Identification of Caveolin-1 Domain Signatures via Graphlet Analysis of Single Molecule Super-Resolution Data

Ismail M. Khater[1], **Fanrui Meng**[2], **Ivan Robert Nabi**[2*#], **Ghassan Hamarneh**[1*]

[1]Medical Image Analysis Lab, School of Computing Science, Simon Fraser University, Burnaby, BC V5A 1S6, Canada

[2]Department of Cellular and Physiological Sciences, Life Sciences Institute, University of British Columbia, Vancouver, BC V6T 1Z3, Canada

*Correspondence: irnabi@mail.ubc.ca (IRN), hamarneh@sfu.ca (GH)

*Equal contribution

[#]Lead Contact

ABSTRACT

MOTIVATION

Network analysis and unsupervised machine learning processing of single molecule localization microscopy (SMLM) of caveolin-1 (Cav1) antibody labeling of prostate cancer cells identified biosignatures and structures for caveolae and three distinct non-caveolar scaffolds (S1A, S1B and S2) (Khater et al., 2018). To obtain further insight into low-level molecular interactions within these different structural domains, we now introduce graphlet decomposition over a range of proximity thresholds and show that frequency of different subgraph (k=4 nodes) patterns for machine learning approaches (classification, identification, automatic labeling, etc.) effectively distinguishes caveolae and scaffold blobs.

RESULTS

Caveolae formation requires both Cav1 and the adaptor protein CAVIN1 (also called PTRF). As a supervised learning approach, we applied a wide-field CAVIN1/PTRF mask to CAVIN1/PTRF-transfected PC3 prostate cancer cells (PC3-PTRF cells) and used the random forest classifier to classify blobs based on graphlet frequency distribution (GFD). GFD of CAVIN1/PTRF-positive (PTRF+) and -negative (PTRF-) Cav1 clusters showed poor classification accuracy that was significantly improved by stratifying the PTRF+ clusters by either number of localizations or volume. Low classification accuracy (<50%) of large PTRF+ clusters and caveolae blobs identified by unsupervised learning suggests that their GFD is specific to caveolae. High classification accuracy for small PTRF+ clusters and caveolae blobs argues that CAVIN1/PTRF associates not only with caveolae but also non-caveolar scaffolds. At low proximity thresholds (PTs) (50-100 nm), the caveolae groups showed reduced frequency of highly connected graphlets and increased frequency of completely disconnected graphlets. GFD analysis of SMLM Cav1

clusters defines changes in structural organization in caveolae and scaffolds independent of association with CAVIN1/PTRF.

SUPPLEMENTARY INFORMATION

Supplementary material and methods are available at *Bioinformatics* online.

INTRODUCTION

Graphs (or networks) are ubiquitous and powerful mathematical structures used for modeling objects and their interactions. In graph representation, objects are modelled via nodes, and interactions between objects are modelled via connections or edges between the nodes. Graph analysis methods have been successfully applied to the analysis of social, computer, brain, and biological networks (Newman, 2003, Bassett and Sporns, 2017, Krivitsky et al., 2009, Costa et al., 2008, Zitnik and Leskovec, 2017, Brown et al., 2014). Point clouds refer to datasets where a collection of objects or nodes are presented without connections. Many methods have been deployed to extract features from point clouds, either via hand-crafted features or automatically derived features via deep learning, for different machine learning tasks such as segmentation and classification (Gumhold et al., 2001, Qi et al., 2017).

Graphlets and motifs are two related concepts used to describe small patterns or subnetworks that occur in large networks. Graphlets, small connected non-isomorphic induced subgraphs of a large network, were introduced by Pržulj et al (Pržulj et al., 2004) as measures of local network structure. Network motifs, on the other hand, are interconnected patterns significantly and highly recurrent in real-world networks compared with random networks (Milo et al., 2002). Consequently,

graphlets and motifs are regarded as the basic building blocks for large networks and graphs. Decomposing large networks into their base graphlets and calculating occurrence frequencies (i.e., GFD), provide robust features to differentiate between networks and discover their underlying signatures.

Graphlet patterns can vary based on the number of the nodes k and their connectivity in each subgraph. The number of possible connectivity patterns of a graphlet increases exponentially with k and hence increases the complexity of enumerating the graphlets in big graphs. Typically, graphlet patterns are extracted for k=2, 3, 4, and 5 nodes. Moreover, finding some of the patterns are NP-hard problems (e.g. finding the clique, or fully inter-connected set of vertices, of maximum size) and require efficient approximate counting algorithms (Ahmed et al., 2015, Ahmed et al., 2017). Many methods have been proposed to find graphlet patterns in networks (Marcus and Shavitt, 2012, Hočevar and Demšar, 2014, Ahmed et al., 2015, Shervashidze et al., 2009). In particular, the parallel graphlet decomposition method (Ahmed et al., 2015, Ahmed et al., 2017) is very fast, scalable and supports extracting macro and micro graphlets features and counting both connected and disconnected graphlets.

Graphlets have been widely used for a myriad of applications such as network similarity, network alignment, biological and protein networks, prediction and classification, image segmentation, etc. (Pržulj et al., 2004, Pržulj, 2007, Dutta and Sahbi, 2017, Kollias et al., 2012, Bressan et al., 2017, Zhang et al., 2013, Shin et al., 2016, Shervashidze et al., 2009). Recently, graphlets/motifs are exploited as a high-order connectivity patterns information for network clustering (Benson et al., 2016), and fast local graph clustering and community detection tasks (Yin et al., 2017).

Recently, we described a graph-based complex network analysis clustering method (Khater et al., 2018) to analyze the 3D point cloud of single molecule localization microscopy (SMLM) super resolution microscopy data. The method consists of an integrated pipeline that includes network construction, multiple blink correction, noise filtering, and cluster segmentation, analysis and identification using machine learning. Application of unsupervised learning network analysis and machine learning to point clouds of Cav1, the coat protein of cell surface caveolae invaginations, reported a modular structure for caveolae, similar to that determined by cryoEM (Stoeber et al., 2016, Ludwig et al., 2016), as well as three non-caveolar scaffolds, including a previously unreported hemispherical scaffold domain (Khater et al., 2018). Caveolae invagination requires the adaptor protein CAVIN1 (also known as Polymerase I and Transcript Release Factor Protein PTRF and referred to in this manuscript as CAVIN1/PTRF and, for brevity, PTRF). Here we apply supervised machine learning approaches, leveraging GFD features and a CAVIN1/PTRF TIRF mask to form training data, to identify CAVIN1/PTRF-associated Cav1 domains and use graphlet analysis to classify and compare the Cav1-labeled domains. We show that graphlets are highly efficient classifiers that distinguish caveolae from scaffolds. Comparison of unsupervised and supervised learning, the latter using a CAVIN1/PTRF mask, suggests that CAVIN1/PTRF associates not only with caveolae but also scaffold domains.

MATERIALS AND METHODS

We summarize the abbreviations and acronyms used in this paper in Table 1.

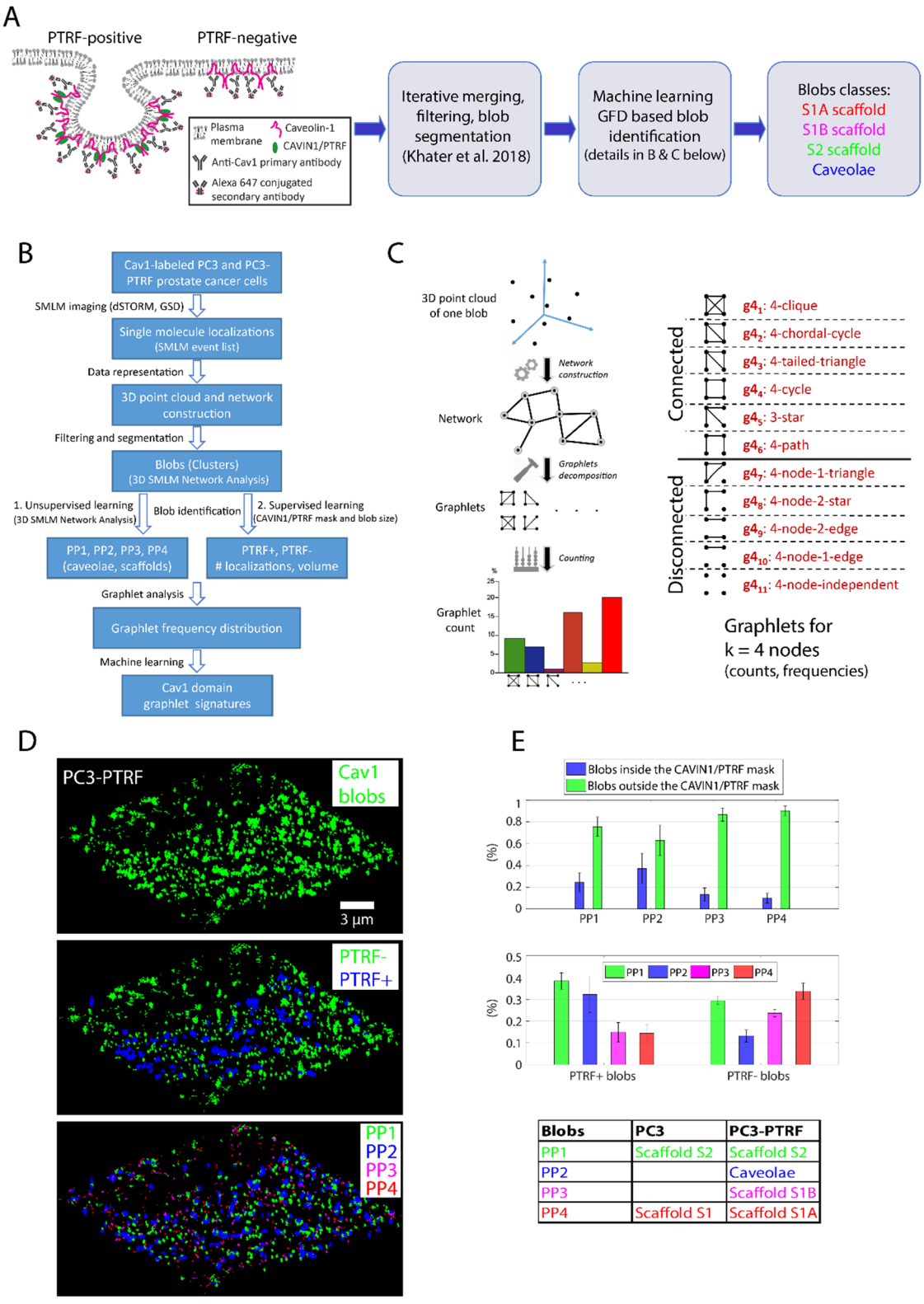
Term/acronym	Explanation
PC3	Prostate cancer cell lines
PC3-PTRF	CAVIN1/PTRF transfected PC3 cells
Cav1	Caveolin-1 protein
PTRF	Polymerase I and Transcript Release Factor protein. Also known as CAVIN1

PTRF+	PTRF-positive (blobs inside the CAVIN1/PTRF mask)
PTRF-	PTRF-negative (blobs outside the CAVIN1/PTRF mask)
PT	Proximity Threshold
GSD	Ground State Depletion
TIRF	Total Internal Reflection Fluorescence
SMLM	Single Molecule Localization Microscopy
GFD	Graphlet Frequency Distribution
Acc	Accuracy (classification evaluation measure)
Spe	Specificity (classification evaluation measure)
Sen	Sensitivity (classification evaluation measure)
AUC	Area Under ROC Curve (classification evaluation measure)
ROC	Receiver Operating Characteristic (classification evaluation measure)
t-SNE	t-distributed Stochastic Neighbor Embedding. It is a dimensionality reduction and visualization technique based on machine learning algorithm

Table 1: The list of abbreviations and acronyms used in this paper.

Figure 1A,B shows an overview of the framework for GFD classification of SMLM Cav1-labeled blobs. SMLM event lists of anti-Cav1 labeling of CAVIN1/PTRF transfected PC3 prostate cancer cells are converted into 3D point cloud representation of blink locations in 3-dimensional space. Following network construction from the Cav1 point clouds, multiple blinking from single fluorophores and closely associated fluorophores is corrected, noise/background is filtered out and clusters identified, as previously described (Khater et al., 2018). We then applied our established unsupervised learning approach based on topological, shape and network features to identify Cav1 clusters (blobs). We previously learned four groups of Cav1 blobs and, through matching the learned groups from both PC3 and PC3-PTRF populations, we identified the blobs types, (Table in Fig. 1E), as PP1 (S2 Scaffolds), PP2 (caveolae), PP3 (S1B scaffolds) or PP4 (S1A scaffolds) (Khater et al., 2018). Alternatively, we used a wide field CAVIN1/PTRF mask for supervised learning in which blobs were classed as either “in mask” or “out mask”. Every blob is processed as shown in Figure 1C, where a network is constructed for every blob and the different graphlet patterns are extracted for k=4 nodes (Fig. 2B). The graphlet patterns are then counted and used to find the GFD for the connected and disconnected graphlet patterns of every blob. Figure 1D shows

representative images of all Cav1 clusters, Cav1 clusters in (PTRF+) and out (PTRF-) of the CAVIN1/PTRF mask and unsupervised learning classification of blobs as PP1, PP2, PP3 or PP4 (Fig. 2A). Quantification of blob distribution showed that all four classes of Cav1 blobs were present in and out of the CAVIN1/PTRF mask, with caveolae (PP2) more present in the PTRF+ blobs, small S1A (PP4) and S1B (PP3) scaffolds less present and hemispherical S2 scaffolds (PP1) present amongst both PTRF+ and PTRF- blobs (Fig. 1E). In light of the heterogenous distribution of the various classes of Cav1 blobs in and out of the CAVIN1/PTRF mask, we decided to classify the blobs using random forest classification based on GFD features.



Blobs	PC3	PC3-PTRF
PP1	Scaffold S2	Scaffold S2
PP2		Caveolae
PP3		Scaffold S1B
PP4	Scaffold S1	Scaffold S1A

Figure 1. SMLM Cav1 blob identification via graphlet analysis. A. An overview of the proposed method. The Cav1 protein is labeled for SMLM imaging. The CAVIN1/PTRF mask is used to obtain the ground truth labels for blobs to define PTRF+ and PTRF- classes. The SMLM

data is processed using (Khater et al., 2018) to obtain the segmented Cav1 blobs and their classes to train the supervised machine learning model. The GFD features for the Cav1 blobs are used to train a machine learning classifier for blob identification. **B.** The proposed analysis framework. Cav1 protein is labeled in PC3 prostate cancer cells and CAVIN1/PTRF-transfected PC3 cells (PC3-PTRF cells). SMLM events are represented as a 3D point cloud of Cav1 blinks and *3D SMLM Network Analysis* computational pipeline (Khater et al., 2018) used for cluster analysis of labeled Cav1 proteins, including modules to correct for multiple blinking of single fluorophores, filtering, segmenting, and identifying the blobs using unsupervised learning (Khater et al., 2018). We also used the wide-field CAVIN1/PTRF TIRF mask to classify Cav1 blobs for the supervised learning task. Graphlet features were extracted for every blob network at various PTs. We used a parallel graphlet decomposition library (Ahmed et al., 2015) to extract feature vector for every network. Feature vectors from all the blobs were used for machine learning classification and identifying biosignatures for the different Cav1 structures. **C.** To get graphlet features for a single blob, a network is constructed for the 3D point cloud of a single blob and graphlets are then counted in every network. GFD features are derived from connected ($g4_1$, $g4_2$, $g4_3$, $g4_4$, $g4_5$, $g4_6$) and disconnected ($g4_7$, $g4_8$, $g4_9$, $g4_{10}$, $g4_{11}$) graphlet patterns for k=4 nodes. GFD features can be extracted for connected graphlets alone, disconnected graphlets alone, and for both connected and disconnected graphlets. We showed the process of extracting graphlet features for some of the Cav1 blobs in **Figure 2B**. **D.** 3D views of a PC3-PTRF cell showing all Cav1 blobs, PTRF+ Cav1 blobs in-mask (blue) with remaining blobs, out-mask, labeled as PTRF- (green), and blobs classified based on unsupervised learning as PP1, PP2, PP3 and PP4 blobs (Khater et al., 2018). **E.** Percentages of blobs for each group in- and out-mask and distribution of blobs within and without the CAVIN1/PTRF mask are shown.

To extract the graphlet features, we constructed networks for every blob based on the PT between each pair of molecules within the blob. The nodes represent molecules in the network, where interaction between the neighbouring molecules is represented as an edge (Fig.1C, Fig. 2B). We leveraged the parallel graphlet decomposition method (Ahmed et al., 2015, Ahmed et al., 2017) to extract graphlet features from the blobs' networks. The parallel graphlet decomposition method extracts macro (global) motif counts for the motifs with *k=2, k=3* and *k=4* nodes, micro-motif counts, and the GFD features. As the micro and macro features are highly dependent on the number of nodes in the network, we based our analysis on GFD features with *k=4* nodes (Fig. 1C) so as not to bias our classification model between small and large blobs and analyzed GFD features regardless of the number of nodes in their corresponding networks. The extracted features are then used to classify the blobs. See Supplementary Methods for details.

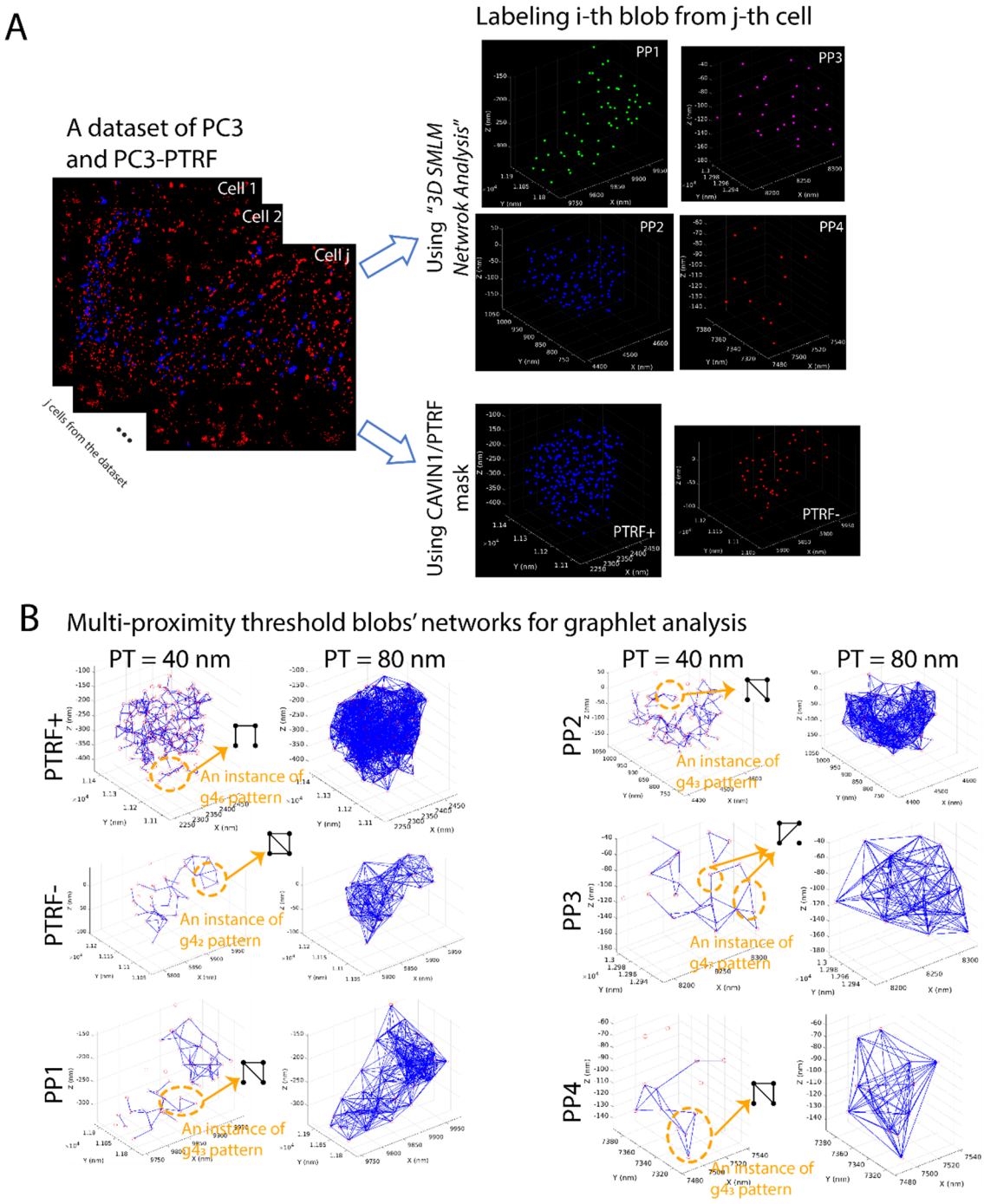


Figure 2. The process of extracting the graphlet features from the various blob types. A. The segmented and labeled blobs from PC3 and PC3-PTRF cells in the dataset are used for feature extraction. We show a representative blob from each group. The blobs are labeled based on the CAVIN1/PTRF mask into PTRF+ and PTRF- classes or labeled based on the unsupervised labeling method (Khater et al., 2018) into four classes (i.e. PP1, PP2, PP3, and PP4). **B.** Networks at PTs of 40 and 80 nm are shown and examples of graphlets in the networks are given. Graphlet features for k = 2, 3, 4 nodes were extracted and counted from the blobs. The GFD features were extracted for k = 4 nodes.

RESULTS

We first applied GFD analysis to the 4 classes of Cav1 domains previously identified in PC3-PTRF cells (PP1: S2 scaffolds; PP2: caveolae; PP3: S1A scaffolds; PP4: S1B scaffolds) based on topology, shape and network features (Khater et al., 2018). As can be seen in Figure 3A, frequency distributions for the 11 k=4 graphlets vary amongst the four classes of blobs, shown here at PT=80 nm. Given that we describe each blob with a vector of features and that this vector is high (multi) dimensional, visualizing these features is not straightforward as with 2 or 3-dimensional vectors. In our work, the vectors are 11-dimensional capturing the distribution of the frequency of the 11 graphlets used in this paper. A powerful and well-known approach for visualizing high-dimensional vectors is t-distributed Stochastic Neighbor Embedding, which is commonly referred as t-SNE (Maaten and Hinton, 2008). t-SNE works by optimally projecting high dimensional feature vectors to lower dimensional 2D so they can be easily displayed. The projection is optimal in that it minimizes the difference (measured via the Kullback–Leibler divergence) between the conditional probability distributions (probability of picking neighbouring high-dimensional feature vector j given reference vector i) before and after the projection and hence retains the local structure and reveal the important global structure In t-SNE plots, each point represents the feature vector of a blob (after projection to 2D space). t-SNE visualization for the GFD features clearly show distinct groups for caveolae (PP2), hemispherical S2 scaffolds (PP1) and overlap between the similar and smaller S1A (PP4) and S1B (PP3) scaffolds.

One-versus-one classification between the four groups (i.e. for M=4 groups, we will have $N = \frac{M\times(M-1)}{2} = 6$ classifiers) was determined across PTs from 20-300 nm, the minimal PT corresponds

to the merge threshold applied and the maximum to the largest blob size. Caveolae (PP2) showed a high classification accuracy, sensitivity, specificity and area under ROC curve (AUC) against all three scaffold groups (Fig. 3B). Classification accuracy was highest versus S1 scaffolds (PP3, PP4) that were most poorly distinguished from each other (Fig. 3B,C). Progressive increases in classification accuracy between caveolae (PP2) and S2 scaffolds and then S1 scaffolds is highlighted by the graph showing average classification accuracy over the 20-300 nm PT range (Fig. 3C). Use of GFD feature based classification has therefore corroborated Cav1 cluster classification based on topology, size and shape features (Khater et al., 2018). We also used the one-versus-all classification (Fig. S1B) and used the voting to combine the results of all the classifiers to get the multi-class classification accuracy (Fig. S1C). The results correspond to the one-versus-one classification results. Moreover, the multi-class classification results show that the classification loss (which measures the predictive inaccuracy of classification) is <10% when the PT is between 50 – 160 nm. GFD features best discriminate and identify Cav1 domains in the 50 – 160 nm PT range.

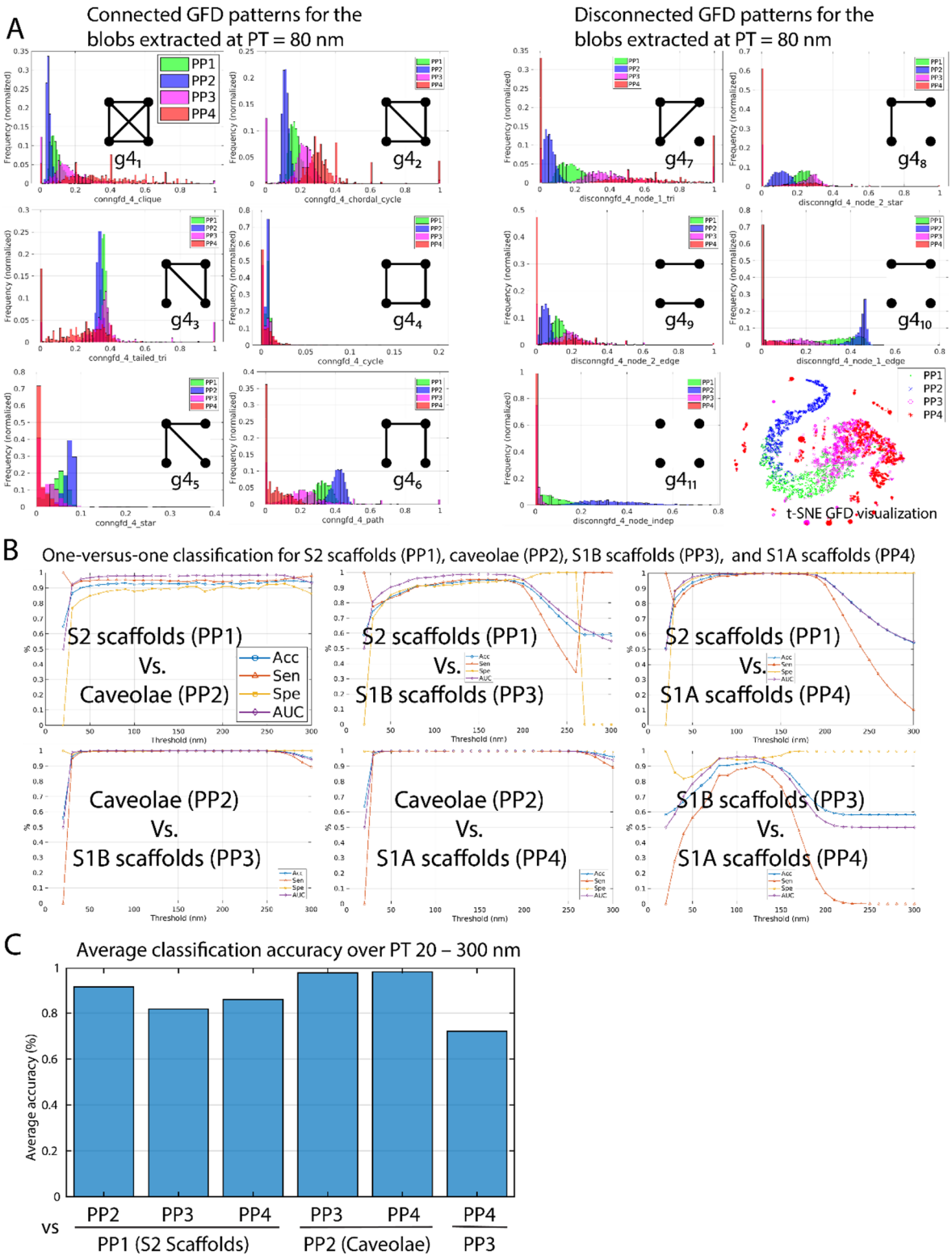


Figure 3. Graphlet biosignatures for the Cav1 domains via machine learning. A. GFD features of S2 scaffolds (PP1), caveolae (PP2), S1B scaffolds (PP3) and S1A scaffolds (PP4) extracted using the *3D SMLM Network Analysis* pipeline. Histograms of the connected and

disconnected GFD features extracted from the networks at PT=80 nm are shown. t-SNE visualization of the GFD features shows a 2D view of the projected features that depicts how the blobs from different classes are distributed. **B.** Machine learning classification of the four groups using the GFD features extracted at various PTs. In the classification task, we used the one-versus-one strategy for the four groups multi-labeled data to classify the blobs with four classes. In total, we have six classifiers that show all the pairs of groups of classification results. The binary classification shows the performance of classifying every group against the other groups at various PTs. We show the classification results for the binary classification task by reporting the accuracy (Acc), sensitivity (Sen), specificity (Spe), and area under ROC curve (AUC). We also showed the result of using one-versus-all and the multi-class classification results in the Supplementary **Figure S1C.** The bar plot shows the average classification accuracy over the used PTs for the binary classifiers in (**B**) above. High classification accuracy suggests that the groups are distinct and the classifier can discriminate and identify them. Low classification accuracy suggests that the classifier misclassify some of the blobs and unable to discriminate them due to the similarity and overlapping of their GFD features.

We then applied GFD classification to PTRF+ and PTRF- blobs. Figure 4A shows the processing of the CAVIN1/PTRF mask by applying the morphological closing operation to remove small holes in the mask and superposition of segmented Cav1 blobs on the CAVIN1/PTRF mask for correspondence analysis. Using this method, Cav1 blobs touching the mask are “in” mask and others “out” mask. Binary GFD pattern analysis shows a high degree of overlap between the PTRF+ and PTRF- classes for all graphlets. GFD similarity between the two classes is highlighted by extensive overlap in the t-SNE visualization and the low (~65%) classification accuracy between the two groups (Fig. 4B,C).

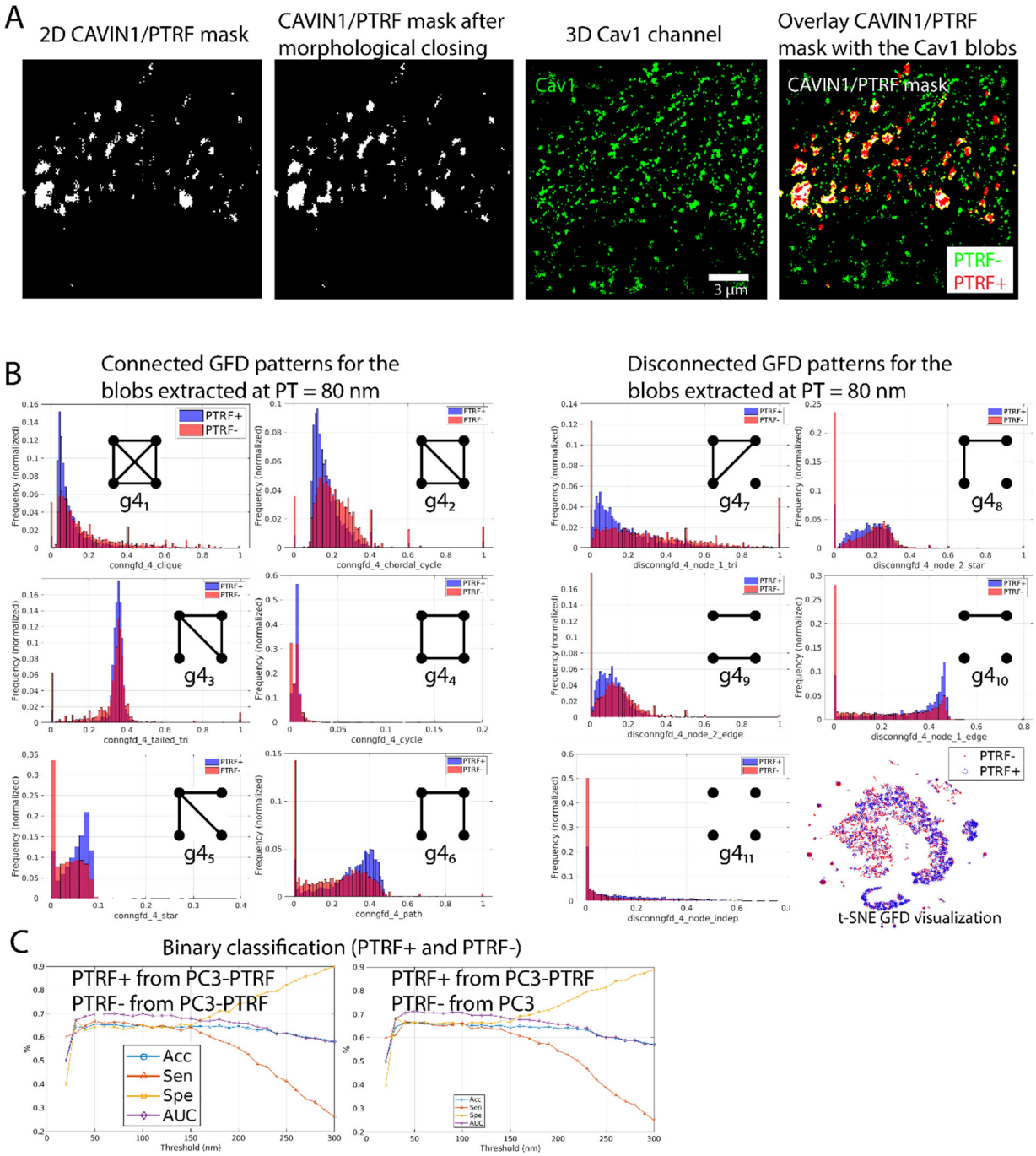


Figure 4. Classification results using GFD features at various PTs for PTRF+ and PTRF- blobs. **A.** A wide-field CAVIN1/PTRF mask (white) is used to label blobs as PTRF-positive (PTRF+) or CAVIN1/PTRF-negative (PTRF-) (supervised labeling) or to assign the learned and matched groups as S2 scaffolds (PP1), caveolae (PP2), S1B scaffolds (PP3) and S1A scaffolds (PP4) blobs (unsupervised labeling) as in **Figure 1D**. **B.** GFD features for the blobs when using the CAVIN1/PTRF mask to label the blobs into PTRF+ and PTRF- are shown as histograms of connected and disconnected GFD features extracted from the networks at PT=80 nm. t-SNE visualization of the GFD features shows that the blobs are not well separated when using the

CAVIN1/PTRF mask to label the blobs. **C.** We used the CAVIN1/PTRF mask to label the blobs as PTRF+ and PTRF- then applied machine learning classification after extracting GFD features. We show the classification results for the binary classification task by reporting the accuracy (Acc), sensitivity (Sen), specificity (Spe), and area under ROC curve (AUC).

CAVIN1/PTRF is required for caveolae formation and known to associated with caveolae (Hansen et al., 2013, Hill et al., 2008, Liu and Pilch, 2008). Caveolae are larger than scaffolds (Khater et al., 2018) and volume and number of nodes (molecular localizations) analysis of the four classes of blobs defined by unsupervised learning identified a clear size demarcation between caveolae (PP2) and larger S2 scaffolds (PP1) at $2.48x10^6$ nm^3 and 60 nodes (Fig. 5A). This is consistent with Khater et al, (Khater et al., 2018), where we reported that caveolae contained 109±52 predicted molecular localizations and had a minimum number of localizations of ≅ 60. Correlation between number of nodes and volume showed that 60 nodes clearly segregated caveolae (PP2) (blue) domains from the three scaffold domains and that $2.48x10^6$ nm^3 selected for the vast majority of caveolae but included some S2 scaffolds (green). Large PTRF+ blobs ($\geq 2.48x10^6$ nm^3) showed high classification accuracy with PTRF- blobs from PC3-PTRF cells as well with all Cav1 blobs from PC3 cells, that lack CAVIN1/PTRF and caveolae (Fig. 5B). Similarly, PTRF+ ≥60 node blobs showed high classification accuracy (i.e. >90%) with PTRF- blobs from PC3-PTRF cells and PC3 blobs; importantly, PTRF+ ≥60 node blobs were accurately classified as distinct from ≥60 node blobs from PC3 cells highlighting that GFD classification is not based on cluster size. PTRF+ <60 node clusters showed poor classification accuracy (i.e. <65%) with PTRF- nodes. The improved classification accuracy due to volume and number of node stratification is more clearly shown in the t-SNE plot (Fig. 5B) and the graph analysis of average classification accuracy over PT 20-300 nm (Fig. 5C). These results suggest that while association with the mask

effectively identifies large caveolae, smaller Cav1 clusters (i.e. scaffolds) cannot be distinguished based on their association with the CAVIN1/PTRF mask.

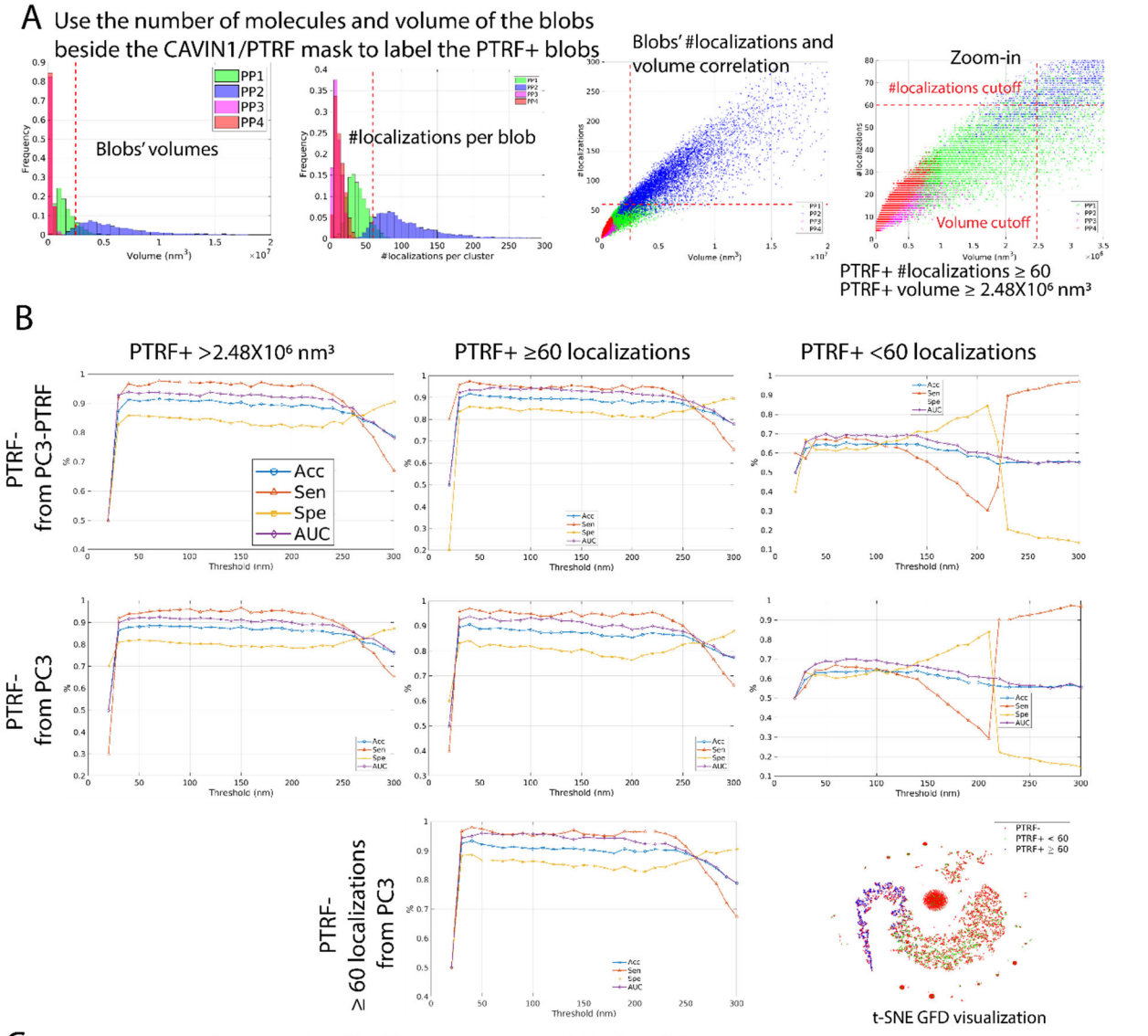
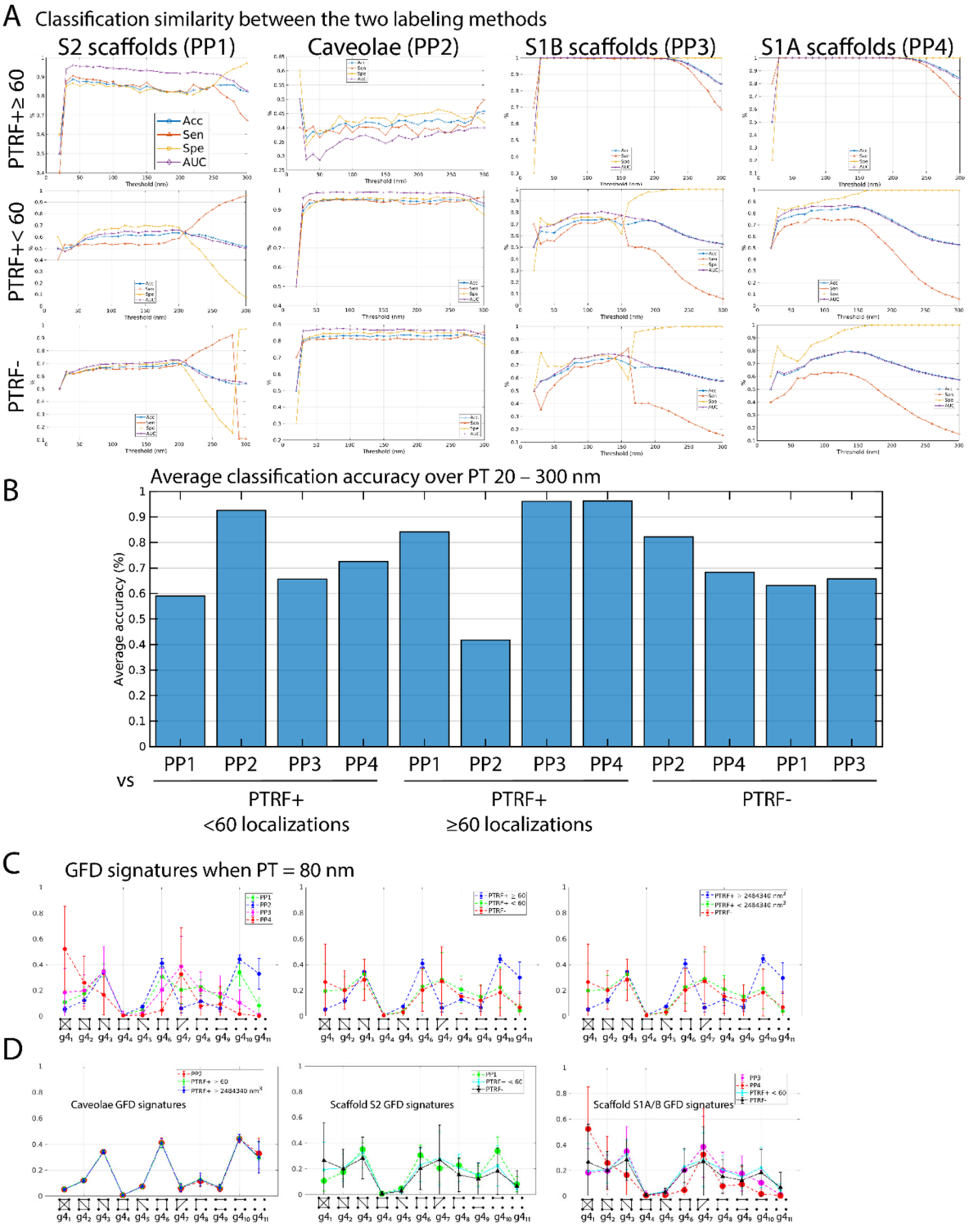


Figure 5. Application of other learned features (number of localizations, volume) to improve blob identification using the CAVIN1/PTRF mask. A. The volume and number of localizations for the blobs labeled using the unsupervised labeling method (Khater et al., 2018). The linear relationship (with 94% correlation) between the volume and number of localizations features shows that either can be used along with CAVIN1/PTRF mask to identify PTRF+ blobs. **B.** Using the CAVIN1/PTRF mask to further label PTRF+ blobs, based on either number of localizations (PTRF+ ≥60 and PTRF+ <60) or volume (PTRF+ $\geq 2.48\times10^6$ nm^3 and PTRF+ $<2.48\times10^6$ nm^3) and then, extracting GFD features for machine learning classification to automatically classify the blobs at various PTs. **C.** The bar plot shows average classification accuracy of the blobs over the used PTs for the binary classifiers in (**B**) above. Number of localizations and volume cut-offs improves blob identification.

We then compared classification similarity between the supervised and unsupervised learning approaches (Fig. 6). Caveolae (PP2) showed low classification accuracy (i.e. <45%) with the large PTRF+ ≥60 nodes group and high classification with small PTRF+ and PTRF- clusters. Conversely, S2 (PP1), S1B (PP3) and S1A (PP4) scaffolds showed high classification accuracy with large PTRF+ clusters and reduced classification accuracy with small PTRF+ and PTRF- clusters (Fig. 6 A,B). Plotting GFD for each graphlet shows excellent matching for caveolae GFD signatures from PP2 and PTRF+ blobs greater than $2.48x10^6$ nm^3 or greater than 60 nodes (Fig. 6 C,D). Small PTRF+ blobs (<60 nodes) closely match PTRF- blobs and show significant reduction in one connected and two disconnected graphlets. Similar analysis of the unsupervised learning groups shows that larger hemispherical S2 scaffolds (PP1) are most similar to caveolae (PP2) and that the smaller S1B (PP3) and S1A (PP4) scaffolds most closely match the PTRF- group (Fig. 6 C,D).



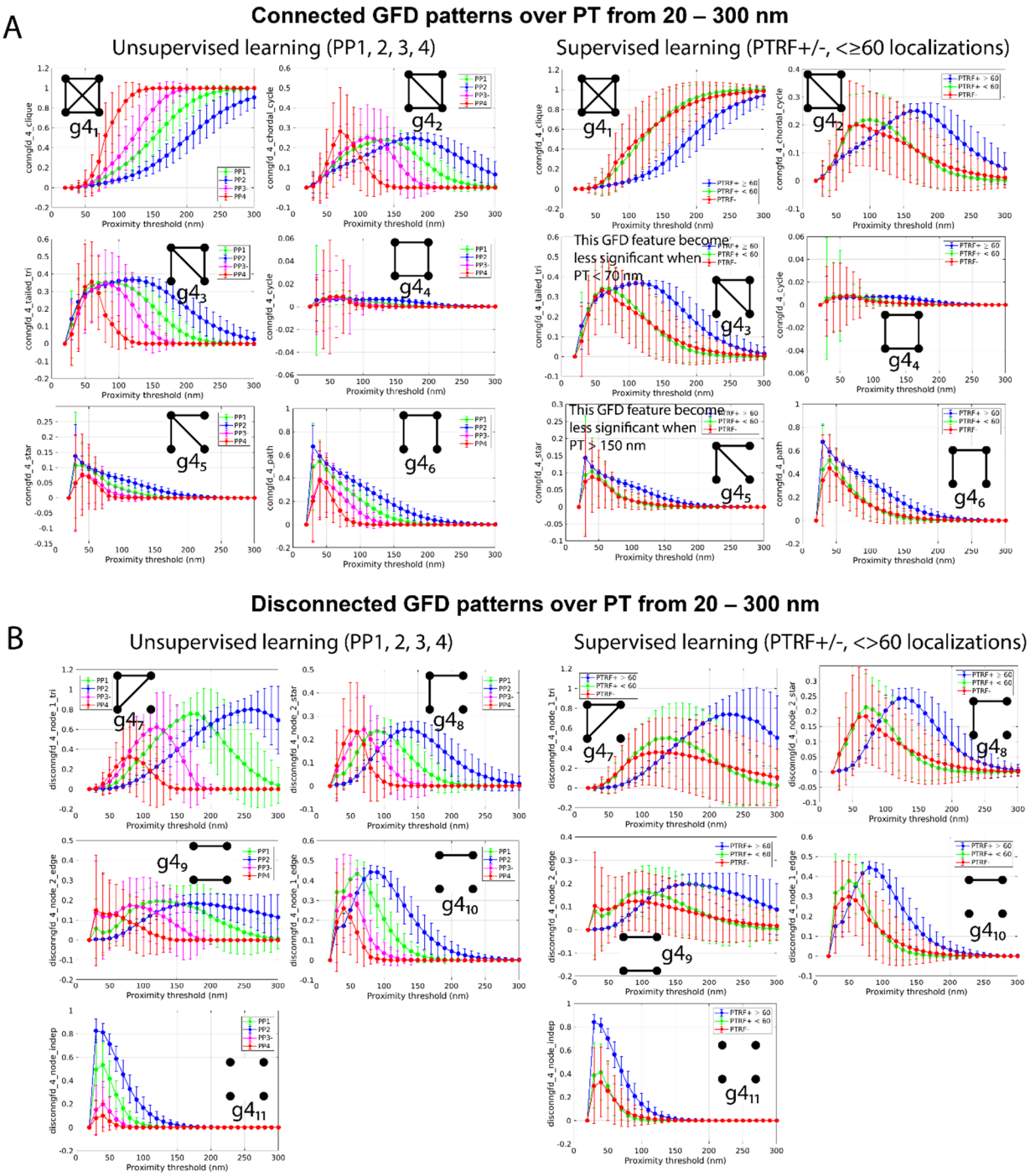
The PTRF+ ≥ 60 and PP2 groups have almost identical GFD signatures

The PTRF+ < 60 and PTRF- groups have similar GFD signatures

The PTRF+ < 60 and PTRF- groups can have more than one sub-category

Figure 6. Using GFD biosignatures for machine learning classification of the Cav1 blobs. A. Identifying the Cav1 domains via classification of all combinations of blob groups using unsupervised learning-based blob labels {PP1, PP2, PP3, PP4} or CAVIN1/PTRF mask-based supervised learning {PTRF+ ≥60, PTRF+ <60, PTRF-}. The classification results show similar and dissimilar groups based on extracted GFD features at various PTs. For example, low classification accuracy of PP2 and PTRF+≥60 suggests that the GFD features of these two groups are very similar and represent the same group of Cav1 domains (i.e. caveolae). **B.** Bar plot shows average classification accuracy of blobs over the used PTs for the binary classifiers in (**A**) above. **C.** GFD signatures for the blobs at PT=80 nm using the different labeling methods. Using either number of localizations or volume cut-offs to label PTRF+ blobs leads to similar GFD biosignatures. **D.** Depicting GFD biosignatures for blobs from the various groups shows that the similar Cav1 domains have similar feature trends. For example, PP2 blobs have almost identical features to PTRF+≥60 and PTRF+≥2.48×10^6 nm^3 and correspond to caveolae. GFD feature trends represents a biosignature for each Cav1 domain.

Caveolae show a high frequency of 4-path ($g4_6$) graphlets while small scaffolds show increased frequency of 4-clique ($g4_1$) and 4-chordal-cycle ($g4_2$) at low PTs that increase for all domains with increasing PT in relation to blob size (Fig. 7A). Conversely, 4-node-independent are more frequent in large blobs at low PT and disappear from all blobs at high PT values. This highlights the high interconnectivity of nodes within Cav1 domains (Fig. 7B). Interestingly, all groups show a similar frequency of 4-tailed triangle ($g4_3$) at low PT whose decreasing frequency at higher PTs corresponds to blob size. Further, 4-cycle graphlets are absent from all Cav1 domains (Fig. 7A). In contrast, we observe a high frequency (>60%) of 4-node-1-triangle ($g4_7$) disconnected and 4-tailed-triangle ($g4_3$) connected graphlet patterns in all Cav1 domains that vary in proportion to blob size.



The GFD signatures for caveolae are always different than the rest of the groups in almost all the PTs

Figure 7. GFD discriminatory features for the different Cav1 blobs over various PTs. The change in GFD features over various PTs using two different labeling strategies for the Cav1 blobs. The mean and standard deviation of (**A**) connected and (**B**) disconnected features have different values based on the used PT. Some GFD features are not informative and unchanged with different PTs (e.g. $g4_4$). Some features show that the similar groups have the same trends over the different PTs (e.g. PTRF- and PTRF+ <60).

DISCUSSION

We show here that graphlet analysis of single molecule super-resolution data can identify and discriminate Cav1 domains. Studying the frequencies of the various graphlet patterns in the different Cav1 domains' networks represent compressed descriptors to training the machine learning approaches to identifying the biological structures automatically and effectively. The different Cav1 domains have distinct GFD patterns and features suggestive of altered molecular stoichiometry within these biological structures. Application of GFD analysis to SMLM data requires reduction of the millions of blinks obtained by SMLM to obtain representative molecular localizations. Each Cav1 molecule can be labeled by multiple primary and secondary antibodies, with the latter carrying multiple fluorophores. To do this, we merge blinks derived from individual Cav1 proteins by iteratively combining blinks within a defined 20 nm distance, generating a predicted Cav1 localization (Khater et al., 2018). Choice of merge threshold will necessarily impact the number of predicted localizations studied and use of a fixed 20 nm threshold will not necessarily take into consideration the reduced axial resolution of cylindrical lens-based 3D SMLM. Nevertheless, application of a 20 nm to the SMLM data set obtained from PC3-PTRF cells effectively identified and distinguished caveolae from smaller scaffold structures (Khater et al., 2018). Importantly, the caveolae identified showed structural similarities to those predicted by cryoEM (Stoeber et al., 2016, Ludwig et al., 2016).

The current study extended that work to include a wide-field mask for CAVIN1/PTRF, a required adaptor for caveolae formation (Hill et al., 2008). The demonstration here that GFD-based machine learning classifies caveolae Cav1 blobs with large (>60 nodes) CAVIN1/PTRF-positive (i.e.

PTRF+) blobs confirms that the 3D SMLM Network Analysis pipeline, as applied here, can enable structural analysis of Cav1 domains. The low classification accuracy of PTRF+ and PTRF- blobs contrasts the high classification accuracy of PP1, PP2, PP3 and PP4 groups. Classification accuracy improved upon stratification of the PTRF+ blobs based on size, either number of nodes or volume. This suggests that smaller S1 and/or S2 scaffolds also interact with the wide field CAVIN1/PTRF mask. While resolution of the wide-field CAVIN1/PTRF mask is significantly reduced relative to the SMLM-characterized Cav1 blobs, this data suggests that CAVIN1/PTRF is not exclusively associated with caveolae, but may also interact with Cav1 scaffolds.

High classification sensitivity (or high true positive rate) (Fig. 3B, 5B) means that the classifier correctly identifies the PTRF+ blobs, while at the same time lower specificity (true negative rate) means that PTRF- blobs are harder to identify and the majority of the mis-classified blobs are from negative blobs. The specificity and sensitivity show that positive blobs have less mis-predictions. The models with high sensitivity and low specificity are good for catching actual positive cases.

Machine learning-based classification of the various Cav1 domains with graphlet analysis helped us to: 1) identify the Cav1 domains and draw their biosignatures; and 2) find the best range of PTs to identify significant molecular interactions (classification loss and mis-prediction is <10%) in each of the Cav1 domains. Classification accuracy is reduced at PTs below 50 and above 150 nm due to the less discriminatory GFD features at these ranges of PTs (Fig. S1C). This suggests that molecular interaction should not be studied at an arbitrary spatial scale since there is a critical range of spatial distances within which class-specific molecular interaction occurs.

Frequency of some molecular interaction patterns in certain structures distinguishes them from other structures (i.e. every biological structure type has unique patterns of its molecular interactions). For instance, the highly connected 4-clique ($g4_1$) and 4-chordal-cycle ($g4_2$) GFD patterns (Fig. 3A, Fig. 7A) have higher frequencies in both S1A and S1B scaffolds compared with S2 scaffold and caveolae. At the same time, the less highly connected 3-star ($g4_5$) and 4-path ($g4_6$) patterns have lower frequencies in both S1A and S1B scaffolds compared with S2 scaffolds and caveolae. The high frequency of 4-clique ($g4_1$) patterns is indicative of saturation of molecular interactions and suggest that S1A and S1B scaffolds contain a higher frequency of stable patterns of strong intermolecular interactions. Indeed, S1A scaffolds correspond to SDS-resistant Cav1 homo-oligomers detected biochemically (Monier et al., 1995, Sargiacomo et al., 1995). Interestingly, the 4-cycle ($g4_4$) closed chain pattern is essentially absent in all Cav1 domains at all proximity thresholds (Fig. 7A) suggesting that a planar square pattern of molecular interaction is not associated with Cav1 domains. This is consistent with the reported presence of regular hexagonal profiles in the polyhedral cage of the Cav1 caveolar coat observed by cryoEM (Ludwig et al., 2016).

The t-SNE visualization in Figures 3A and 5B show that the projected GFD features depict more clustering behaviour amongst the blobs from similar classes. The separation of the blob classes in Figure 3A is better than Figure 5B and much better than Figure 4B. Also, it shows that the clustering behaviour of the PTRF+ blobs is much clearer after stratification based on the number of molecular localizations (Fig. 5B). The overlap of the GFD features for some of the PTRF+ $\geq$60 blobs with some of the PTRF- blobs may be due to the commonality of GFD features that appear in blobs from the different Cav1 domains.

This study represents the novel application of graphlet analysis to point clouds generated from SMLM data sets. Graphlets represent features that can be used to classify biological structures, define biosignatures associated with specific biological structures and identify molecular interaction motifs associated with specific biological structures. Extension of this approach to biological structures other than Cav1 domains is warranted.

COMPETING FINANCIAL INTEREST

An international patent PCT/CA2018/051553 covering the material presented in the manuscript has been submitted by the authors: "Methods for Analysis of Single Molecule Localization Microscopy to Define Molecular Architecture", US Patent Application No. 62/594,642, Dec 5, 2018.

ACKNOWLEDGMENTS

Supported by grants from the CIHR (PJT-156424, PJT-159845), NSERC, Prostate Cancer Canada and CFI/BCKDF (IRN, GH) and a China Scholarship Council doctoral fellowship (FM). We thank Dr. Keng Chou (Chemistry, UBC) for helpful comments and discussion.

REFERENCES

AHMED, N. K., NEVILLE, J., ROSSI, R. A. & DUFFIELD, N. Efficient graphlet counting for large networks. 2015 2015. IEEE, 1-10.

AHMED, N. K., NEVILLE, J., ROSSI, R. A., DUFFIELD, N. G. & WILLKE, T. L. 2017. Graphlet decomposition: Framework, algorithms, and applications. *Knowledge and Information Systems,* 50**,** 689-722.

BASSETT, D. S. & SPORNS, O. 2017. Network neuroscience. *Nature neuroscience,* 20**,** 353.

BENSON, A. R., GLEICH, D. F. & LESKOVEC, J. 2016. Higher-order organization of complex networks. *Science,* 353**,** 163-166.

BRESSAN, M., CHIERICHETTI, F., KUMAR, R., LEUCCI, S. & PANCONESI, A. Counting graphlets: Space vs time. 2017 2017. ACM, 557-566.
BROWN, C. J., MILLER, S. P., BOOTH, B. G., ANDREWS, S., CHAU, V., POSKITT, K. J. & HAMARNEH, G. 2014. Structural network analysis of brain development in young preterm neonates. *Neuroimage,* 101**,** 667-680.
COSTA, L. D. F., RODRIGUES, F. A. & CRISTINO, A. S. 2008. Complex networks: the key to systems biology. *Genetics and Molecular Biology,* 31**,** 591-601.
DUTTA, A. & SAHBI, H. 2017. High order stochastic graphlet embedding for graph-based pattern recognition. *arXiv preprint arXiv:1702.00156*.
GUMHOLD, S., WANG, X. & MACLEOD, R. S. Feature Extraction From Point Clouds. 2001 2001.
HANSEN, C. G., SHVETS, E., HOWARD, G., RIENTO, K. & NICHOLS, B. J. 2013. Deletion of cavin genes reveals tissue-specific mechanisms for morphogenesis of endothelial caveolae. *Nat Commun,* 4**,** 1831.
HILL, M. M., BASTIANI, M., LUETTERFORST, R., KIRKHAM, M., KIRKHAM, A., NIXON, S. J., WALSER, P., ABANKWA, D., OORSCHOT, V. M., MARTIN, S., HANCOCK, J. F. & PARTON, R. G. 2008. PTRF-Cavin, a conserved cytoplasmic protein required for caveola formation and function. *Cell,* 132**,** 113-24.
HOČEVAR, T. & DEMŠAR, J. 2014. A combinatorial approach to graphlet counting. *Bioinformatics,* 30**,** 559-565.
KHATER, I. M., MENG, F., WONG, T. H., NABI, I. R. & HAMARNEH, G. 2018. Super Resolution Network Analysis Defines the Molecular Architecture of Caveolae and Caveolin-1 Scaffolds. *Scientific Reports,* 8**,** 9009.
KOLLIAS, G., MOHAMMADI, S. & GRAMA, A. 2012. Network similarity decomposition (nsd): A fast and scalable approach to network alignment. *IEEE Transactions on Knowledge and Data Engineering,* 24**,** 2232-2243.
KRIVITSKY, P. N., HANDCOCK, M. S., RAFTERY, A. E. & HOFF, P. D. 2009. Representing degree distributions, clustering, and homophily in social networks with latent cluster random effects models. *Social networks,* 31**,** 204-213.
LIU, L. & PILCH, P. F. 2008. A critical role of cavin (polymerase I and transcript release factor) in caveolae formation and organization. *J Biol Chem,* 283**,** 4314-22.
LUDWIG, A., NICHOLS, B. J. & SANDIN, S. 2016. Architecture of the caveolar coat complex. *J Cell Sci,* 129**,** 3077-83.
MAATEN, L. V. D. & HINTON, G. 2008. Visualizing data using t-SNE. *Journal of machine learning research,* 9**,** 2579-2605.
MARCUS, D. & SHAVITT, Y. 2012. Rage–a rapid graphlet enumerator for large networks. *Computer Networks,* 56**,** 810-819.
MILO, R., SHEN-ORR, S., ITZKOVITZ, S., KASHTAN, N., CHKLOVSKII, D. & ALON, U. 2002. Network motifs: simple building blocks of complex networks. *Science,* 298**,** 824-827.
MONIER, S., PARTON, R. G., VOGEL, F., BEHLKE, J., HENSKE, A. & KURZCHALIA, T. V. 1995. VIP21-caveolin, a membrane protein constituent of the caveolar coat, oligomerizes in vivo and in vitro. *Molecular biology of the cell,* 6**,** 911-927.
NEWMAN, M. E. J. 2003. The structure and function of complex networks. *SIAM review,* 45**,** 167-256.

PRŽULJ, N. 2007. Biological network comparison using graphlet degree distribution. *Bioinformatics,* 23**,** e177-e183.
PRŽULJ, N., CORNEIL, D. G. & JURISICA, I. 2004. Modeling interactome: scale-free or geometric? *Bioinformatics,* 20**,** 3508-3515.
QI, C. R., SU, H., MO, K. & GUIBAS, L. J. 2017. Pointnet: Deep learning on point sets for 3d classification and segmentation. *Proc. Computer Vision and Pattern Recognition (CVPR), IEEE,* 1**,** 4.
SARGIACOMO, M., SCHERER, P. E., TANG, Z., KÜBLER, E., SONG, K. S., SANDERS, M. C. & LISANTI, M. P. 1995. Oligomeric structure of caveolin: implications for caveolae membrane organization. *Proceedings of the National Academy of Sciences,* 92**,** 9407-9411.
SHERVASHIDZE, N., VISHWANATHAN, S. V. N., PETRI, T., MEHLHORN, K. & BORGWARDT, K. Efficient graphlet kernels for large graph comparison. 2009 2009. 488-495.
SHIN, K., ELIASSI-RAD, T. & FALOUTSOS, C. CoreScope: Graph Mining Using k-Core Analysis—Patterns, Anomalies and Algorithms. 2016 2016. IEEE, 469-478.
STOEBER, M., SCHELLENBERGER, P., SIEBERT, C. A., LEYRAT, C., HELENIUS, A. & GRÜNEWALD, K. 2016. Model for the architecture of caveolae based on a flexible, net-like assembly of Cavin1 and Caveolin discs. *Proceedings of the National Academy of Sciences,* 113**,** E8069-E8078.
YIN, H., BENSON, A. R., LESKOVEC, J. & GLEICH, D. F. Local higher-order graph clustering. 2017 2017. ACM, 555-564.
ZHANG, L., HAN, Y., YANG, Y., SONG, M., YAN, S. & TIAN, Q. 2013. Discovering discriminative graphlets for aerial image categories recognition. *IEEE Transactions on Image Processing,* 22**,** 5071-5084.
ZITNIK, M. & LESKOVEC, J. 2017. Predicting multicellular function through multi-layer tissue networks. *Bioinformatics,* 33**,** i190-i198.

Supplementary information

Identification of Caveolin-1 Domain Signatures via Graphlet Analysis of Single Molecule Super-Resolution Data

Ismail M. Khater[1], **Fanrui Meng**[2], **Ivan Robert Nabi**[2*#], **Ghassan Hamarneh**[1*]

[1]Medical Image Analysis Lab, School of Computing Science, Simon Fraser University, Burnaby, BC V5A 1S6, Canada

[2]Department of Cellular and Physiological Sciences, Life Sciences Institute, University of British Columbia, Vancouver, BC V6T 1Z3, Canada

*Correspondence: irnabi@mail.ubc.ca (IRN), hamarneh@sfu.ca (GH)

*Equal contribution

[#]Lead Contact

MATERIAL AND METHODS

Cell culture, immunofluorescent labeling and SMLM image acquisition

PC3-PTRF cells were cultured in RPMI-1640 medium (Thermo-Fisher Scientific Inc.) complemented with 10% fetal bovine serum (FBS, Thermo-Fisher Scientific Inc.) and 2 mM L-Glutamine (Thermo-Fisher Scientific Inc.). Cells were fixed with 3% paraformaldehyde (PFA) permeabilized with 0.2% Triton X-100 in PBS/CM, and blocked with PBS/CM containing 10% goat serum (Sigma-Aldrich Inc.) 1% bovine serum albumin (BSA; Sigma-Aldrich Inc.) prior to labeling with the primary polyclonal rabbit anti-caveolin-1 (BD Transduction Labs Inc.; Cat. No. 610060) and mouse monoclonal anti-CAVIN1/PTRF (BD Transduction Labs Inc; Cat. No. 611258) antibodies for >12 h at 4°C followed by secondary antibody (Alexa Fluor 647-conjugated goat anti-rabbit; Alexa488-conjugated goat anti-mouse (Thermo-Fisher Scientific Inc.) for 1 h at room temperature, and pot-fixed with 3% PFA, as described previously (Khater et al., 2018).

GSD super-resolution imaging of 10 PC3-PTRF and 10 PC3 cells was performed on a Leica SR GSD 3D system using a 160x objective lens (HC PL APO 160x/1.43, oil immersion), a 642 nm laser line and a EMCCD camera (iXon Ultra, Andor), as described previously (Khater et al., 2018). The GSD event list was exported and processed by merge analysis at 20 nm, filtering, network analysis to identify Cav1 clusters and machine learning to define Cav1 clusters (PP1, PP2, PP3, PP4). A wide-field TIRF image of Alexa488-labeled CAVIN1/PTRF was acquired for each of 10 cells in parallel and used as a mask to identify Cav1 clusters. We processed the CAVIN1/PTRF mask by applying the morphological closing operation to rub off the small holes in the mask and then superimposed on the segmented Cav1 blobs to identify Cav1 blobs touching the mask as “in mask” and the others as “out mask”.

Data

- 10 SMLM prostate cancer (PC3) cells (14,491 blobs)
- 10 SMLM CAVIN1/PTRF-transfected PC3 (PC3-PTRF) cells (10,866 blobs)
- 10 widefield TIRF CAVIN1/PTRF masks that correspond the cells from PC3-PTRF population.

Table S1: Dataset details after pre-processing. It shows the number of extracted blobs from each population.

	PTRF+	**PTRF-**	**Total**
PC3	-	14,491	14,491
PC3-PTRF	2136	8,730	10,866
PC3-PTRF (≥60)	857	10,009	10,866
PC3-PTRF (<60)	1,279	9,587	10,866

Graphlet feature extraction

Given a cell with N blobs, let $blob_i$ be the i-th blob with 3D coordinates (X_i Y_i Z_i) of molecule localizations. $|blob_i| = l_i \times 3$, where l_i is the number of molecules in $blob_i$. The input to the network construction of $blob_i$ are (X_i Y_i Z_i) and a set of proximity thresholds $\{PT_1, PT_2,\ldots,PT_T\}$ and the output are $blob_i$'s networks $\{G_i^1, G_i^2, \ldots, G_i^T\}$ at the different thresholds, where G_i^t is composed of a set of nodes V_i and edges E_i^t, i.e. $G_i^t = (V_i, E_i^t)$. V_i represents the molecules of $blob_i$ and E_i^t is the set of edges connecting all pairs of molecules interacting within PT_t nm.

We extract D graphlet features for each of the resulting $N \times T$ networks from all the blobs at all thresholds using parallel graphlet decomposition (Ahmed et al., 2015, Ahmed et al., 2017). We then identify blob biosignatures by extracting the most discriminative graphlet features and corresponding proximity thresholds that best discriminates between the blob classes. Parallel graphlet decomposition extracts three kinds of features: The macro (global) motif counts for the motifs with $k = 2, 3,$ *and* 4 nodes; the micro motif counts (extracted per each edge in the network); and the Graphlet Frequency Distribution (GFD) features. Given that the micro and macro features

are dependent on the number of nodes in the network and since the number of molecules per blob (l_i) varies across the blobs, to avoid biasing our classification model to rely on number of molecules, we based our analysis on the GFD features (with $k = 4$ nodes), since GFDs capture relative values (Fig. 1C)

The 4-node graphlets have 11 patterns (Fig. 1C) that can be arranged into connected and disconnected graphlets to extract GFD features (Rossi et al., 2017). The connected graphlets have 6 features ($g4_1$, $g4_2$, $g4_3$, $g4_4$, $g4_5$, $g4_6$) and disconnected graphlets have 5 features ($g4_7$, $g4_8$, $g4_9$, $g4_{10}$, $g4_{11}$). The GFD features can be extracted for the connected graphlets alone, the disconnected graphlets alone, and for both the connected and disconnected graphlets. For example, the GFD for the connected graphlet $g4_1$ can be calculated by using the equation $GFD_{g4_1} = \frac{count(g4_1)}{\sum_{i=1}^{6} count(g4_i)}$, where $count(g4_1)$ is the total number of recurrences of the pattern $g4_1$ in the network. That is, the GFD finds the frequency of a given pattern in a network relative to the other patterns.

Machine learning classification of the blobs

The supervised machine learning task (classification) requires the feature vector for every blob as well as its class (ground truth GT label). The blobs classes were obtained using two different methods. Let *Y1* be the classes obtained using the first blob's labelling method and *Y2* are the classes when using the second labelling method. $|Y1| = |Y2| = n$, $y_{i1} \in$ [*PTRF+, PTRF-*] and $y_{i2} \in$ [*PP1, PP2, PP3, PP4*]. Where y_i is the class label for $blob_i$. For the PC3 population, the blobs can take only one label, PTRF-.

Our goals from using the classification model are: 1) to automatically identify the blobs, 2) to build a learning model that can be used to label the blobs of unknown samples, and 3) to show that the

blobs have differences and use the extracted features to identify the different blobs (Cav1 domains) signatures.

In all the classification experiments, we used the 10-fold cross validation strategy. We used sampling to obtain a balanced number of blobs for the classification task. For example, the total number of positive blobs is 857. The same number is sampled from the negative blobs to have a balanced dataset.

We divided the binary classification into the following based on the source of the extracted blobs (Fig. 4A,C):

1) Classifying the blobs from PC3-PTRF data (positive VS negative blobs taken from the same population).
2) Classifying the blobs from both populations. Positive blobs taken from PC3-PTRF and the negative blobs taken from PC3. Where the PC3 population have pure PTRF- blobs.
3) Classifying the blobs from both populations positive blobs taken from PC3-PTRF and the negative blobs taken from PC3 (Fig. 5B), where the minimum number of molecule per blob chosen to be ≥ 60 (excluding all the small blobs). This experiment shows that the graphlets features are very good features to discriminate between the blobs classes even when the number of molecules in the negative class is the same as the number of molecules in the positive class. Which means that there are intrinsic properties in the negative features that made the classifier discriminate them using the subtle graphlets features.

We used the random forest classifier to classify the blobs based on the GFD features. The classification accuracy, specificity, and the sensitivity are reported at each proximity threshold.

Multi-class Classification

Sensitivity and specificity are generally used to measure the performance of the binary classifiers. We find the sensitivity and the specificity measures for the multi-class classification from the confusion matrix (Fig. S1A) as $Sensitivity = \frac{TP}{(TP+FN)}$ and $Specificity = \frac{TN}{(TN+FP)}$ and calculate the values per class as seen in the confusion matrix (Fig. S1A), where it shows an example of finding the evaluation measures for PP3 class. Where, the TP is the true positive, FN is false negative, TN is true negative, and FP is the false positive. For PP3 class, $FP_{PP3} = E_{PP1PP3} + E_{PP2PP3} + E_{PP4PP3})$, and $FN_{PP3} = E_{PP3PP1} + E_{PP3PP2} + E_{PP3PP4})$, $TN_{PP3} = all\ excluding\ TP_{PP3}, FP_{PP3},\ and\ FN_{PP3}$. The same idea applied to calculate the sensitivity and specificity for the other classes (i.e. PP1, PP2, and PP4). The accuracy measure for the multi-class classification is calculated using the correctly classified blobs divided by the total number of blobs. Formally, using the confusion matrix (Fig. S1A), $Accuracy = \frac{TP_{PP1}+TP_{PP2}+TP_{PP3}+TP_{PP4}}{Total\ number\ of\ blobs}$.

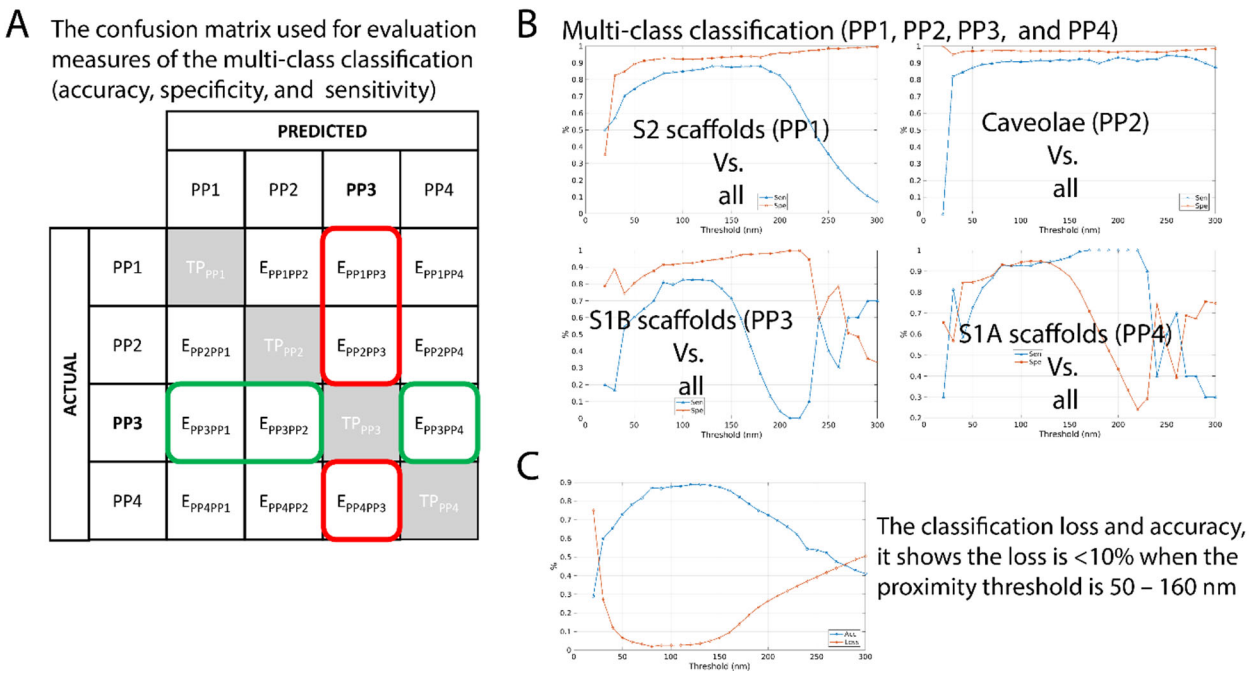


Figure S1. Multi-class classification results. A. The confusion matrix used for the classification performance calculations for the multi-class classification task. **B.** The result of

classifying the multi-labeled blobs using one-versus-all strategy. The specificity and sensitivity results were reported when using one class versus all by using the confusion matrix in (**A**). **C.** The results of the multi-class classification accuracy when using a voting scheme. It shows that the best range of PTs to identify the different Cav1 domains is 50 – 160 nm.

REFERENCES

AHMED, N. K., NEVILLE, J., ROSSI, R. A. & DUFFIELD, N. Efficient graphlet counting for large networks. 2015 2015. IEEE, 1-10.

AHMED, N. K., NEVILLE, J., ROSSI, R. A., DUFFIELD, N. G. & WILLKE, T. L. 2017. Graphlet decomposition: Framework, algorithms, and applications. *Knowledge and Information Systems,* 50**,** 689-722.

KHATER, I. M., MENG, F., WONG, T. H., NABI, I. R. & HAMARNEH, G. 2018. Super Resolution Network Analysis Defines the Molecular Architecture of Caveolae and Caveolin-1 Scaffolds. *Scientific Reports,* 8**,** 9009.

ROSSI, R. A., ZHOU, R. & AHMED, N. K. 2017. Estimation of graphlet statistics. *arXiv preprint arXiv:1701.01772*.