Automatic sorting of point pattern sets using Minkowski functionals

Joshua Parker,[1,*] Eilon Sherman,[2] Matthias van de Raa,[3] Devaraj van der Meer,[3] Lawrence E. Samelson,[4] and Wolfgang Losert[1]

[1]*Department of Physics, University of Maryland, College Park, Maryland 20740, USA*

[2]*Racah Institute of Physics, The Hebrew University of Jerusalem, Jerusalem, Israel*

[3]*Physics of Fluids Group, MESA and Institute for Nanotechnology, and J. M. Burgers Centre for Fluid Dynamics,*
University of Twente, The Netherlands

[4]*Center for Cancer Research, The National Institutes of Health, Bethesda, Maryland 20892, USA*

Point pattern sets arise in many different areas of physical, biological, and applied research, representing many random realizations of underlying pattern formation mechanisms. These pattern sets can be heterogeneous with respect to underlying spatial processes, which may not be visually distiguishable. This heterogeneity can be elucidated by looking at statistical measures of the patterns sets and using these measures to divide the pattern sets into distinct groups representing like spatial processes. We introduce here a numerical procedure for sorting point pattern sets into spatially homogenous groups using functional principal component analysis (FPCA) applied to the approximated Minkowski functionals of each pattern. We demonstrate that this procedure correctly sorts pattern sets into similar groups both when the patterns are drawn from similar processes and when the second-order characteristics of the pattern are identical. We highlight this routine for distinguishing the molecular patterning of fluorescently labeled cell membrane proteins, a subject of much interest in studies investigating complex spatial signaling patterns involved in the human immune response.

PACS number(s): 87.10.Vg, 87.18.Vf

I. INTRODUCTION

A. Motivation: Why study point patterns?

Spatial points patterns naturally arise in many areas of research in both the physical and life sciences, including ecology [1,2], crime statistics [3], epidemeology [4], economics [5], seismology [6], material science [7], and astronomy [8]. Whether the points represent molecules, trees, cell phone users, or entire galaxies, the spatial distributions of point patterns belie the underlying stochastic processes that govern the pattern's formation.

A new area of point pattern analysis involves studying the molecular patterning of proteins on the surfaces of cells. Due to photoactivated localization microscopy (PALM) [9], a new super-resolution microscopy technique, cell biologists are now able to measure the spatial distribution of fluorescently tagged membrane proteins and determine the response of the molecules to different stimuli (see Fig. 1). By fixing the cells on a slide and exposing them to laser light, researchers can activate molecules one by one in multiple cells, locating the center to within 20 nm. This new technique has resulted in a wealth of new point pattern data representing different molecules and surface treatments, and quantitative analysis of these patterns can contribute much to understanding protein-protein and protein-membrane interactions [10,11]. From a theoretical standpoint, each pattern is a pure realization of an underlying spatial process and can be used to characterize that process. From a practical perspective, however, it takes many experimental realizations with finite systems to discern the underlying structure. Furthermore, if the point interactions are complex or the patterns are formed in complicated environments (such as the membrane of a cell), the amount of data needed to confidently quantify a process becomes large and cumbersome to analyze. This gives rise to the need to be able to confidently divide large sets of patterns, sorting the patterns into smaller, homogenous groups that can be analyzed further. In addition to simplifying analysis, this type of sorting can also provide researchers with quick information about the homogeneity of a process and the experimental parameters that affect this homogeneity.

B. Current methods for sorting patterns

The standard method of sorting pattern sets is as follows: For each pattern, one calculates a list of numerical summary characteristics (e.g., index of dispersion, Clark-Evans index). These can be regarded as the "coordinates" of a pattern, to which distance-based clustering algorithms can be applied [12]. This approach presents the researcher with the task of deciding which characteristics to use, how to compare them (normalizing, z cores, etc.) as well as how many: too few may result in missing information, too many could result in redundancy. This adds nuance to the sorting, limiting the statistical conclusions that can be drawn and making trustable automation of the sorting procedure for large pattern sets difficult to accomplish.

A more robust sorting technique has been developed, where patterns are sorted using functional principal component analysis (FPCA) on smoothed second-order functionals of the patterns. This routine treats the point set as a set of functions, $\{a_i(t)\}$, such as the pairwise correlation function, $g(r)$ [13]. The coordinates of each pattern are then calculated by finding the eigenfunctions and corresponding eigenvalues of the equation:

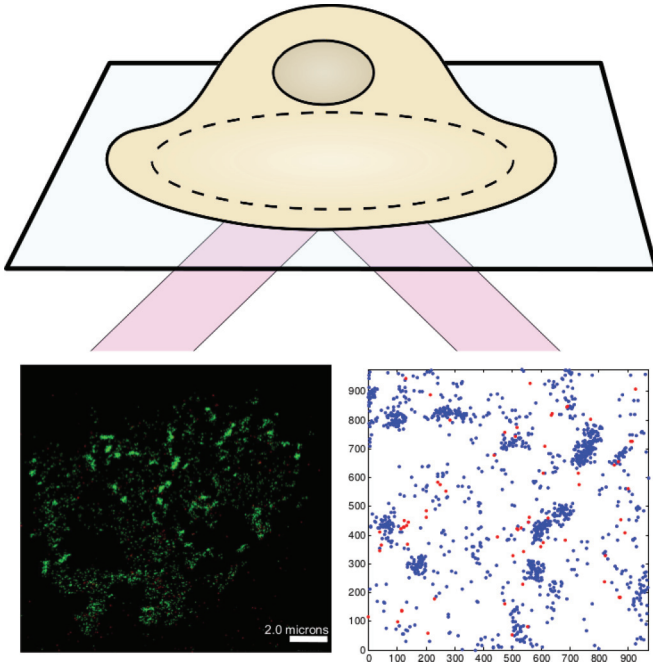$$\int v(s,t)w_i(s)ds = \lambda_i w_i(t). \tag{1}$$

---

*jmparker@umd.edu

FIG. 1. (Color online) Using photoactived localization microscopy, the fluorescently labeled proteins are localized by fitting a point spread function to the stochastically photoactived molecules; the final pattern represents thousands of fluorescent images.

Here $v(s,t)$ is the variance-covariance function of the set of functionals $a(r)$, defined as

$$v(s,t) = (N-1)^{-1} \sum_{i=0}^{N} [a_i - \bar{a}_i(t)][a_i(s) - \bar{a}_i(s)]. \quad (2)$$

The "score" of the ith pattern on the jth principal component is then $\int a_i(t)w_j(t)dt$ (see Refs. [14,15]). Like standard PCA, the eigenvalues form a positive decreasing set whose truncated sum represents the total variance encapsulated in the included principal components. For automation, one can simply set a threshold for the amount of variance to be included, which in turn prescribes the number of coordinates to be used. This feature removes the arbitrariness of sorting patterns via the standard method, making FPCA an easily automatable way of quantifying the difference between patterns. Illian *et al.* [13] showed that this routine was robust to point location uncertainty approaching 20% of the window size [16].

However, spatial processes can create patterns with more structure than second-order functionals can measure. The Neyman-Scott process (NS), introduced to study galaxy clustering, involves randomly distributed parent points generating clusters of varying size. The complexity of the parent-daughter interaction gives rise to families of NS processes with the same pairwise correlation function [12,17,18], despite underlying spatial differences in the patterns.

Baddeley and Silverman [19] also introduced a cell process which is built by partitioning a domain into cells of equal size which are then filled with a varying number of uniformly distributed points. Though the process is rather regular, they showed analytically that their process was indistinguishable from a Poisson process when considering only second-order functionals of the pattern [19], meaning that higher-order functionals must be used to resolve this ambiguity.

C. From points to disks

In this paper, we apply the proximity measure of FPCA to the approximated Minkowski functionals of point patterns [20]. These functionals are calculated by centering a disk on each point and analyzing the topology of this secondary pattern of overlapping disks as a function of the radius. Since the overlap can be very complex, involving all possible combinations of individual points, these functionals depend on all orders of interaction simultaneously. This makes them a more complete "fingerprint" for pattern comparison [20,21]. These functionals have enjoyed marked success in astrophysics [22], soft matter [23], and fluid turbulence [24].

For completeness, we first explain the Minkowski functionals and how they are applied to point pattern analysis. We then demonstrate the sorting procedure by clustering sets of patterns of both synthetically generated data and biological data representing the spatial distributions of membrane proteins. Using both agglomerative and divisive clustering algorithms, we show that this procedure outperforms FPCA clustering with second-order functionals, and in general we demonstrate it to be a viable method for automatically sorting point pattern sets.

II. OUTLINING THE PROCEDURE

A. Minkowski functional analysis of point patterns

The first step in two-dimensional (2D) Minkowski functional analysis [20,21] is to turn a point pattern into a "secondary pattern" by centering a disk of radius r at the center of every point (see Fig. 2). In this paper, we deal only with 2D patterns, but our procedure is easily generalizable to patterns of higher dimensions.

If the radius is large enough, some of these disks will overlap. By combining the overlapping disks, a pattern of differently shaped objects is formed. The total *area*, A, of this collection of objects is then just the total area of the disks excluding any overlapping area. This is the first Minkowski measure, which is related to the functional that can be used to calculate the fractal dimension of a point pattern, if one assumes a power law scaling of A with the disk radius [25]. The second Minkowski measure, the total *perimeter*, P, of the pattern is the perimeter of all of the shapes, which is again reduced from the perimeter of the individual disks because of overlaps. The *Euler number*, χ, is the final Minkowski measure, defined as the total number of distinct shapes or *components* in the window (created by the overlapping disks) minus the number of holes.

By calculating each of these measures first for small radii, where the disks do not overlap, and growing the radius after each calculation until the entire pattern window is covered, the three Minkowski functionals, $A(r), P(r), \chi(r)$, are approximated. Because at each radius, the Minkowski
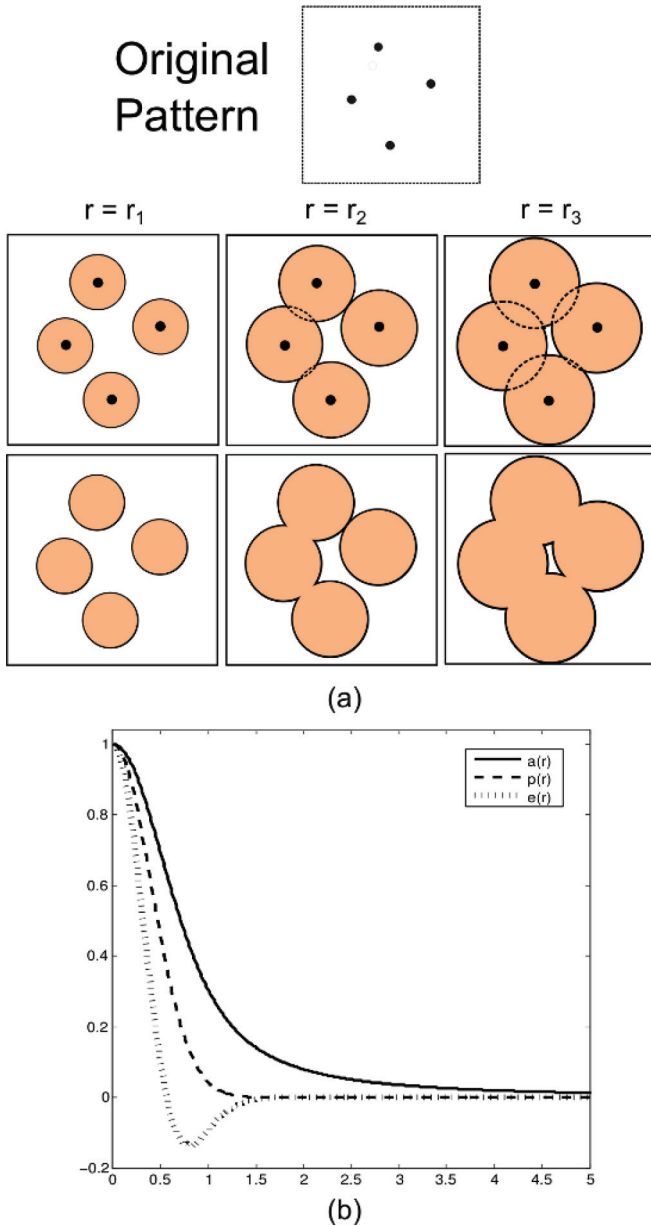
(a)

(b)

FIG. 2. (Color online) (a) The Minkowski functionals are calculated by imposing disks on the point pattern. This new secondary structure can be characterized using topological measures, which vary for different radii. (b) The three reduced Minkowski functionals for a 2D Poisson (random) process. These functionals are unitless due to the normalization by the same measure one would expect for a set of nonoverlapping disks.

measures depend on the locations of all of the points simultaneously, these functionals include information about every type of spatial structure present in the pattern, completely characterizing it (a consequence of Hadwiger's theorem from integral geometry, see Ref. [26]). This feature makes the Minkowski functionals a more complete measure of the underlying point interactions, including information from all possible groupings of points.

When comparing patterns, one actually uses the *reduced* Minkowski functionals, namely, the Minkowski functionals

for the pattern divided by what is expected for a set of nonoverlapping disks. These are given by

$$a(r) = \frac{A(r)}{\pi N r^2}, \tag{3}$$

$$p(r) = \frac{P(r)}{2\pi N r}, \tag{4}$$

$$e(r) = \frac{\chi(r)}{N}. \tag{5}$$

The functionals for a Poisson process are shown in Fig. 2(b). The analysis in this paper relies exclusively on these reduced functionals, so we do not differentiate between the two.

B. Sorting the patterns

Our aim is to automatically sort patterns by performing FPCA on their approximated Minkowski functionals, clustering the patterns with their individual scores on the principal components. We do the same with the pairwise correlation function so that we can directly compare our method with that of Ref. [13]. For each pattern set, we use enough principal components to account for 95% of the variation. For the Minkowski functionals, we calculate the principal component scores individually for the *area*, *perimeter*, and *Euler number* and then concatenate the scores into a larger vector. Then, we use these scores as coordinates, applying two different clustering algorithms.

(i) *Ward's method* [27]. An agglomerative technique which seeks to minimize the total intercluster variance of the distances between objects. We chose this method because it is well known to the pattern analysis community and allows us to directly compare our method with that of Illian *et al.* [13].

(ii) *Fast weighted modularity* [28,29]. To implement this routine, we first calculate the pairwise Euclidean distance between all patterns, D_{ij}, and transform our pattern set into a weighted graph with edge weights:

$$W_{ij} = \max(D_{ij}) - D_{ij}. \tag{6}$$

Then, this algorithm aims to maximize "modularity" of this weighted network by dividing the set into groups where the average edge weight between members of the same group is higher than that between members of different groups. We chose this method because it is a global clustering routine with large popularity in the cluster analysis literature and because the software implementation is able to work with very large data sets (millions of objects).

By utilizing both cluster analysis algorithms, we can verify whatever results we obtain and more completely demonstrate the efficacy of our sorting method using the Minkowski functionals.

III. TESTING OUR SORTING METHOD

We now apply this procedure to three different data sets. These sets of patterns highlight three possible situations in which one would sort patterns: (i) comparing different systems, (ii) varying a parameter in an experiment, and (iii) comparing different components of a bidisperse system [16]. Each set is comprised of two groups which have an *a priori* cluster structure. We then apply the sorting technique, which allows us to calculate the percentage that is misclassified (P_{MC}) by

looking at the fraction of patterns that are assigned to a group that is dominated by a different pattern type.

A. Data set 1: Two Strauss processes

The *Strauss process* [30] is a germ-grain pattern simulation model specified by two parameters, a radius $r \in \mathbb{R}^+$ and an interaction parameter $\gamma \in [0,1]$. The interaction parameter determines if grains of radius r will be allowed to overlap during the formation of the pattern [see Fig. 3(a)].

If γ is small, there is a strong repulsion between grains, where $\gamma = 0$ yields a hard-core process. If γ is close to 1, the repulsion is weak, where $\gamma = 1$ yields a completely
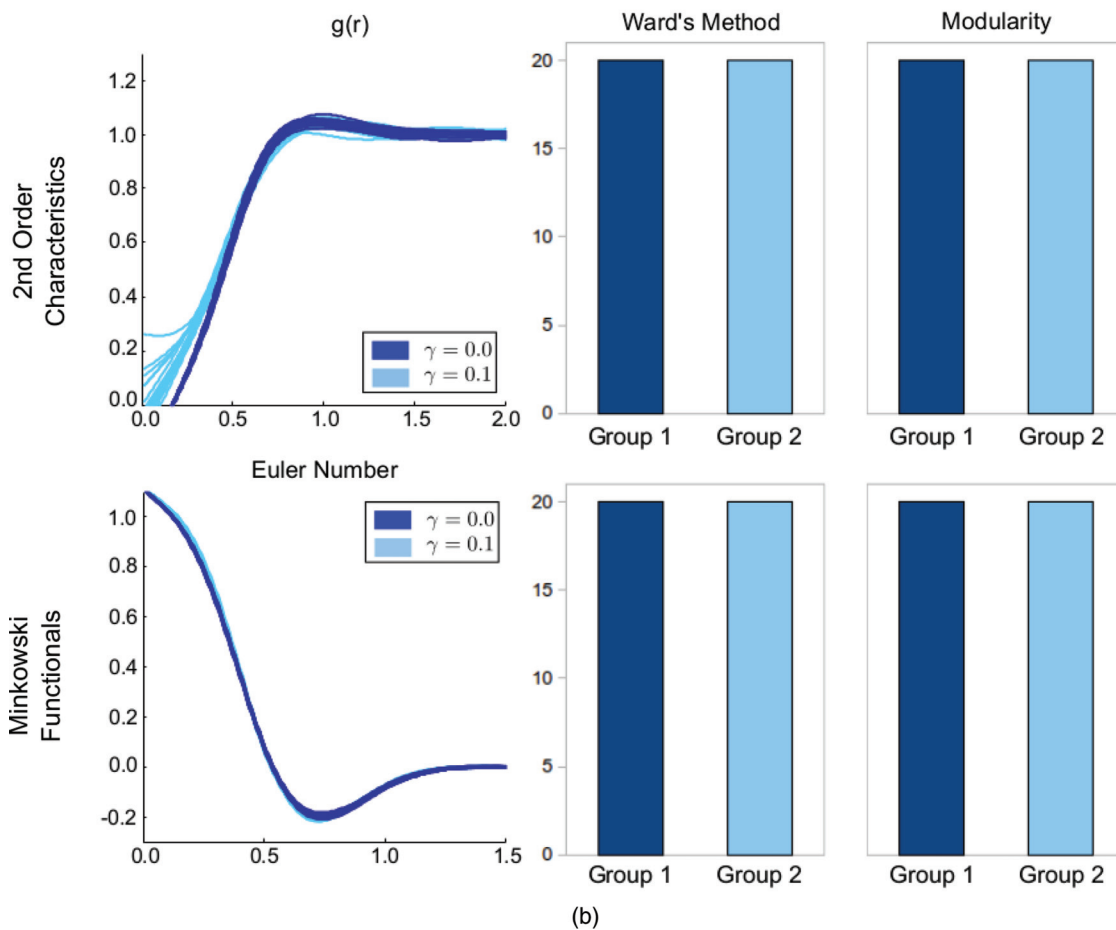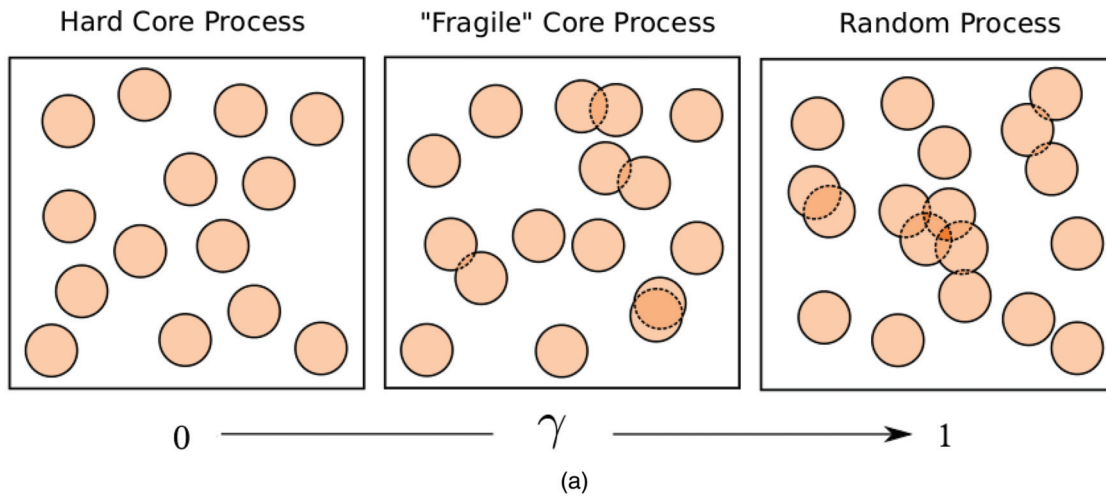

FIG. 3. (Color online) (a) The regularity of a Strauss process is completely determined by γ, the interaction parameter. For $\gamma = 0$, no overlaps are allowed. For $\gamma = 1$, all overlaps are allowed. (b) The leftmost panes display the $g(r)$ and $\chi(r)$ for the 40 simulated Strauss processes. On the right are the results of using FPCA scores to divide the pattern set into two groups. As can be seen, both $g(r)$ and the Minkowski functionals can perfectly separate the set into two groups corresponding to different values of γ.

random process. In Ref. [13], it was reported that even for comparison of pattern sets with similarly strong repulsion ($\gamma = 0.0$ and $\gamma = 0.1$) the pairwise correlation function was able to effectively distinguish different Strauss processes. We

here repeat this test with 20 patterns each, fixing the number of points at $N = 1000$ and letting $r = 0.025$. We also fix the number intensity to be $\lambda = (2r)^{-2}$, which forces interaction between the points.
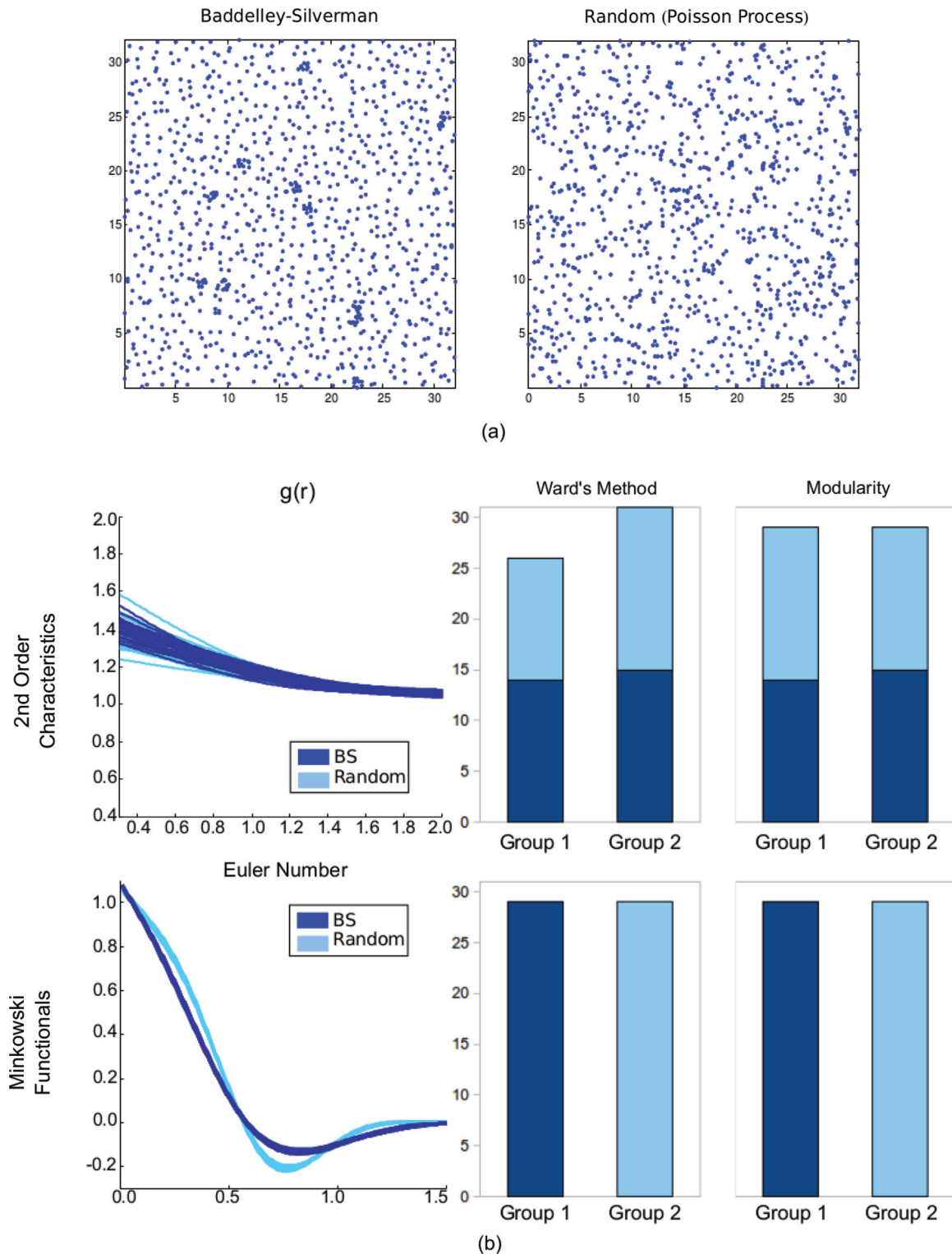

(a)

(b)

FIG. 4. (Color online) (a) A Baddeley-Silverman (left) process side-by-side with a Poisson process (right). Despite the visible differences, the pairwise correlation functions are identical. (b) The leftmost panes display the $g(r)$ and $\chi(r)$ for the 58 patterns simulated. On the right are the results of using FPCA scores to divide the pattern set into two groups. As can be seen, FPCA sorting with $g(r)$ creates two perfectly heterogenous groups, while FPCA sorting with the Minkowski functionals groups the patterns correctly.

As can be seen in Fig. 3(b), both $g(r)$ and the Minkowski functionals are able to distinguish the two Strauss processes, separating the pattern set into two homogenous groups. This is to be expected for $g(r)$, as second-order interactions dominate the process, and is consistent with the findings of Ref. [13].

B. Data set 2: Baddeley-Silverman vs random

The *Baddeley-Silverman process* [19] is built by partitioning a domain into a grid and moving from box to box, distributing N points in each box uniformly. N itself is a random number, taking on the values 0, 1, and 10 with probabilities $1/10$, $8/9$, and $1/90$, respectively. This causes the
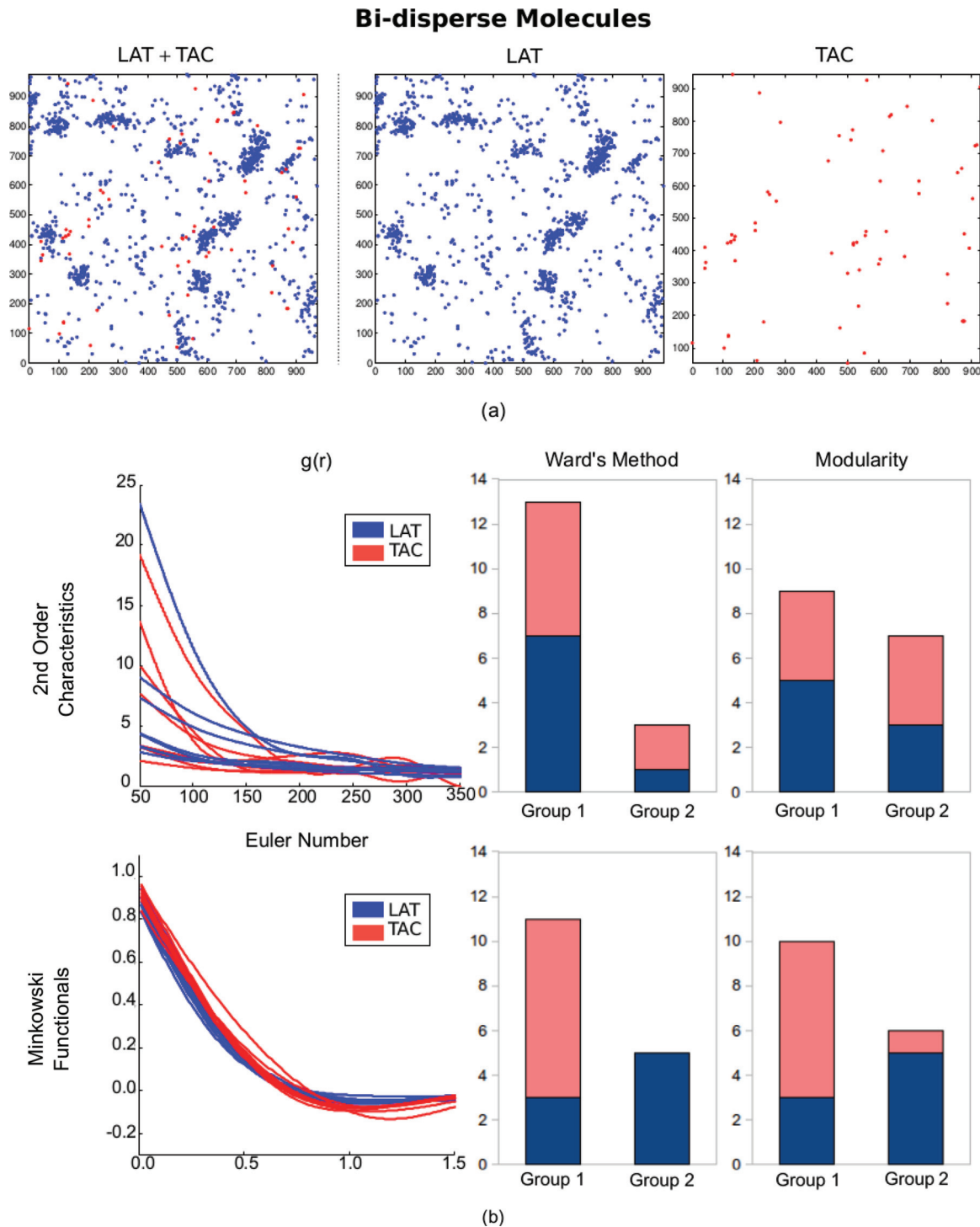

FIG. 5. (Color online) (a) The two proteins are both dispersed in the cell membrane. LAT (blue) and TAC (red) proteins are separately tagged, allowing them to be visualized separately. (b) The leftmost panes display the $g(r)$ and $\chi(r)$ for the 16 molecular patterns. On the right are the results of using FPCA scores to divide the pattern set into two groups. As can be seen, using Minkowski functionals with FPCA improves the differentiation of the two sets.

process to be rather regular, but with some strong clustering occurring every now and then (see Fig. 6).

Since $E[N] = \text{Var}[N] = 1$, it can be shown that the Baddeley-Silverman process shares all of the same second-order characteristics as a random (Poisson) process. In Ref. [20], the Minkowski functionals were shown to be able to distinguish these two processes. Therefore, we expect to see proper sorting when using $a(r)$, $p(r)$, and $e(r)$, and failure using $g(r)$.

We simulated 29 Baddeley-Silverman processes and 29 Poisson processes, fixing the point number $N = 1024$, using both the pairwise correlation function and the Minkowski functionals, and sorting the pattern sets into two groups. As can be seen in Fig. 4(b), the pairwise correlation function fails to sort the patterns correctly, creating two heterogenous groups ($P_{MC} > 40$). However, the Minkowski functionals successfully divide the pattern set into two homogenous groups.

C. Data set 3: Bidisperse patterns of intercellular proteins

For an application to an experimental data set, we look at super-resolution images of two proteins residing at the membrane of immune cells [see Fig. 5(a)].

One protein under study is LAT, short for "linker for activation of T cells", a naturally occurring protein crucially involved in the reactions that regulate T-cell antigen-dependent activation, a critical event in the adaptive immune response. LAT proteins have been seen to form clusters on the membrane with potentially complicated hierarchies [11]. However, the membrane of the cell can have a first-order effect on the molecular patterning of membrane proteins. It has been found that the location of other membrane protein clusters often correlates with how close the membrane is to the surface and anticorrelates with regions of high membrane fluctuations [31].

Another protein, TAC (the α chain of the interleukin-2 receptor), can also be localized and differentiated from LAT by tagging with a different fluorescent molecule and using two different lasers with different wavelengths. TAC is a membrane protein that does not form clusters, instead distributing uniformly in regions where protein-membrane interactions have not excluded proteins. This means that TAC can serve as a membrane marker when studying the clustering of other proteins. Since LAT and TAC are part of separate signaling pathways, they also do not interact biochemically [32]. Therefore, upon sorting, we should get two homogenous groups representing the two different molecules.

Applying FPCA on the approximated pairwise correlation functions of these data sets again yields strongly heterogeneous groups ($P_{MC} \approx 50\%$). This is visible in the pair-correlation functions [Fig. 5(b)], where the individual patterns exhibit large variability. In contrast, because the Minkowski functionals consider more than just second-order interactions, the *Euler number* is able to visibly distinguish the molecular patterns, and the pattern sorting is improved ($P_{MC} \approx 25\%$).

Further success is achieved if we look at how FPCA sorting with each functional performs on its own (see Fig. 6). When only using the *area*, the sorting is identical to the sorting based on all three Minkowski functionals. However, sorting the LAT-TAC protein pattern set improves when just using the *perimeter*, and we achieve almost perfect classification when
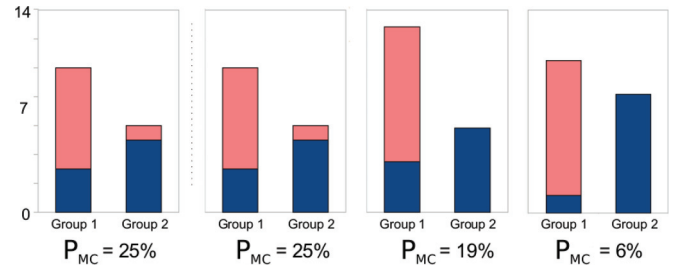

FIG. 6. (Color online) From left to right: Sorting with all three functionals, the area, the perimeter, and the Euler number. Considering the individual functionals (sorting using weighted modularity), the Euler number outperforms the other two, only misclassifying one pattern.

using the *Euler number*, only misclassifying one pattern. This is not surprising, since the area and perimeter are constrained to be smooth, positive, and monotonically decreasing, and thus cannot vary as much while the Euler number can vary more wildly.

IV. CONCLUSIONS AND DISCUSSION

In this work, we have introduced the procedure to automatically sort point pattern sets using the approximated Minkowski functionals and FPCA. Using Strauss processes with strong repulsion, we have shown that this method can accurately sort point pattern sets drawn from very similar processes. Further, this method also distinguishes Baddeley-Silverman processes from Poisson processes, a task which the pairwise correlation function perfectly fails to accomplish. We then found that when looking at experimental point patterns representing proteins, FPCA sorting using the Minkowski functionals outperformed FPCA sorting with the pairwise correlation function. This sufficiently demonstrates that the Minkowski functionals can successfully quantify the differences between pattern sets showing complex behavior.

We also found that FPCA sorting using only the Euler number strongly outperforms the other two. While mathematically the three functionals do completely classify a pattern, the area and perimeter may only be slightly different for different spatial processes. This means that error introduced when approximating the functionals numerically may blur these differences, resulting in improper sorting. Since the Euler number is allowed to vary more dramatically as the disks combine and holes form, it can visually distinguish very similar pattern sets and therefore leads to better sorting.

Though we have presented this technique as a way to sort patterns into distinct sets for further analysis, the sorting itself can serve as an analysis tool. We are currently working to apply this tool to examine how the presence of different chemical cues effect the clustering of LAT proteins, as well as how T-cell activation perturbs the patterns. Because of the Minkowski functional's ability to robustly characterize a pattern, we can treat the membership of a pattern in a particular group as a sign of similarity between it and its comembers. We can therefore look at group statistics to determine what experimental variables change the molecular patterns and to what degree, allowing for systematic large-scale investigations of the membrane proteins and their response to different stimuli.

[1] R. Law, J. Illian, D. F. R. P. Burslem, G. Gratzer, C. V. S. Gunatilleke, and I. A. U. N. Gunatilleke, J. Ecol. **97**, 616 (2009).

[2] T. Wiegand and K. Moloney, Oikos **104**, 209 (2004).

[3] G. O. Mohler, M. B. Short, P. J. Brantingham, F. P. Schoenberg, and G. E. Tita, J. Am. Stat. Assoc. **106**, 100 (2011).

[4] A. C. Gatrell, T. C. Bailey, P. J. Diggle, and B. S. Rowlingson, Trans. Inst. Br. Geogr. New Ser. **21**, 256 (1996).

[5] R. Bivand and S. Szymanski, Regional Science and Urban Economics **30**, 203 (2000).

[6] Y. Ogata, Pure Appl. Geophys. **155**, 471 (1999).

[7] S. J. L. Billinge, Z. Kristallogr. **219**, 117 (2004).

[8] W. Kendall, J. Am. Stat. Assoc. **98**, 248 (2003).

[9] E. Betzig, G. Patterson, R. Sougrat, O. Lindwasser, S. Olenych, J. Bonifacino, M. Davidson, J. Lippincott-Schwartz, and H. Hess, Science **313**, 1642 (2006).

[10] E. Lachmanovich, D. E. Shvartsman, Y. Malka, C. Botvin, Y. I. Henis, and A. M. Weiss, J. Microsc. **212**, 122 (2003).

[11] E. Sherman, V. Barr, S. Manley, G. Patterson, L. Balagopalan, I. Akpan, C. K. Regan, R. K. Merrill, C. L. Sommers, J. Lippincott-Schwartz, and L. E. Samelson, Immunity **35**, 705 (2011).

[12] J. Illian, A. Penttinen, H. Stoyan, and D. Stoyan, *Statistical Analysis and Modeling of Spatial Point Patterns*, Statistics in Practice (Wiley & Sons, New York, 2008).

[13] J. Illian, E. Benson, J. Crawford, and H. Staines, in *Case Studies in Spatial Point Process Modeling*, Lecture Notes in Statistics, Vol. 185, edited by A. Baddeley, P. Gregori, J. Mateu, R. Stoica, D. Stoyan, P. Bickel, P. Diggle, S. Fienberg, U. Gather, I. Olkin, and S. Zeger (Springer, New York, 2006), pp. 135–150.

[14] J. O. Ramsay and B. W. Silverman, *Functional Data Analysis*, 2nd ed., Springer Series in Statistics (Springer, New York, 2005).

[15] J. O. Ramsay, G. Hooker, and S. Graves, *Functional Data Analysis with R and MATLAB*, 1st ed. (Springer, New York, 2009).

[16] See Supplemental Material at http://link.aps.org/supplemental/10.1103/PhysRevE.88.022720 for more information regarding how to incorporate spatial uncertainty into point pattern analysis and specifics regarding software and computational methods used.

[17] J. P. Diggle, *Statistical Analysis of Spatial Point Patterns* (Academic Press, New York, 1983), pp. 1–148.

[18] A. Tscheschel and D. Stoyan, Comput. Stat. Data Anal. **51**, 859 (2003).

[19] A. J. Baddeley and B. W. Silverman, Biometrics **40**, 1089 (1984).

[20] K. R. Mecke and D. Stoyan, Biom. J. **47**, 473 (2005).

[21] J. Schmalzing, M. Kerscher, and T. Buchert, in *Dark Matter in the Universe*, Proceedings of the International School of Physics "Enrico Fermi," Course CXXXII, Varenna, 1995, edited by S. Bonometto, J. Primack, and A. Provenzale (IOP Press, Amsterdam, 1996), p. 281.

[22] M. Kerscher and A. Tikhonov, Astron. Astrophys. **509**, A57 (2010).

[23] S. Gallier, Propellants, Explos., Pyrotech. **34**, 97 (2009).

[24] E. Calzavarini, M. Kerscher, D. Lohse, and F. Toschi, J. Fluid Mech. **607**, 13 (2008).

[25] The Minkowski functionals can also be defined for fractal geometric sets; see Ref. [26] for additional discussion of the fractal scaling properties of the Minkowski functionals.

[26] K. R. Mecke, Lect. Notes Phys. **554**, 111 (2000).

[27] J. Ward, J. Am. Stat. Assoc. **58**, 236 (1963).

[28] A. Clauset, M. E. J. Newman, and C. Moore, Phys. Rev. E **70**, 066111 (2004).

[29] M. E. J. Newman, Phys. Rev. E **70**, 056131 (2004).

[30] D. J. Strauss, Biometrika **62**, 467 (1975).

[31] K. Lam Hui, C. Wang, B. Grooman, J. Wayt, and A. Upadhyaya, Biophys. J. **102**, 1524 (2012).

[32] R. Robb, W. Greene, and C. Rusk, J. Exp. Med. **160**, 1126 (1984).