**Title**
Multitype Point Process Analysis of Signaling Proteins in the Signaling Complexes of T cells



**Author**
Piao, Juan

**Publication Date**
2021

Peer reviewed|Thesis/dissertation

UNIVERSITY OF CALIFORNIA

Los Angeles

# Multitype Point Process Analysis of Signaling Proteins in the Signaling Complexes of T cells

A thesis submitted in partial satisfaction
of the requirements for the degree
Master of Science in Statistics

by

Juan Piao

2021

ABSTRACT OF THE THESIS

Multitype Point Process Analysis of Signaling Proteins in the Signaling Complexes of T cells

by

Juan Piao
Master of Science in Statistics
University of California, Los Angeles, 2021
Professor Frederic Paik Schoenberg, Chair

Studying the spatial organization of the molecules in T cell following activation of the T cell antigen receptor can improve our understanding of the association of the spatial structure of the molecules with T cell activation states. In particular, it has been found that the formation of the signaling complex is tightly related to proper signal transduction and T cell activation during the immune response. The purpose of this work is to discover the relationship between the spatial structure of the molecules in the signaling complex and the state of T cell by extracting biological knowledge from cellular imaging data using spatial point process methods.

In this thesis, we present discoveries on spatial distributions and attributes of proteins in microcluster and non-microcluster areas of three activated T cells and compare the differences between the spatial distributions in microcluster and non-microcluster. This thesis attempts to propose some possible biological hypotheses based on strong statistical evidence discovered from spatial point pattern analysis.

The thesis of Juan Piao is approved.

Jingyi Jessica Li

Vivian Lew

Frederic Paik Schoenberg, Committee Chair

University of California, Los Angeles

2021

*To Lin and Minna*

# Table of Contents

# List of Figures

# List of Tables

# Acknowledgments

First of all, I would like to express my heartfelt gratitude to my thesis advisor Professor Frederic Paik Schoenberg for his patient guidance, kind encouragement and helpful advice throughout my master study. I must give my great appreciation to Professor Vivian Lew, who led me into Statistics and gave me valuable support and guidance as a mentor whenever I needed. Also, I would like to thank Professor Jessica Jingyi Li who taught me so many things in class and helped me editing this thesis. I would like to offer my special thanks to Dr.Valarie Barr for her kind permission to use the data for this thesis and for providing detailed information on the data collection method and other background knowledge.

Finally, I wish to thank my family for their unwavering support and encouragement throughout my years of study.

# CHAPTER 1

# Introduction

T cells play important role in immune response system. When there are bacteria, viruses or other infectious agents present in human body, T cells recognize foreign antigens expressed and activates the T cell antigen receptor (TCR) in response. In the process, new signaling complexes are formed in TCR and numerous proteins are recruited to the signaling complexes [13]. The formation of signaling complexes is a crucial step in T cell activation as it is tightly linked to the actin cytoskeleton, which plays an essential role in intracellular trafficking, protein-protein interactions, and regulating the process of forming the complexes [9] [8].

Although little is known about the effect of spatial organization and the interactions of the signaling proteins in T cell on the regulation T cell activation, there is a growing body of evidence suggesting that there might be structure-function relationships in TCR [9] [20] [2]. The advancements in fluorescence microscopy imaging technology made it possible to the observation of the cellular structure at nanoscale [21]. In particular, single-molecule localization microscopy and direct Stochastic Optical Reconstruction Microscopy (dSTORM) are used to obtain the locations of individual molecules of various proteins in T cells [25]. In this study, we use point pattern analysis method to study the spatial organizations of the signaling proteins found in signaling complexes and compare this with the spatial organization of proteins outside the signaling complexes to discover if there are statistically significant differences between them.

# CHAPTER 2

# Methodology

## 2.1 Method to obtain the imaging data

The dataset and the details of the data collecting method was provided by Dr.Valarie Barr [10] [11]. T cells are activated by contact with a stimulatory antigen receptor activating surface. Since the signaling complexes are only formed on the contact sites between a cell and the surface, the locations of the molecules in signaling complexes are considered on a 2D surface, and therefore only the x-y positions of the molecules are of interest. The molecular positions or localizations of six proteins, phosphorylated T cell receptor zeta chain (TCRz), phosphorylated Linker for Activated T cells (LAT), phosphorylated Zeta Chain Associated Protein (ZAP-70) kinase, phosphorylated SH2-domain-containing leukocyte protein of 76 kDa (SLP-76), phosphorylated Phospholipase C gamma 1 (PLCg1) and CD45, were determined using the MadSTORM method [25] on Jurkat T cells stably expressing LAT-YFP. Each protein was labeled with antibodies conjugated to the fluorescent dye Alexa 647. The following antibodies from BD Pharmingen (San Jose CA) were used; mouse anti-human phospho-LAT (pY226) (cat# 558363, RRID AB_647281), mouse anti-CD247 (pY142) (cat# 558402, RRID AB_647307) and mouse anti-human CD45 (cat# 555480, RRID AB_395872). These antibodies from Cell Signaling Tech (Danvers, MA) were also used; rabbit anti-phospho-SLP-76 (pY145) (cat# 14770, RRID AB_2798604), rabbit anti-phospho-ZAP-70 (pY319) (cat# 2717, RRID AB_2218658) and rabbit anti-phospho-PLCg1 (pY783) (cat# 14008, RRID AB_2728690). The most likely location of each protein was determined by using the MadSTORM method. Imaging was performed as previously described with the following modifications: before staining, a TIRF image was taken of LAT-YFP followed by

a 10,000 frame series of the unstained cells and a full series of 10,000 images was taken after every bleach step to confirm bleaching. Thunderstorm software was used to detect localizations and previously described MatLab codes were used for drift correction and alignment [25]. The first series of unstained cells was used to determine the positions of fiducial markers which were removed from the all data sets. ThunderSTORM software was then used to remove all localizations with a total number of photons less than 1000. The YFP TIRF image was overlaid on the rendering of the data set of phospho-TCRz staining and microcluster and non-microcluster areas were hand drawn on the YFP image. These Regions of Interest were transferred to the rendered phospho-TCRz image, their locations were extracted and were used to produce masks that divided the data sets into microcluster and non-microcluster areas. We used the x-y locations of the proteins and the types of the proteins to create marked point process data for microcluster and non-microcluster areas.

## 2.2 Statistical method

### 2.2.1 Exploratory Analysis

Point pattern analysis is a method used to study events that can be localized in space or in space and time. The point pattern can have marks attached to the points in addition to the location/time information. Marked point patterns can have categorical or numerical marks to either characterize the type of the point or provide numerical measurement of some aspects of the data [6]. Point process analysis use statistical method such as exploratory analysis and parametric model-fitting to explore, analyze and model point patterns [7]. Exploratory analysis of point process includes investigating the first order effect and second order effect of the point process [3]. Point process parametric model-fitting aims to find a mathematical system to describe the underlying behavior of the point pattern, and additional covariates can be included in the model to help explain the underlying behavior [7]. Point processes now have a wide range of applications in the fields such that ecology, epidemiology, astronomy and crime research [4].

Spatial point patterns can be influenced by first order effects and second order effects. First order effects studies the intensity of the point pattern and the variation of intensity on large scale due to the change of the underlying area's structure [3] [16]. On the other hand, second order effects describe the distribution of distances between the neighbouring points that shows spatial dependency of the point pattern on small-scale locations [4].

The density of the point pattern measures the first order effects of the underlying point process. Non-parametric kernel density estimation is used to estimate the variation of intensity throughout the study area. Density plots were made using kernel smoothing method with isotropic Gaussian density and the same default bandwidth (one-eighth of the shortest side length of the enclosing rectangle) applied to each type [4].

G function is one of the methods to identify second order effects of the data by measuring the cumulative frequency distribution of nearest neighbour distances of the points [18]. G function calculates the ratio of the number of points with nearest-neighbour distances less than a given radius $r$ to the total number of points in the point process [18]. The theoretical value of the G function of a homogeneous Poisson point process with intensity $\lambda$ is $G(r) = 1 - \exp(-\lambda\pi r^2)$. If the estimated G function value is greater than the theoretical value, then the point pattern is clustered; if the estimated G function value is lower than the theoretical value, then the point pattern is dispersed. The shape of the function offers information about the way the points are distributed in a certain area. If the curve of the function increases sharply at a short range of distance, then it indicates the points are clustered in that area. On the other hand, if the curve increases slowly up to the distance where most points are spread and then increases sharply, then it means the points are evenly spread. In this way, G-function can be helpful in investigating the interaction between the points.

Cross-type J-functions are used to inspect the dependency of points between different types [17]. The J-function for the pair of points of type $i$ and $j$ , $J_{ij}(r)$, is calculated by taking the ratio of $1 - G_{ij}(r)$ to $1 - F_j(r)$, where $G_{ij}(r)$ measures the nearest distance from a point of type $i$ to type $j$ points, and $F_j(r)$ measures the nearest distance from an empty location to type $j$ points [24] [4]. The theoretical value of $J_{ij}(r)$ is 1 for all $(i, j)$ pairs. When

$i = j$, the theoretical value of $J_{ii}(r)$ is achieved when $G_{ii}(r) = F_i(r)$, meaning type $i$ point follows complete spatial randomness (CSR). When $i \neq j$, the benchmark of $J_{ij}(r)$ happens when $G_{ij}(r) = F_j(r)$, meaning the presence of type $i$ points are independent of the presence of type $j$ points. If the empirical curve of the J function is above the benchmark, then it indicates dispersion of different types of points in the point process. If the empirical curve of J function is blow the benchmark, then it indicates clustering between types. Envelopes were created using Monte Calro simulations under the null hypothesis of CSR for each type $i$ and complete spatial randomness and independence (CSRI) for each pair of type $(i, j)$ [4].

Ripley's K-function is a widely used second-order statistics that calculates the expected number of points in a distance $r$ around an arbitrary point [19] [6]. L-function is a scaled version of K-function that stabilizes the variance of K-function [12]. The empirical K-L function values can be compared with their theoretical values under the assumption of CSR to detect the characteristic of the point process such as clustering, random or dispersion in relation to CSR. These concepts of K-L function can also be extended to multitype point processes to measure the spatial relationships between two types $i$ and $j$ at each distance $r$ [4]. We used centered cross type L function to compare the empirical value with the benchmark value of 0. When the observed L function value $L_{ij}(r)$ is higher than the theoretical benchmark, the point process shows more type $j$ points are around the type $i$ point at the given distance $r$, indicating clustering. When the empirical values are below the theoretical line, the type $j$ points are less than expected around type $i$ points at distance $r$, indicating dispersion. To test if the deviation from the theoretical values are statistically significant, envelopes obtained through Monte Carlo simulations with permutation of labels were used to indicate the confidence bounds. The null hypothesis tested is CSRI for between-types and CSR for within-types. Each simulation was run 19 times to test the null hypothesis at significance level 0.05.

### 2.2.2 Modeling

After gaining understandings of the spatial relationships between the proteins through exploratory analysis, Poisson models and Gibbs models were applied to model the trend of the intensity and the pairwise interaction patterns among the points.

For Poisson model, linear, quadratic and cubic trend were considered and each combined with additive or interactive protein type term. It models the intensity function $\lambda(u, m)$ as a log linear function of location $u = (x, y)$ and the protein type m. The estimates of the optimal parameters were obtained by maximizing the likelihood using the Berman-Turner device applied to multitype patterns [5].

Gibbs models further incorporates pairwise interaction effects to the Poisson process. We considered multitype Strauss and Strauss-hard core for the interaction term to model the mark-dependent pairwise-interactions among the six types of proteins [15] [7].

Following [7], we assume a multitype point process is represented by,

$$\mathbf{x} = \{(x_1, m_1), ..., (x_n, m_n)\}, \quad x_i \in W, \quad m_i \in \mathcal{M}, \quad i = 1, 2, \ldots, n(\mathbf{x}),$$

where $x_i$ and $m_i$ are the location and mark of the $i$th point respectively, $\mathcal{M}$ is the set of possible types and $W$ is the observation window in $\mathbb{R}^2$.

Multitype Strauss models are based on the distances between the points of type $i$ and type $j$, and it assumes the pair of points will interact if it is within the interaction radius $r_{ij}$, but will have no interaction if the points are further apart than the interaction radius. A Multitype Strauss process uses a set of parameters such as intensity, interaction radius and interaction parameters to characterize the point process. The intensity parameter $\beta_m(x_i)$ represents the first-order trend for the type $m$ proteins at location $x_i$, and the interaction parameter $\gamma_{m,m'}$ represents the second-order interaction between points of type $m$ and type $m'$.

The pairwise interaction terms for the model require symmetry and are defined as:

$$c_{m,m'}(x_i, x_j) = \begin{cases} 1 & \text{if } \|x_i - x_j\| > r_{m,m'} \\ \gamma_{m,m'} & \text{if } \|x_i - x_j\| \leq r_{m,m'} \end{cases}, \tag{2.1}$$

where $r_{m,m'}$ is the interaction radius between the pair of type $m$ and type $m'$ and $\gamma_{m,m'}$ is the interaction parameter. All the terms are symmetric, which satisfy $r_{m,m'} = r_{m',m}$ and $\gamma_{m,m'} = \gamma_{m',m}$. The interaction radius matrix was estimated by observations from the G-function values of intertype and intratype points.

A Strauss-hard core process combines a hard core process with a Strauss process. A hard core process assumes no points are allowed within a radius of a hard core distance around each point. This concept can be extended to multitype process, and an additional set of hard core distances parameter is introduced to model the hard core process between each pair of types.

The pairwise interaction terms for the Strauss-hard core process are defined as :

$$c_{m,m'}(x_i, x_j) = \begin{cases} 1 & \text{if } \|x_i - x_j\| > r_{m,m'} \\ \gamma_{m,m'} & \text{if } h_{m,m'} \leq \|x_i - x_j\| \leq r_{m,m'} \\ 0 & \text{if } \|x_i - x_j\| < h_{m,m'} \end{cases}, \tag{2.2}$$

where $h_{m,m'}$ is the hard core distance and satisfies $h_{m,m'} = h_{m',m}$.

The conditional intensity for a multitype Strauss process or a multitype Strauss hard-core process at location $u$ for type $m$ is defined as:

$$\lambda((u,m)|\mathbf{x}) = \beta_m(u) \left( \prod_{i=1}^{n(\mathbf{x})} c_{m,m_i}(u, x_i) \right),$$

where $n(\mathbf{x})$ is the number of points in the point process.

Given the conditional intensity, the log pseudolikelihood of the model is defined as:

$$\log PL = \sum_{i=1}^{n(\mathbf{x})} \log \lambda((u_i, m_i)|\mathbf{x}) - \sum_{m \in \mathcal{M}} \int_W \lambda((u,m)|\mathbf{x}) du \quad .$$

Then the parameters for a multitype Gibbs model can be obtained by maximizing the pseudolikelihood using Berman-Turner device [5].

To reduce the number of parameters to estimate, we used the minimum interpoint distances as the estimated hard core distances. Strauss interaction parameter, $\gamma_{m,m'}$, quantifies between-type interactions and within-type interactions. $\gamma_{m,m'} = 1$ indicates homogeneous Poisson process and $\gamma_{m,m'} = 0$ indicates hard core process. $\gamma_{m,m'} > 1$ implies positive associations between type $m$ and type $m'$ and $\gamma_{m,m'} < 1$ indicates inhibition between the type $m$ and type $m'$.

Models were compared via their corresponding Akaike Information Criterion(AIC) scores [1]. AIC is defined as:

$$AIC = -2 \log PL + 2p,$$

where $p$ is the number of parameters in the model and $PL$ is the model's maximum pseudolikelihood. The model with higher maximum pseudolikelihood and fewer parameters will result in a lower overall AIC score, and therefore the model with a lower AIC score is considered as more optimal.

Throughout this analysis, R software (version 4.0.4) was used, and 'spatstat' package was used for the exploratory analysis and modeling of point pattern data [6].

# CHAPTER 3

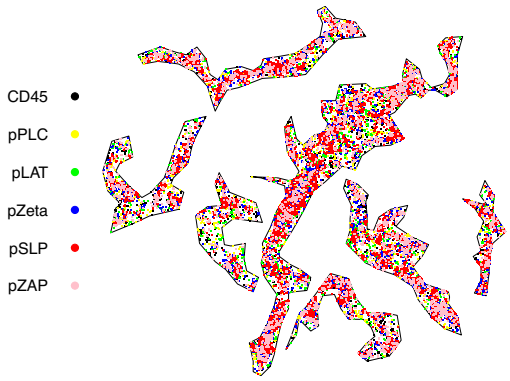
# Results

## 3.1 Exploratory Analysis

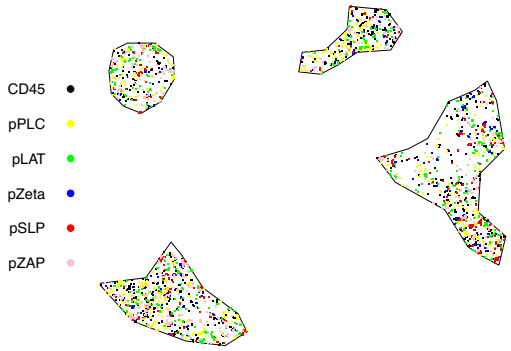
The experiment was done on three T cells which we will call Cell1, Cell2 and Cell3 respectively. In each cell, a segment from microcluster and a segment from non-microcluster were observed and analyzed. 59407 and 4921 proteins were observed from Cell1's microcluster and non-microcluster respectively. 21325 and 4244 proteins were observed from Cell2's microcluster and non-microcluster respectively. 21466 and 7279 proteins were observed from Cell3's microcluster and non-microcluster respectively.

Table 3.1 provides the estimated intensity of each protein type by assuming constant intensity in the process. In microclusters of Cell1, Cell2 and Cell3, the intensities of pZeta, pSLP and pZAP were significantly higher than the other types. While in non-microclusters, the intensities of all six types of proteins were roughly on the same scale. In all three cells, the intensities of all the proteins were higher in microcluster than in non-microcluster except for CD45 which had approximately the same intensities in both microcluster and non-microcluster. This might indicate that the number of CD45 remained approximately the same as before after the activation of T cell while other types of proteins were recruited to the signaling complexes after the activation of T cells.

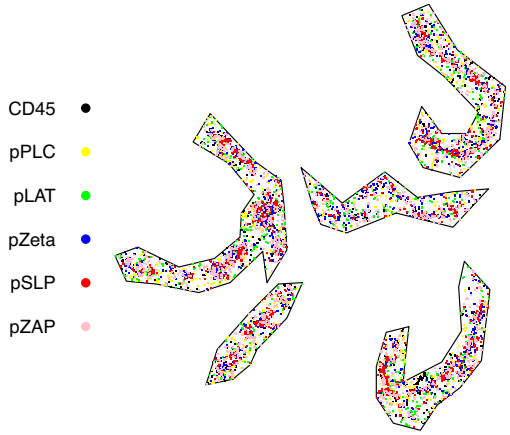
Figure 3.1 shows the spatial distribution of the point process for each segment in Cell1, Cell2 and Cell3. All the point processes had irregular observation windows, and each window consisted of several numbers of polygons. Visually, the overall number of points in microcluster areas were higher than the overall densities in non-microcluster areas as shown in Figure 3.1, indicating the proteins were more clustered in microclusters than in non-microclusters.



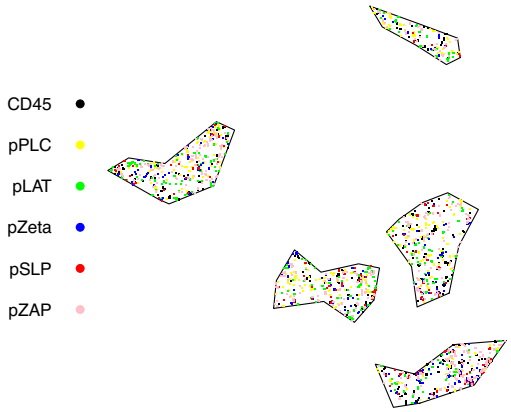

(a) Cell1 : Microcluster.



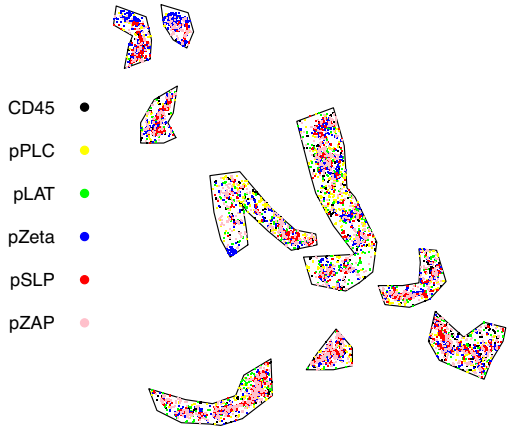

(b) Cell1 : Non-microcluster.



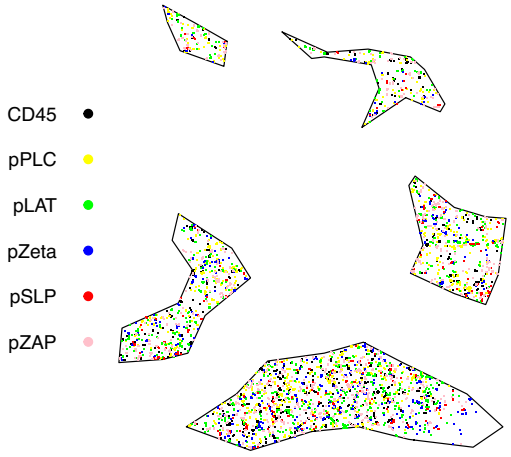

(c) Cell2 : Microcluster.




(d) Cell2 : Non-microcluster.

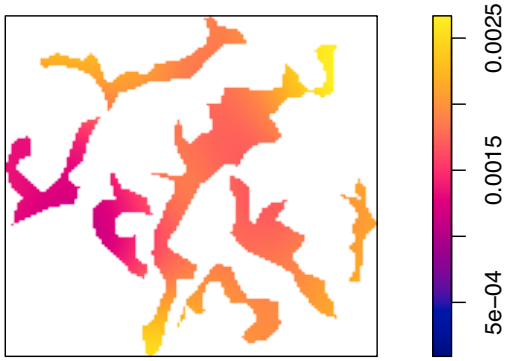



(e) Cell3 : Microcluster.

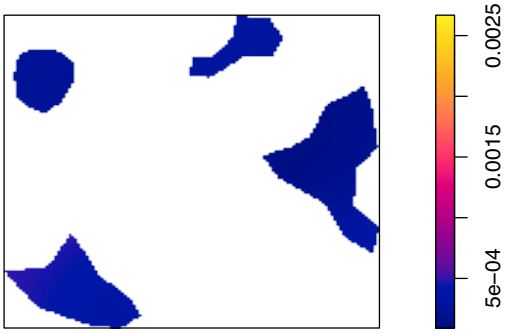



(f) Cell3 : Non-microcluster.

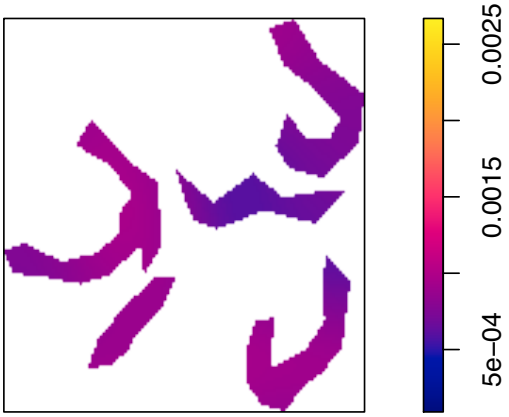
Figure 3.1: Cell1, Cell2 and Cell3's microcluster and non-microcluster point process plot



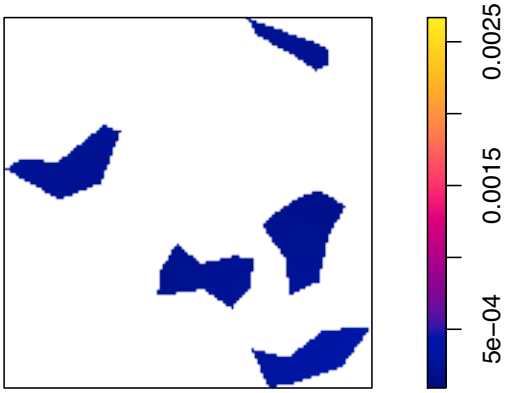

(a) Cell1 : Microcluster.



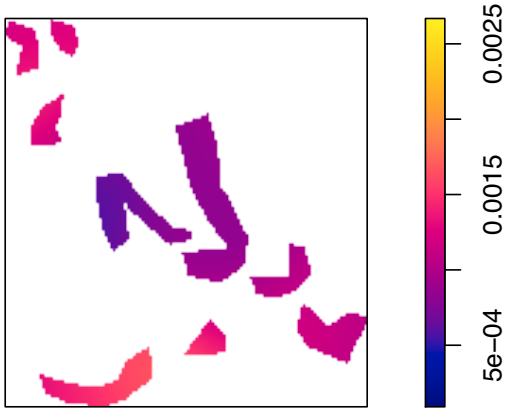

(b) Cell1 : Non-microcluster.



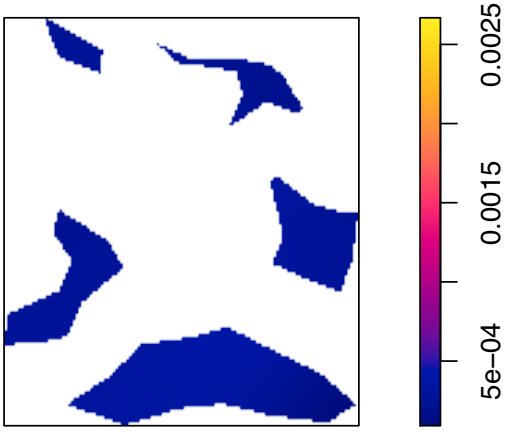

(c) Cell2 : Microcluster.




(d) Cell2 : Non-microcluster.

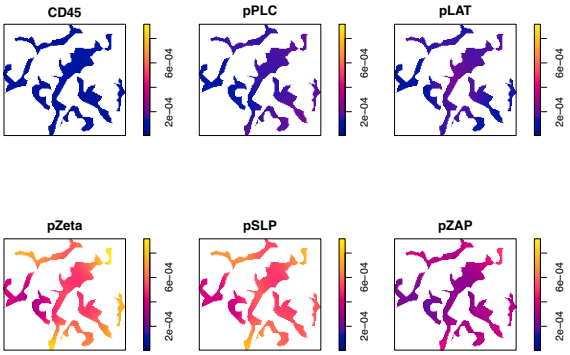



(e) Cell3 : Microcluster.

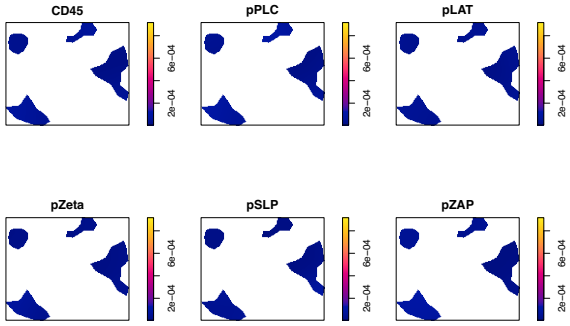



(f) Cell3 : Non-microcluster.

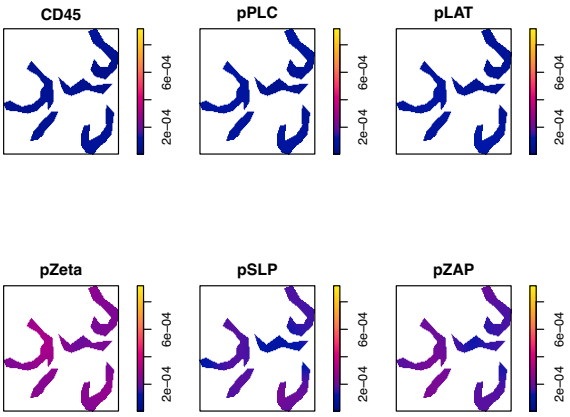
Figure 3.2: Microcluster and non-microcluster densities of Cell1, Cell2 and Cell3



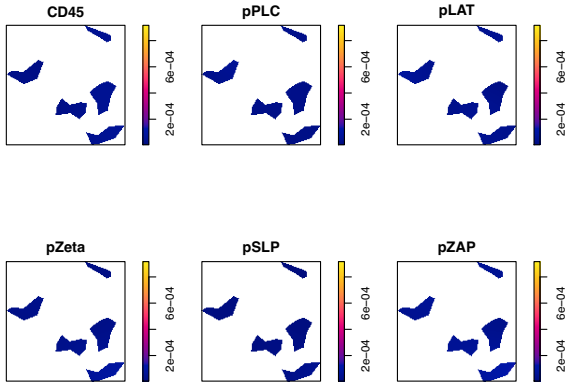

(a) Cell1 : Microcluster.



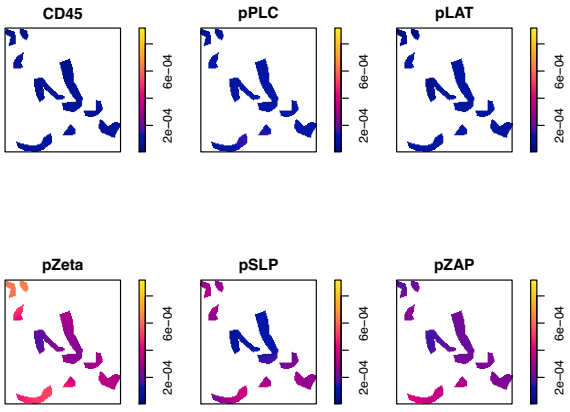

(b) Cell1 : Non-microcluster.



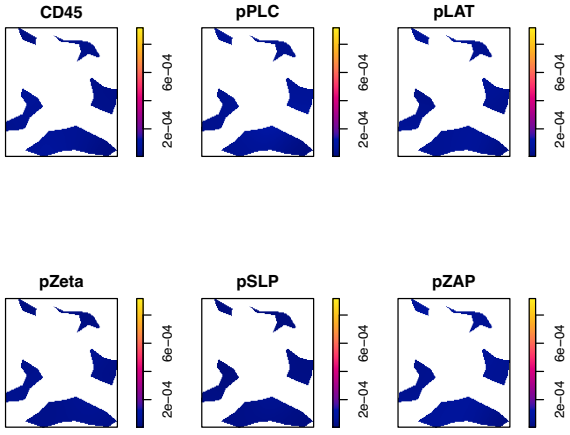

(c) Cell2 : Microcluster.




(d) Cell2 : Non-microcluster.




(e) Cell3 : Microcluster.




(f) Cell3 : Non-microcluster.

Figure 3.3: Cell1, Cell2 and Cell3's microcluster and non-microcluster densities split by protein type

| Protein type | Cell 1 | | Cell 2 | | Cell 3 | |
|---|---|---|---|---|---|---|
| | **Microcluster** | **Non-micro** | **Microcluster** | **Non-micro** | **Microcluster** | **Non-micro** |
| CD45 | 7.966432e-05 | 5.048166e-05 | 5.369623e-05 | 4.236690e-05 | 6.022332e-05 | 4.597895e-05 |
| pPLC | 1.580863e-04 | 6.155858e-05 | 7.598807e-05 | 3.377583e-05 | 1.076602e-04 | 5.760756e-05 |
| pLAT | 1.551358e-04 | 4.787889e-05 | 7.817207e-05 | 4.442640e-05 | 8.042791e-05 | 5.427964e-05 |
| pZeta | 5.809129e-04 | 4.206805e-05 | 2.503690e-04 | 4.401450e-05 | 3.551907e-04 | 3.779301e-05 |
| pSLP | 5.569049e-04 | 4.140223e-05 | 1.613523e-04 | 2.030080e-05 | 2.044860e-04 | 2.402228e-05 |
| pZAP | 3.143713e-04 | 5.447662e-05 | 1.834182e-04 | 6.484489e-05 | 2.396244e-04 | 5.875513e-05 |

Table 3.1: Intensity of each type of protein in each cell's microcluster and non-microcluster segment.

Figure 3.2 depicts the corresponding overall density distribution for each segment. In three microclusters, the maximum intensity was approximately $2.6 \times 10^{-3}$ and the minimum intensity was $6 \times 10^{-4}$, whereas, in three non-microclusters, the maximum and minimum intensities were $5 \times 10^{-4}$ and $1 \times 10^{-4}$ respectively. This showed the range of intensities in microclusters was generally wider than the range of intensities in non-microclusters. In addition, in each subplot, the intensities varied across different polygons since some polygons had high intensities indicated by the color yellow and some polygons had low intensities indicated by dark blue color. This showed the intensity was inhomogeneous within each segment.

Figure 3.3 illustrates the density distributions of the point process split by protein type in each segment of Cell1, Cell2 and Cell3. Generally, the density distributions varied by type, but some types of proteins shared similar density distributions. For example, in Cell1's microcluster, pZeta, pSLP and pZAP had similar density distributions, indicating there might be spatial associations between these proteins since a high density of one of these proteins indicated a high density of other types. The range of the intensity was approximately on the same scale in both microclusters and non-microclusters for CD45, but for other proteins, the range of intensities was wider in microclusters compared to the range of intensities in non-microclusters. The increased intensity of the proteins in microclusters might be the re-

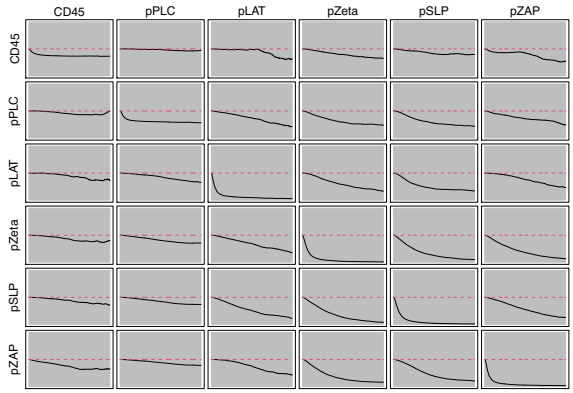
sult of the recruited signaling proteins into the signaling complexes following the activation of the T cells.

Figure 3.4 shows the result of cross-type J-function with Monte Carlo simulation envelopes based on CSR(within-type) or CSRI(between-type). In microclusters, the proteins in some columns showed clustered patterns around points of its same type or around other types of points. But with the empirical curves inside the randomization envelopes, the clustering patterns were not statistically significant. In the last three columns of Figure 3.4a, Figure 3.4c and Figure 3.4e, pZeta, pSLP and pZAP have shown clustering patterns around points of its same type or around other types of proteins. The only exception is that when pZeta, pSLP and pZAP were around points of CD45 type, the observed J function followed the theoretical curve closely. In the first column of each microcluster's cross-type J function plot, we observed protein type CD45 showed no significant clustering or regular pattern either around points of its same type or around other types of proteins.

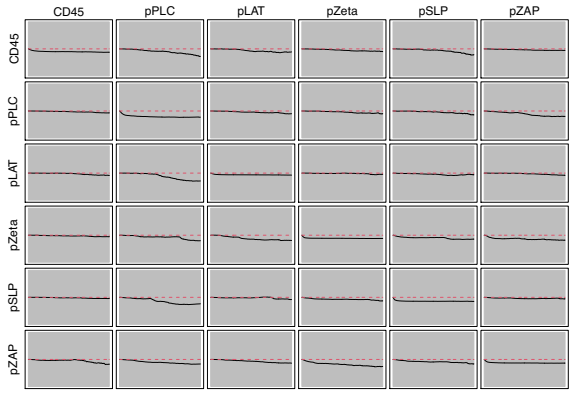
In non-microclusters, the empirical distribution followed the theoretical Poisson distribution closely and there was no evidence of departure from the null hypothesis of CSR or CSRI.

Figure 3.5 illustrates the cross-type L-function with envelopes obtained from the Monte Carlo random labeling test of the null hypothesis. There were some similar patterns shown among the microclusters, as well as some similar patterns shown among the non-microclusters of the three cells.

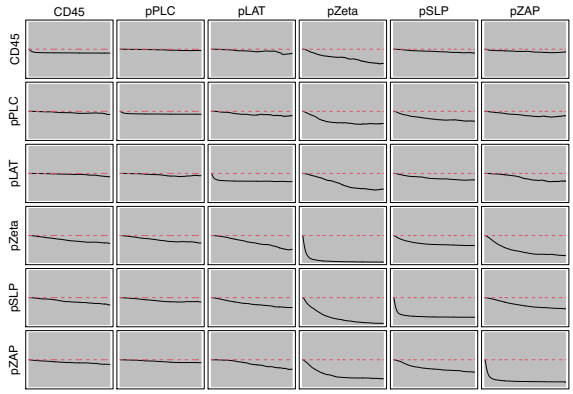
In each cell's microcluster segment, each individual panel showed similar patterns with other individual panels within the same column. The columns of pZeta, pSLP and pZAP from Cell1, Cell2 and Cell3 showed that there was significant clustering at most scales in each panel with the exception that, CD45 and pZAP pair in Cell1 (Figure 3.5a), which was first clustered up to 1000nm, then the observed L-function entered the grey envelope and left it again showing dispersion. The columns of CD45, pPLC and pLAT from the three microclusters showed more complicated behavior. In the microcluster of Cell1, the columns of pLAT and pPLC showed significant clustering at almost all distances, except



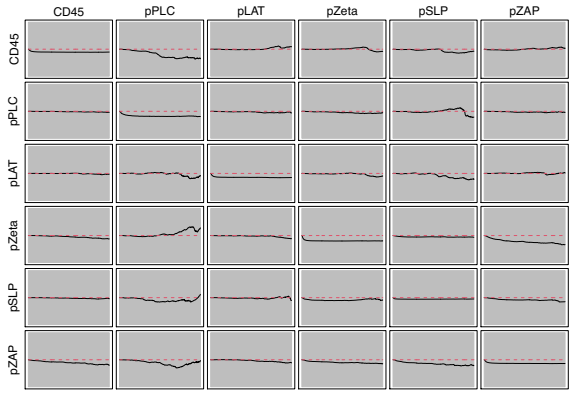

(a) Cell1 : Microcluster.



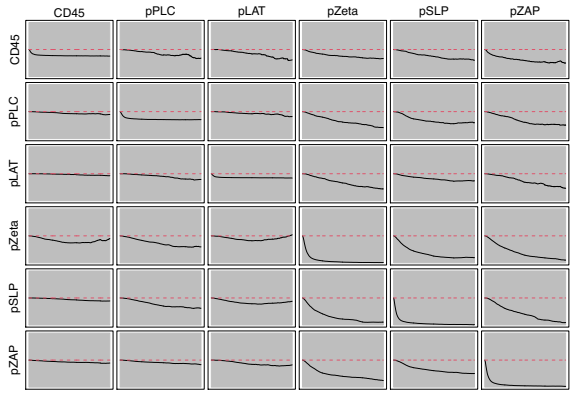

(b) Cell1 : Non-microcluster.



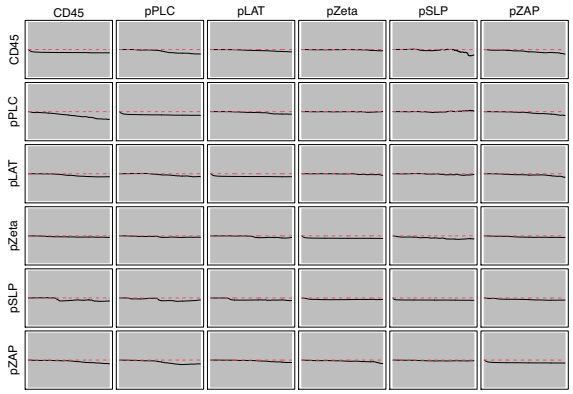

(c) Cell2 : Microcluster.




(d) Cell2 : Non-microcluster.

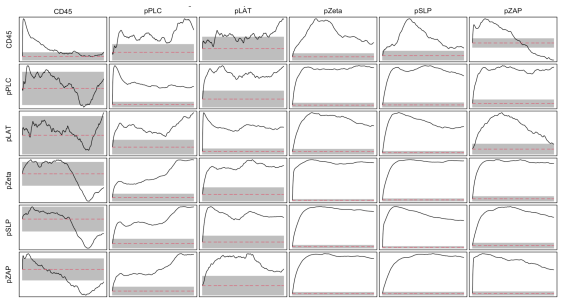



(e) Cell3 : Microcluster.

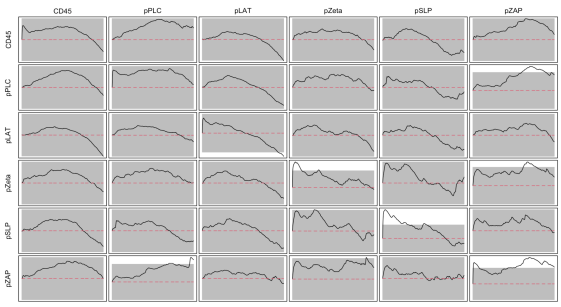



(f) Cell3 : Non-microcluster.

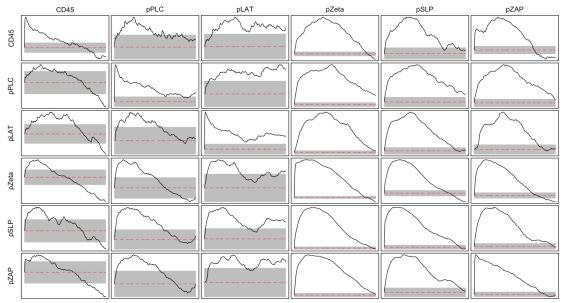
Figure 3.4: Cross-type J-function and global envelopes of 19 simulations of CSR or CSRI, for each combination of types for the microcluster and non-microcluster in Cell1, Cell2 and Cell3.



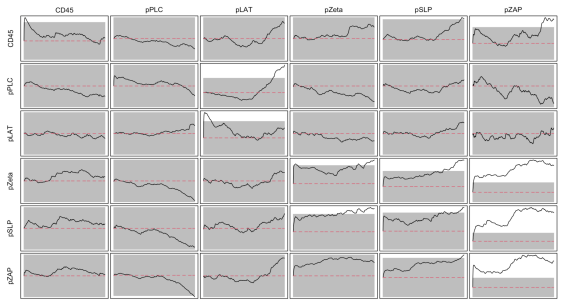

(a) Cell1 : Microcluster.



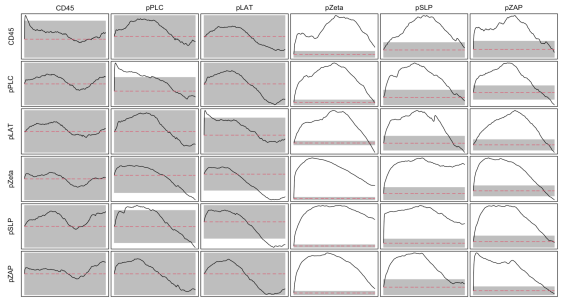

(b) Cell1 : Non-microcluster.



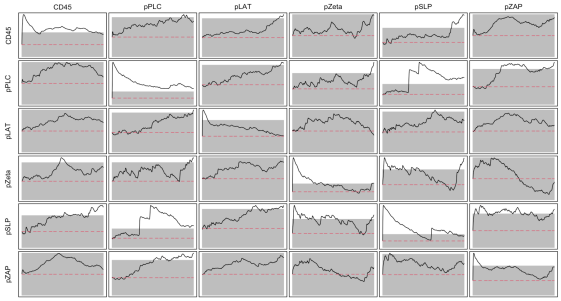

(c) Cell2 : Microcluster.




(d) Cell2 : Non-microcluster.




(e) Cell3 : Microcluster.




(f) Cell3 : Non-microcluster.

Figure 3.5: Centred L-function and global envelopes of 19 simulations of CSR, for each pair of types assuming constant intensity of each type for point patterns in the microcluster and the non-microcluster in Cell1, Cell2 and Cell3. Solid lines show empirical estimate. Grey shading represents the envelope of the summary functions based on simulations of a uniform Poisson process with the same estimated intensity.

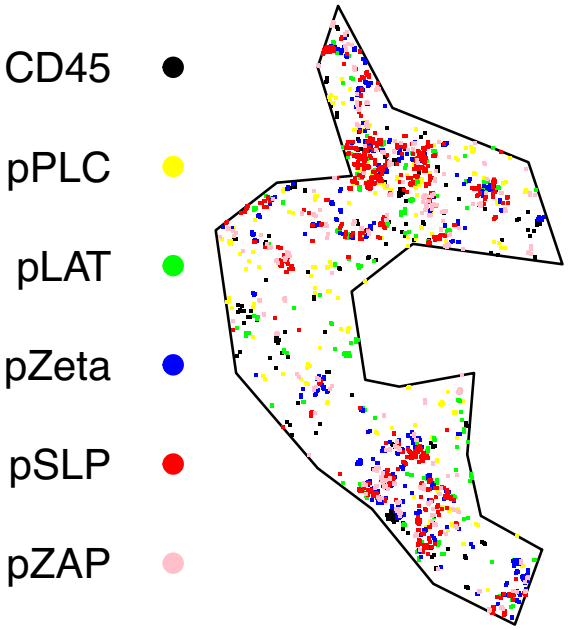
for the pLAT and CD45 pair. But in the microcluster of Cell2, most pairs in CD45, pPLC and pLAT column showed clustering up to some distances but in some ranges were within the grey envelopes. In Cell3's microcluster segment, most pairs showed the patterns that the empirical centred L function increased at first and then gradually dropped below the theoretical line, but most pairs' L-function values were within the boundary of the envelope.

In each cell's non-microcluster segment, most pairs did not show significant clustering or dispersion. In Cell1, the pZAP and pPLC pair, the pZAP and pZeta pair, and the pPLC and pZAP pair were clustered at some further distances. The Centred L-function for type pLAT, pZeta and pSLP showed there was some dispersion or clustering for a very short distance. In Cell2, pZAP paired with any of pZeta, pSLP and pZAP showed significant clustering at most distances. For other pairs, no significant clustering or dispersion was shown at most distances. In Cell3, the diagonal panels showed that the within protein types were all clustered at first and then gradually dropped. For other pairs, no significant clustering or dispersion at most distances except for the pSLP and CD45 pair, the pSLP and pPLC pair, and the pSLP and pZeta pair where it all showed significant clustering at longer distances.

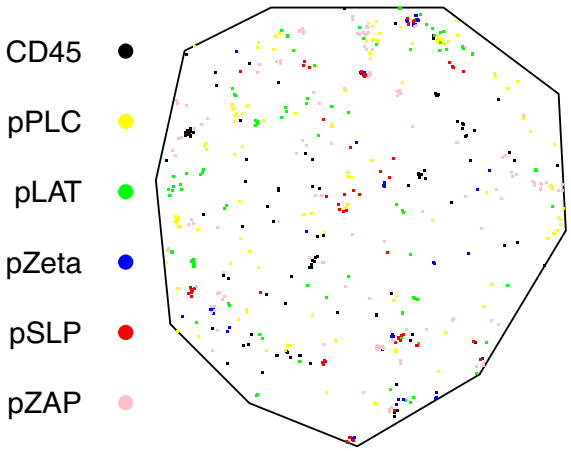
## 3.2 Model fitting

To reduce the computational load, one polygon was each selected from Cell1's microcluster and non-microcluster for model-fitting. The subset selected from Cell1's microcluster is shown in Figure 3.6a and the subset selected from Cell1's non-microcluster is shown in Figure 3.6b. The corresponding density plots split by the type of protein are shown in Figure 3.7a and Figure 3.7b respectively.

Figure 3.8 shows the cross-type J-function of the subsets. In both microcluster and non-microcluster subset, the empirical lines did not significantly deviate from the theoretical lines, indicating no significant pairwise interactions presented in the subsets. While the cross-type J-function of the subset of the non-microcluster exhibited similar pattern with the whole dataset (Cell1's non-microcluster) as shown in Figure 3.4, the cross-type J-function of the




(a) The Subset of Cell1 : Microcluster.



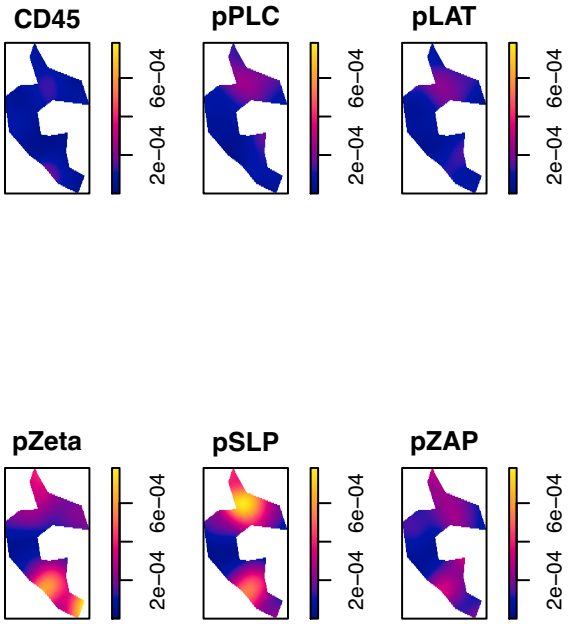

(b) The Subset of Cell1 : Non-microcluster.

Figure 3.6: The point process plots for the subset of Cell1's microcluster and the subset of Cell1's non-microcluster.

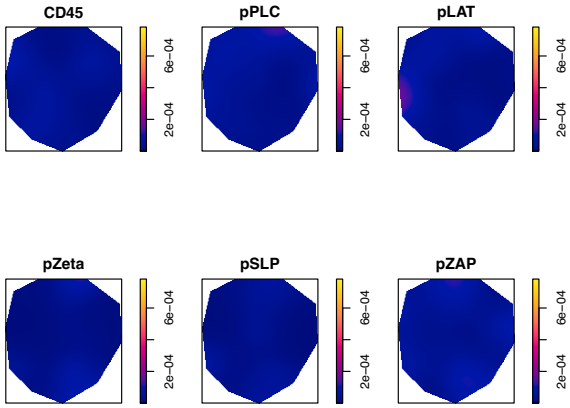
subset of the microcluster was very different compared to the cross-type J-function of the whole dataset (Cell1's microcluster). Unlike in the whole dataset, the observed crosstype J-function values in the microcluster subset was not below the theoretical values. The difference between the subset and the whole dataset could indicate that the spatial distributions of the point process were not homogeneous across different polygons within the cell.

Figure 3.9 illustrated the cross-type L-function of the subsets. In microcluster, most pairs showed significant clustering at most distances, except for pZeta and pPLC pair. In non-microcluster, all the within-type proteins showed significant clustering, and a few between-type pairs showed significant clustering or inhibition at some distances. Comparing the cross-type L-function of the subsets and that of the corresponding whole dataset, we observed more pairs showed significant clustering in both the microcluster and the non-microcluster subsets.

G functions were used to exam the distribution of the nearest neighbor distances and

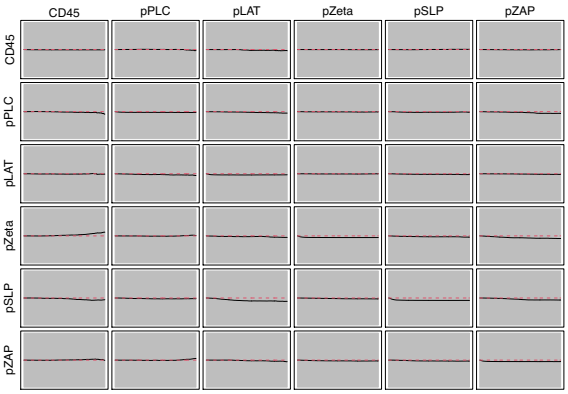



(a) The Subset of Cell1 : Microcluster.

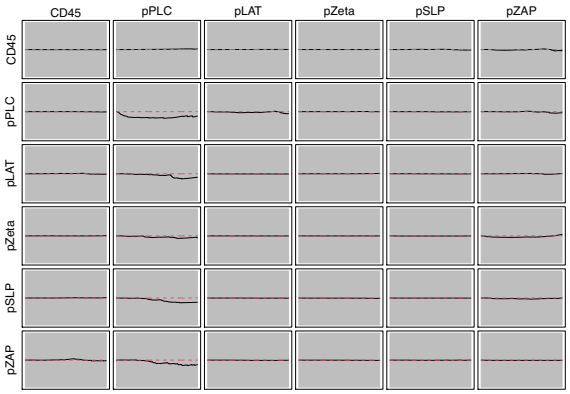



(b) The Subset of Cell1 : Non-microcluster.

Figure 3.7: Density plot split by protein type for the subset of Cell1's microcluster and the subset of Cell1's non-microcluster

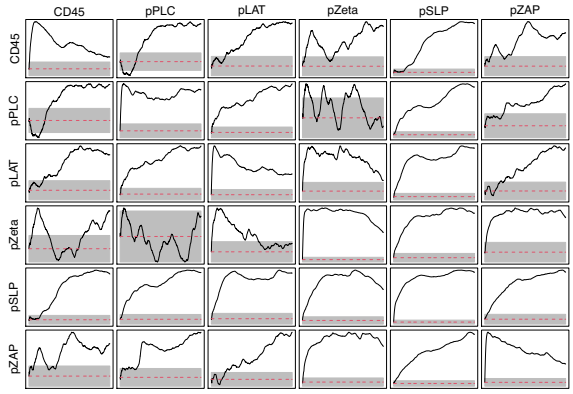



(a) The Subset of Cell1 : Microcluster.

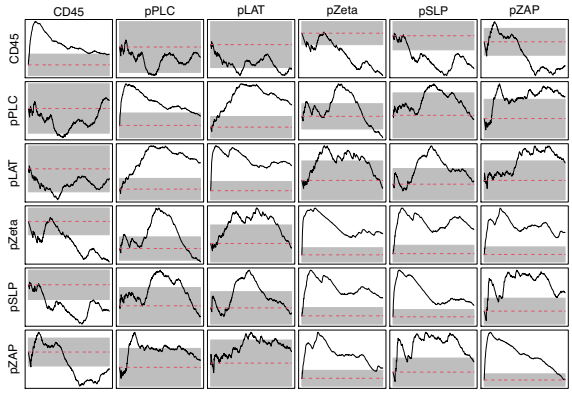



(b) The Subset of Cell1 : Non-microcluster.

Figure 3.8: Cross-type J-function and global envelopes of 19 simulations of CSR or CSRI, for each combination of types for the subset of microcluster and non-microcluster in Cell1




(a) The Subset of Cell1 : Microcluster.




(b) The Subset of Cell1 : Non-microcluster.

Figure 3.9: Centred L-function and global envelopes of 19 simulations of CSR, for each pair of types assuming constant intensity of each type for point patterns in the subset of microcluster and the non-microcluster in Cell1. Solid lines show empirical estimate. Grey shading represents the envelope of the summary functions based on simulations of a uniform Poisson process with the same estimated intensity

the interaction radius of the points. Figure 3.10 and Figure 3.11 showed the cumulative distribution of the nearest neighbor distance from each point of type $i$ to type $j$ points for the subset of microcluster and non-microcluster data respectively. The plots in the diagonal panels of 3.10 and Figure 3.11 show the empirical curves for the nearest neighbor distances were all above the curves of the Poisson process, indicating clustering within the same type of proteins. In addition, all the estimated curves increased rapidly at some shorter distances. These showed proteins of the same types were more clustered within short scales. From the cross-type G functions between different types of proteins, we observed the estimated G functions were either slightly below or closely follow the theoretical Poisson curve for most regions in both plots. This indicated the proteins of different types might be either independent or slightly dispersed from each other.

We fitted various inhomogeneous Poisson models and Multitype Strauss/ hard-core models. From the AIC scores of Poisson models and Strauss models applied to these multitype point processes (see Table 3.2), Strauss models generally had better fits than Poisson models.

Among the different Poisson models, the models with cubic trend and interaction with mark types had the best fit in both microcluster and non-microcluster. Among the different Strauss models, Strauss-hard-core model fit better than Strauss model. In particular, Strauss-hard-core with cubic trend term had the lowest AIC. Considering the Strauss-hard-core with quadratic trend term had only slightly higher AIC but fewer parameters, we selected the Strauss-hard-core with quadratic trend model for both microcluster and non-microcluster to proceed with the analysis.

| **Model type** | **Formula** | **AIC** | |
|---|---|---|---|
| | | **Microcluster** | **Non-micro** |
| Additional linear | ppm(X∼polynom(x,y,1)+marks) | 54988.16 | 13690.07 |
| Linear with interaction | ppm(X∼polynom(x,y,1)*marks) | 54844.93 | 13655.07 |
| Additional quadratic | ppm(X∼polynom(x,y,2)+marks) | 54601.50 | 13662.44 |
| Quadratic with interaction | ppm(X∼polynom(x,y,2)*marks) | 54325.86 | 13575.07 |
| Additional cubic | ppm(X∼polynom(x,y,3)+marks) | 54534.88 | 13611.21 |
| Cubic with interaction | ppm(X∼polynom(x,y,3)*marks) | 54050.94 | 13443.02 |
| Strauss with quadratic trend | ppm(X∼polynom(x,y,2), MultiStrauss()) | 25725.44 | 4216.64 |
| Strauss-hard-core with quadratic trend | ppm(X∼polynom(x,y,2), MultiStraussHard()) | 25006.05 | 4181.15 |
| Strauss-hard-core with cubic trend | ppm(X∼polynom(x,y,3), MultiStraussHard()) | 24988.47 | 4172.84 |

Table 3.2: AIC for different models

| | CD45 | pPLC | pLAT | pZeta | pSLP | pZAP |
|---|---|---|---|---|---|---|
| CD45 | 0.2768 | **-0.0383 ± 0.0074** | -0.0011 ± 0.0071 | -0.0152 ± 0.0082 | **0.0222 ± 0.0059** | **-0.0250 ± 0.0084** |
| pPLC | | **0.1795 ± 0.0123** | **-0.0206 ± 0.0079** | 0.0026 ± 0.0037 | **0.0212 ± 0.0013** | -0.0071± 0.0056 |
| pLAT | | | 0.5392 | -0.0031 ± 0.0036 | **0.0271 ± 0.0040** | **-0.0411 ± 0.0047** |
| pZeta | | | | 0.2114 | **0.0379 ± 0.0011** | **0.0361 ± 0.0015** |
| pSLP | | | | | 0.2890 | **0.0383 ± 0.0015** |
| pZAP | | | | | | 0.4344 |

Table 3.3: The estimated interaction parameters and the corresponding standard errors of Strauss-hard-core model fitted for microcluster data, listed in the form of mean ± standard error. 11 out of 21 parameters were statistically significant, which were marked in bold typeface. 5 of the coefficients did not have a standard error due to singularity.

| | CD45 | pPLC | pLAT | pZeta | pSLP | pZAP |
|---|---|---|---|---|---|---|
| CD45 | **0.3048 ± 0.0304** | 0.0114 ± 0.0447 | -0.0189 ± 0.0334 | -0.0214 ± 0.0273 | -0.0620 ± 0.0509 | 0.0496 ± 0.0351 |
| pPLC | | 0.1170 ± 0.1047 | 0.0357 ± 0.0789 | -0.0823 ± 0.0519 | 0.0755 ± 0.0813 | 0.0032± 0.0827 |
| pLAT | | | 0.2700 ± 0.2136 | 0.0941 ± 0.1096 | -0.0669 ± 0.1068 | -0.0978 ± 0.0962 |
| pZeta | | | | **0.3786 ± 0.1058** | 0.0623 ± 0.0921 | 0.0162 ± 0.0586 |
| pSLP | | | | | 0.4903 ± 0.3788 | 0.0052 ± 0.0767 |
| pZAP | | | | | | **0.1912± 0.0531** |

Table 3.4: The estimated interaction parameters and the corresponding standard errors of Strauss-hard-core model fitted for non-microcluster data, listed in the form of mean ± standard error. 3 out of 21 parameters were statistically significant, which were marked in bold typeface.

Table 3.3 showed the output of the log quadratic model with a multitype Strauss hard-core process of six types of proteins in the microcluster subset with AIC of 25006.05, and 3.4 showed the corresponding output for the non-microcluster subset with AIC of 4181.15. The tables showed the logarithms of the estimated interaction parameters $\gamma$ and the corresponding standard errors in the fitted models. In the model fitted for microcluster, the standard errors for 11 interaction parameters were relatively small, indicating statistical significance. In the model fitted for non-microcluster, 3 out of 21 interaction parameters were significant. This showed more pairwise interactions in microcluster than in non-microcluster were detected when modeled with Strauss hard-core.

One of the requirements for fitting Strauss model is to supply the estimation of the interaction radii. There is no analytical way to calculate the optimum values of the interaction radii [4], and a common method is to observe G function values to determine the pairwise interaction radii. The following matrix was obtained from observation of cross-type G function and was used as the interaction radii to fit the Multitype Strauss/hard-core model for the microcluster subset.

$$r_{ij} = \begin{bmatrix} 60 & 150 & 140 & 120 & 120 & 140 \\ 150 & 60 & 120 & 120 & 120 & 100 \\ 140 & 120 & 30 & 140 & 120 & 120 \\ 120 & 120 & 140 & 15 & 100 & 100 \\ 120 & 120 & 120 & 100 & 20 & 80 \\ 140 & 100 & 120 & 100 & 80 & 25 \end{bmatrix}$$

The following matrix was used as the interaction radii to fit the Multitype Strauss hard-core model for the non-microcluster subset.

$$r_{ij} = \begin{bmatrix} 120 & 200 & 250 & 300 & 250 & 150 \\ 200 & 140 & 200 & 250 & 200 & 150 \\ 250 & 200 & 100 & 200 & 200 & 150 \\ 300 & 250 & 200 & 40 & 150 & 150 \\ 250 & 200 & 200 & 150 & 50 & 150 \\ 150 & 150 & 150 & 150 & 150 & 120 \end{bmatrix}$$

The hard-core distance matrix is required for fitting Strauss hard-core model. The hard-core distance matrix was estimated by using the minimum nearest-neighbor distance between the corresponding types of points.

The hard-core distance matrix used to fit the Multitype Strauss hard-core model for the microcluster subset is shown below.

$$h_{ij} = \begin{bmatrix} 0.00 & 1.22 & 2.55 & 0.86 & 1.99 & 0.45 \\ 1.22 & 0.00 & 1.56 & 0.76 & 0.32 & 2.52 \\ 2.55 & 1.56 & 0.00 & 0.70 & 1.51 & 3.53 \\ 0.86 & 0.76 & 0.70 & 0.00 & 0.15 & 1.72 \\ 1.99 & 0.32 & 1.51 & 0.15 & 0.00 & 0.74 \\ 0.45 & 2.52 & 3.53 & 1.72 & 0.74 & 0.00 \end{bmatrix}$$

The hard-core distance matrix used to fit the Multitype Strauss hard-core model for the

non-microcluster subset is shown below.

$$h_{ij} = \begin{bmatrix} 0.00 & 6.79 & 4.21 & 24.53 & 6.55 & 13.85 \\ 6.79 & 0.00 & 6.05 & 7.06 & 3.30 & 8.94 \\ 4.21 & 6.05 & 0.00 & 9.57 & 6.21 & 1.64 \\ 24.53 & 7.06 & 9.57 & 0.00 & 1.10 & 1.50 \\ 6.55 & 3.30 & 6.21 & 1.10 & 0.00 & 11.43 \\ 13.85 & 8.94 & 1.64 & 1.50 & 11.43 & 0.00 \end{bmatrix}$$

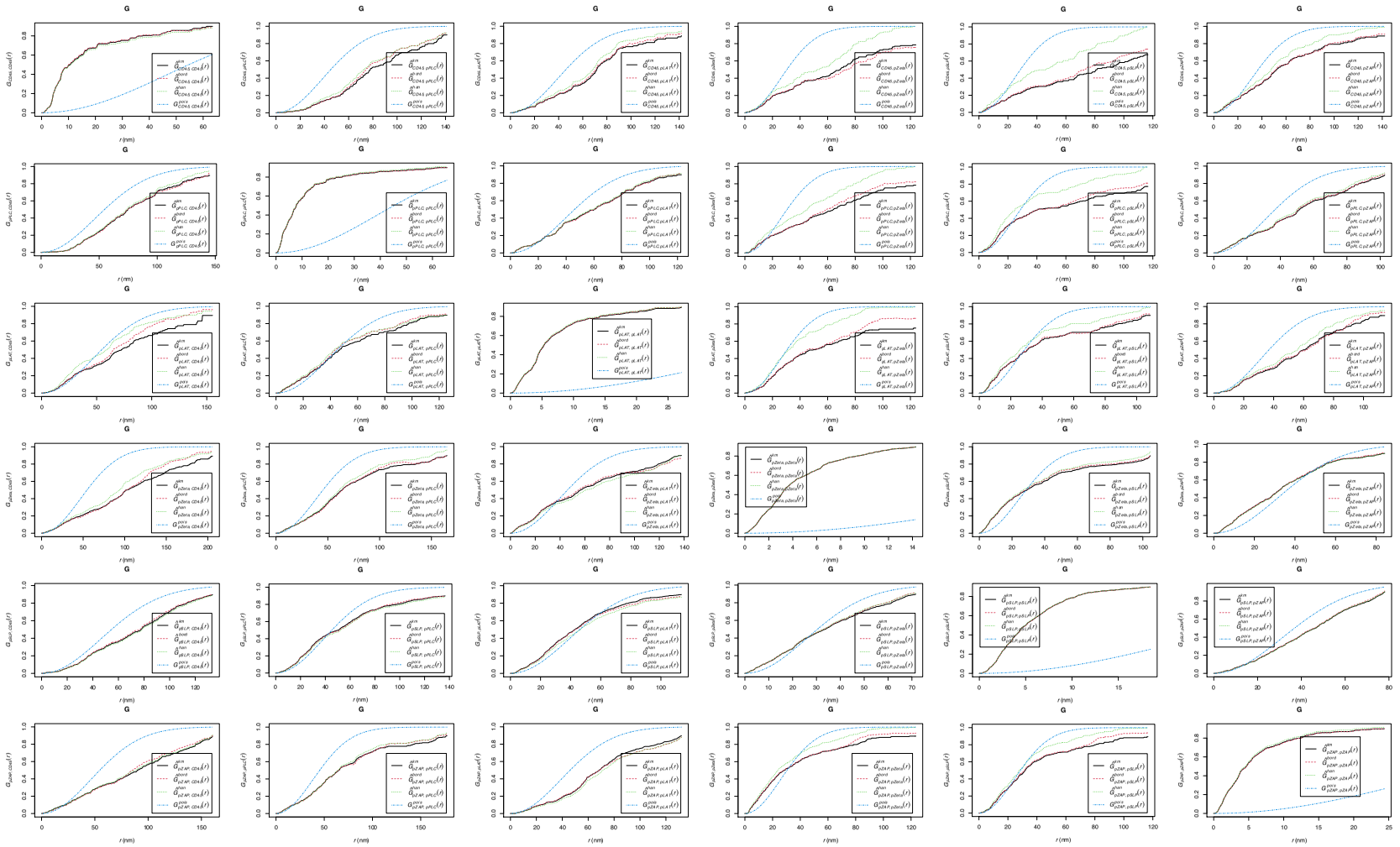
The values of the hard-core radii indicated that for the microcluster subset, all the points had their nearest-neighbor within 4nm; and for the non-microcluster subset, all the points had their nearest-neighbor within 25nm. In both matrices, the nearest neighbor of the same type of points was 0. These may suggest that the points were more clustered in microcluster than in non-microcluster and points of the same types were more clustered compared to the different types of points.

The plots of the fitted interactions were shown in Figure 3.12. Each panel showed the value of the pairwise interaction parameter between the given types. The estimates of within-type interaction parameters on the diagonal panels of both plots were all above 1, suggesting positive associations within the same type of proteins. The off-diagonal panels in both plots showed the interactions between different types of proteins were not very strong. More specifically, in both plots, the between-type dependencies among pZeta, pSLP and pZAP were all slightly above 1, indicating weak positive associations between these types. In microcluster, with the exception of the pairs of pSLP with other types of proteins, all the rest of the pairs have parameters that were slightly below 1. However, since all the off-diagonal between-type parameters in both plots were very close to unity, the attraction or inhibition between different types of proteins was not very strong.

The fitted trend of each type of protein in microcluster and non-microcluster subset are shown in Figure 3.13 and Figure 3.15 respectively, demonstrating the predicted trend of the intensity for each type superimposed by the observed points. The fitted trend in microcluster captured the clustering on the top of the polygon but did not capture the clustering on

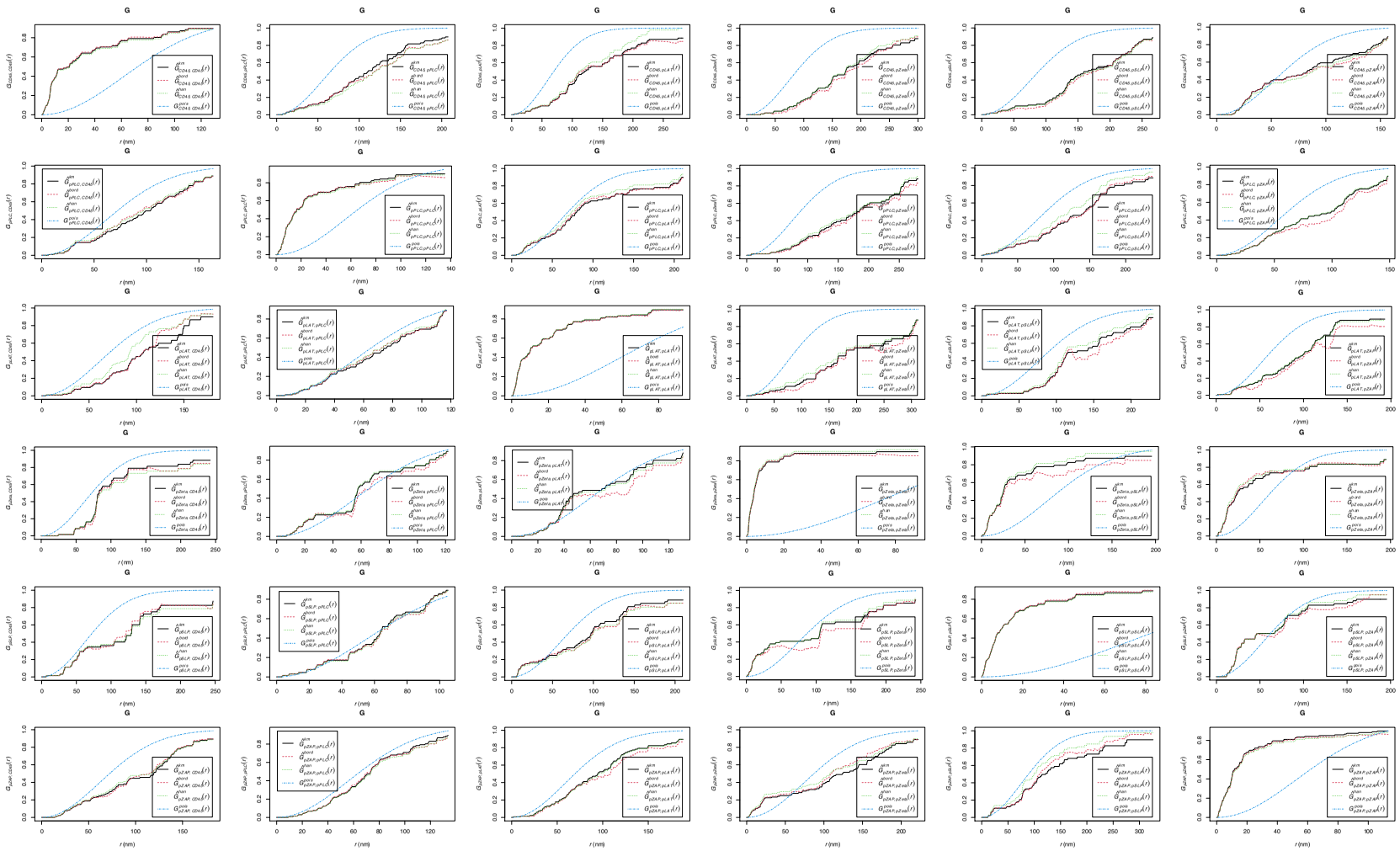
the bottom of the polygon (see Figure 3.7a). The fitted trend in non-microcluster predicted higher intensities in the middle part and gradually lowers outwards, but the observed intensity (see Figure 3.7b) for some types such as pPLC or pZAP were more clustered near the outer edge.

The fitted conditional intensity of each type of protein in microcluster and non-microcluster subset are shown in Figure 3.14 and Figure 3.16 respectively. In microcluster, the model predicted low intensity for most protein types except for pZeta, where it predicted higher conditional intensities near the top of the polygon. In non-microcluster, the model predicts higher conditionalintensities in some areas of CD45, pPLC, pLAT and pZAP.



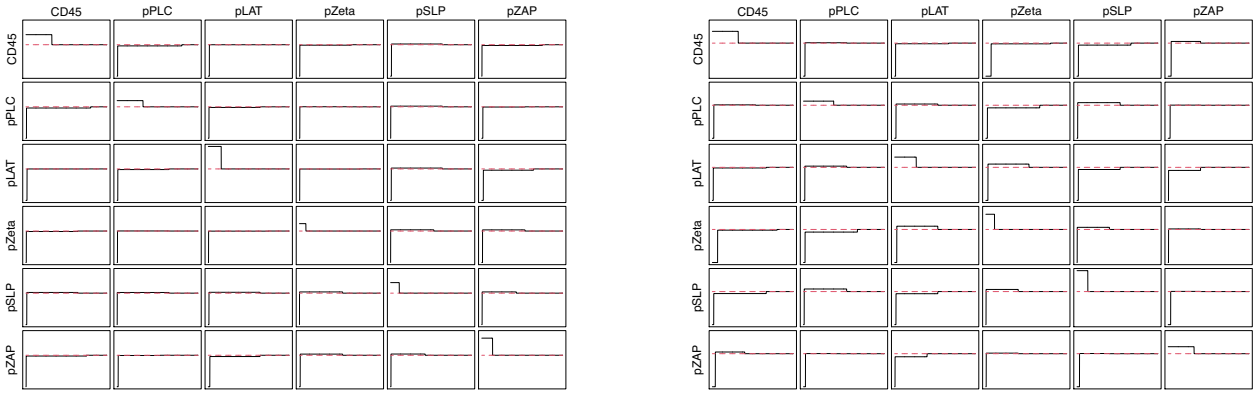
(a) (b) (c) (d) (e) (f)

Figure 3.10: Crosstype G functions $G_{ij}$ for each pair of protein types in i-th row and j-th column for the subset of Cell1's microcluster data. The rows and columns are organized in the following order (a) CD45; (b) pPLC; (c) pLAT; (d) pZeta; (e) pSLP; (f) pZAP. In each subplot, the black line represents the Kaplan-Meier estimate G function, the red dashed line represents the border corrected (reduced sample) estimate G function, the green dotted line represents the Hanisch-style estimator of the G function, and the blue line represents the theoretical Poisson value.



(a) (b) (c) (d) (e) (f)

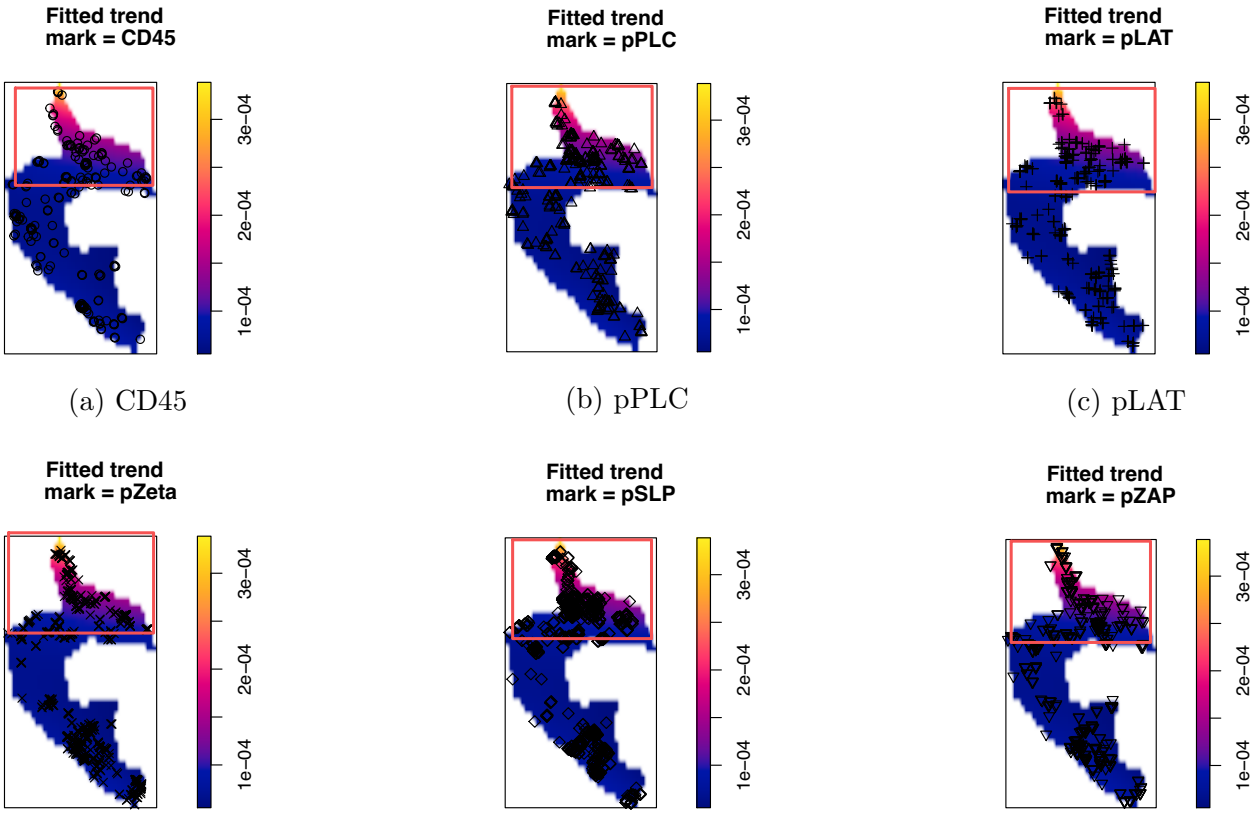
Figure 3.11: Crosstype G functions $G_{ij}$ for each pair of protein types in i-th row and j-th column for the subset of Cell1's non-microcluster data. The rows and columns are organized in the following order (a) CD45; (b) pPLC; (c) pLAT; (d) pZeta; (e) pSLP; (f) pZAP. In each subplot, the black line represents the Kaplan-Meier estimate G function, the red dashed line represents the border corrected (reduced sample) estimate G function, the green dotted line represents the Hanisch-style estimator of the G function, and the blue line represents the theoretical Poisson value.




(a) The Subset of Cell1's Microcluster.

(b) The Subset of Cell1's Non-microcluster.

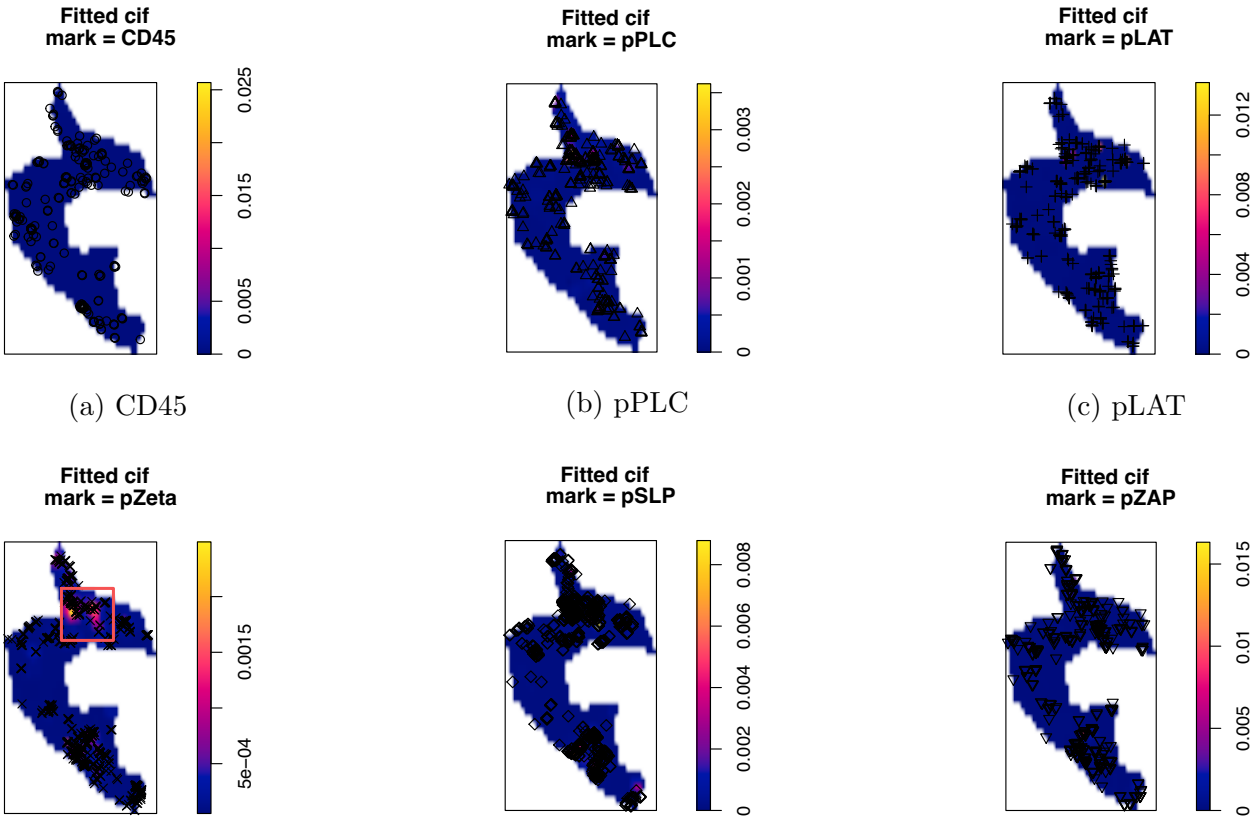
Figure 3.12: Fitted interaction parameter plot for the subset of microcluster and non-microcluster in Cell1. The black line indicates the fitted interaction parameter and the red line indicates value 1.




(a) CD45 (b) pPLC (c) pLAT

(d) pZeta (e) pSLP (f) pZAP

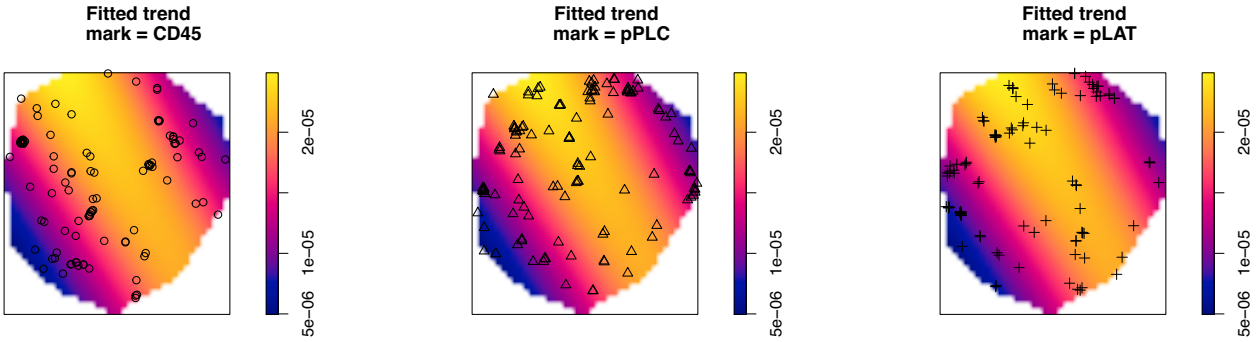
Figure 3.13: Fitted trend for each type of protein for the subset of Cell1's microcluster. In each subplot, the region in the rectangle box shows the area where the model predicted higher intensity for the trend.



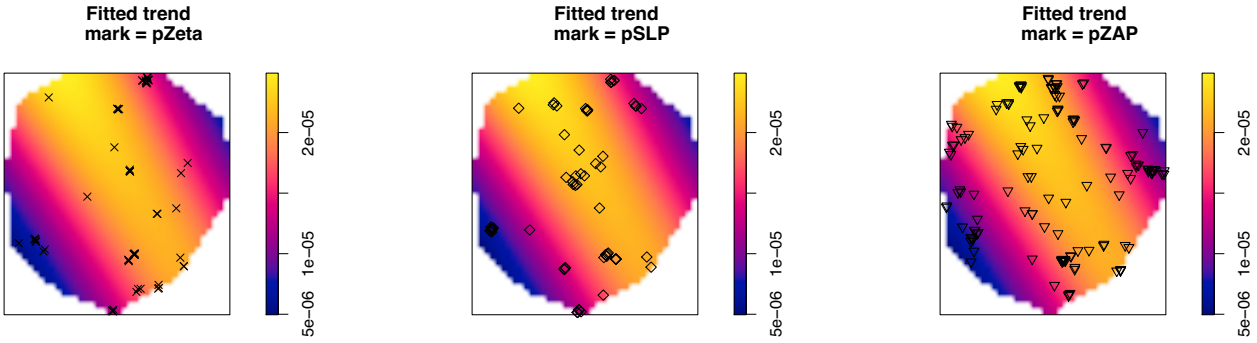

(a) CD45 (b) pPLC (c) pLAT

(d) pZeta (e) pSLP (f) pZAP

Figure 3.14: Fitted conditional intensity for each type of protein for the subset of Cell1's microcluster. (d) The highlighted part shows where the model predicted higher conditional intensity.

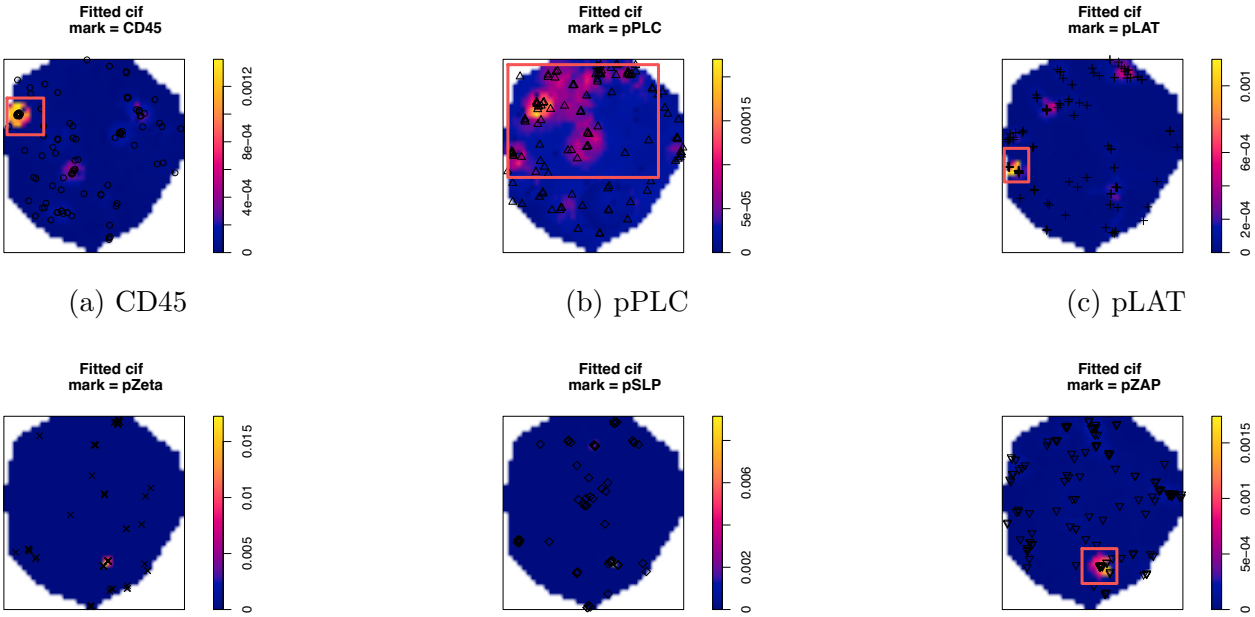



(a) CD45 (b) pPLC (c) pLAT




(d) pZeta (e) pSLP (f) pZAP

Figure 3.15: Fitted trend for each type of protein for the subset of Cell1's non-microcluster.




(a) CD45 (b) pPLC (c) pLAT

(d) pZeta (e) pSLP (f) pZAP

Figure 3.16: Fitted conditional intensity for each type of protein for the subset of Cell1's non-microcluster. The highlighted boxes show where the model predicted higher conditional intensity.

# CHAPTER 4

# Discussions

The analysis of multitype point patterns of the proteins can help the understanding of the first-order effect of the point process as well as the interactions patterns within or between the different types of proteins.

In the exploratory analysis, we used cross-type L, J function to characterize the point processes of Cell1, Cell2 and Cell3's microcluster and non-microcluster and compare the similarities and the differences between microclusters and non-microclusters. First of all, the three cells' L, J functions of the microcluster data were very distinguishable from that of the non-microcluster data. In all three cells' microclusters, more pairs had their observed values deviated from the theoretical benchmark in J-function. But in all three cells' non-microclusters, the observed values followed the theoretical J-function values closely. For all three cells' L-function in microcluster, the panels in the pZeta, pSLP and pZAP exhibited significant clustering around points of its same type or around other types of points. But in all the three cells' L-function in non-microcluster, most panels were within the envelope showing no significant clustering or dispersion. Hence, comparing three cells' microcluster to the non-microcluster, we observed more clustering of proteins in microclusters.

Although the three cells' cross-type L, J functions share some similarities, there are also some differences between the three cells. For example, in microclusters, the L-function of CD45, pPLC and pLAT exhibited different patterns around points of its same type or around other types of points in each cell. As in Cell3, most pairs are within the envelope at most distances while in Cell1, only a few panels are in the envelope at most distances. The differences could be due to the number or shape of the polygons in each cell. The number

or shape of the polygons could affect the observed values of the L or J function due to the edge effects.

CD45 often showed different characteristics compared to the other types of proteins. In each cell, the intensity of CD45 was roughly on the same scale across microcluster and non-microcluster, while other types of proteins had significant increases of their intensity in microcluster compared to that of non-microcluster. In addition, the J-function values for the pair of points involving CD45 followed the theoretical values closely in microclusters while other pairs showed some interactions. These findings may suggest that the distribution of CD45 exhibited no significant difference in microclusters compared to the distribution in non-microclusters and therefore the activation of T cell did not have a notable effect on CD45. This agrees with the earlier research finding that CD45 is a protein that should be excluded from the signaling complexes [14].

We have used Poisson process to model the trend of the point process and Gibbs process to model the within-type and between-type dependencies. AIC test of the various models showed that Gibbs model with Strauss-hard core interaction term outperforms the Poisson model without interaction terms, indicating the strong presence of interactions within the point pattern. In addition, not only the presence of pairwise interactions were detected, but also the quantification of the strength of the interactions was achieved. But compared to non-microcluster, microcluster subset had more statistically significant interaction parameters, indicating more pairwise interactions in microcluster were presented.

In the modeling process, we have only used the data from one polygon from each of the microcluster and non-microcluster datasets to reduce the computational load. But the selection of the polygon might lead to a possibly biased results. Figure 3.3 illustrated the varying distribution patterns for each type of protein across different polygons. Furthermore, comparing the cross-type J-function or the cross-type L-function between the subsets and the whole dataset, we observed the subset did not share all the characteristics of the whole dataset. This means there were inhomogeneity across different polygons and a subset of the data in a polygon may not be representative of the whole dataset. To have a further

understanding of the point pattern's behavior within each polygon, we can analyze and fit models separately for the data in each polygon and compare the results.

For future improvements, the goodness-of-fit of the model can be evaluated by comparing the first-order or second-order statistics between the observed data and the simulated data from the model. For example, we can repeatedly simulate realizations under the fitted model and calculate the L, J-functions to obtain the simulated envelope and evaluate the fitness of the observed data with respect to the simulated envelope. This approach provides a flexible way to evaluate the model fit but requires high computational loads. Other possible improvements may include incorporating possible covariates into the model to help explain the characteristic of the point process.

All in all, Biological imaging of T-cells combined with methods of Point pattern analysis is helpful in gaining various insights into the underlying processes that caused an observed point pattern. The presented methodological workflow is helpful where only the locations of the proteins are observed without other available information. Another potential application of point process in Biology is that to apply spatial point processes in extending existing computational methods for single-cell RNA-sequencing data to single-cell spatial transcriptome data [22] [23]. Therefore, point process analysis could be a useful tool to analyze the biological imaging data and develop biological insights.

# Bibliography


[1] Hirotugu Akaike. A new look at the statistical model identification. *IEEE Transactions on Automatic Control*, 19(6):716–723, 1974.

[2] James Allison and L Lanier. Structure, function, and serology of the t-cell antigen receptor complex. *Annual review of immunology*, 5:503–40, 02 1987.

[3] A Baddeley. *Analysing Spatial Point Patterns in R*. 01 2008.

[4] Adrian Baddeley, Ege Rubak, and Rolf Turner. *Spatial Point Patterns: Methodology and Applications with R*. Chapman and Hall/CRC Press, London, 2015.

[5] Adrian Baddeley and Rolf Turner. Practical maximum pseudolikelihood for spatial point patterns. *Australian —& New Zealand Journal of Statistics - AUST N Z J STAT*, 42:283–322, 09 2000.

[6] Adrian Baddeley and Rolf Turner. spatstat: An R package for analyzing spatial point patterns. *Journal of Statistical Software*, 12(6):1–42, 2005.

[7] Adrian Baddeley and Rolf Turner. Modelling spatial point patterns in r. In *Case Studies in Spatial Point Pattern Modelling. Lecture Notes in Statistics 185, 23–74*. Springer, 2006.

[8] Lakshmi Balagopalan, Robert Kortum, Nathan Coussens, Valarie Barr, and Lawrence Samelson. The linker for activation of t cells (lat) signaling hub: From signaling complexes to microclusters. *The Journal of biological chemistry*, 290, 09 2015.

[9] Mira Barda-Saad, Alex Braiman, Rachel Titerence, Stephen Bunnell, Valarie Barr, and Lawrence Samelson. Dynamic molecular interactions linking the t cell antigen receptor to the actin cytoskeleton. *Nature immunology*, 6:80–9, 02 2005.

[10] Valarie Barr. email, September 2020.

[11] Valarie Barr. email, June 2021.

[12] Julian Besag. Contribution to the discussion on dr. ripley's paper. *Journals of the Royal Statistical Society*, B(39):193–195, 1977.

[13] Stephen Bunnell, David Hong, Julia Kardon, Tetsuo Yamazaki, Jane Mcglade, Valarie Barr, and Lawrence Samelson. T cell receptor ligation induces the formation of dynamically regulated signaling assemblies. *The Journal of cell biology*, 158:1263–75, 10 2002.

[14] Stephen Bunnell, David Hong, Julia Kardon, Tetsuo Yamazaki, Jane Mcglade, Valarie Barr, and Lawrence Samelson. T cell receptor ligation induces the formation of dynamically regulated signaling assemblies. *The Journal of cell biology*, 158:1263–75, 10 2002.

[15] Illian J, Penttinen A, Stoyan H, and Stoyan D. *Statistical Analysis and Modelling of Spatial Point Patterns*. John Wiley and Sons, Chichester, 2008.

[16] Daniel Knitter and Oliver Nakoinz. *Point Pattern Analysis as Tool for Digital Geoarchaeology: A Case Study of Megalithic Graves in Schleswig-Holstein, Germany*, pages 45–64. 09 2018.

[17] M.N.M. Lieshout and Adrian Baddeley. A nonparametric measure of spatial interaction in point patterns. *Statistica Neerlandica*, 50:344–361, 11 1996.

[18] David O'Sullivan and David Unwin. *Geographic Information Analysis and Spatial Data*, pages 1 – 31. 03 2010.

[19] B. Ripley. The second-order analysis of stationary point processes. *Journal of Applied Probability*, 13:255–266, 06 1976.

[20] Jérémie Rossy, David Williamson, Carola Benzing, and Katharina Gaus. The integration of signaling and the spatial organization of the t cell synapse. *Frontiers in immunology*, 3:352, 11 2012.

[21] Eilon Sherman, Valarie Barr, Suliana Manley, George Patterson, Lakshmi Balagopalan, Itoro Akpan, Carole Regan, Robert Merrill, Connie Sommers, Jennifer Lippincott-Schwartz, and Lawrence Samelson. Functional nanoscale organization of signaling molecules downstream of the t cell antigen receptor. *Immunity*, 35:705–20, 11 2011.

[22] Dongyuan Song and Jingyi Li. Pseudotimede: inference of differential gene expression along cell pseudotime with well-calibrated p-values from single-cell rna sequencing data. *Genome Biology*, 22, 04 2021.

[23] Tianyi Sun, Dongyuan Song, Wei Vivian Li, and Jingyi Li. scdesign2: a transparent simulator that generates high-fidelity single-cell gene expression count data with gene correlations captured. *Genome Biology*, 22, 05 2021.

[24] John W. Tukey. *Exploratory Data Analysis*. Addison-Wesley, 1977.

[25] Jason Yi, Asit Manna, Valarie Barr, Jennifer Hong, Keir Neuman, and Lawrence Samelson. madstorm: a super-resolution technique for large-scale multiplexing at single molecule accuracy. *Molecular biology of the cell*, 27, 10 2016.