and the symmetry enforced by an equal number of interface terms, fig. (7), the MPL framework is so far not capable of exactly inferring the parameter set from which a pattern configuration roots. Still, the implemented parameter learning procedure correctly outputs a set of parameters that is likely to be responsible for the generation of the pattern configuration under consideration. Inspecting the contour landscape of the PL, the parameter tendencies which most likely generated the image can be read of, fig. 6, 4.
It could be demonstrated that the MCMC framework is able to generate pattern configurations for multiple cell types, on the basis of pairwise interactions. Controlling the set of parameters **J** and thus the magnitude of repulsion and affinity from one cell type towards the other one, it is possible to generate a desired pattern, fig. 11.

The next result will discuss a full decoding-encoding cycle, fig. 1, on two out of the 41 TNBC tissue samples. With the MPL framework the parameter set **J** of tissue samples 12, characterised as mixed type, and 4, characterised as compartmentalized type, are inferred. For the sake of simplicity the simulation depicts a cell type selection that classifies a cell either as immune or tumor type. This analysis reveals the direction (affinity or repulsion) and magnitude of inter the cellular connections between those two types. The inferred set of parameters is then used to generate a pattern configuration for each sample respectively. The pathological image of the tissue of patient 12, see fig. 14, and patient 4, see fig. 18, is qualitatively compared to the configuration generated by the simulation, which is fed by the previously inferred parameter set **J** for both patients respectively, fig. 16, 20. Similar to the probability matrix **P** in eq. (17), fig. 15, 19 compares the probability of finding an immune or tumor cell within the local neighborhood of a reference immune cell to the simulated configurations, fig. 17, 21. This quantifies the success of the encoding - decoding cycle. The very general differentiation between only two cell types may not raise a lot of attention to a biologist, since a more detailed classification would be needed to deliver answers to the complex interplay within a tumor environment. We may therefor consider these simulations as a proof of concept.

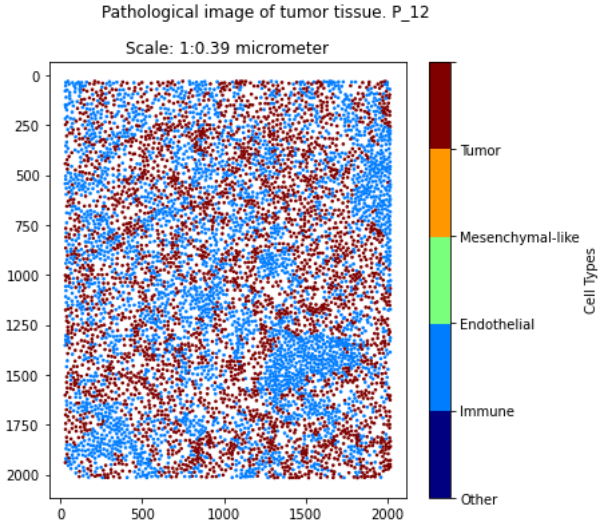

Figure 14: Original pathological image of patient P12 with general (immune, tumor) differentiation.

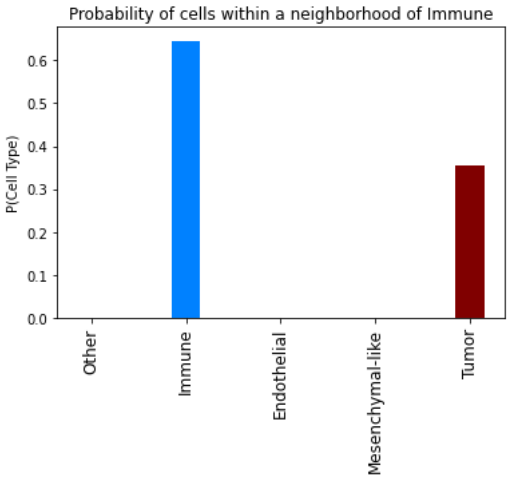

Figure 15: Probability of finding different cell types in the neighborhood of immune cells. The analysis is done fore the pathological image of patient P12.

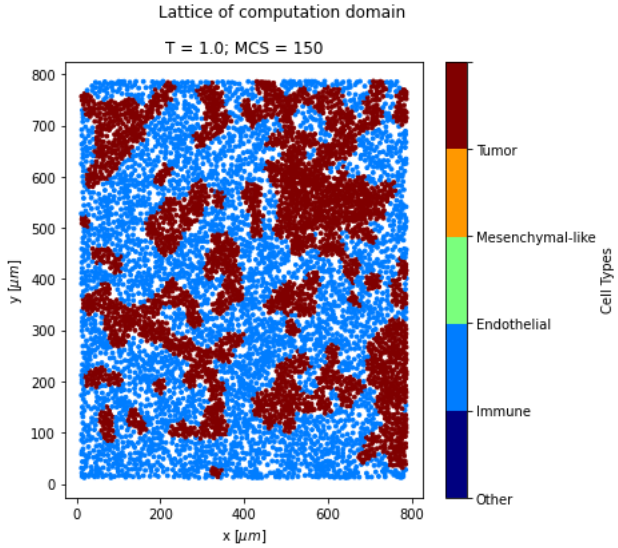

Figure 16: Simulated configuration of patient P12 with connectivity parameters $\mathbf{J_{12}} = \begin{bmatrix} 1.10 & 0.37 \\ 0.37 & -3.24 \end{bmatrix}$.

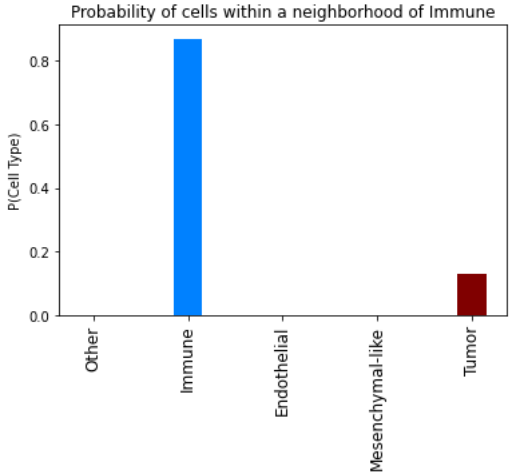

Figure 17: Probability of finding different cell types in the neighborhood of immune cells. The analysis is done fore the simulated configuration of patient P12.

Comparing the pathological image 14 and the simulated configuration 16 of patient 12, we can see that the general structure is similar. However, in the pathological image, the tumor environment looks more heterogeneous, than in the simulation, where we can see the formation of small compartments. This observation is supported by comparing the probabilities of a tumor or immune cell being in the local neighborhood of an immune

cell: In the original image these probabilities are almost balanced, where as in the simulation immune cells are much more likely to be surrounded by other immune cells. This means that the inferred parameters J_{11}, J_{22} over estimate self affinity: diagonal elements are small compared to off diagonal elements.

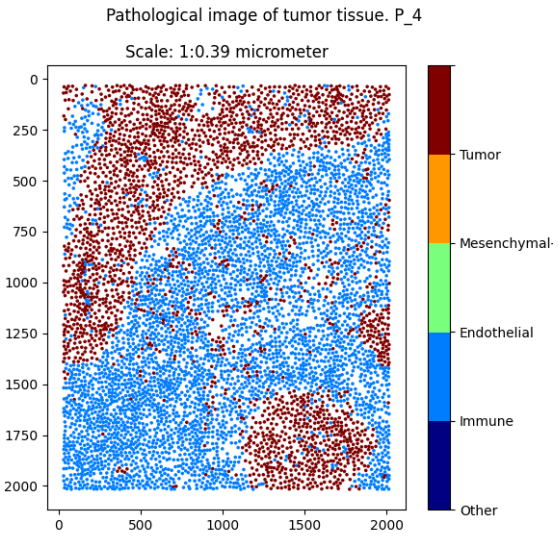

Figure 18: Original pathological image of patient P4 with general(immune, tumor) differentiation.

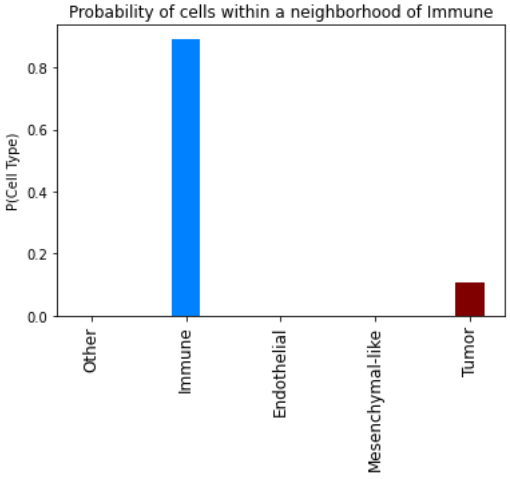

Figure 19: Probability of finding different cell types in the neighborhood of immune cells. The analysis is done fore the pathological image of patient P4.

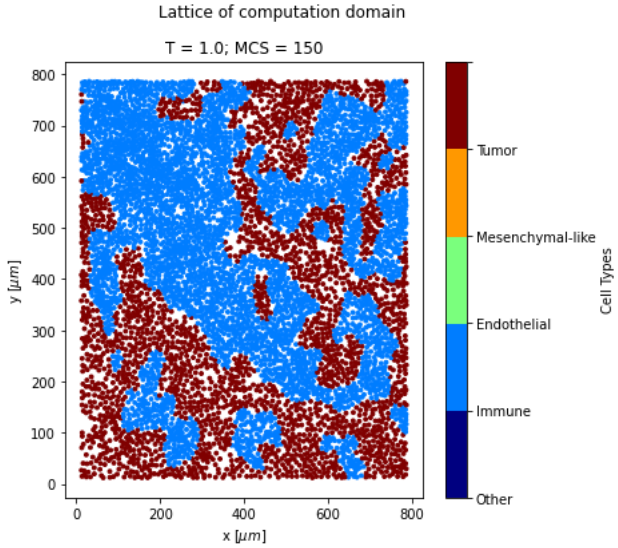

Figure 20: Simulated configuration for patient P4 with connectivity parameters $\mathbf{J_4} = \begin{bmatrix} 1.37 \cdot 10^{-4} & 1.93 \cdot 10^{3} \\ 1.93 \cdot 10^{3} & 1.93 \cdot 10^{3} \end{bmatrix}$.

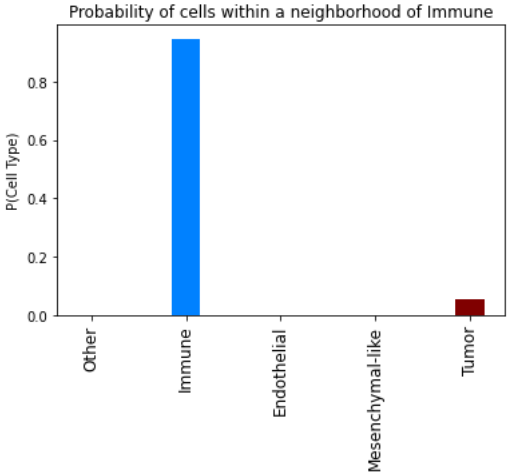

Figure 21: Probability of finding different cell types in the neighborhood of immune cells. The analysis is done fore the simulated configuration of patient P4.

For patient 12 we again find rough agreement between the pathological image 18 and the simulated configuration 20. A qualitative difference is the immune- tumor interface region. This boundary regions are much more prominent in the simulation than in the original image, suggesting that the repulsion to the other cell type is under estimated: The off diagonal elements in $\mathbf{J}_4$ are small compared to the diagonal elements. Compar-

ing the probability of finding an immune or tumor cell within the local neighborhood of an immune cell, shows that this qualitative difference is almost not reflected in the quantitative analysis.

Performing the discussed simulations comes with two computational difficulties: First, inferring the optimal parameters is very time consuming. The samples under considerations consist of $\approx$ 6000 individual cells. To find a parameter set that most likely generated this image, the optimizer has to scan through all cells several times. Doing so, the computation of the local partition function can easily experience overflow errors, which is the second computational challenge. From expression (7) we can see that the local partition function becomes big for an increasing number of local neighbors or different states (cell types) encountered. Also, if the optimizer proposes a parameter $|J_{s_i s_j}| >> 0$, this will result in a local partition which exceeds the computers memory capacity. To tackle these two challenges the simulation for the encoding- decoding cycle on real data is computed on a high performance computer. In order to counteract overflow errors we add a Gaussian prior, centered around 0 with a standard deviation of 1, to the pseudo likelihood function (7). This penalises the optimisers suggestions for too big parameters and constraints the inferred parameters to be in an order of magnitude, which is feasible for the computers memory capacity.

4 Conclusion

This work provides a quantitative approach to infer the guiding parameters within tumor organisation. Inspired from the well explored Potts model, the general idea is to assume pairwise inter cellular interactions as the underlying driving forces in the organisation of the tumor environment. The MCMC framework delivers a generative model that is capable of producing a macroscopic pattern configuration from a given choice of inter cellular interactions. The main effort of this project was to present a yet unexplored and novel method to infer these key parameters that guide the large scale pattern formation. The MPL framework provideds a quantitative approach to infer the magnitude of pairwise inter cellular connections. This approach reduces the number of possible parameters guiding the formation of the tumor environment to a reasonable number, which scales quadratic with the number of different cell types encountered.

The limitations of the provided parameter inference method are intrinsic to the formulation of the model. As shown in eq. (13) the maximum likelihood for a set of parameters is symmetric with respect to any translation. Also, the interface region, which is the region where a reference cell is surrounded by equally many cells of its own type as of another type leads to the fact that connectivities of equal size are the preferred choice of the method. These symmetries forbid finding a unique solution of inter cellular connections that are responsible for the formation of a certain configuration.

Being aware of the methods limitations, it is still possible to infer a set of parameters that could be responsible for the formation of the analyzed image, see fig. 6, 4. As this work has shown, several versions of inter cellular connectivities can result in a similar macroscopic configuration. For a configuration, which consists of two different cell types the MPL framework is capable of suggesting one out of many scenarios correctly.

The author also celebrates the simulations done on an actual pathological image as a success. Comparing the qualitative features of the pathological images under consideration to the simulated configurations, we find some discrepancies. More precisely, the parameter inferring method is not capable of correctly inferring a set of parameters that resembles the underlying truth of connectivities. Still, even in the eyes of a critical observer, the simulation and the original image are comparable and show general similarities. Also the strategies to perform the computation in a feasible time are worth calling a success. To increase the models accuracy, s.t. the simulations would more accurately resemble the tumor environment observed the best chance is to work on a more sophisticated prior function for the pseudo likelihood, see eq. (7). Such an expression is omitted in all results presented, except from the simulations shown on real data, sec. 3.3.4. This is the only time when a Gaussian prior, centered around zero with a standard deviation of one is added in order to perform computations that do not overflow in memory.

Acknowledgement

I want to thank Jean Hausser [2] who was my mentor at SciLifeLab [3] during the six months this voluntary project was done. Thanks to all the stimulating and encouraging conversations with him I learned to appreciate unexpected results, because that is when a scientist learns the most.

References

[1] D. Hanahan, R. A. Weinberg, The hallmarks of cancer, Cell 100 (1) (2000) 57–70. doi:10.1016/s0092-8674(00)81683-9.

[2] Y. Hart, H. Sheftel, J. Hausser, P. Szekely, N. B. Ben-Moshe, Y. Korem, A. Tendler, A. E. Mayo, U. Alon, Inferring biological tasks using pareto analysis of high-dimensional data, Nature methods 12 (3) (2015) 233.

[2] https://www.hausserlab.org
[3] https://www.scilifelab.se

[3] O. Shoval, H. Sheftel, G. Shinar, Y. Hart, O. Ramote, A. Mayo, E. Dekel, K. Kavanagh, U. Alon, Evolutionary trade-offs, pareto optimality, and the geometry of phenotype space, Science 336 (6085) (2012) 1157–1160.

[4] L. Keren, M. Bosse, D. Marquez, R. Angoshtari, S. Jain, S. Varma, S.-R. Yang, A. Kurian, D. V. Valen, R. West, et al., A structured tumor-immune microenvironment in triple negative breast cancer revealed by multiplexed ion beam imaging, Cell 174 (6) (2018). doi:10.1016/j.cell.2018.08.039.

[5] D. Mallet, L. De Pillis, A cellular automata model of tumor–immune system interactions, Journal of Theoretical Biology 239 (3) (2006) 334–350. doi:10.1016/j.jtbi.2005.08.002.
URL https://linkinghub.elsevier.com/retrieve/pii/S0022519305003334

[6] D. Chandler, Introduction to modern statistical mechanics, Oxford Univ. Press, 2009.

[7] M. Scianna, L. Preziosi, Cellular potts models: multiscale extensions and biological applications, CRC Press, Taylor Francis Group, 2013.

[8] F. Graner, J. A. Glazier, Simulation of biological cell sorting using a two-dimensional extended potts model, Physical Review Letters 69 (13) (1992) 2013–2016. doi:10.1103/physrevlett.69.2013.

[9] A. L. M. Levada, N. D. A. Mascarenhas, A. Tannús, Pseudo-likelihood equations for potts model on higher-order neighborhood systems: A quantitative approach for parameter estimation in image analysis, Brazilian Journal of Probability and Statistics 23 (2) (2009) 120–140. doi:10.1214/08-bjps018.

[10] S. Song, B. Si, J. M. Herrmann, X. Feng, Local autoencoding for parameter estimation in a hidden potts-markov random field, IEEE Transactions on Image Processing 25 (5) (2016) 2324–2336. doi:10.1109/tip.2016.2545299.

[11] M. T. M. Julien Stoehr, Richard Everitt, A review on statistical inference methods for discrete markov random fields (2017). doi:hal-01462078.

Appendix

4.1 Local neighborhoods

The local neighborhood of order k of cell i is indicated as n_i^k. Strong self repulsion from one cell type $s_i = A$ towards the other cell type $s_j = B$ results in a pattern where cell type A is surrounded by cells of a different kind B. Considering a first order neighborhood system $k = 1$, it is possible to construct a pattern of maximal self repulsion if the local neighborhood of cell i with $s_i = A$ is surrounded by cells $j \in n_i^1$ of a different type $s_j = B$. For higher order neighborhood systems $k > 1$ a global optimum of self repulsion is reached if the reference cell i also has cells of its own kind within its local neighborhood, see fig. 2.

Table 2: The local neighborhood systems of a first and second order neighborhood configuration that globally results in the maximal self repulsion possible.

[1:]

A	B	A
B	**A**	B
A	B	A

[2:]

B	B	B
A	**A**	A
B	B	B

4.2 Simulation specifications

All discussed simulations and analysis was programmed by the author. The code is available at the authors GitHub account: https://github.com/david-alber. The rispository "Parameter Inference In Tumor Environment" provides three custom libraries *methodslib1*, *mclib*, *mplib* which are imported in the main script *patientParamInf*. These are the backbone of the data preprocessing, Monte Carlo generative framework and the maximum pseudo likelihood parameter inference method. Other dependencies are the Python standard libraries *pandas*, *matplotlib*, *numba*, *scipy*, *numpy*, *tqdm*, *scikit-image*. The workflow of the main script is as follows:

- Set the simulation parameters: patient, depiction, cellTypes_orig, limiter, runs, N.
- Load original images and preprocess them in the selected depiction of cell types.
- Infer connectivity parameters **J** from the original image.
- Generate a simulated configuration of a tumor environment with the inferred parameter set.

A detailed description of specific functions and imput parameters will be given in the README

file on GitHub.
It is possible to infer parameters of small systems $\mathcal{O}(10^3)$ cells with a household computer within minutes. For larger systems, i.e: $\mathcal{O}(10^4)$ cells, a high performance computer or computation cluster is advised. Factors that influence the computation run time are the size of the local neighborhood of each cell, the total number of cells in the configuration, the number of free parameters to infer, which scales with $\frac{q(q+1)}{2}$ independent cell types.