S1 Supplementary Materials

S1.1 Experimental Methods

For details, see [2] and associated Supplementary Materials.

S1.2 Cluster identification

Our approach [2] is a version of hierarchical distance based clustering [3]. It is based on the method of Espinoza and coworkers [1] who first adapted it to the analysis of nano-gold labeled membrane proteins.

Hierarchical clustering relies on a single length parameter L, which induces a pattern of connections between the points of a given set. Two points are directly connected if their distance is less than L. Connected groups result by transitivity, similarly to connected graphs: a point that is directly connected to a member of a connected group is indirectly connected to all members of the group. The connected groups amount to a partition of the point set and define the clusters. The number of clusters induced in a given image decreases as L increases. Espinoza and coworkers investigated an optimal length parameter L^* for membrane bound receptors and proposed an intrinsic clustering distance d_I, the value of L that maximizes the number of clusters containing at least two points[1].

We calculated the number of clusters as a function of the length parameter, $N_C(L)$ for each image. This is a decreasing staircase function that equals the number of points when L is small and approaches 1 for large L. For a set of randomly distributed points, $N_C(L)$ follows a universal curve (blue line in Fig. S1B). Similarly to nearest neighbor distributions, the short distance portion of $N_C(L)$ is consistent with having the same number of particles distributed randomly and uniformly in a smaller area (Fig. S1B confined particles dashed line). This smaller area is comparable to the total area occupied only by clusters. If there was a perfect **scale separation** between the distribution of clusters and that of particles within clusters, the $N_C(L)$ curve would have a **plateau** (see Fig. S1B red curve), and the spatial distribution of particles within clusters would be consistent with random placement. We therefore choose the theoretical optimal value of L to correspond to the length scale where the plateau starts. Empirically, we identify the optimal L as two times the value of L (the factor of two adds a margin of surety) corresponding to the point where the second derivative of $N_C(L)$ is zero.

Domain reconstruction algorithm and cluster analysis

For each identified cluster, we constructed a geometric footprint (shape) around the member points to provide us with an estimate of the area and perimeter for the cluster. This footprint can be identified with an underlying physical support, such as a lipid raft or aggregation of proteins.

We have computed the area and perimeter of cluster footprints / domains using the domain reconstruction algorithm in [4]. The process involves drawing a connected

[1] We use "optimal length parameter" (L^*) to distinguish from the specific choice of [1].

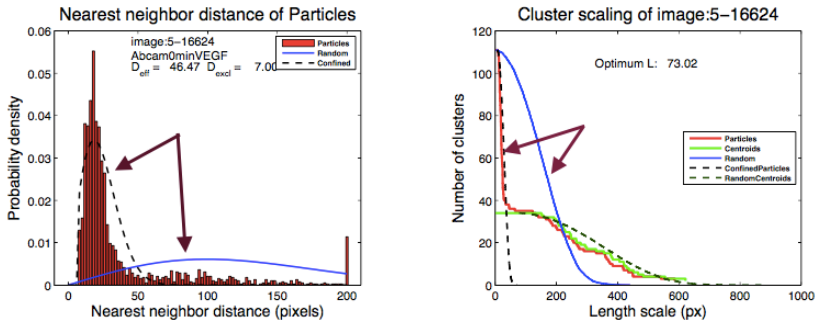

Fig. S1: The short distance portion of nearest neighbor distributions and cluster scaling curves are consistent with particles distributed randomly and uniformly in a smaller area, corresponding to the total cluster enclosed area.

graph containing all the points in a cluster; identification of an external contour on the graph; and defining the footprint by contour inflation. Fig. S2 illustrates in detail how the contours were determined.

Model for the origin of clusters

Domain hypothesis. The reasons for receptor clustering (in the absence of direct binding) are not completely understood, but it is generally accepted that it is a consequence of the physical features of the cell membrane, which influence the movement of receptors. Static images of receptors appear to show accumulations of receptors associated with small membrane regions of a distinct consistency, as illustrated in Figs. (1D,2A). The membrane is studded with small domains, typically a few to a few tens of nanometers across, which cover a significant but relatively small fraction of the membrane area. However, not all such regions contain receptors, and not all receptors or clusters are located in regions that are visually identifiable. When comparing with cluster size distributions, we assume that receptor clusters are actually groups of receptors localized in single microdomains. The distribution of the observed cluster sizes should then be consistent with placing a number of receptors into these pre-existing attractive domains.

Domain number, area and density enrichment factors estimated from cluster footprints.

All clusters of size 2+ identified with domains, all singles assumed free. In a first approximation we identify the area occupied by clusters of at least two particles with the putative attractive domains. Denote the corresponding footprint area $A_C^{(2+)}$, number of clusters, $N_C^{(2+)}$ and particles $N_P^{(2+)}$; the number of singleton particles is $N_S = N_P - N_P^{(2+)}$. The empirical estimates for the area fraction $f_A^{(\mathrm{exp})} = A_C^{(2+)}/A_{\mathrm{total}}$,

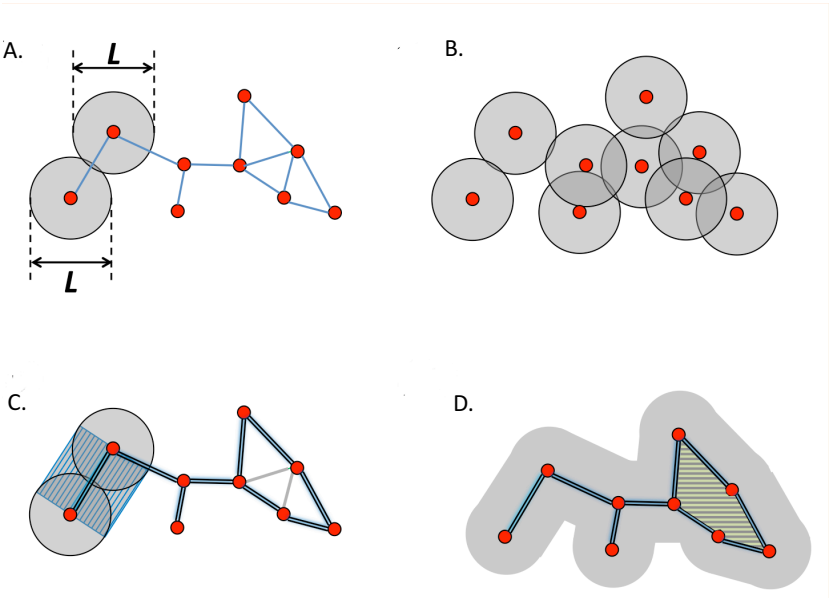

Fig. S2: Schematic of the domain reconstruction algorith (from [2]). (A) The points in a cluster form a connected graph, where edges connect points whose distance is less than the length scale L. Circles of diameter L centered on two points intersect if and only if the points are connected in the sense described above. (B) We want to define the footprint based on the union of all the circles of diameter L, centered on the points in the cluster. (C) We first identify the outer contour of the cluster graph (double blue shaded lines). We pad the area defined by the circles by creating rectangles along the edges of the contour graph. (D) The reconstructed region is the union of the inside of the contour graph (if any), and the circles and padding rectangles around the vertices and edges of the contour graph [4].

number of domains $N_D^{(\mathrm{exp})} = N_C^{(2+)}$ and population fraction $f_P^{(\mathrm{exp})} = N_P^{(2+)}/N_P$ can be used to estimate the attractiveness or enrichment factor for the domains, as the ratio between the density of particles inside domains versus in the rest of the observed area.

$$\alpha^{(\mathrm{exp})} = \frac{N_P^{(2+)}}{A_C^{(2+)}} \left(\frac{N_S}{A_{\mathrm{total}} - A_C^{(2+)}} \right)^{-1} = \frac{f_P^{(\mathrm{exp})}}{f_A^{(\mathrm{exp})}} \left(\frac{1 - f_P^{(\mathrm{exp})}}{1 - f_A^{(\mathrm{exp})}} \right)^{-1} \tag{S1}$$

Free particles and low occupancy domains. Confinement of particles is likely a stochastic process, where each of the N_D domains receive a number of the N_P particles. Some domains will be empty, and others will have a single particle. Particles in single occupancy domains are indistinguishable from free particles. A statistical model placing the N_{PD} confined particles into N_D boxes will predict the expected number of domains by occupancy k based on the total number of domains N_D and confined

particles N_{PD}, as $\left\langle n_{\mathrm{box}}^{(k)} \right\rangle = N_D \cdot \mathrm{PMF}_{\mathrm{box}}(k)$ based on the per-box probability mass function (PMF). The corresponding per-particle PMF provides the distribution of particles by clusters of occupancy k; the two PMFs are related by the expected number of particles by domain of occupancy k, $\left\langle n_{\mathrm{particle}}^{(k)} \right\rangle = k \left\langle n_{\mathrm{box}}^{(k)} \right\rangle$.

$$\mathrm{PMF}_{\mathrm{particle}}(k) = \frac{k \cdot N_D}{N_{PD}} \cdot \mathrm{PMF}_{\mathrm{box}}^{(k)}, \quad k = 1, 2, \cdots N_{PD}, \tag{S2}$$

In particular, the number of empty domains is $n_{\mathrm{empty}} = \left\langle n_{\mathrm{box}}^{(0)} \right\rangle = N_D \mathrm{PMF}_{\mathrm{box}}^{(0)}$ (no simple relation to the particle PMF), and the number of domains with a single particle is the same as the number of confined particles in single occupancy domains, $n_{\mathrm{singles}} = \left\langle n_{\mathrm{box}}^{(1)} \right\rangle$. The relation to experimentally observable quantities (number of single particles (confined or not), and number of clusters of size 2 or larger, as well as the number or particles in them is:

$$\begin{aligned} N_S &= N_P - N_{PD} + N_D \cdot \mathrm{PMF}_{\mathrm{box}}(1), \\ N_P^{(2+)} &= N_{PD} - N_D \cdot \mathrm{PMF}_{\mathrm{box}}(1), \\ N_D^{(2+)} &= N_D \left(1 - \mathrm{PMF}_{\mathrm{box}}(0) - \mathrm{PMF}_{\mathrm{box}}(1)\right) \end{aligned} \tag{S3}$$

Domain based model

Assume there are N_D domains that have approximately the same area and physical properties. We can infer the distribution of the number of particles in each domain, and the corresponding distribution of particle numbers. Since these models are expected to fit approximately, we prefer conceptually and mathematically simpler approaches. N_P particles in N_D enriched domains. The core idea is that a number N_P of particles are placed randomly into identical domains (sometimes referred to as "boxes"). From the perspective of one domain, each particle may fall into the domain with a probability $p_{\mathrm{bino}} = 1/N_D$. The probability that the domain receives exactly k particles is the binomial PMF (probability mass function) with N_P drawings and success probability p_{bino}.

$$P_k^{(\mathrm{box})} = p_{\mathrm{bino}}^k \cdot (1 - p_{\mathrm{bino}})^{N_P - k} \frac{N_P!}{k!\,(N_P - k)!} = \mathrm{binopdf}(k; N_P, p_{\mathrm{bino}}) \tag{S4}$$

Since we only have access to clusters, which we identify with non-empty domains, it is practical to look at the distribution of particles by cluster size, i.e., the number of total particles in a domain. The probability that one specific particle falls into a box with $k-1$ other particles is the same as the PMF for boxes that contain exactly $k-1$ of the other N_{P-1} particles. Our main observable is the number of particles by box size $n_k^{(\mathrm{particle})}$.

$$\begin{aligned} P_k^{(\mathrm{particle})} &= \mathrm{binopdf}(k-1; N_P - 1, p_{\mathrm{bino}}), \\ n_k^{(\mathrm{particle})} &= N_P \cdot \mathrm{binopdf}(k-1; N_P - 1, p_{\mathrm{bino}}) \end{aligned} \tag{S5}$$

The above is normalized as a PMF: $\sum_{k=1}^{N_P} P_k^{(\text{particle})} = 1$. We estimate the expected number of clusters of size k, $n_k^{(\text{box})}$, by dividing the expected number of particles by k. It is useful to estimate the number of empty domains n_{empty} (which cannot be observed) and the number of domains with a single particle.

$$\begin{aligned} n_k^{(\text{box})} &= N_D \cdot \text{binopdf}(k; N_P, \frac{1}{N_D}), \\ n_{\text{empty}} &= \frac{(N_D-1)^{N_P}}{(N_D)^{N_P-1}}, \\ n_{\text{singles}} &= \frac{N_P(N_D-1)^{N_P-1}}{(N_D)^{N_P-1}} \end{aligned} \tag{S6}$$

Finally, the number of boxes of size 2 and higher and the corresponding number of particles are directly comparable to observed quantities:

$$\begin{aligned} N_C^{(2+)} &= N_D \left(1 - \left(1 + \frac{N_P-1}{N_D}\right) \cdot \left(1 - \frac{1}{N_D}\right)^{N_P-1}\right), \\ N_P^{(2+)} &= N_P \left(1 - \left(\frac{N_D-1}{N_D}\right)^{N_P-1}\right) \end{aligned} \tag{S7}$$

One type of domain plus singletons

A plausible explanation for the accumulation of receptors in domains is a mechanism characterized by an enrichment factor α that results in a proportionally higher average particle density inside domains compared to the rest of the membrane area. The particles in a portion of the membrane of area A_{total} will be divided among domains and the rest of the image. Denote the aggregate area of all domains $A_D = aN_D$, where a is the average area of one domain, and the rest by $A_{\text{free}} = A_{\text{total}} - A_D$. The number of particles in each sector will be

$$N_{PD} = N_P \frac{\alpha A_D}{A_{\text{total}} + (\alpha-1)A_D}, \quad N_{P,\text{free}} = N_P \frac{A_{\text{free}}}{A_{\text{total}} + (\alpha-1)A_D} \tag{S8}$$

Define area and population fractions (trapped in domains vs. total).

$$\begin{aligned} f_A &= \frac{A_D}{A_{\text{total}}} = \frac{\alpha N_D}{A_{\text{total}}}, \\ f_P &= \frac{N_{PD}}{N_P} = \frac{\alpha A_D}{A_{\text{total}} + (\alpha-1)A_D} = \frac{\alpha N_D \alpha}{A_{\text{total}} + (\alpha-1)N_D\alpha} \end{aligned} \tag{S9}$$

Singletons (particles that are in clusters of size 1) are experimentally indistinguishable from free particles outside domains. The observed number of singles is then $N_S = N_{P,\text{free}} + n_{C,\text{singles}}$ and we can use this to estimate the total number of particles in

clusters of size 2 or higher:

$$N_S = N_P \left(1 - f_P \left(1 - \left(1 - \frac{1}{N_D}\right)^{N_{PD}-1}\right)\right),$$
$$N_P^{(2+)} = N_P - N_S = N_{PD} \left(1 - \left(1 - \frac{1}{N_D}\right)^{N_{PD}-1}\right) \tag{S10}$$

The number of empty domains is $n_{\text{empty}} = N_D(1-1/N_D)^{N_{PD}}$. The number of clusters of size 2 or higher is the same as we derived previously (replace $N_P \to N_{PD} = N_P f_P$):

$$N_C^{(2+)} = n_{2+} = N_D \left(1 - \left(1 + \frac{N_{PD}-1}{N_D}\right) \cdot \left(1 - \frac{1}{N_D}\right)^{N_{PD}-1}\right) \tag{S11}$$

The above could be used to compare directly to the observed values of $\{N_P^{(2+)}, N_S\}$. The number of particles by cluster size for clusters of two or more remains as in the previous case:

$$\left\langle n_{\text{particle}}^{(k)} \right\rangle = N_{PD}\ \text{binopdf}(k-1; N_{PD}-1, \frac{1}{N_D}), \qquad k \geq 2 \tag{S12}$$

Two types of domain plus singletons

Some of the distributions observed in the image set are not well fit in a model that had one type of domain. We could have multiple types of domains, resulting from different physical mechanisms. For two types of domain, we assume that a fraction f_S of all particles (so $f_S N_P$ free particles) are outside domains and (always) appear as singles outside clusters. The remaining particles are split among two types of domains with lower and higher affinity, labelled L and H; denote by f the proportion of particles in the high affinity domains. The particles in each type of domain (H, L) are distributed consistent with a binomial distribution with success probability p_H and p_L, respectively; N_{DH}, N_{DL} represent the number of each type of domain (in the image).

$$N_{PH} = (1-f_S) f N_P, \qquad N_{DH} = \frac{1}{p_H}$$
$$N_{PL} = (1-f_S)(1-f) N_P, \qquad N_{DL} = \frac{1}{p_L} \tag{S13}$$

The number of singles is the sum of the free particles and the single particle domains of both types:

$$N_S = f_S N_P + N_{PH}(1-p_H)^{N_{PH}-1} + N_{PL}(1-p_L)^{N_{PL}-1} \tag{S14}$$

The number of particles in domains with exactly k particles work out to:

$$n_k^{(\text{particle})} = N_{PH}\ \text{binopdf}(k-1; N_{PH}-1, p_H) + N_{PL}\ \text{binopdf}(k-1; N_{PL}-1, p_L), \quad \forall k \geq 1 \tag{S15}$$

The corresponding number of domains:

$$n_k^{(\mathrm{domain})} = N_{DH}\ \mathrm{binopdf}(k; N_{PH}, p_H) + N_{DL}\ \mathrm{binopdf}(k; N_{PL}, p_L), \quad \forall k \geq 1 \tag{S16}$$

The number of particles in domains with 2 or more particles:

$$N_P^{(2+)} = N_{PH}\left(1 - (1-p_H)^{N_{PH}-1}\right) + N_{PL}\left(1 - (1-p_L)^{N_{PL}-1}\right) \tag{S17}$$

Number of domains with 2 or more particles:

$$N_D^{(2+)} = N_{DH}\left(1 - (1-p_H)^{N_{PH}-1}\right) + N_{DL}\left(1 - (1-p_L)^{N_{PL}-1}\right) \tag{S18}$$

References

1. Espinoza, F.A., Oliver, J.M., Wilson, B.S.: Using hierarchical clustering and dendrograms to quantify the clustering of membrane proteins. Bull. Math. Biol. **74**(1), 190–211 (2011)
2. Güven, E., Wester, M.J., Wilson, B.S., Edwards, J.S., Halász, A.M.: Characterization of the experimentally observed clustering of vegf receptors. In: Češka, M., Šafránek, D. (eds.) Computational Methods in Systems Biology: 16th International Conference, CMSB 2018, Brno, Czech Republic, September 12–14, 2018 Proceedings, Lecture Notes in Bioinformatics 11095. pp. 75–92. Springer (2018). https://doi.org/10.1007/978-3-319-99429-1_5
3. Jain, A., Murty, M., Flynn, P.: Data clustering: A review. ACM Computing Surveys **31**(3), 264–323 (September 1999)
4. Pryor, M.M., Steinkamp, M.P., Halasz, A.M., Chen, Y., Yang, S., Smith, M.S., Zahoransky-Kohalmi, G., Swift, M., Xu, X., Hanien, D., et al.: Orchestration of ErbB3 signaling through heterointeractions and homointeractions. Molecular biology of the cell **26**(22), 4109–4123 (2015)