Identifying Intermolecular Interactions in Single-Molecule Localization Microscopy

Xingchi Yan[1,2,3], Polly Y. Yu[3], Arvind Srinivasan[1,2], Sohaib Abdul Rehman[1,2], and Maxim B. Prigozhin[1,2,✉]

[1]Department of Molecular and Cellular Biology, Harvard University, Cambridge, MA 02138, USA

[2]John A. Paulson School of Engineering and Applied Sciences, Harvard University, Cambridge, MA 02138, USA

[3]NSF-Simons Center for Mathematical and Statistical Analysis of Biology, Harvard University, Cambridge, MA 02138, USA

Intermolecular interactions underlie all cellular functions, yet visualizing these interactions at the single-molecule level remains challenging. Single-molecule localization microscopy (SMLM) offers a potential solution. Given a nanoscale map of two putative interaction partners, it should be possible to assign molecules either to the class of coupled pairs or to the class of non-coupled bystanders. Here, we developed a probabilistic algorithm that allows accurate determination of both the absolute number and the proportion of molecules that form coupled pairs. The algorithm calculates interaction probabilities for all possible pairs of localized molecules, selects the most likely interaction set, and corrects for any spurious colocalizations. Benchmarking this approach across a set of simulated molecular localization maps with varying densities (up to $\sim$50 molecules µm^{-2}) and localization precisions (5 to 50 nm) showed typical errors in the identification of correct pairs of only a few percent. At molecular densities of $\sim$5-10 molecules µm^{-2} and localization precisions of 20-30 nm, which are typical parameters for SMLM imaging, the recall was $\sim$90$\%$. The algorithm was effective at differentiating between non-interacting and coupled molecules both in simulations and experiments. Finally, it correctly inferred the number of coupled pairs over time in a simulated reaction-diffusion system, enabling determination of the underlying rate constants. The proposed approach promises to enable direct visualization and quantification of intermolecular interactions using SMLM.

single-molecule imaging | molecular coupling | inverse problems | probabilistic methods

Correspondence: *maxim_prigozhin@harvard.edu*

Introduction

Biomolecular interactions are fundamental to cellular physiology, governing critical processes such as cell signaling, gene regulation, and enzymatic catalysis. Understanding the spatiotemporal distribution of these interactions within cells is key to elucidating cellular function. However, direct observation of biomolecular interactions is a major technical challenge because these interactions occur on nanometer length scales – far below the diffraction limit of conventional optical techniques. While Förster Resonance Energy Transfer (FRET) (1; 2) can detect molecular colocalization using changes in either the intensity of fluorescence spectra or the excited state lifetimes, this method remains constrained by diffraction-limited spatial resolution. Biochemical approaches like coimmunoprecipitation (3; 4) or proximity labeling (5; 6) are powerful in identifying interaction partners, but they forfeit spatial information altogether. Thus, a critical need and opportunity exist to develop techniques to directly map biomolecular interactions with nanoscale resolution.

Single-molecule localization microscopy (SMLM) (7; 8) offers a promising approach to directly visualize and quantify biomolecular interactions at nanometer resolution. SMLM can determine the locations of individual fluorescently labeled molecules with a precision of 20-30 nm, approaching the length scale of intermolecular interactions. A key advantage of SMLM as compared to FRET is the nanoscale spatial mapping of molecular positions provided by SMLM. If supplemented with quantitative data on molecular binding, these spatial maps would allow probing how intermolecular coupling propensities depend on the heterogeneous local environment within cells. For example, SMLM could elucidate how interaction probabilities vary with local protein densities or how binding equilibria depend on access to protein nanodomains (9; 10; 11; 12; 13). Furthermore, by tracking dynamic interactions triggered by

stimuli over time, SMLM could map spatiotemporal changes in reaction rates. Thus, adding quantitative binding information to super-resolved spatial maps provided by SMLM would result in a powerful approach to elucidate biomolecular interactions at the nanoscale in the complex cellular milieu.

Various methods have been proposed to analyze the spatial relationships between two kinds of molecules. For example, precise intermolecular distance measurements based on iterative localization of nominally identical protein complexes are possible (14; 15; 16). Unfortunately, these approaches cannot assign protein binding states in a spatially heterogeneous ensemble. For such datasets, methods have been developed to measure spatial correlations between molecular maps. For example, correlation-based metrics from diffraction-limited microscopy – Pearson correlation (17), cross correlation (18), and Mander's overlap coefficient (19) – have been modified for use in SMLM (20; 21). However, these correlation-based methods are typically not constrained by the underlying reaction stoichiometry. Other recent methods include the development of an optimal transport approach to measure the distance between two distributions in a pixelated image (22), tessellation-based analysis to access the spatial organization of molecules by Voronoï diagrams (23; 24), and spatial point process based on Ripley's K vector (25) to quantify the relative signal overlap between the two color channels. Although these methods are powerful in providing a relative measure of colocalization, they do not provide an absolute number of pairwise interactions.

Our goal was to count bound molecular pairs in the cell. This capability remains relatively underexplored in SMLM because it is challenging: SMLM localizations have finite precision and multiple sets of interacting pairs may be plausible. Here, we integrated these factors into a probabilistic model with the goal of determining both the absolute number and the fraction of coupled molecular pairs from two-color SMLM datasets. Given the observed inter-fluorophore distance and the corresponding localization precisions, we calculated the likelihood that the fluorescent tags resided within the expected interaction range for a bound complex. We then determined the most probable set of bound pairs in the SMLM image by maximizing the total probability over all putative pairs. Importantly, our approach focused on pairwise interactions by allowing each molecule to couple with at most one partner from the other channel, respecting stoichiometric constraints. Finally, to exclude random colocalizations, we estimated the number of spuriously paired molecules using Monte Carlo simulations of non-interacting particles at the relevant density and subtracted these chance events.

Evaluation of the overall analysis pipeline on simulated datasets with varying densities and localization precisions demonstrated excellent performance. Across the range of conditions tested, the fraction of correctly identified interacting pairs was typically over 95%. At the molecular density of $\sim$5-10 molecules per square micron and localization precision of 20-30 nm – conditions commonly encountered in SMLM measurements – recall exceeded 90%. Notably, our approach reliably distinguished non-interacting and bound proteins in both simulations and SMLM experiments. Furthermore, it accurately deduced the changing numbers of protein complexes that formed over time in a simulated reaction-diffusion system that evolved towards equilibrium; this allowed us to extract the underlying binding rate constants. These results demonstrate the capability of our probabilistic framework to identify molecular interactions using super-resolution microscopy data. This strategy has the potential to significantly advance our understanding of protein coupling at the subcellular level.

Computational Methods

Physical model of molecular interactions captured by SMLM imaging

We consider a prototypical system of two membrane proteins A and B that can bind reversibly: $A + B \rightleftharpoons AB$. Molecules of A and B are labelled with fluorescent tags of orthogonal spectral identities. In an SMLM image, the complex AB appears as a colocalization event between two spectrally distinct fluorophores. However, this colocalization is imperfect. Partly, this is due to the physical separation between the fluorophores (d_{true}) labeling the two proteins. More importantly, d_{true} is a random variable because proteins may populate an ensemble of conformational states, and because dyes are typically attached to their target proteins via flexible tethers of finite length that undergo thermal fluctuations (Figure 1(a)). Furthermore, the detected positions of each fluorophore are also random variables. In the shot-noise-limited regime, the precisions of these localizations scale as $1/\sqrt{N}$ (26), where N is the number of detected photons. So, the observed distance d_{obs} is not only non-zero, but also is not necessarily equal to d_{true} (Figure 1(b)).

Defining the probability that the observed distance is compatible with dimerization

We define the proximity probability as a metric of whether the positions and localization precision values of two emitters are compatible with a colocalization event. Assuming that the localized positions of the fluorophores labeling

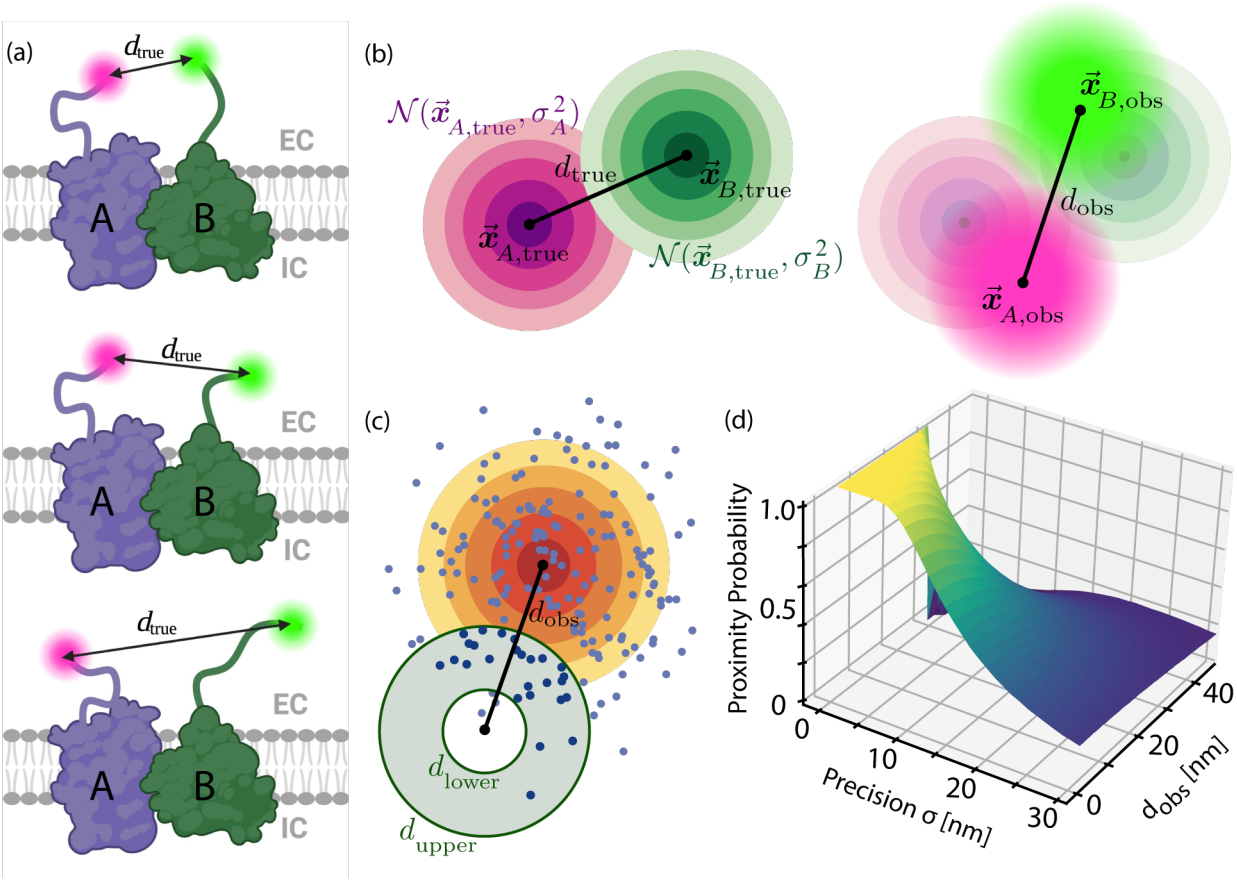

Figure 1. Defining and estimating the proximity probability P_{prox}. (a) Illustration of three conformational states of a complex between two fluorescently labeled membrane proteins, A and B, with a variable distance d_{true} between the dyes, where the random variable d_{true} depends on the conformational state of the proteins and the flexibility of the dye linkers. EC and IC indicate extracellular and intracellular regions, respectively. (b) The observed distance d_{obs} is a random variable. Left: the magenta rings represent the Gaussian distribution centered at the true position $\vec{x}_{A,\text{true}}$ of the fluorophore A with localization precision σ_A. Similarly, the green rings represent the Gaussian distribution at the location of the fluorophore B. The black line denotes the true distance d_{true}. Right: the observed positions $\vec{x}_{A,\text{obs}}$, $\vec{x}_{B,\text{obs}}$ are random variables drawn from these distributions. Magenta and green halos illustrate signals obtained from the two fluorophores. The black line denotes the observed distance d_{obs}. (c) Illustration of how the proximity probability P_{prox} is estimated by sampling points from a Gaussian distribution and counting the fraction of points that land in an annulus. The Gaussian $\mathcal{N}\left(\vec{x}_{A,\text{obs}} - \vec{x}_{B,\text{obs}}, \sigma_A^2 + \sigma_B^2\right)$, represented by the yellow rings, depends on the observed distance (black line) and localization precisions. The annulus, represented by the gray region, depends on a structural model of the complex AB and its fluorophores, where the lower bound d_{lower} and upper bound d_{upper} are constraints on d_{obs} consistent with colocalization. (d) P_{prox} can be estimated for arbitrary d_{obs}, σ_A, σ_B, d_{lower}, and d_{upper}. For $d_{\text{lower}} = 0$, P_{prox} approaches 1 as $d_{\text{obs}}, \sigma_A, \sigma_B \to 0$.

A and B follow Gaussian uncertainties σ_A and σ_B, and are observed to be d_{obs} apart (Figure 1(b)), it can be shown (Supplementary Note 1) that the normalized squared observed distance follows a non-central chi-square distribution:

$$\frac{d_{\text{obs}}^2}{\sigma_A^2 + \sigma_B^2} \sim \chi_2'^2 \left(\frac{d_{\text{true}}^2}{\sigma_A^2 + \sigma_B^2} \right), \tag{1}$$

which fully describes $\mathbb{P}(d_{\text{obs}} | d_{\text{true}}, \sigma_A, \sigma_B)$. However, we are interested in the inverse problem: given d_{obs}, σ_A, and σ_B, can we infer whether d_{true} lies within a range given by a physical model? Rather than determining d_{true} from iterative SMLM measurements as done previously (14; 15; 16), we wish to estimate the proximity probability

$$P_{\text{prox}} = \mathbb{P}\left(d_{\text{lower}} \leq d_{\text{true}} \leq d_{\text{upper}} | \, d_{\text{obs}}, \sigma_A, \sigma_B \right), \tag{2}$$

where d_{lower} and d_{upper} are constraints imposed by a structural model of the macromolecular complex AB and its fluorphores. Given two observed localizations, P_{prox} is the probability that the true distance d_{true} between the fluorophores lies within $[d_{\text{lower}}, d_{\text{true}}]$. When d_{true} lies within $[d_{\text{lower}}, d_{\text{true}}]$, this proximity can originate from two scenarios: (1) molecular coupling, when the proteins A and B are bound in a complex AB, or (2) transient

background pairing, when the uncoupled A and B diffuse close to each other by chance (Figure S3). We call the former “coupling”, and the latter “background pairing”. We refer to both cases jointly as “pairings" since they are indistinguishable at the level of individual colocalization events.

Monte Carlo estimation of the proximity probability

We approximate P_{prox} by Monte Carlo sampling (Supplementary Note 2). Given observed positions $\vec{\boldsymbol{x}}_{A,\text{obs}}$, $\vec{\boldsymbol{x}}_{B,\text{obs}}$ and localization precisions σ_A, σ_B, each Monte Carlo trial draws N points from $\mathcal{N}(\vec{\boldsymbol{x}}_{A,\text{obs}} - \vec{\boldsymbol{x}}_{B,\text{obs}}, \sigma_A^2 + \sigma_B^2)$. The fraction of points landing in the annulus with inner and outer radii d_{lower} and d_{upper} approximates P_{prox} (Figure 1(c)), which is highest when σ_A, σ_B are small and the centroid of the Gaussian lies within $[d_{\text{lower}}, d_{\text{dupper}}]$. Figure 1(d) shows P_{prox} for a range of values of d_{obs} and $\sigma_A = \sigma_B$, with $d_{\text{lower}} = 0$ and $d_{\text{upper}} = 25$ nm. Consistent with the ideal scenario of point particles localized with infinite precision, $P_{\text{prox}} \to 1$ as $d_{\text{obs}} \to 0$ and σ_A, $\sigma_B \to 0$. In summary, our proposed Monte Carlo method estimates the probability that d_{true} lies within a range that is consistent with AB complex formation.

Identifying pairings through Graph Matching Optimization (GMO)

Once a proximity probability is assigned to each pair of A and B using Eq. 2, we construct a bipartite graph that encodes all plausible coupling configurations. We select the most probable configuration by Graph Matching Optimization (GMO), the main idea of which is to represent a SMLM dataset as a bipartite graph with two sets of nodes, where each node in the set V_A represents a localization in the channel for A and each node in V_B represents a localization for B. We connect the nodes A_i and B_j with an edge if their proximity probability p_{ij} is nonzero, and assign p_{ij} as its edge weight. The most probable configuration is given by a selection of edges that maximizes the sum of p_{ij}.

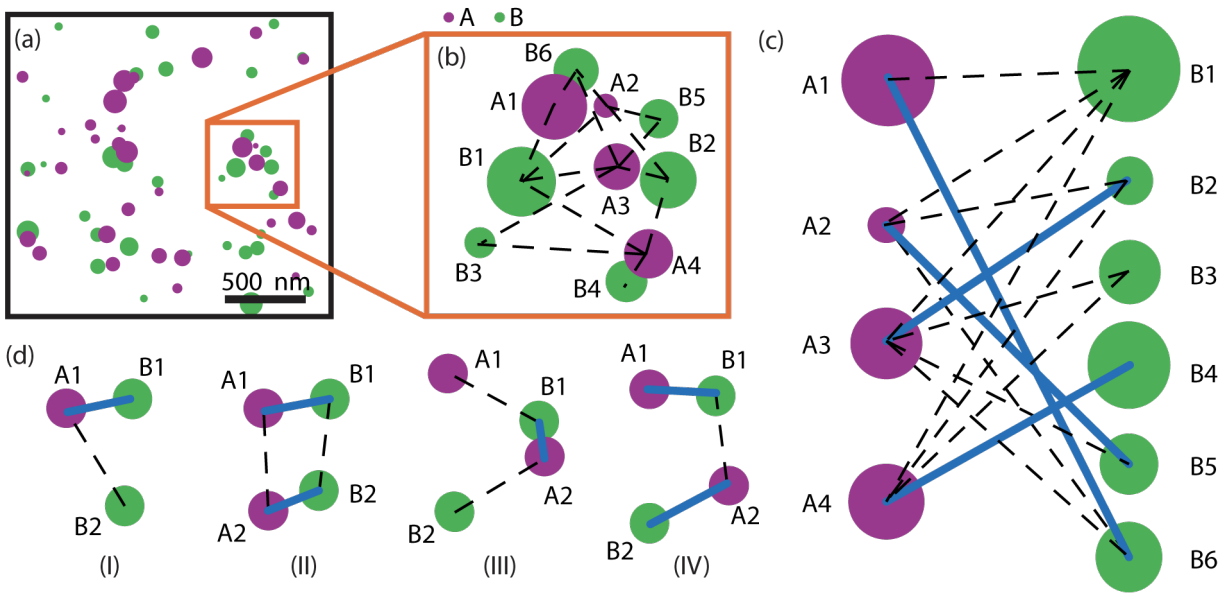

Figure 2. Graph Matching Optimization (GMO) selects the most probable configuration of molecular pairing. (a) A simulated SMLM image of proteins A (magenta) and B (green). Size of marker correlates with localization precision. (b) A connected component of the bipartite graph constructed for the dataset in (a). A node represents a localization of either A or B. An edge (dashed lines) connects A_i and B_j if their proximity probability p_{ij} is positive and their normalized observed distance is less than a data-driven threshold. (c) GMO selects a maximal weight matching (indicated by blue lines), a subgraph that maximizes the sum of p_{ij}, and where each node is selected at most once. The matching represents a possible configuration of molecular pairing. (d) Matchings selected in four example scenarios. I: (A_1, B_1) is chosen over A_1 and B_2 if $p_{11} > p_{12}$; II: (A_1, B_1) and (A_2, B_2) are pairs if $p_{11} + p_{22} > p_{12} + p_{21}$; III: (A_2, B_1) is a pair if $p_{21} > p_{11} + p_{22}$; IV: (A_1, B_1) and (A_2, B_2) are pairs if $p_{11} + p_{22} > p_{21}$.

To save computational time, we only compute p_{ij} when $d_{\text{obs}}/\sqrt{\sigma_A^2 + \sigma_B^2}$ is less than a data-driven threshold C (Supplementary Note 3). Because the distribution in Eq. 1 has a thin tail, for a given distribution of localization precision values, C can be selected such that $P(d_{\text{obs}}/\sqrt{\sigma_A^2 + \sigma_B^2} \leq C)$ is arbitrarily close to one. For example,

with $C = 4$, $P(d_{\text{obs}}/\sqrt{\sigma_A^2 + \sigma_B^2} \leq 4 | d_{\text{true}} = 10, \sigma_A = \sigma_B = 15) = 99.96\%$, accounting for most relevant interactions (Figure S1(a)).

In summary, from a two-channel SMLM dataset (Figure 2(a)), we construct a weighted bipartite graph with two sets of vertices V_A and V_B, and the set of all plausible edges E based on P_{prox} (Figure 2(b)). Each edge e is weighted by the proximity probability $w(e) = p_{ij}$ of the two localizations.

A possible configuration is represented by a *matching* M in the graph (V_A, V_B, E, w). A matching is a vertex-disjoint subset of edges; put simply, a matching pairs nodes such that each node is used at most once. In our context, (A_i, B_j) is selected for the matching M if and only if the localizations A_i and B_j are paired. For example, in the matching shown in Figure 2(c), the (blue) edges represent pairs. We emphasize that not every pairing is necessarily a molecular coupling event as the molecules could be near each other by random chance. We account for this effect in the next section.

The most probable configuration is given by a matching that maximizes the total sum of proximity probabilities. Thus, the objective is to find a set of edges M^* that maximizes the sum of edge weights p_{ij}, subject to the constraint that each node is selected at most once. This can be achieved by the following combinatorial optimization problem (27):

$$
\begin{aligned}
&\text{Maximize} && \sum_{e \in E} w(e)\chi(e) && \\
&\text{subject to:} && \chi(e) \in \{0,1\}, && \text{for all } e \in E, \\
& && \sum_{e=(i,j)\in E} \chi(e) \leq 1, && \text{for all } i \in V_A, \\
& && \sum_{e=(i,j)\in E} \chi(e) \leq 1, && \text{for all } j \in V_B,
\end{aligned}
$$

where each node $i \in V_A$ represents a molecule of A, $j \in V_B$ represents a molecule of B, and an edge e may connect the nodes i and j, with $\chi(e) = 1$ indicating that this edge is selected for the matching M^*. The last two constraints ensure that each node is selected by at most one edge. Figure 2(d) shows several simple examples of this optimization step. In all cases, the selected matchings (blue edges) maximize the sums of proximity probabilities p_{ij}.

Iterative Monte Carlo Estimation of Molecular Couplings and Background Pairings (iMEC)

The matching obtained by GMO represents the most probable configuration of molecular pairings, which consists of both bona fide couplings and pairings by random chance, i.e., "background pairings" (Figure S3). We estimate the number of background pairs, and thus the number of coupling events, by iterative estimation. We call this process Iterative Monte Carlo Estimation of Molecular Couplings and Background Pairings (iMEC).

To illustrate the method, consider the extreme scenario where A and B do not interact at all ($K_{\text{binding}} = 0$) but the densities of both species are sufficiently high. If all localizations from A and B are uniformly distributed, GMO would return a non-empty matching. We would interpret all these edges as background pairs.

More generally, consider a SMLM dataset with N_A localizations in the spectral channel for species A and N_B localizations in the spectral channel for species B, where GMO returns a maximal weight matching M^* with N^*_{pairs} edges. We assume $N^*_{\text{pairs}} \approx N^*_{\text{coupled}} + N^*_{\text{bg}}$, where N^*_{coupled} is the number of true molecular coupling events and N^*_{bg} is the number of background pairings. If N^*_{coupled} is known, we can estimate N^*_{bg} by simulating a spatial Poisson process using $N_A - N^*_{\text{coupled}}$ copies of A and $N_B - N^*_{\text{coupled}}$ copies of B and applying GMO. However, to infer the unknown quantity N^*_{coupled}, we estimate N^*_{bg} by an iterative procedure (Figure 3(a)). By estimating N^*_{bg}, we can infer N^*_{coupled}.

For the first iteration of this iterative procedure, we assume $N^0_{\text{coupled}} = 0$, and simulate a spatial Poisson process with all N_A localizations of A and N_B localizations of B. GMO provides an initial estimate N^1_{bg}, from which the number of putative true couplings can be inferred as $N^1_{\text{coupled}} = N^*_{\text{pairs}} - N^1_{\text{bg}}$. If $N^1_{\text{bg}} \geq N^*_{\text{pairs}}$, the number of pairs can be explained by chance alone. Otherwise, if $N^*_{\text{pairs}} > N^1_{\text{bg}}$, then there are at least $N^*_{\text{pairs}} - N^1_{\text{bg}}$ real coupling events. In the second iteration, we exclude the putative true couplings by applying GMO to a spatial Poisson process with $N_A - N^1_{\text{coupled}}$ and $N_B - N^1_{\text{coupled}}$ localizations to re-estimate the number of background pairs N^2_{bg}. We then iterate this process, each time reducing the pool of potential background pairings by the number of couplings inferred in the previous interation round (Figures 3(b) and 3(c)). The process can be stopped after it meets a convergence criterion, or, in our case, after a fixed number of iterations. Details and pseudo-code are available in Supplementary Note 4.

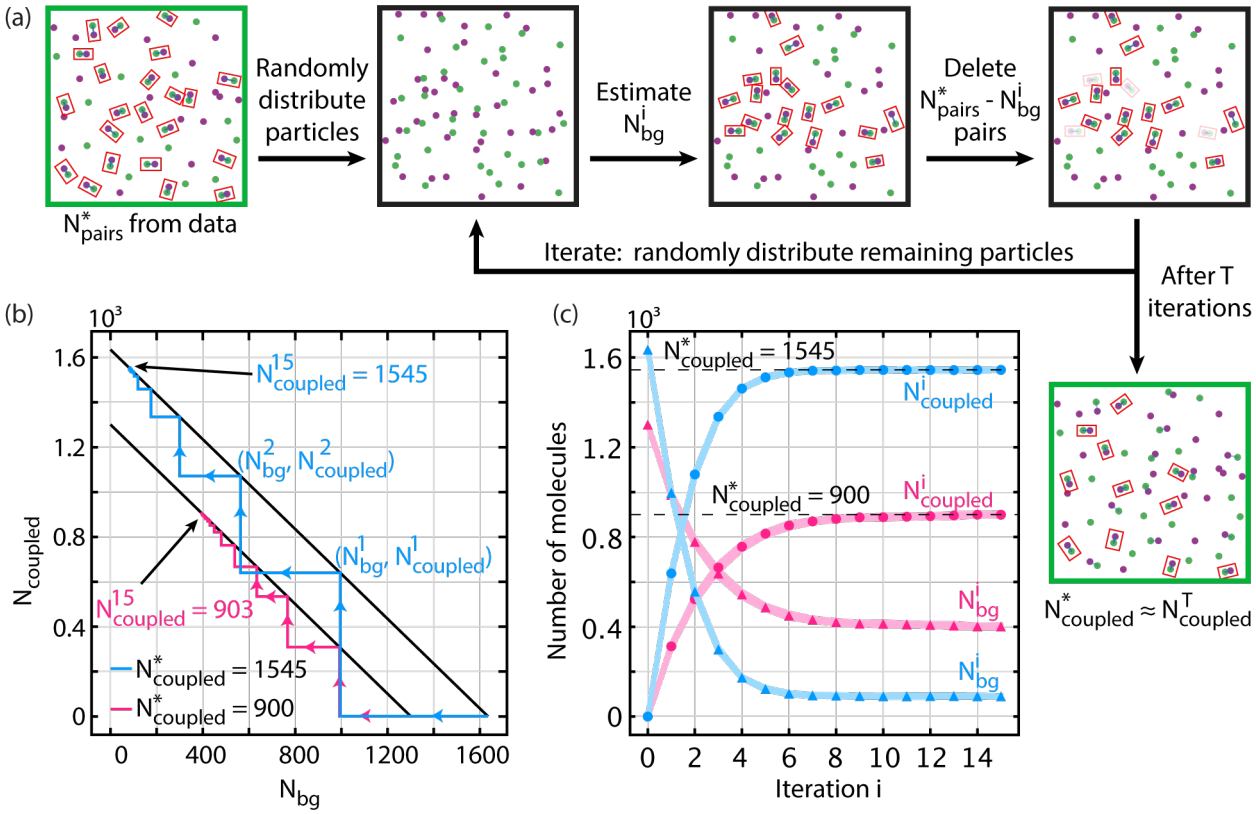

Figure 3. Iterative Monte Carlo Estimation of Molecular Couplings and Background Pairings (iMEC) estimates the number of background pairs among a configuration of molecular pairing. (a) Process overview of iMEC: starting with the most probable configuration with N^*_{pairs} pairings from GMO, distribute the localizations in an experimentally imaged region uniformly random to estimate the number of background pairs N^i_{bg}. For the next iteration, repeat the process with $2 \times N^i_{\text{bg}}$ fewer localizations, which is equivalent to deleting N^i_{bg} pairs from the dataset. The number of couplings N^*_{coupled} is estimated successively by $N^i_{\text{coupled}} = N^*_{\text{pairs}} - N^i_{\text{bg}}$. (b) Two example progressions shown on a plot of N^i_{coupled} versus N^i_{bg}. Of the 2000 localizations from A and 2000 localizations from B, the number of true molecular couplings were $N^*_{\text{coupled}} = 1545$ (blue) and 900 (pink). The lines of slope -1 denote the conserved sum $N^*_{\text{pairs}} = N^i_{\text{coupled}} + N^i_{\text{bg}}$. Arrows show iterative progression to convergence. After 15 iterations, the estimates were $N^{15}_{\text{coupled}} = 1545$ (blue) and 903 (pink). (c) N^i_{bg} and N^i_{coupled} for the examples shown in (b), averaged over 10 Monte Carlo trials.

Figure 3(c) shows N^i_{coupled} and N^i_{bg} converging to N^*_{coupled} and N^*_{bg}, respectively, after 7 iterations. If $\mathbb{E}[N^i_{\text{bg}}]$ is non-increasing, the algorithm converges. Indeed, in all simulations, convergences similar to those shown in Figure 3(c) were observed.

Results

The algorithm described above was evaluated across a range of possible scenarios. First, the performance was benchmarked via Monte Carlo simulations across varying experimental conditions – localization precision and density values. Next, the method was validated using experimental and simulated data of both non-interacting and interacting particles at equilibrium. Finally, we tested our approach on simulated non-equilibrium reaction-diffusion dynamics with the goal of capturing temporal changes in molecular couplings.

Algorithm performance across varying localization precision and density

We evaluated the algorithm's performance over a range of localization precision values, σ_i, and molecular densities, ρ, using simulated data with equal numbers of species A, B, and AB. Two metrics were assessed: recall rate, measuring the fraction of true positives detected, and error rate, quantifying accuracy in estimating the number of true couplings. These two metrics evaluate the two facets of the algorithm: identifying likely couplings through GMO (recall), and inferring the number of true couplings via iMEC (error). The Methods section contains a description of

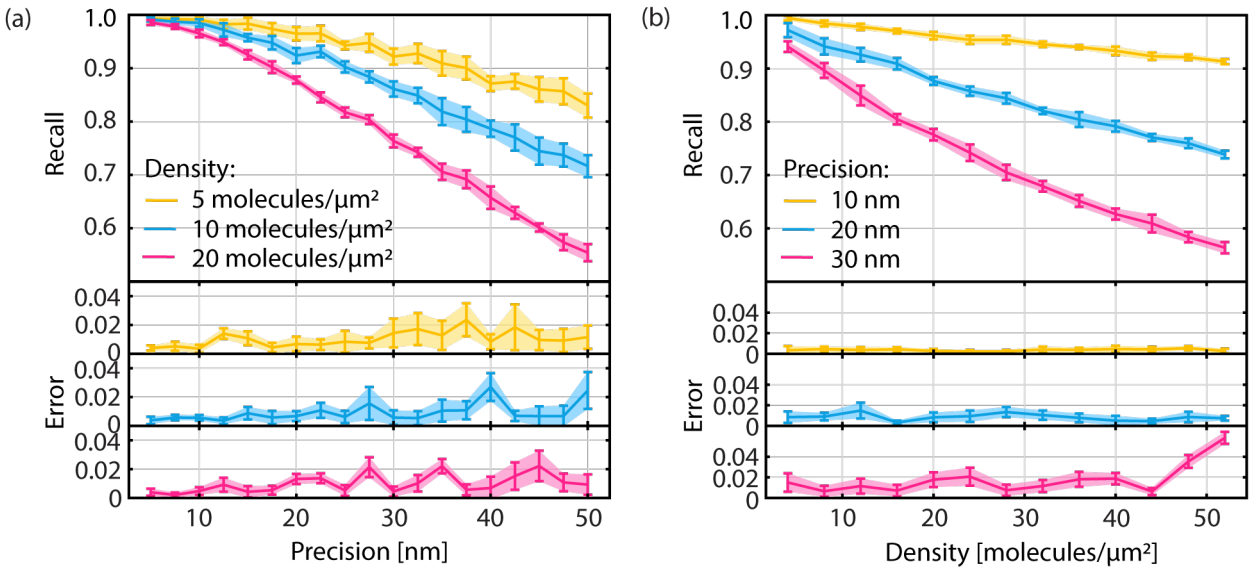

Figure 4. Algorithm performance on simulated datasets. (a) Recall rate and error rate as a function of localization precision. Precise localizations lead to better performance. (b) Recall rate and error rate as a function of molecular density. Recall is better at lower densities, while error remains mostly unchanged. Error bars indicate ± 1 standard deviation across 10 trials of iMEC.

how we generated the simulated datasets; full implementation details and parameters can be found in Supplementary Notes 6 and 7.

At the density of $\rho = 5$ molecules/μm^2, recall rates remained above 80% even when localization precision reached up to 50 nm (Figure 4(a)). At 20 molecules/μm^2, recall was above 80% for localization precision values below 25 nm (Figure 4(a)). We also did an analogous analysis across densities (Figure 4(b)). At the localization precision $\sigma_A = \sigma_B = 10$ nm, recall was greater than 90% for densities up to 50 molecules/μm^2. Retaining recall greater than 80% at $\sigma_A = \sigma_B = 30$ nm required molecular densities below 15 molecules/μm^2. These results demonstrate robust detection of molecular pairs with GMO at typical SMLM imaging conditions.

Estimation errors were less than 5% across all conditions (Figure 4), with minimal dependence on density from 4-50 molecules/μm^2 (Figure 4(b)). Errors increased slightly at lower localization precision but remained below 4% even at a marginal localization precision of 50 nm (Figure 4(a)). These results demonstrate reliable estimation of the number of true pairs by iMEC. We also found that iMEC consistently outperformed the naive approach of selecting colocalizations based on minimal distances; see Supplementary Note 8 for details.

In summary, the algorithm achieved accurate detection of molecular couplings under typical SMLM experimental conditions (localization precision 5-50 nm and densities $\leq$ 50 molecules/μm^2), which establishes its suitability for application to experimental SMLM datasets. As expected, recall decreased as localization precision deteriorated and as density increased.

Algorithm validation at equilibrium using simulations and experiments

We validated the algorithm by asking whether it could accurately quantify the fraction of molecular couplings at equilibrium. Experimental SMLM data were acquired using two protein populations. The first population consisted of HaloTag linked to the N terminus of a β2 adrenergic receptor (Halo-β2AR) and SNAP-tag linked to a CaaX box sequence (SNAP-CaaX) that attaches to the plasma membrane. This population acted as a non-interacting negative control. The second protein population consists of a transmembrane helix with HaloTag on the N terminus and SNAP-tag on the C terminus (Halo-TM-SNAP), which acts as a positively interacting control (Figure 5(a)). We hypothesized that the Halo-TM-SNAP data would show significant colocalization between the spectral channels, while the Halo-β2AR/SNAP-CaaX pair would mimic randomly distributed non-interacting proteins. Confocal imaging confirmed expression and membrane localization of both constructs (Figure 5(b)). SMLM imaging (Figure 5(c)) and subsequent analysis by the GMO and iMEC pipeline showed significantly more coupled pairs for the positive control ($21 \pm 5\%$) compared to the negative control ($6 \pm 2\%$) (Figure 5(d)). These results demonstrate successful quantification of equilibrium interactions from experimental data. Details of experimental and imaging protocols can be found in Supplementary Note 9; parameters used for analysis are available in Supplementary Note 7.

We also analyzed simulations of the reaction $A + B \rightleftharpoons AB$ at equilibrium, where the number of couplings was

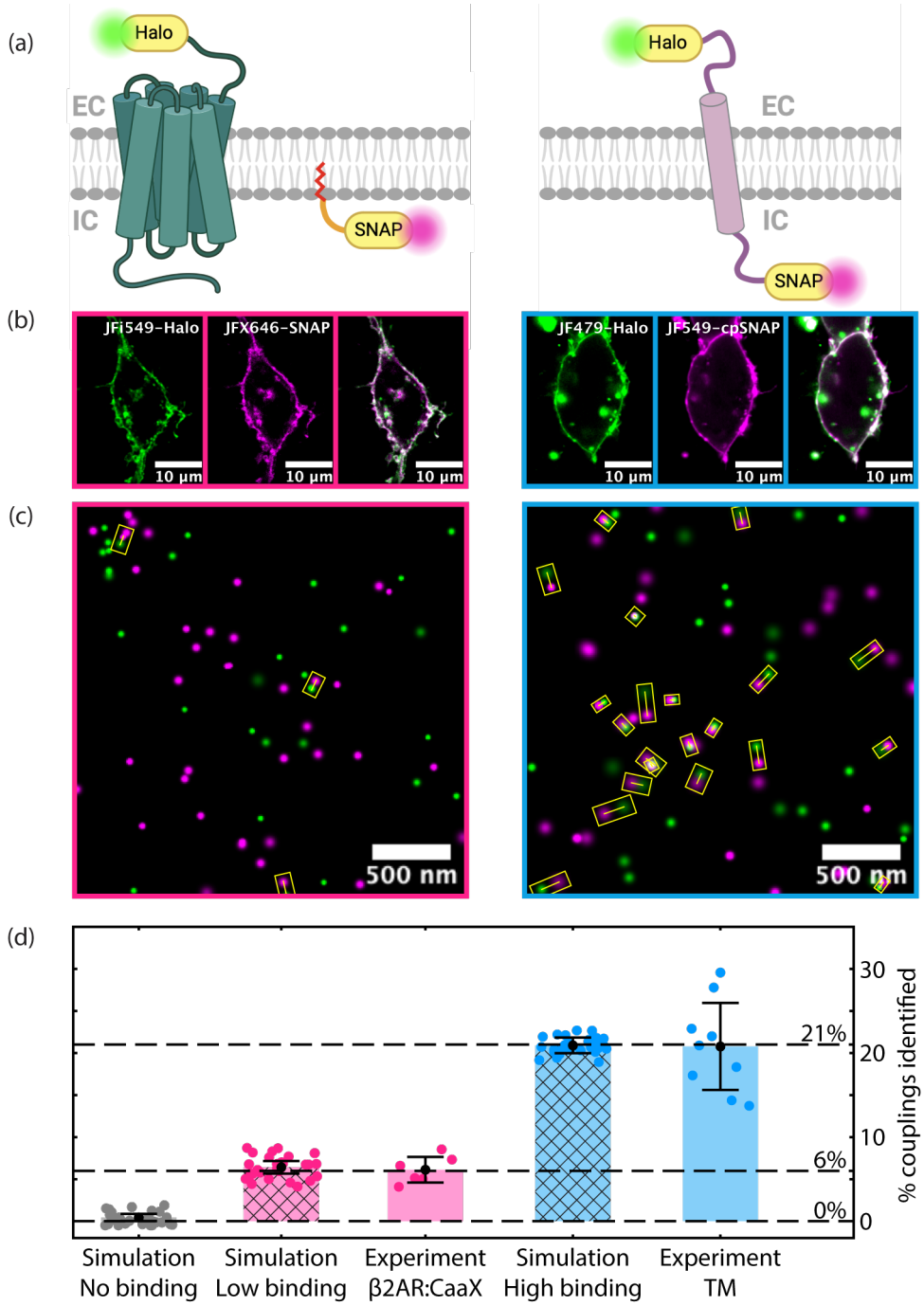

Figure 5. Validation of the algorithm on simulated and experimental SMLM data. (a) Schematics of Halo-β2AR/SNAP-CaaX (negative control) and Halo-TM-SNAP (positive control) – two test systems of membrane proteins. Halo-TM-SNAP was expected to show significantly more colocalization events than Halo-β2AR/SNAP-CaaX. (b) Confocal images of Halo-β2AR/SNAP-CaaX (left) and Halo-TM-SNAP (right) expressed in HEK 293FT cells. Confocal imaging demonstrates localization of constructs to the plasma membrane but does not provide information about molecular coupling. (c) Sample regions of experimental SMLM images for Halo-β2AR/SNAP-CaaX (left) and Halo-TM-SNAP (right). Paired molecules are highlighted with yellow rectangles. (d) Percentage of identified colocalizations for Halo-TM-SNAP, Halo-β2AR/SNAP-CaaX, and simulated datasets with different binding strengths. In the simulations, coupling percentages are set at 0% for no binding, 6% for Halo-β2AR/SNAP-CaaX (low-binding), and 21% for Halo-TM-SNAP (high-binding), the latter two derived from mean values calculated from experimental data. Each dot represents either a cell or a simulated dataset; bars represent averages over all cells/simulated datasets. Error bars indicate ± 1 standard deviation. Dashed horizontal lines denote the ground truth for no-binding, low-binding, and high-binding scenarios.

dictated by the equilibrium constant K_{eq} (Supplementary Note 12). Our algorithm reliably recovered the true number of AB complexes across low ($\sim$0.005) and high ($\sim$0.07) K_{eq} values (Figure 5(d)). In addition, we investigated the algorithm's performance across a broad range of densities (at different K_{eq} values) and found that the output of the algorithm matched the theoretical expected values (Figure S16). Finally, we evaluated and demonstrated the consistency of our algorithm across different density ratios between the two channels, a common challenge in colocalization analysis (17; 25; 24) (Supplementary Note 13). Taken together, these results validated our pipeline's capability of accurately quantifying molecular couplings at equilibrium from both simulated and experimental SMLM data.

Algorithm validation on non-equilibrium dynamics

Next, we tested whether our method could analyze data from cells in non-equilibrium states. Recent advances in time-resolved techniques, like time-resolved cryo-vitrification (28; 29; 30; 31), allow precise stimulation and fixation of cells at defined timepoints followed by super-resolution imaging (Figures 6(a) and 6(b)). Applying our pipeline to these static images can connect them into a dynamic sequence, revealing temporal information on molecular binding.

To evaluate the method's utility for reaction kinetics, we simulated a model of the reaction $A + B \rightleftharpoons AB$ (details in Supplementary Note 6B), extracted the positions of molecules at select timepoints to generate simulated SMLM datasets (Figure 6(c)), and analyzed them using our algorithm. We accurately reproduced the number of complexes across timepoints and densities, with $\sim$10% average error (Figure 6(d)). The slightly higher error relative to those of Figure 4 likely stemmed from fewer total molecules in these simulations. Errors were also larger at earlier timepoints, again due to fewer complexes. In summary, our algorithm captured the dynamics in non-equilibrium simulations, demonstrating its potential for analyzing time-resolved SMLM measurements of cellular processes.

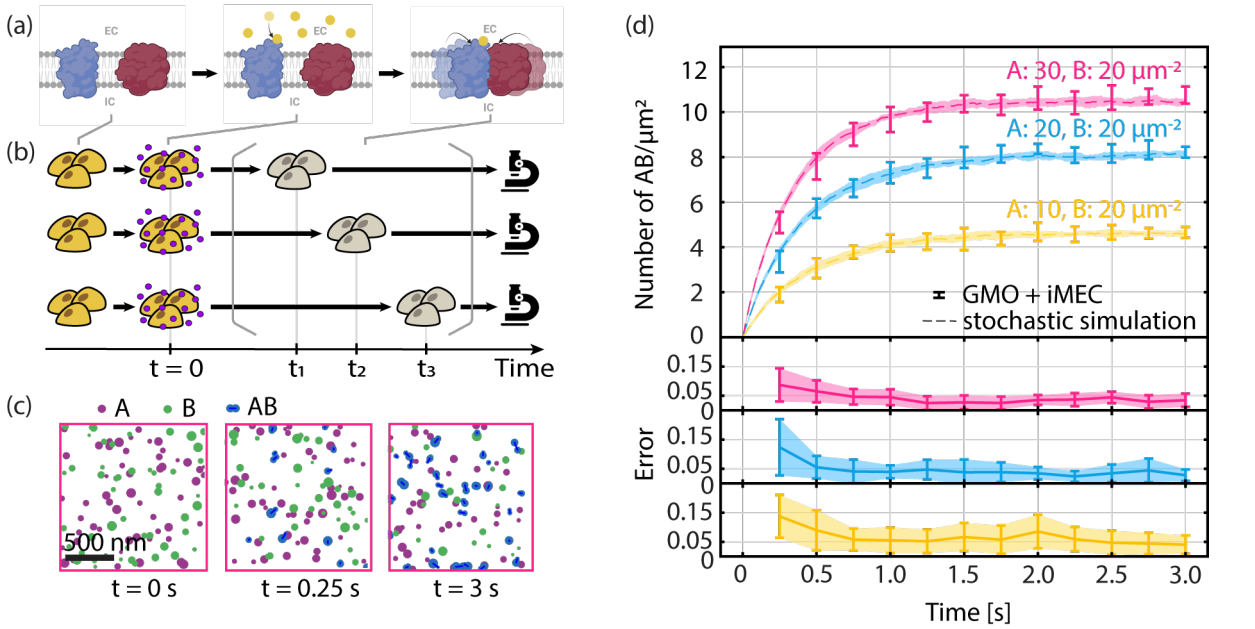

Figure 6. Validation of the algorithm on simulated non-equilibrium binding dynamics. (a) Schematic of a hypothetical biological system: following ligand stimulation at $t = 0$, membrane protein A (blue) becomes active and can bind to membrane protein B (red). EC and IC indicate extracellular and intracellular regions, respectively. (b) Schematic of a hypothetical experimental protocol: ensembles of cells are stimulated with ligand at $t = 0$ and are fixed at times t_1, t_2, t_3 for SMLM imaging. (c) Snapshots from a simulation modeling protein binding, $A + B \rightleftharpoons AB$. AB shown indicate true binding. Snapshots were taken at $t = 0, 0.25$ s, and 3 s after ligand addition. (d) Top: Density of AB identified by the algorithm at various time points (each time point consists of 10 simulated datasets). The variations in the shaded regions originate from stochastic simulations, while the variations in the error bars are attributed to the inference algorithm. Bottom: Error rate at various time points. Error bars indicate ± 1 standard deviation.

We also estimated the rate constants by fitting the density of coupled pairs to the concentration of AB over time (Supplementary Note 14). Our estimates of $\hat{k}_{on} = 0.0505$ molecules^{-1}μm^{2}s^{-1} and $\hat{k}_{off} = 0.862$ s^{-1} closely match the theoretical input values of $k_{on} = 0.0549$ molecules^{-1}μm^{2}s^{-1} and $k_{off} = 1$ s^{-1}. This and the results shown in Figure 6(d) together demonstrate the method's potential to quantitatively track reaction progression.

Discussion and Conclusion

Super-resolution microscopy enables nanoscale mapping of cellular components but cannot directly discern functional binding states. We proposed a pipeline that bridges this gap by statistically deducing protein interactions from single-molecule localization data. Performance assessment showed strong recall and small error rates under typical SMLM conditions. The technique was also able to differentiate between positive and negative experimental controls – a result that was supported by Monte Carlo simulations. When applied to non-equilibrium reaction-diffusion simulations, the approach reliably recovered the temporal evolution of molecular coupling. Taken together, these results validate the algorithm's capacity to infer bona fide couplings from SMLM data in both steady-state and transient cellular processes.

The key innovation of our approach is the ability to estimate the absolute number of molecular couplings. Our algorithm consists of three parts. First, we compute the proximity probabilities between pairs of molecules, indicating potential interactions. Second, GMO identifies the most probable configuration of pairings by maximizing the sum of proximity probabilities. Finally, iMEC estimates the subset of pairings arising from true interactions rather than random colocalization. Notable strengths of our approach include: (1) elucidation of position-dependent, molecular-scale interactions undetectable by diffraction-limited techniques; (2) the use of a biophysical probabilistic model to provide a robust foundation for statistical inference; (3) rigorous correction for chance colocalizations; and (4) demonstrated accuracy on equilibrium and kinetic binding data across a broad range of densities and localization precision values.

However, limitations also exist. While the algorithm can estimate the total number and percentage of true interactions, it cannot accurately determine whether a specific interacting pair represents a true coupling or a background pairing. In addition, the approach relies on the quality of the input SMLM data. Factors that limit SMLM performance, including drift, optical aberrations, incomplete labeling, premature photoactivation or photobleaching, also constrain the quality of algorithm's output. Accurate detection of molecular coupling is particularly challenging because colocalization probability scales with the square of the fraction of detected molecules (Supplementary Note 11C). Moreover, complex spatial distributions of molecules inside cells typically deviate from a uniform distribution, which makes the Monte-Carlo-based inference prone to error. Addressing these limitations is an opportunity for future work. Additional future directions could explore algorithmic alternatives to GMO and iMEC, expand the approach to handle 3D SMLM images, analyze multi-channel SMLM datasets beyond two colors, and address heterogeneous and higher-order interactions.

In summary, the broadly applicable framework presented here infers protein binding from single-molecule localization data, enabling quantification of the spatial relationships of molecular coupling. This versatile new platform has the potential to elucidate the intricate protein interaction landscapes governing diverse cellular functions. We envision the algorithm finding biological applications in probing dynamic protein interactions underlying key cellular processes including transmembrane signaling, gene regulation, and enzymatic catalysis.

Materials and Methods

Implementation of GMO. In constructing the weighted bipartite graph for a dataset, we only considered pairs of localizations that are not too far apart. We first determined the threshold parameter based on the distribution of localization precision values (Supplementary Note 3). For pairs of localizations meeting this threshold, we estimated the proximity probability p_{ij} (Supplementary Note 2). Using these probabilities, the bipartite graph was built, and the edge weights were taken to be $\lfloor 10^5 p_{ij} \rfloor$. The resulting graph was fed into NetworkX's function `max_weight_matching` (32), which then returned a graph matching that maximized the sum of proximity probabilities.

Implementation of iMEC. iMEC was implemented in Python. Detailed description can be found in the corresponding section in the main text and Supplementary Note 4.

Generating simulated datasets. To simulate a system at equilibrium, we instantiated molecules of A, B, and AB uniformly in space, and symmetrically split any AB into a localization for A and a localization for B, at a random distance drawn from $\mathrm{Unif}(0, d_{\mathrm{true,max}})$, where $d_{\mathrm{true,max}} = 10$ nm. See Supplementary Note 6A for implementation details. Localization precision values were either set at a fixed value, or drawn from the experimental localization precision distribution in Supplementary Note 6C with a cutoff at 40 nm. All parameters used are available in Supplementary Note 7.

Generating simulated non-equilibrium binding dynamics. The non-equilibrium dynamics in Figure 6 were generated by a stochastic particle-based reaction-diffusion simulation, implemented using ReaDDY (33). Implementation details, including parameters, are available in Supplementary Note 6B.

Cell culture and transfection. Two cell lines were used: Human Embryonic Kidney 293FT (HEK293FT) cells and HEK293 cells with a CRISPR/Cas9 knockout of the G-protein Gs (HEK293 ΔGs). Cells were cultured in treated cell culture flasks with Dulbecco's Modification of Eagle's Medium (DMEM) with 4.5 g/L glucose, L-glutamine, & sodium pyruvate. Penicillin-Streptomycin solution was added to prevent bacterial contamination. Plasmid transfections were done in either 6-well or 12-well plates using either Lipofectomene 3000 or Polyplus JetOptimus according to manufacturer protocol. Details about the cell lines and protocols are available in Supplementary Note 9A. A full list of plasmids used can be found in Supplementary Note 9A.

Confocal imaging. Cells were seeded onto glass-bottom dishes and labeled with fluorescent dyes. Identities of dyes used in specific experiments can be found in relevant figures. For fluorescent labeling, a dye solution in cell media was prepared at a final concentration of 2 μM, and cells were labeled by incubation in dye solution for 15 minutes. Cells were subsequently washed to remove non-specifically bound dye. Dishes were imaged on a Zeiss LSM 880 confocal microscope using a 63x / 1.40 NA oil objective. Full labeling and imaging parameters are available in Supplementary Note 9B. Confocal images were adjusted for brightness and contrast using Fiji software.

SMLM imaging. Cells were labeled with super-resolution compatible fluorescent dyes with a protocol modified from "Confocal Imaging." After labeling and washing, cells were replated onto pre-cleaned glass coverslips. Once cells adhered, they were fixed with a solution of 4% PFA in PBS. Coverslips were mounted onto glass slides with a drop of Fluoromount-G as the mounting medium. Coverslips were sealed prior to imaging. Samples were imaged on a Zeiss Elyra microscope using a 63x/1.40 NA oil objective. Widefield image stacks were collected in two emission channels corresponding to red and far red dyes. Over the course of imaging, dyes were stochastically activated with a 405 nm laser. Details can be found in Supplementary Note 9C.

Processing of SMLM images. Single-molecule localizations were processed in Zeiss Zen software. Localizations were grouped to aggregate localizations from a single dye molecule spread over multiple consecutive frames. They were then filtered, drift-corrected, and clusters were identified and subsequently removed using DBSCAN (34), implemented using the `sklearn.cluster.DBSCAN` package, with parameters `eps` $= 75$ nm and `min_samples` $= 10$. Finally, the image was divided into smaller subregions, of which any dense subregions were excluded from analysis. See Supplementary Note 10A for details.

Analyzing SMLM images. Each subregion was analyzed using the GMO+iMEC pipeline (Figure S4), and the percentage of couplings in the subregion was calculated using Eq. 5 below. Each dot in Figure 5(d) represents the average percentage of couplings for each cell. See details in Supplementary Note 10B.

Evaluation metrics. Two metrics were used to evaluate the performances of our methods. *Recall* was used to measure the percentage of true couplings successfully retrieved (true positive or TP) relative to false negatives (FN) in GMO:

$$\text{Recall} = \frac{\text{TP}}{\text{TP} + \text{FN}}. \tag{3}$$

We used the *error rate* to measure the performance of iMEC:

$$\text{Error} = \frac{|N^T_{\text{coupled}} - N^*_{\text{coupled}}|}{N^*_{\text{coupled}}}, \tag{4}$$

where N^T_{coupled} was the output from iMEC after T iterations, and N^*_{coupled} was the number of true couplings in the simulated dataset. To compare datasets with different densities (e.g., in Figure 5(d)), we defined for each region

$$\%\text{ of couplings} = \frac{N^T_{\text{coupled}}}{\min(N_A, N_B)}, \tag{5}$$

where N_A was the number of localizations corresponding to A, and N_B was the number of localizations corresponding to B.

Acknowledgements

This research was supported by the Aramont Fellowship Fund for Emerging Science Research, NIH Research Project Grant R01 GM146791 and R21 GM146127 and startup funds from Harvard University. X.Y. and P.Y.Y. were supported by NSF-Simons Center for the Mathematical & Statistical Analysis of Biology (DMS-174269) and the Harvard Quantitative Biology Initiative. A.S. was supported in part by Harvard Qbio Student Award and Simmons Award from Harvard Center for Biological Imaging. The authors thank the Harvard Center for Biological Imaging (RRID:SCR_018673) for infrastructure and support. The computations in this paper were run on the FASRC Cannon cluster supported by the FAS Division of Science Research Computing Group at Harvard University. The authors thank Asuka Inoue at Tohoku University for providing the HEK293 ΔGs cells. The authors thank Johannes Broichhagen at Leibniz-Forchungsinstitut for Molecular Pharmocology for providing the Halo-TM-SNAP DNA construct. The authors thank Simon Merminod, Bridget Queenan, Samuel Kou, Jeremy Conway, Ami Thakrar, Daphne-Eleni Archonta and Jenny Hong for helpful discussions. We thank Rachelle Gaudet, Daniel Needleman and Douglas Richardson for feedback on the manuscript.

Author Contributions

X.Y., and M.B.P. conceived the project; X.Y. developed the method and associated software; P.Y.Y. developed the stochastic simulation and associated software; A.S. performed the experiments and collected data; X.Y., P.Y.Y., A.S., and M.B.P. analyzed data; A.S. and S.A.R. contributed to single-molecule image processing and analysis. X.Y., P.Y.Y., A.S., and M.B.P. wrote the manuscript. M.B.P. supervised the research.

Competing Interests

The authors declare no competing interest.

Bibliography

1. T Forster, Energiewanderung und fluoreszenz. *Naturwissenschaften* **33**, 166–175 (1946).
2. T Ha, et al., Probing the interaction between two single molecules: fluorescence resonance energy transfer between a single donor and a single acceptor. *Proceedings of the National Academy of Sciences* **93**, 6264–6268 (1996).
3. DP Lane, LV Crawford, T antigen is bound to a host protein in SY40-transformed cells. *Nature* **278**, 261–263 (1979).
4. MR Green, J Sambrook, Molecular cloning: A laboratory manual 4th. *Cold Spring Harbor Laboratory Press* **I,II,III** (2012).
5. HW Rhee, et al., Proteomic mapping of mitochondria in living cells via spatially restricted enzymatic tagging. *Science* **339**, 1328–1331 (2013).
6. TC Branon, et al., Efficient proximity labeling in living cells and organisms with TurboID. *Nature Biotechnology* **36**, 880–887 (2018).
7. E Betzig, et al., Imaging intracellular fluorescent proteins at nanometer resolution. *Science* **313**, 1642–1645 (2006).
8. MJ Rust, M Bates, X Zhuang, Sub-diffraction-limit imaging by stochastic optical reconstruction microscopy (STORM). *Nature Methods* **3**, 793–796 (2006).
9. T Sungkaworn, et al., Single-molecule imaging reveals receptor–G protein interactions at cell surface hot spots. *Nature* **550**, 543–547 (2017).
10. D Calebiro, T Sungkaworn, Single-molecule imaging of GPCR interactions. *Trends in Pharmacological Sciences* **39**, 109–122 (2018).
11. SE Anton, et al., Receptor-associated independent cAMP nanodomains mediate spatiotemporal specificity of GPCR signaling. *Cell* **185**, 1130–1142 (2022).
12. N Scheefhals, M Westra, HD MacGillavry, mGluR5 is transiently confined in perisynaptic nanodomains to shape synaptic function. *Nature Communications* **14**, 244 (2023).
13. CJ Obara, et al., Motion of VAPB molecules reveals ER–mitochondria contact site subdomains. *Nature* pp. 1–8 (2024).
14. LS Churchman, Z Ökten, RS Rock, JF Dawson, JA Spudich, Single molecule high-resolution colocalization of Cy3 and Cy5 attached to macromolecules measures intramolecular distances through time. *Proceedings of the National Academy of Sciences* **102**, 1419–1423 (2005).
15. LS Churchman, H Flyvbjerg, JA Spudich, A non-gaussian distribution quantifies distances measured with fluorescence localization techniques. *Biophysical Journal* **90**, 668–671 (2006).
16. S Niekamp, et al., Nanometer-accuracy distance measurements between fluorophores at the single-molecule level. *Proceedings of the National Academy of Sciences* **116**, 4275–4284 (2019).
17. SV Costes, et al., Automatic and quantitative measurement of protein-protein colocalization in live cells. *Biophysical Journal* **86**, 3993–4003 (2004).
18. JW Comeau, S Costantino, PW Wiseman, A guide to accurate fluorescence microscopy colocalization measurements. *Biophysical Journal* **91**, 4611–4622 (2006).
19. S Bolte, FP Cordelières, A guided tour into subcellular colocalization analysis in light microscopy. *Journal of Microscopy* **224**, 213–232 (2006).
20. S Malkusch, et al., Coordinate-based colocalization analysis of single-molecule localization microscopy data. *Histochemistry and Cell Biology* **137**, 1–10 (2012).

21. MB Stone, SL Veatch, Steady-state cross-correlations for live two-colour super-resolution localization data sets. *Nature Communications* **6**, 7347 (2015).
22. C Tameling, et al., Colocalization for super-resolution microscopy via optimal transport. *Nature Computational Science* **1**, 199–211 (2021).
23. F Levet, et al., A tessellation-based colocalization analysis approach for single-molecule localization microscopy. *Nature Communications* **10**, 2379 (2019).
24. AL Ejdrup, et al., A density-based enrichment measure for assessing colocalization in single-molecule localization microscopy data. *Nature Communications* **13**, 4388 (2022).
25. T Lagache, et al., Mapping molecular assemblies with fluorescence microscopy and object-based spatial statistics. *Nature Communications* **9**, 698 (2018).
26. RE Thompson, DR Larson, WW Webb, Precise nanometer localization analysis for individual fluorescent probes. *Biophysical Journal* **82**, 2775–2783 (2002).
27. EL Lawler, *Combinatorial optimization: networks and matroids*. (Courier Corporation), (2001).
28. S Watanabe, et al., Ultrafast endocytosis at mouse hippocampal synapses. *Nature* **504**, 242–247 (2013).
29. D Kontziampasis, et al., A cryo-em grid preparation device for time-resolved structural studies. *IUCrJ* **6**, 1024–1031 (2019).
30. VP Dandey, et al., Time-resolved cryo-em using spotiton. *Nature Methods* **17**, 897–900 (2020).
31. GF Kusick, et al., Synaptic vesicles transiently dock to refill release sites. *Nature Neuroscience* **23**, 1329–1338 (2020).
32. AA Hagberg, DA Schult, PJ Swart, Exploring network structure, dynamics, and function using networkx in *Proceedings of the 7th Python in Science Conference*, eds. G Varoquaux, T Vaught, J Millman. (Pasadena, CA USA), pp. 11 – 15 (2008).
33. M Hoffmann, C Fröhner, F Noé, Readdy 2: Fast and flexible software framework for interacting-particle reaction dynamics. *PLoS Computational Biology* **15**, e1006830 (2019).
34. M Ester, HP Kriegel, J Sander, X Xu, , et al., A density-based algorithm for discovering clusters in large spatial databases with noise in *kdd*. Vol. 96, pp. 226–231 (1996).

Supplementary Information Content

Supplementary Information

Supplementary Note 1: Derivation of the distribution for the observed distance between two fluorophores

Here we derive the distribution of the observed distance between two fluorophores in 2D SMLM. We show that the normalized observed distance squared and the normalized true distance squared follow a non-central chi-square relationship:

$$\frac{d_{\text{obs}}^2}{\sigma_A^2+\sigma_B^2} \sim \chi_2'^2\left(\frac{d_{\text{true}}^2}{\sigma_A^2+\sigma_B^2}\right). \tag{S1}$$

Consider two spectrally distinct fluorophores A and B, whose true positions are (A_x, A_y) and (B_x, B_y) respectively. Assume their observed positions $O_A = (O_{Ax}, O_{Ay})$ and $O_B = (O_{Bx}, O_{By})$ are drawn from Gaussian distributions centered at their true positions. For simplicity, we assume that the localization uncertainties are radially symmetric. Let the localization precision values for A and B be respectively σ_A and σ_B (1).

From $O_A \sim \mathcal{N}((A_x, A_y), \sigma_A^2)$ and $O_B \sim \mathcal{N}((B_x, B_y), \sigma_B^2)$, we rescale to obtain

$$\frac{(O_{Ax}-O_{Bx})-(A_x-B_x)}{\sqrt{\sigma_A^2+\sigma_B^2}} \sim \mathcal{N}(0,1) \quad \text{and} \quad \frac{(O_{Ay}-O_{By})-(A_y-B_y)}{\sqrt{\sigma_A^2+\sigma_B^2}} \sim \mathcal{N}(0,1). \tag{S2}$$

Converting to chi-square distributions, we get

$$\frac{(O_{Ax}-O_{Bx})^2}{\sigma_A^2+\sigma_B^2} \sim \chi_1'^2\left(\frac{(A_x-B_x)^2}{\sigma_A^2+\sigma_B^2}\right) \quad \text{and} \quad \frac{(O_{Ay}-O_{By})^2}{\sigma_A^2+\sigma_B^2} \sim \chi_1'^2\left(\frac{(A_y-B_y)^2}{\sigma_A^2+\sigma_B^2}\right), \tag{S3}$$

where $\chi_k'^2(\lambda)$ is the non-central chi-square distribution with degree of freedom k and noncentrality parameter λ. Note that the noncentrality parameter $\lambda_{\sigma_A,\sigma_B,d_{\text{true}}}$ depends on σ_A, σ_B as well as d_{true}. Then the square of the observed distance, normalized by $\sigma_A^2+\sigma_B^2$, follows a non-central chi-square distribution with degree of freedom $k=2$ (2):

$$\frac{d_{\text{obs}}^2}{\sigma_A^2+\sigma_B^2} = \frac{(O_{Ax}-O_{Bx})^2}{\sigma_A^2+\sigma_B^2} + \frac{(O_{Ay}-O_{By})^2}{\sigma_A^2+\sigma_B^2} \sim \chi_2'^2\left(\frac{(A_x-B_x)^2}{\sigma_A^2+\sigma_B^2} + \frac{(A_y-B_y)^2}{\sigma_A^2+\sigma_B^2}\right) = \chi_2'^2\left(\frac{d_{\text{true}}^2}{\sigma_A^2+\sigma_B^2}\right). \tag{S4}$$

Fig. S1(a-b) show examples of the distributions of the normalized observed distance.

From Fig. S1, we see that as localization precision σ_i increases, the mean of $d_{\text{obs}}/\sqrt{\sigma_A^2+\sigma_B^2}$ decreases. As d_{true} increases, the mean of $d_{\text{obs}}/\sqrt{\sigma_A^2+\sigma_B^2}$ also increases, and the distribution starts to resemble a Gaussian. For any fixed σ_A and σ_B, the monotonic dependence of $\mathbb{E}[d_{\text{obs}}/\sqrt{\sigma_A^2+\sigma_B^2}]$ on the normalized true distance (Fig. S1(c)) can be obtained using the exact formula for the mean of a non-central chi distribution (3). For completeness: the mean of $\chi_k'(\lambda)$ is given by

$$\mathbb{E}\chi_k'(\lambda) = \sqrt{2}\exp\left(-\frac{\lambda}{2}\right)\frac{\Gamma(\frac{k+1}{2})}{\Gamma(\frac{k}{2})}\text{hyp1f1}\left(\frac{k+1}{2}, \frac{k}{2}, \frac{\lambda}{2}\right)$$

where $\Gamma(z)$ is the gamma function,

$$\Gamma(z) = \int_0^\infty t^{z-1}e^{-t}\,dt,$$

and hyp1f1 is the confluent hypergeometric function 1F1,

$$\text{hyp1f1}(a,b,x) = \sum_{j=0}^{\infty}\frac{\Gamma(a+j)/\Gamma(a)}{\Gamma(b+j)/\Gamma(b)j!}x^j.$$

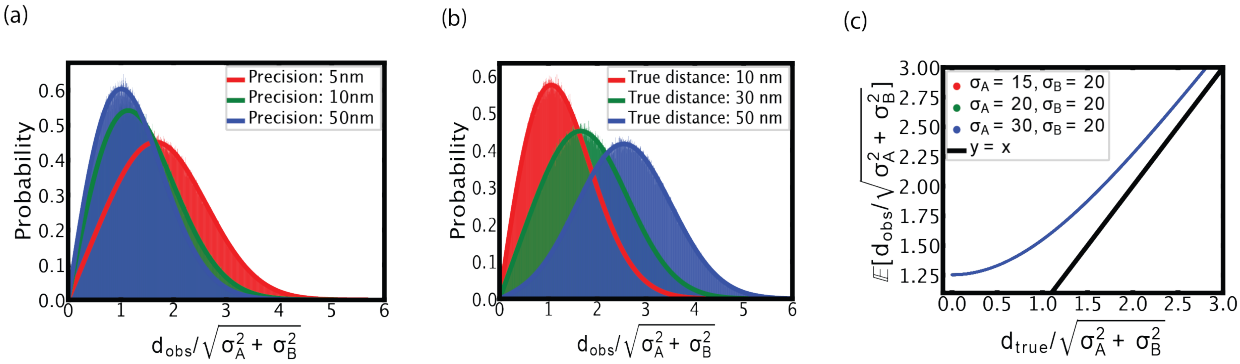

Figure S1. Theoretical distributions and histograms (10^6 sampled points) of $d_{\text{obs}}/\sqrt{\sigma_A^2+\sigma_B^2}$ simulated using Eq. (S4) for (a) $d_{\text{true}} = 10$ nm and localization precisions $\sigma_A = \sigma_B = 5, 10, 50$ nm, as well as (b) $\sigma_A = \sigma_B = 15$ nm and $d_{\text{true}} = 10, 30, 50$ nm. The simulated histograms align with the theoretical non-central chi-distributions characterized by a degree of freedom $k = 2$. (c) The mean of the random variable $d_{\text{obs}}/\sqrt{\sigma_A^2+\sigma_B^2}$ is independent of σ_A and σ_B.

Supplementary Note 2: Estimating the proximity probability of a pair

In this section, we describe a procedure to estimate the proximity probability P_{prox} in Eq. (2) of the main text. As in the previous section, suppose emitters A and B are located at (A_x, A_y) and (B_x, B_y) respectively, with observed positions at O_A, O_B, and localization precisions σ_A, σ_B.

From Eq. (S2) above, let

$$Z_x = \frac{(O_{Ax} - O_{Bx}) - (A_x - B_x)}{\sqrt{\sigma_A^2+\sigma_B^2}} \quad \text{and} \quad Z_y = \frac{(O_{Ay} - O_{By}) - (A_y - B_y)}{\sqrt{\sigma_A^2+\sigma_B^2}},$$

so that $Z_x, Z_y \sim \mathcal{N}(0,1)$. Then the proximity probability P_{prox} is given by

$$\begin{aligned}
&\mathbb{P}(d_{\text{lower}} \le d_{\text{true}} \le d_{\text{upper}} \,|\, d_{\text{obs}}, \sigma_A, \sigma_B) \\
&= \mathbb{P}(d_{\text{lower}}^2 \le |A_x - B_x|^2 + |A_y - B_y|^2 \le d_{\text{upper}}^2 \,|\, d_{\text{obs}}, \sigma_A, \sigma_B) \\
&= \mathbb{P}(d_{\text{lower}}^2 \le (O_{Ax} - O_{Bx} - \sqrt{\sigma_A^2+\sigma_B^2}Z_x)^2 + (O_{Ay} - O_{By} - \sqrt{\sigma_A^2+\sigma_B^2}Z_y)^2 \le d_{\text{upper}}^2). \qquad \textbf{(S5)}
\end{aligned}$$

The random variables in Eq. (S5) are Z_x and Z_y. We can estimate P_{prox} by generating many Z_x, Z_y, and counting how frequently the event in Eq. (S5) occurs. This process, whose pseudocode can be found in Algorithm 1, is computationally inexpensive, because it involves low-dimensional sampling. Typically $N = 10^5$ trials is sufficient to obtain a reliable estimate of the proximity probability.

Algorithm 1 Monte Carlo procedure to estimate the proximity probability of two emitters

Input: σ_A, σ_B, O_A, O_B, d_{lower}, d_{upper}, number of MC trials N
Output: Proximity probability P_{prox}

```
m ← 0
for i ← 1 to N do
    Sample Z_x ~ N(0,1), Z_y ~ N(0,1)
    if d_lower^2 ≤ (O_Ax − O_Bx − √(σ_A^2+σ_B^2) Z_x)^2 + (O_Ay − O_By − √(σ_A^2+σ_B^2) Z_y)^2 ≤ d_upper^2 then
        m ← m + 1
    end if
end for
return P_prox ≈ m/N
```

Fig. S2 shows the effect of observation parameters on estimating P_{prox}. In (a-c), the left panels show the true positions of the two emitters, with the black bar indicating d_{obs}; the right panels show the Gaussians $\mathcal{N}(O_A - O_B, \sigma_A^2+\sigma_B^2)$, and green rings marking the annulus $0 = d_{\text{lower}} \le r \le d_{\text{upper}}$. The proximity probability is maximized when most of the mass of $\mathcal{N}(O_A - O_B, \sigma_A^2+\sigma_B^2)$ lies within the annulus (Fig. S2(a,d)). The examples showcased in Fig. S2 are computed as follows, with all units in nm:

$$\mathbb{P}\,(\,0 \le d_{\text{true}} \le 20 |\, d_{\text{obs}} = 5, \sigma_A = \sigma_B = 5) \approx 96\%,$$
$$\mathbb{P}\,(\,0 \le d_{\text{true}} \le 25 |\, d_{\text{obs}} = 10, \sigma_A = 15, \sigma_B = 10) \approx 56\%,$$
$$\mathbb{P}\,(\,0 \le d_{\text{true}} \le 25 |\, d_{\text{obs}} = 50, \sigma_A = 15, \sigma_B = 25) \approx 13\%.$$

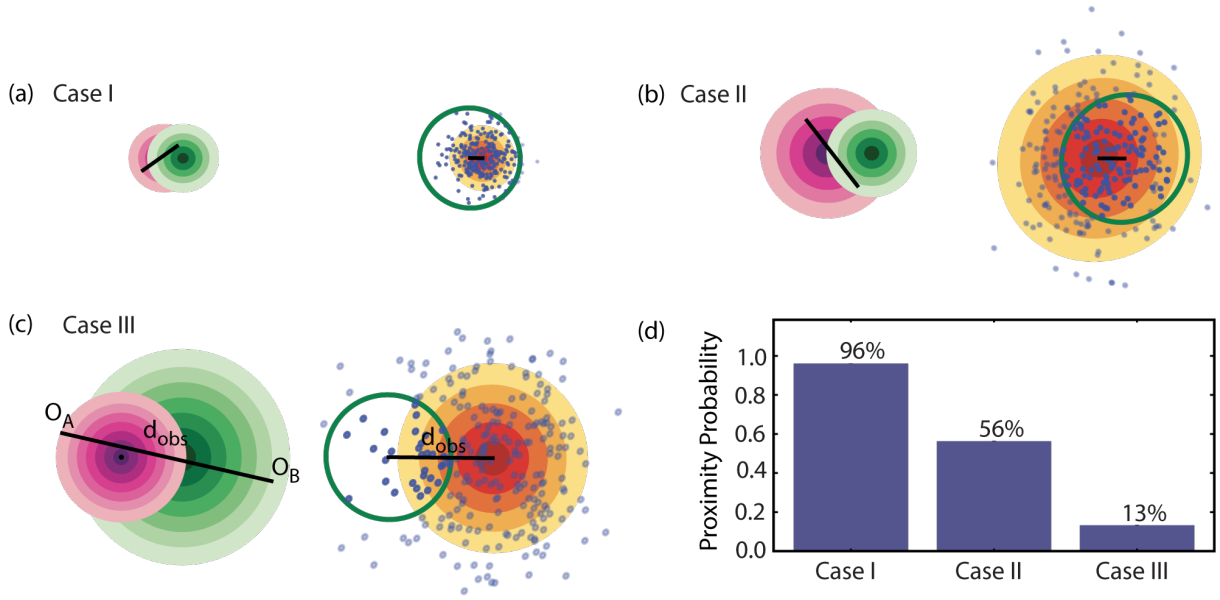

Figure S2. Monte Carlo procedure for estimating the proximity probability P_{prox} in three scenarios (not drawn to scale). (a-c) Pink (respectively green) rings represent the distribution for $O_A \sim \mathcal{N}((A_x, A_y), \sigma_A^2)$ (respectively $O_B \sim \mathcal{N}((B_x, B_y), \sigma_B^2)$); black bars mark observed distances d_{obs}. Yellow rings represent $\mathcal{N}(O_A - O_B, \sigma_A^2 + \sigma_B^2)$, from which points were drawn. The fraction of points that land inside the green ring estimates P_{prox}. (d) The approximated P_{prox} is largest when the localization uncertainties are small and the observed displacement lies in the annulus.

Supplementary Note 3: Data-driven selection of a threshold parameter for d_{obs}

Intuitively, if A and B are extremely far away, i.e., $d_{\text{obs}} \gg d_{\text{true}}$, we would not expect them to pair up. By considering only emitters that are reasonably close by, we cut down the computational time for Graph Matching Optimization (GMO). In this section, we quantify this threshold, and provide a heuristic for how to select the threshold parameter for d_{obs} based on the statistics of the localizations in a SMLM image.

In principle, GMO selects a matching from all possible pairs. However, using all possible pairs to construct the bipartite graph would be computationally expensive, and with diminishing return for the performance of GMO. For example, with N_A molecules A and N_B molecules B, there can be up to $N_A N_B/2$ possible pairs. Selecting the maximum weight matching from the input bipartite graph, as implemented in NetworkX (4; 5), is computationally expensive, with time complexity that scales as $O(n^3)$, where n denotes the number of nodes. Thus, when building the bipartite graph for GMO, we only consider pairs that are within a threshold distance away from each other. We do this by cutting out the thin tail of the distribution in Eq. (S4) and considering only pairs where $d_{\text{obs}}/\sqrt{\sigma_A^2 + \sigma_B^2} \le C$ for some thresholding parameter C.

This truncation of the tail has minimal effect on the distribution. For example, for the distributions in Fig. S1(a), choosing $C = 4$ captures the bulk of the distributions:

$$\mathbb{P}(d_{\text{obs}}/\sqrt{\sigma_A^2 + \sigma_B^2} \le 4 \,|\, d_{\text{true}} = 10, \sigma_A = \sigma_B = 5) = 99.13\%,$$
$$\mathbb{P}(d_{\text{obs}}/\sqrt{\sigma_A^2 + \sigma_B^2} \le 4 \,|\, d_{\text{true}} = 10, \sigma_A = \sigma_B = 10) = 99.87\%,$$
$$\mathbb{P}(d_{\text{obs}}/\sqrt{\sigma_A^2 + \sigma_B^2} \le 4 \,|\, d_{\text{true}} = 10, \sigma_A = \sigma_B = 50) = 99.96\%.$$

Similarly, for the distributions in Fig. S1(b):

$$\mathbb{P}(d_{\text{obs}}/\sqrt{\sigma_A^2+\sigma_B^2} \le 4 \,|\, d_{\text{true}} = 10, \sigma_A = \sigma_B = 15) = 99.93\%,$$
$$\mathbb{P}(d_{\text{obs}}/\sqrt{\sigma_A^2+\sigma_B^2} \le 4 \,|\, d_{\text{true}} = 30, \sigma_A = \sigma_B = 15) = 99.13\%,$$
$$\mathbb{P}(d_{\text{obs}}/\sqrt{\sigma_A^2+\sigma_B^2} \le 4 \,|\, d_{\text{true}} = 50, \sigma_A = \sigma_B = 15) = 93.05\%.$$

Clearly, we cannot blindly choose $C = 4$ for any SMLM image. To choose C, one must first estimate d_{true} for the two interacting molecule partners. Such an estimate may be obtained from fitting the distributions in Eq. (S1) from experiments such as iterative localization of the identical molecular complex (6; 2; 7). Next, one would plot the distribution of localization precision values of the SMLM image, and select a threshold σ_* such that it captures most of the precision values, e.g., $\sigma_i \le \sigma_*$ captures the 90th percentile of all σ_A, σ_B from the dataset. One can then simulate the distribution of $d_{\text{obs}}/\sqrt{2\sigma_*^2}$ according to Eq. (S4). From the resulting distribution similar to those in Fig. S1(a-b), C can be chosen such that $\{d_{\text{obs}}/\sqrt{\sigma_A^2+\sigma_B^2} \le C\}$ captures the bulk of the distribution.

In principle, a larger value of C would capture more possible pairings. However, small increases in C might significantly increase the computational time due to the increased density of the graph. Choices must be made based on the dataset being analyzed, and the goal of the analysis.

Supplementary Note 4: Estimating the true number of molecular couplings and the number of pairings from transient proximity

Because SMLM captures the observed positions of emitters at a fixed time, we cannot distinguish between molecular coupling and transient proximity. For example, if the SMLM image is a snapshot of Fig. S3 at t_2, we cannot distinguish between the two scenarios. We devised a process we call Iterative Monte Carlo Estimation of Molecular Couplings and Background Pairings (iMEC) as part of our colocalization algorithm, to estimate how many pairs are true couplings in a given SMLM dataset, and how many are due to transient proximity. More precisely, given N_A and N_B emitters, of which there are N^*_{pairs} pairings identified by GMO, Algorithm 2 estimates how many of the N^*_{pairs} pairings are transient, in the sense of Fig. S3(b). Thus, the number of true molecular couplings is what remains after subtracting the transient pairs.

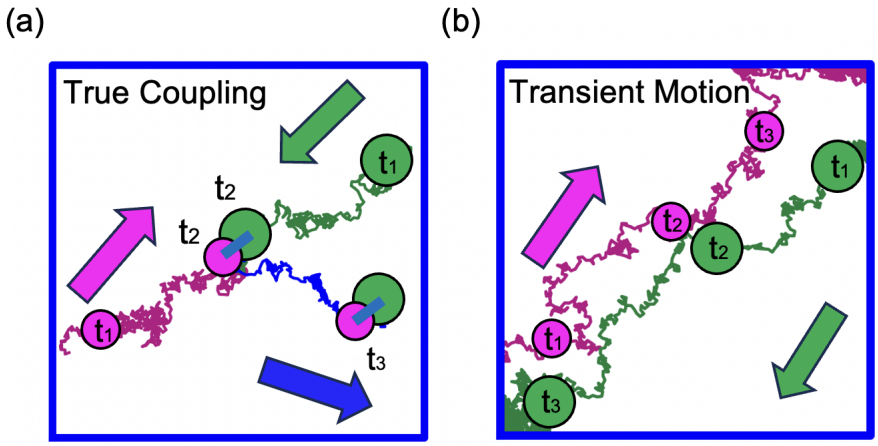

Figure S3. Stochastic Brownian trajectories of two interacting partners. (a) Coupling occurs between two molecules at time t_2, after which they proceed together until t_3 along the trajectory in blue. (b) Two molecules experience transient proximity during diffusion. A measurement only at time t_2 cannot distinguish between the two scenarios.

We assume a system where A and B diffuse freely. If a molecule of A and a molecule of B are sufficiently close, they are capable of binding (Fig. S3(a)). However, if the system does not pass the activation energy barrier for the binding reaction to occur, the two molecules may simply diffuse away at a later time (Fig. S3(b)). A comprehensive stochastic model elucidating the pivotal role of colocalization in determining the probability of protein-protein interactions is detailed in the works (8; 9) by Batada et al..

In this work, we take a slightly different approach. In iMEC (Algorithm 2), we successively estimate the number of pairings among molecules that we assume cannot interact, with a smaller number of molecules at the start of each iteration. The inputs to Algorithm 2 are the number of signals corresponding to A (N_A), the number of signals corresponding to B (N_B), the number of pairs identified by GMO (N^*_{pairs}), and the area of the region being analyzed (S) . In addition we specify the number of iterations per trial (T), and the number of trials (M).

In any given trial, for the first iteration, we assume all N_A and N_B molecules are non-interacting ($N^0_{\text{coupled}} = 0$), and estimate the number of pairs N^1_{bg}. All pairs identified at this step are considered background pairs since we assumed non-interacting particles. Then we set $N^1_{\text{coupled}} = N^*_{\text{pairs}} - N^1_{\text{bg}}$. For the second iteration, since we assume there are at least N^1_{coupled} coupled pairs, we repeat the process by assuming that the remaining $N_A - N^1_{\text{coupled}}$ molecules of A and $N_B - N^1_{\text{coupled}}$ molecules of B are non-interacting. After each iteration, we update the estimated number of background pairs N^i_{bg}, and thus obtain an estimated number of true couplings $N^i_{\text{coupled}} = N^*_{\text{pairs}} - N^i_{\text{bg}}$. We stop the trial after T iterations, and obtain estimates of N^T_{bg} and N^T_{coupled}. The output of Algorithm 2 is the average over M trials: $N^*_{\text{coupled}} \approx \langle N^T_{\text{coupled}} \rangle$.

An example of $\langle N^t_{\text{bg}} \rangle$ and $\langle N^t_{\text{coupled}} \rangle$ is shown in Fig. 3 of the main text. In the main text, we suppress the notation for multiple MC trials for simplicity, i.e., N^T_{bg} in the main text is in actuality $\langle N^T_{\text{bg}} \rangle$, and likewise for the number of molecular couplings.

In principle, it is possible that $N^i_{\text{bg}} > N^*_{\text{pairs}}$ at the end of an iteration, so $N^i_{\text{coupled}} < 0$ by definition. This reflects the situation where there are more background pairs expected than those identified by GMO. In such a case, the next iteration starts with $N_A - N^i_{\text{coupled}}$ and $N_B - N^i_{\text{coupled}}$ molecules of A and B respectively. Note that here we instantiate more molecules than in the dataset. If this occurs at the last iteration, we set $N^T_{\text{coupled}} = 0$. However, this is unlikely to occur in a dataset for which molecular couplings are expected.

Algorithm 2 Iterative Monte Carlo procedure to estimate the number of molecular couplings.

Input: N_A, N_B, N^*_{pairs}, area S, number of iterations T per trial, number of MC trials M.
Output: $\langle N^T_{\text{bg}} \rangle$, $\langle N^T_{\text{coupled}} \rangle$
$S_{\text{bg}} \leftarrow 0$
$S_{\text{coupled}} \leftarrow 0$
  for $m \leftarrow 1$ **to** M **do**
    $N^0_{\text{coupled}} \leftarrow 0$
    for $i \leftarrow 1$ **to** T **do**
      Simulate $N_A - N^{i-1}_{\text{coupled}}$ and $N_B - N^{i-1}_{\text{coupled}}$ particles over area S uniformly random.
      Use pairing algorithm to find the most probable graph matching to estimate N^i_{bg}.
      $N^i_{\text{coupled}} \leftarrow N^*_{\text{pairs}} - N^i_{\text{bg}}$
    end for
    $S_{\text{bg}} \leftarrow S_{\text{bg}} + N^T_{\text{bg}}$
    $S_{\text{coupled}} \leftarrow S_{\text{coupled}} + N^T_{\text{coupled}}$
  end for
  $\langle N^T_{\text{bg}} \rangle \leftarrow S_{\text{bg}}/M$
  $\langle N^T_{\text{coupled}} \rangle \leftarrow S_{\text{coupled}}/M$
  if $\langle N^T_{\text{coupled}} \rangle < 0$ **then**
    Set $\langle N^T_{\text{coupled}} \rangle = 0$
  end if
return $\langle N^T_{\text{bg}} \rangle$, $\langle N^T_{\text{coupled}} \rangle$

Supplementary Note 5: Pipeline to estimate the number of molecule couplings in an SMLM image

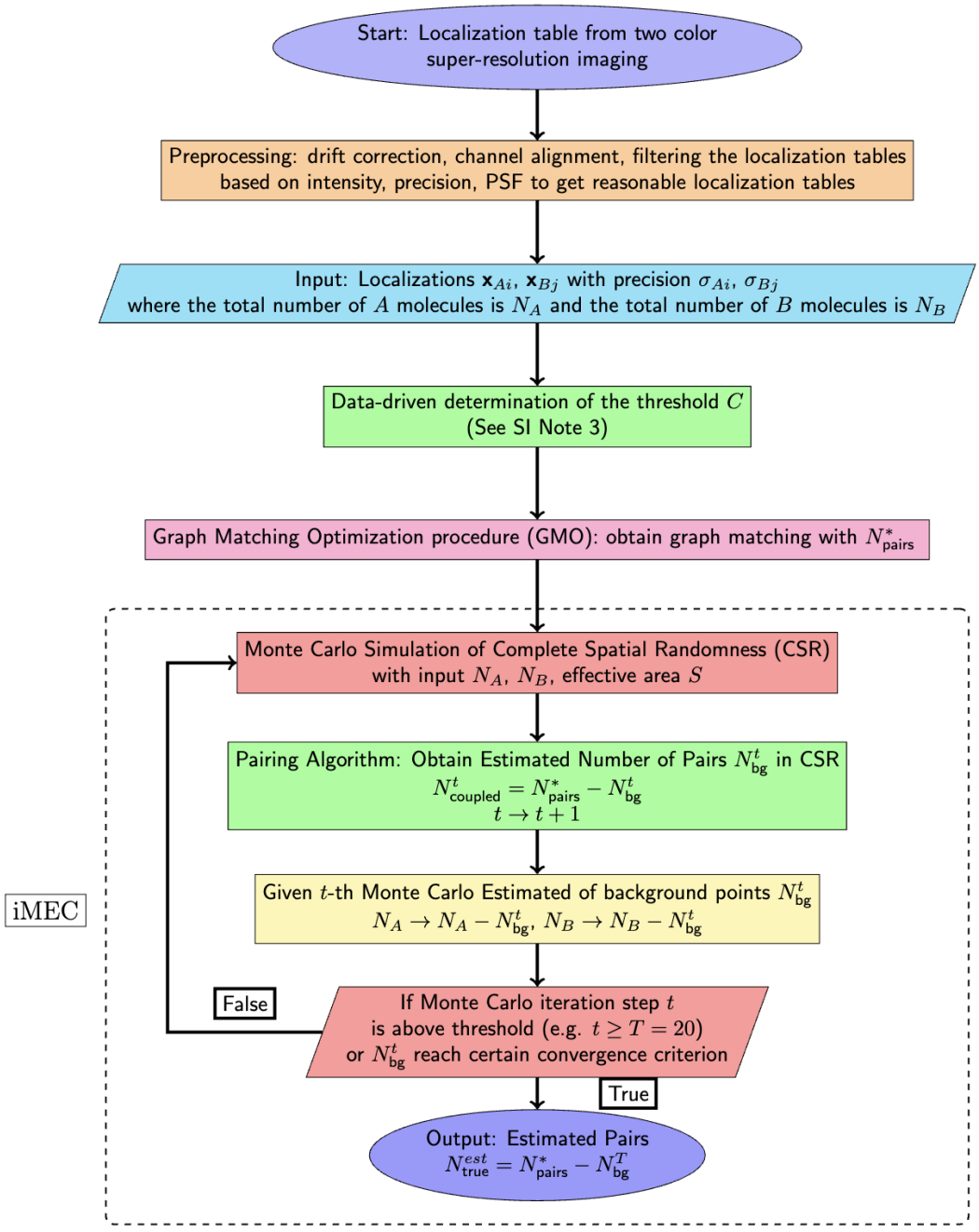

Figure S4. Full pipeline to detect molecular pairings and count the number of true couplings in a given SMLM localization table. The core components include a graph matching optimization procedure (GMO) and an iterative Monte Carlo estimation of molecular couplings (iMEC).

Supplementary Note 6: Generating simulated SMLM datasets

Many of our results were based on simulated SMLM datasets. To generate a simulated dataset, we instantiated the positions of molecules A, B, and AB either from a uniform distribution, or from a stochastic reaction-diffusion simulation.

A. Generating two-channel signals

To generate the localizations in the two channels from the positions of A, B, and AB, we kept the localizations for A and B, and split each AB into two localizations. For each AB located at (x, y), we sampled from the uniform distribution a stochastic distance $d_{\text{obs}} \sim \text{Unif}(0, 10\,\text{nm})$, and an angle $\theta \sim \text{Unif}(0, 2\pi)$. We placed signals corresponding to A and B at

$$\left(x + \frac{d_{\text{obs}}}{2}\cos\theta, y + \frac{d_{\text{obs}}}{2}\sin\theta\right) \quad \text{and} \quad \left(x - \frac{d_{\text{obs}}}{2}\cos\theta, y - \frac{d_{\text{obs}}}{2}\sin\theta\right)$$

respectively. For the localization uncertainty of each signal, we either set it as a constant (e.g., for Fig. 4 of the main text; see Tables S3), or drew from the distribution in Fig. S6.

B. Stochastic simulation

The stochastic simulations reported in Fig. 6 of the main text were generated using a particle-based reaction–diffusion model, implemented using the package ReaDDY, version 2.0.12 (10).

We set the simulation box to be $\sqrt{\text{Area}} \times \sqrt{\text{Area}} \times 1$ nm, with periodic boundaries. This effectively simulates the system as if it is 2-dimensional for Area $\gg 1$. Diffusion is modeled by $\frac{d\vec{\boldsymbol{x}}_i(t)}{dt} = \sqrt{2D_i}\vec{\boldsymbol{\xi}}_i(t)$, where $\vec{\boldsymbol{x}}_i(t)$ is the position of particle i, with diffusion coefficient D_i, and $\{\vec{\boldsymbol{\xi}}_i\}_i$ are independent Gaussian random variables.

Each specified reaction occurs with probability $p(\lambda; dt) = 1 - e^{-\lambda dt}$, where λ plays a similar role to a rate constant. For a reaction involving a single reactant, e.g., $AB \to A + B$, the propensity constant is given by the macroscopic rate constant: $\lambda_{\text{off}} = k_{\text{off}}$. For binary reactions, e.g., $A + B \to AB$, a reaction would occur with probability $p(\lambda; dt)$ only if the reacting molecules, A and B in this case, are no more than d_{educt} apart, where the educt distance d_{educt} is a chosen parameter. For this simulation, we chose $d_{\text{educt}} = 30$ nm. Note that for binary reactions, $k_{\text{on}} \neq \lambda_{\text{on}}$ in general. For $d_{\text{educt}} = 30$ nm, the relationship for $\lambda_{\text{on}} \leq 25\,\text{s}^{-1}$ was numerically determined to be

$$k_{\text{on}} = 2.72 \times 10^{-3}\lambda_{\text{on}} + 4.54 \times 10^{-4},$$

as shown in Fig. S5. For $\lambda_{\text{on}} = 20\,\text{s}^{-1}$, the corresponding rate constant is $k_{\text{on}} = 0.0549\ \mu\text{m}^2\text{molecule}^{-1}\text{s}^{-1}$.

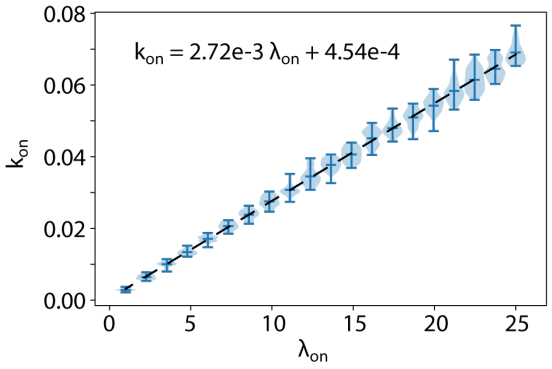

Figure S5. Conversion between the rate constant k_{on} and the propensity constant λ_{on} for $A + B \to AB$, at $d_{\text{educt}} = 30$ nm.

We simulated the system with initially only A and B, but no AB. At the densities stated in Fig. 6(c) (pink 30 $A/\mu\text{m}^2$, 20 $B/\mu\text{m}^2$; blue 20 $A/\mu\text{m}^2$, 20 $B/\mu\text{m}^2$; yellow 10 $A/\mu\text{m}^2$, 20 $B/\mu\text{m}^2$), the particles were uniformly distributed in the simulation box. We ran the simulation, and extracted the positions of A, B, and AB at various time points. The parameters for the stochastic simulations are presented in Table S1.

Table S1. Stochastic Simulation Parameters in Fig. 6

Timestep dt	0.5×10^{-3} s
Total Simulation Time	3 s
Area	100 μm^2
λ_{on}	20 s^{-1}
λ_{off}	1 s^{-1}
d_{educt}	30 nm
Diffusion coefficient D_A	0.3 μm^2/s
Diffusion coefficient D_B	0.3 μm^2/s
Diffusion coefficient D_{AB}	0.01 μm^2/s
Densities of (A, B)	(30,20), (20,20) or (10,20) molecules/μm^2

C. An experimental distribution of localization precisions

The average experimental precision in our datasets typically ranged from 15 to 30 nm. In several of our simulated SMLM datasets, we used localization precision values drawn from the experimental distribution in Fig. S6, with $\sigma_i \leq 40\,\text{nm}$ to model localizations with moderate to high signal-to-noise ratios. We found that the localization precision distributions between different cells (under the same experimental conditions) were remarkably similar. Fig. S6 is representative of the typical distributions observed; experimental protocol can be found in Supplementary Note 9.

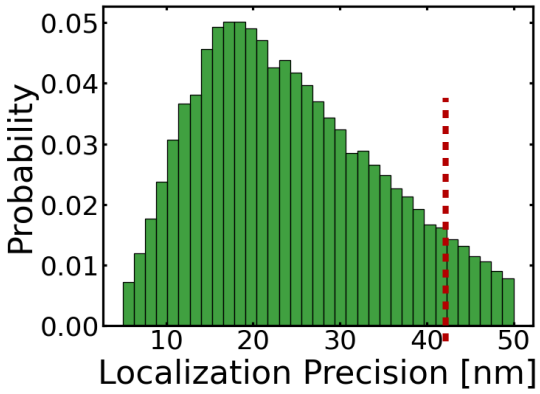

Figure S6. An experimental localization precision distribution with the peak at approximately 20 nm. The cutoff localization precision used in all simulations was 40 nm.

Supplementary Note 7: Parameters used for various analyses

For Figures 3-6 in the main text, to produce the results, we ran GMO and iMEC. This section consists of tables of parameters used in the analysis.

Table S2. Simulation Parameters for Fig. 3(b-c)

Area S	100 μm^2
Number of AB	1545 or 1100
Number of background A N_A	455 or 900
Number of background B N_B	455 or 900
True distance d_{true}	Sampled from Unif(0,10 nm)
Localization Precision σ_A, σ_B	Sampled from Fig. S6
d_{lower}	0 nm
d_{upper}	20 nm
Search distance factor threshold C from SN 3	4
Number of sampling points to estimate probability N from Algorithm 1	10^5
Number of iterations T to estimate N^*_{bg} from Algorithm 2	15
Number of MC trials M in each estimation of N^i_{bg} in Algorithm 2	10

Table S3. Simulation Parameters in Fig. 4

	Fig. 4(a)	Fig. 4(b)
Density ρ	$\rho =$ 5, 10 or 20 (see plot)	Varying (see x-axis)
Area S	$S =$ 100 μm^2	100 μm^2
Number of AB	$\rho S/2$	$\rho S/2$
Number of background A	$\rho S/2$	$\rho S/2$
Number of background B	$\rho S/2$	$\rho S/2$
True distance d_{true}	Sampled from Unif(0,10 nm)	Sampled from Unif(0,10 nm)
Localization Precision σ_A, σ_B	Varying (see x-axis)	$\sigma_A = \sigma_B =$ 10, 20 or 30 nm (see plot)
d_{lower}	0 nm	0 nm
d_{upper}	20 nm	20 nm
Search distance factor threshold C from SN 3	4	4
Number of sampling points N to estimate probability from Algorithm 1	10^5	10^5
Number of iterations T to estimate N^*_{bg} from Algorithm 2	25	25
Number of MC trials M in each estimation of N^i_{bg} in Algorithm 2	10	10

Table S4. Simulation Parameters in Fig. 5(d)

Percentage of couplings c	$c =$ 0%, 6% or 21%
Density ρ_A	$\rho_A =$ 5 μm^{-2}
Density ρ_B	$\rho_B =$ 5, 10, 15 μm^{-2}
Area S	$S =$ 100 μm^2
Number of AB	$\rho_A S c$
Number of background A	$\rho_A S(1-c)$
Number of background B	$\rho_B S(1-c)$
K_{eq}	$\min(\rho_A, \rho_B)c/(1-c)^2 \rho_A \rho_B$
True distance d_{true}	Sampled from Unif(0,10 nm)
Localization Precision σ_A, σ_B	Sampled from Fig. S6
d_{lower}	0 nm
d_{upper}	20 nm
Search distance factor threshold C from SN 3	4
Number of sampling points to estimate probability N from Algorithm 1	10^5
Number of iterations T to estimate N^*_{bg} from Algorithm 2	25
Number of MC trials M in each estimation of N^i_{bg} in Algorithm 2	10

Table S5. Simulation Parameters in Fig. 6(c-d)

Density ρ	$\rho_A = 30, 20, 10$, $\rho_B = 20$ μm^{-2}
Area S	100 μm^2
Number of AB at time t	From stochastic simulation
Number of background A at time t	From stochastic simulation
Number of background B at time t	From stochastic simulation
True distance d_{true}	Sampled from Unif(0,10 nm)
Localization Precision σ_A, σ_B	Sampled from Fig. S6
d_{lower}	0 nm
d_{upper}	20 nm
Search distance factor threshold C from SN 3	4
Number of sampling points to estimate probability N from Algorithm 1	10^5
Number of iterations T to estimate N^*_{bg} from Algorithm 2	25
Number of MC trials M in each estimation of N^i_{bg} in Algorithm 2	10

Supplementary Note 8: Comparing GMO's performance against an alternative method

Most colocalization methods focus on evaluating spatial correlations between the two channels, and do not directly address colocalization between two specific signals (11; 12; 13; 14; 15). In order to compare our method (GMO) against an alternative, we designed an algorithm based on the ideas in (16). We called this the *Minimal Pairwise Distance Method* (MinDist), as it aims to minimize distances between pairs.

A. Colocalization by the Minimal Pairwise Distance Method

MinDist sorts all pairwise distances d_{obs} (within a threshold d_{thres}), and selects colocalized pairs starting with the smallest distances. After each pair has been selected, their corresponding signals are no longer available for further colocalization, so we delete their pairwise distances to any other localizations from the list. The pseudo-code is provided in Algorithm 3.

Algorithm 3 Minimal Pairwise Distance Method (MinDist)

Input: d_{thres}, list of observed positions $\{O_{A_1},\ldots,O_{A_N}\}$, $\{O_{B_1},\ldots,O_{B_M}\}$
Output: list of pairs $C=\{(i,j)\}$ of putative colocalizations

```
I ← [ ]
D ← [ ]
C ← [ ]
for n ← 1 to N, m ← 1 to M do
    if d_nm = dist(O_An, O_Bm) < d_thres then
        Add d_nm to D
        Add (n, m) to I
    end if
    Sort I and D by ascending order of D
    while I is not empty do
        Add the first entry (n, m) ∈ I to C
        Delete all entries (n, j) from I and the corresponding entries d_nj from D
        Delete all entries (i, m) from I and the corresponding entries d_im from D
    end while
end for
return C
```

Although MinDist is very intuitive, it has two obvious drawbacks. 1) It does not take into account localization precision values, and 2) it is unclear how d_{thres} should be chosen. The purpose of d_{thres} is to speed up computation. As $d_{\text{thres}} \to \infty$, all possible pairs are being considered. For finite d_{thres}, as it increases, more putative pairs are being considered, leading to increases in both true and false positives.

B. Parameters used in GMO and MinDist

We compared GMO and MinDist in two scenarios: short interaction distance ($15 \leq d_{\text{true}} \leq 25$ nm) and long interaction distance ($55 \leq d_{\text{true}} \leq 65$ nm). Parameters used in each scenarios are presented in Table S6.

As noted earlier, a larger value of d_{thres} in MinDist in general increases both the number of true positives (TP) and false positives (FP). To ensure a fair comparison, we set d_{thres} so that FP for GMO is nearly identical to that of MinDist (see right-panels of Figs. S7(a) and S8(a)).

Table S6. Parameters used in the comparison of GMO to MinDist

	Small interaction distance	**Large interaction distance**
Density ρ	Varying (see x-axis of Fig. S7)	Varying (see x-axis of Fig. S7)
Area S	$S = 100$	100
Number of AB	ρS	ρS
Number of background A	0	0
Number of background B	0	0
True distance d_{true}	Sampled from Unif(15, 25 nm)	Sampled from Unif(55, 65 nm)
Localization Precision σ_A, σ_B	Sampled from Fig. S6	Sampled from Fig. S6
d_{lower}	10 nm	45 nm
d_{upper}	30 nm	75 nm
Search distance factor threshold C from SN 3	4	4
Number of sampling points to estimate probability	10^5	10^5
d_{thres} in minimal pairwise distance method	85 nm	90 nm

C. Results from comparing GMO to MinDist

We evaluated the performance of the two methods by comparing the number of true positives per micron (TP), the number of false positives per micron (FP), the TP ratio of GMO to MinDist, recall, and precision, where

$$\text{Recall} = \frac{\text{TP}}{\text{TP}+\text{FN}} \quad \text{and} \quad \text{Precision} = \frac{\text{TP}}{\text{TP}+\text{FP}}.$$

In all these metrics, GMO outperformed MinDist.

Consider first the case of short interaction distance, where we sampled d_{true} from $[15, 25]$ nm uniformly. Fig. S7(a) plots TP and FP for GMO and MinDist. We chose d_{thres} to minimize the difference in FP. GMO clearly outperformed MinDist by at least 10%, as demonstrated by the ratio of TPs in Fig. S7(b). Moreover, this performance improved slightly as the density of localization signals increased. Fig. S7(c-d) show recall and precision of the two algorithms. GMO's recall was constantly $\sim$10% better than MinDist's, while precision was marginally better.

We also considered the scenario where the interaction distance was large. For example, (13) reported an interaction distance of $\sim$60 nm between PSD95 and Homer, which are involved in the functioning of synapses. For long interaction distance, we uniformly sampled d_{true} from [55, 65] nm. Fig. S8(a) shows TP and FP for both GMO and MinDist, where again, d_{thres} for MinDist was chosen to minimize the difference in FP's. Here GMO performed significantly better than MinDist at all densities, by $\sim$40-50% (Fig. S8(b)). In terms of recall, Fig. S8(c) shows that GMO outperformed MinDist by $\sim$25% across all densities, and GMO's precision was increasingly better than MinDist's at higher density (Fig. S8(d)).

At all interaction distances, GMO constantly outperformed MinDist both in the number of true positives discovered and recall, across all densities of localizations. The outperformance was more significant for longer interaction distances. This is unsurprising as MinDist prioritizes *shorter* pairwise distances for its colocalizations.

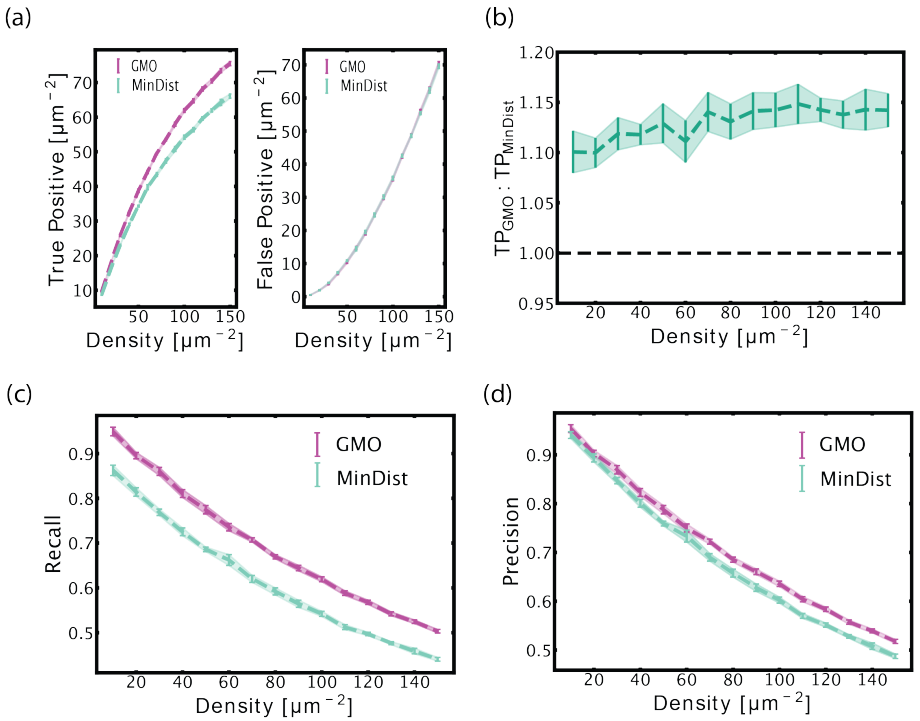

Figure S7. Performance of GMO versus MinDist at short interaction distance. (a) True Positives (TP) and False Positives (FP) discovered at different densities. (b) The ratio of TPs from the two methods, where TP_{GMO}:$\text{TP}_{\text{MinDist}} > 1$ means GMO outperforms MinDist at identifying TPs. In terms of (c) recall and (d) precision, GMO outperformed MinDist. Error bars indicate ± 1 standard deviation over 10 simulated datasets.

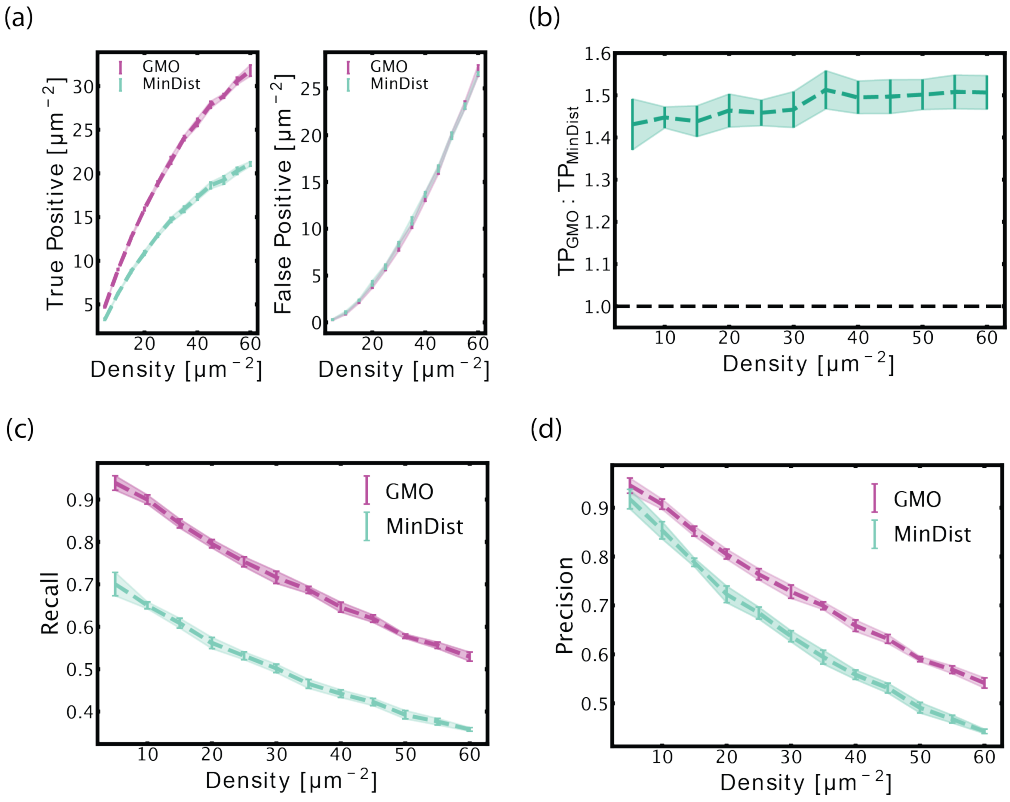

Figure S8. Performance of GMO versus MinDist at long interaction distance. (a) True Positives (TP) and False Positives discovered at different densities. (b) The ratio of TPs from the two method, where TP_{GMO}:$\text{TP}_{\text{MinDist}} > 1$ means GMO does better picking up TPs. In terms of (c) recall and (d) precision, GMO outperformed MinDist. Error bars indicate ± 1 standard deviation over 10 simulated datasets.

Supplementary Note 9: Experimental materials and methods

A. Cells and DNA constructs

A.1. Cell lines. Two cell lines were used: Human Embryonic Kidney 293FT (HEK293FT) cells and HEK293 cells with a CRISPR/Cas9 knockout of the G-protein Gs (HEK293 ΔGs). HEK293 ΔGs cells were received as a gift from Dr. Asuka Inoue at Tohoku University (17; 18; 19).

A.2. Cell culture and transfection. Cells were cultured in Dulbecco's Modification of Eagle's Medium (DMEM) with 4.5 g/L glucose, L-glutamine & sodium pyruvate (Corning CAT #10-013-CV). DMEM (500 mL) was supplemented with 50 mL of Fetal Bovine Serum (FBS) (VWR CAT #89510-186) and 5 mL of 100x Penicillin-Streptomycin solution (Corning CAT #30-002-CI). Media was sterilized by filtration with a 0.2 µm PES bottle filter (CAT #10040-436). Culture media for live-cell confocal imaging was prepared the same way except that Phenol-Red-free FluoroBrite DMEM (Fisher Scientific CAT #A1896701) was used.

Cells were cultured in 25 cm^2, 75 cm^2, and 175 cm^2 treated culture flasks (VWR #10062-872, GenClone #25209, Corning #353136, VWR #10062-864) using standard sterile techniques. Passaging was done using a solution of 0.05% Trypsin + 0.53 mM EDTA in Phosphate-Buffered Saline (PBS) without calcium and magnesium (Corning CAT #21-040-CV). A 10x stock of this buffer was made by diluting 2.5% trypsin (ThermoFisher #15090046) and powdered EDTA (CalbioChem CAT #4050) to 0.5% trypsin + 5.3 mM EDTA, pH = 7.4. Stock solution was frozen in 1 mL aliquots at -20°C in 15 mL conical centrifuge tubes. Before use, the 1 mL aliquot was thawed by adding 9 mL of PBS directly to the tube.

Confluency of 80% was targeted for optimal transfection. Cells were seeded either in 6-well or 12-well plates for transfection. Catalog numbers for 6, 12, 24, 48, and 96 well plates used at various points in these experiments are VWR CAT #10062-892, 10062-894, 10062-896, 10062-898, 10062-900. Cells were transfected using either Lipofectomene 3000 (ThermoFisher CAT #L3000008) or Polyplus JetOptimus (Polyplus-Sartorius CAT #101000006) according to manufacturer's protocols. DNA amounts were scaled according to the surface area of the well plate used. For example, the following amounts of DNA were used for a 6-well plate or a 35 mm diameter dish:

- Halo-TM-SNAP: 1 µg
- Halo-β2AR: 1 µg
- GαS-SNAP: 1 µg
- SNAP-CaaX: 2.5 µg
- Gβ: 0.5 µg
- Gγ: 0.2 µg

A.3. Plasmids. The following plasmids were used. Full circular sequence maps for all plasmids are provided as downloadable files.

- pHR-CMV-Tet02_Twin-Strep: pHR-CMV-Tet02_Twin-Strep was purchased from Addgene (plasmid #113833).
- Halo-TM-SNAP was received as a gift from Dr. Johannes Broichhagen at Leibniz-Forchungsinstitut for Molecular Pharmocology (20). Amino acid sequence of Halo-TM-SNAP:

```
MVPSSDPLVTAASVLEFGLGISTMETDTLLLWVLLLWVPGSTGDYPYDVPDYAGAQPARSGSEIG
TGFPFDPHYVEVLGERMHYVDVGPRDGTPVLFLHGNPTSSYVWRNIIPHVAPTHRCIAPDLIGMG
KSDKPDLGYFFDDHVRFMDAFIEALGLEEVVLVIHDWGSALGFHWAKRNPERVKGIAFMEFIRPI
PTWDEWPEFARETFQAFRTTDVGRKLIIDQNVFIEGTLPMGVVRPLTEVEMDHYREPFLNPVDRE
PLWRFPNELPIAGEPANIVALVEEYMDWLHQSPVPKLLFWGTPGVLIPPAEAARLAKSLPNCKAV
DIGPGLNLLQEDNPDLIGSEIARWLSTLEISGVDEQKLISEEDLNAVGQDTQEVIVVPHSLPFKV
VVISAILALVVLTIISLIILIMLWQKKPRGAQPARSMDKDCEMKRTTLDSPLGKLELSGCEQGLH
EIIFLGKGTSAADAVEVPAPAAVLGGPEPLMQATAWLNAYFHQPEAIEEFPVPALHHPVFQQESF
TRQVLWKLLKVVKFGEVISYSHLAALAGNPAATAAVKTALSGNPVPILIPCHRVVQGDLDVGGYE
GGLAVKEWLLAHEGHRLGKPGLGGRLEVLFQGVD
```

- Halo-β2AR was purchased from Addgene (plasmid #66994) and transformed into backbone vector #113833. Amino acid sequence of Halo-β2AR:

```
MATGSRTSLLLAFGLLCLPWLQEGSAFPTIPLSGSEIGTGFPFDPHYVEVLGERMHYVDVGPRDG
TPVLFLHGNPTSSYLWRNIIPHVAPSHRCIAPDLIGMGKSDKPDLDYFFDDHVRYLDAFIEALGL
```

EEVVLVIHDWGSALGFHWAKRNPERVKGIACMEFIRPIPTWDEWPEFARETFQAFRTADVGRELI
IDQNAFIEGALPMGVVRPLTEVEMDHYREPFLKPVDREPLWRFPNELPIAGEPANIVALVEAYMN
WLHQSPVPKLLFWGTPGVLIPPAEAARLAESLPNCKTVDIGPGLFLLQEDNPDLIGSEIARWLPG
LAGSGGGGSMGQPGNGSAFLLAPNRSHAPDHDVTQQRDEVWVVGMGIVMSLIVLAIVFGNVLVIT
AIAKFERLQTVTNYFITSLACADLVMGLAVVPFGAAHILMKMWTFGNFWCEFWTSIDVLCVTASI
ETLCVIAVDRYFAITSPFKYQSLLTKNKARVIILMVWIVSGLTSFLPIQMHWYRATHQEAINCYA
NETCCDFFTNQAYAIASSIVSFYVPLVIMVFVYSRVFQEAKRQLQKIDKSEGRFHVQNLSQVEQD
GRTGHGLRRSSKFCLKEHKALKTLGIIMGTFTLCWLPFFIVNIVHVIQDNLIRKEVYILLNWIGY
VNSGFNPLIYCRSPDFRIAFQELLCLRRSSLKAYGNGYSSNGNTGEQSGYHVEQEKENKLLCEDL
PGTEDFVGHQGTVPSDNIDSQGRNCSTNDSLL

- GαS-SNAP was designed in our lab and synthesized by TWIST Biosciences. The gene was sub-cloned into backbone vector #113833 using Gibson Assembly. SNAP-tag placement was chosen according to functional studies with a BRET-based sensor (21). Amino acid sequence of GαS-SNAP:

MGCLGNSKTEDQRNEEKAQREANKKIEKQLQKDKQVYRATHRLLLLGAGESGKSTIVKQMRILHV
NGFNGEGGEEDPQAARSNSDGEKATKVQDIKNNLKEAIETIVAAMSNLVPPVELANPENQFRVDY
ILSVMNSGGGGTRSTGMDKDCEMKRTTLDSPLGKLELSGCEQGLHRIIFLGKGTSAADAVEVPAP
AAVLGGPEPLMQATAWLNAYFHQPEAIEEFPVPALHHPVFQQESFTRQVLWKLLKVVKFGEVISY
SHLAALAGNPAATAAVKTALSGNPVPILIPCHRVVQGDLDVGGYEGGLAVKEWLLAHEGHRLGKP
GLGGTATGGGGSVPDFDFPPEFYEHAKALWEDEGVRACYERSNEYQLIDCAQYFLDKIDVIKQAD
YVPSDQDLLRCRVLTSGIFETKFQVDKVNFHMFDVGGQRDERRKWIQCFNDVTAIIFVVASSSYN
MVIREDNQTNRLQEALNLFKSIWNNRWLRTISVILFLNKQDLLAEKVLAGKSKIEDYFPEFARYT
TPEDATPEPGEDPRVTRAKYFIRDEFLRISTASGDGRHYCYPHFTCAVDTENIRRVFNDCRDIIQ
RMHLRQYELL

- Gβ was purchased from Addgene (plasmid #133856). It was sub-cloned into backbone vector #113833 without the fluorescent tag. Amino acid sequence of Gβ:

MSELDQLRQEAEQLKNQIRDARKACADATLSQITNNIDPVGRIQMRTRRTLRGHLAKIYAMHWGT
DSRLLVSASQDGKLIIWDSYTTNKVHAIPLRSSWVMTCAYAPSGNYVACGGLDNICSIYNLKTRE
GNVRVSRELAGHTGYLSCCRFLDDNQIVTSSGDTTCALWDIETGQQTTTFTGHTGDVMSLSLAPD
TRLFVSGACDASAKLWDVREGMCRQTFTGHESDINAICFFPNGNAFATGSDDATCRLFDLRADQE
LMTYSHDNIICGITSVSFSKSGRLLLAGYDDFNCNVWDALKADRAGVLAGHDNRVSCLGVTDDGM
AVATGSWDSFLKIWN

- Gγ was purchased from Addgene (plasmid #133857). It was sub-cloned into backbone vector #113833 without the fluorescent tag. Amino acid sequence of Gγ:

MASNNTASIAQARKLVEQLKMEANIDRIKVSKAAADLMAYCEAHAKEDPLLTPVPASENPFREKK
FFCAIL

- SNAP-CaaX: pSNAP-CaaX was purchased from Addgene (plasmid #101133). Amino acid sequence of SNAP-CaaX:

MELHRGGGRDIKLTMDKDCEMKRTTLDSPLGKLELSGCEQGLHEIKLLGKGTSAADAVEVPAPAA
VLGGPEPLMQATAWLNAYFHQPEAIEEFPVPALHHPVFQQESFTRQVLWKLLKVVKFGEVISYQQ
LAALAGNPAATAAVKTALSGNPVPILIPCHRVVSSSGAVGGYEGGLAVKEWLLAHEGHRLGKPGL
GPAGCMSCKCVLS

B. Confocal imaging

B.1. Fluorescent labeling of cells for confocal imaging. We used coated glass-bottom dishes directly from the manufacturer with no additional cleaning. A list of dishes used: Matek CAT #P35G-1.5-14-C, VWR CAT #75856-742, Cellvis CAT #d35-14-1.5-n. Cells were seeded onto glass-bottom dishes and labeled with Janelia Fluor (JF) dyes. The dyes used were JF479-HaloTag, JF552-HaloTag, JF552-cpSNAP-tag, JF525-HaloTag, JF549i-SNAP-tag, JF549i-HaloTag, and JFX646-SNAP-tag. Identities of dyes used in specific experiments can be found in figure captions. Long-term, all dye stocks were stored at 500 µM at -20°C, and at 4°C in a dark box for short-term storage.

For cell labeling, a dye solution in cell media (DMEM + 10% FBS + PS) was prepared at a final concentration of 2 µM. Solution volume was sufficient to fully cover the bottom surface of the dish (e.g., 500 µL for a 35 mm dish). For

the labeling reaction, the dye solution was warmed to 37°C and cells were incubated with this solution at 37°C for 15-20 minutes, then washed. The protocol for one wash cycle was as follows:

- Dye/DMEM solution was aspirated.
- PBS at 37°C was added gently and swirled. Enough solution was added to cover the bottom of the dish completely.
- PBS was aspirated.
- Steps 2 and 3 were repeated once.
- Fresh cell media (DMEM + 10% FBS + PS) was added and cells were placed in the incubator.

After a 15-minute incubation, the entire wash cycle was repeated but in the final wash step, phenol-red-free FluoroBrite DMEM was used. Next, confocal imaging was performed.

We found that without an incubation step between wash cycles, background dye remained inside the cell. Additional incubation time allowed this non-specifically bound dye to diffuse out of the cell, which improved imaging contrast.

B.2. Confocal image acquisition. Confocal image acquisition was performed at the Harvard Center for Biological Imaging (HCBI) using a Zeiss LSM 880 confocal microscope with a Zeiss Plan-Apochromat 63x / 1.40 NA oil DIC M27 objective (Zeiss CAT #420782-9900-000). Immersol 518F imaging oil (Microscope World CAT #444960-0000-000) was used. Images were acquired at a pixel depth of 16 bits. Images were acquired in frame scanning mode with line averaging set to 2 per pixel. Pixel dwell time was 12.20 µs for Halo-TM-SNAP with a pixel size of 98 nm, and 16.38 µs for Halo-β2AR/SNAP-CaaX with a pixel size of 88 nm. Green fluorescence was excited with a 488 nm multi-line argon laser. Red fluorescence was excited with a DPSS 561 nm laser. Far-red fluorescence was excited with a HeNe 633 nm laser. The lasers were directed to the objective via a 458/561 multi-beamsplitter for green and red excitation and a 514/561/633 multi beamsplitter for far red excitation. The following laser powers and gain settings were used to acquire images presented in figures:

Halo-β2AR/SNAP-CaaX:
561: laser power 0.056 mW ($\sim$44 kW/cm^2), gain 750, detection window 570-651 nm
633: laser power 0.062 mW ($\sim$39 kW/cm^2), gain 800, detection window 638-755 nm

Halo-TM-SNAP:
488: laser power 0.165 mW ($\sim$173 kW/cm^2), gain 811.4, detection window 490-570 nm
561: laser power 0.282 mW ($\sim$224 kW/cm^2), gain 650, detection window 566-679 nm

The pinhole was set to 1 AiryUnit. The spectral point detector used was a 32 channel GaAsP PMT for green and red emission plus the IR enhanced PMT 2 for far red emission.

B.3. Confocal image processing. Confocal images were adjusted for brightness and contrast using Fiji software. All original images are available upon request.

C. SMLM imaging

C.1. Preparation of mounting surfaces. We mounted coverslips on plain pre-cleaned glass slides, 25 x 75 mm, 1 mm thick (VWR CAT #48300-026). Glass slides were cleaned with methanol (Sigma Aldrich CAT #34860-1L-R) right before mounting to remove any dust. The mounting surface was covered with methanol using a dropper pipette and wiped once in one direction with a KimWipe. Any remaining methanol quickly evaporated, and the slides were ready for mounting. We used 22 x 22 mm No. 1.5 glass coverslips (VWR CAT #48366-227). Coverslips were placed on a Wash-N-Dry coverslip rack (Sigma Aldrich CAT #Z688568) to expose both sides of the surface. The rack was placed in a glass beaker and submerged in 1 M HCl. The beaker was covered with 2 layers of aluminum foil and heated to 60°C for 4-6 hours. Coverslips were then washed twice by immersing the entire coverslip rack into fresh DI water each time. Finally, coverslips were washed one additional time using 100% ethanol. The coverslips were stored in ethanol for up to 1 hour. Next, coverslips were removed from 100% ethanol and allowed to dry inside of a sterile TC hood. One coverslip was placed in each well of a 6 well tissue culture plate and 2 mL of DMEM + 10% FBS + PS were added.

C.2. Fluorescent labeling. Cells seeded in a multi-well plate were labeled with photo-activatable (PA) Janelia Fluor (JF) dyes. We used PA-JF549-HaloTag and PA-JF646-SNAP-tag dyes. Identities of dyes used in specific

experiments can be found in figure captions. All dyes were stored at 500 µM at -20°C for long-term storage. In contrast to confocal dyes, SMLM dyes were not stored at 4°C.

For cell labeling, a dye solution in cell media (DMEM + 10% FBS + PS) was prepared at a final concentration of 2 µM. Solution volume was sufficient to fully cover the bottom surface of the dish (e.g., 500 µL for a 35 mm dish). For the labeling reaction, the dye solution was warmed to 37°C and cells were incubated with this solution at 37°C for 15-20 minutes, then washed. The protocol for one wash cycle was as follows:

- Dye/DMEM solution was aspirated.
- PBS at 37°C was added gently and swirled. Enough solution was added to cover the bottom of the dish completely.
- PBS was aspirated.
- Steps 2 and 3 were repeated once.
- Fresh cell media (DMEM + 10% FBS + PS) was added and cells were placed in the incubator.

After a wash cycle cells were allowed to incubate at 37°C for 15 minutes. We repeated this wash cycle with the 15-minute incubation step 6 total times. Trypsin + EDTA (prepared as described in "B. Cell Culture and Transfection") was added to detach cells. Next, cells were seeded at low confluency (∼25%) onto the HCl-washed coverslips (from "C.1. Preparation of Mounting Surfaces").

C.3. Fixation buffer. Fixation buffer was prepared using 4% w/v granular paraformaldehyde (PFA) (Electron Microscopy Sciences CAT #19208) in PBS (∼50-500 mL total volume). PFA/PBS mixture was heated to 50-60°C using a heating plate while stirring with a stir bar. Initially, the solution was cloudy as not all PFA dissolved. Next, 1 M NaOH in Milli-Q water was added drop-wise until the PFA/PBS solution turned clear. The solution was stirred at 50-60°C until all PFA dissolved, which typically took 5-10 minutes. The pH was adjusted to 7.4 and the solution was frozen at -20°C in 10 mL aliquots in 15 mL conical centrifuge tubes. When needed, aliquots were thawed in water at room temperature and sonicated until completely clear. Once thawed, aliquots were not refrozen.

C.4. Fixation protocol. Media was aspirated from dishes with fluorescently labeled cells (from "C.2. Fluorescent Labeling") and replaced with fixation buffer. Cells were then kept in fixation buffer at room temperature for 15 minutes. During this process the dishes were swirled gently every 5 minutes. After 15 minutes of fixation, cells were washed with room-temperature PBS three times and stored at 4°C if not imaged immediately.

Immediately before imaging, PBS was gently aspirated. 0.1 µm TetraSpek Microspheres (Thermo-Fisher CAT #T7279) were added to the glass coverslip containing cells to act as fiducial markers. The concentration from the manufacturer was reported to be $\sim 1.8 \times 10^{11}$ particles/mL. We diluted Microspheres by a factor of 200 in Milli-Q water and sonicated before each use. This dilution factor led to the desired surface density of Microspheres and was determined empirically by serial dilution and imaging. We aimed for 5-10 Microspheres per field of view. 10 µL of diluted Microspheres were added to each coverslip. Diluted Microspheres were stored at 4°C.

C.5. Mounting coverslips with cells for SMLM imaging. To mount coverslips, a few µL of Fluoromount-G mounting media (ThermoFisher CAT #00-4958-02) were added to the cleaned microscope slide (from "C.1. Preparation of Mounting Surfaces"). Fluoromount-G is extremely viscous. To ensure there were no bubbles, 20 µL of Fluoromount-G were drawn up in a pipette and a small portion was expelled. Next, a drop of Fluoromount-G was added to the microscope slide. A fresh pipette tip was then used to remove any air bubbles within the deposited drop of Fluoromount-G. Coverslips with cells facing the mounting media were mounted on top of Fluoromount-G and allowed to set for 10 minutes. After 10 minutes, the edges of the coverslips were sealed with clear nail polish and allowed to dry for ∼10 minutes. The samples were then moved to a light-proof slide storage box. Samples were stored at 4°C.

C.6. SMLM image acquisition. SMLM image acquisition was performed on a Zeiss Elyra microscope at the HCBI with a Zeiss Plan-Apochromat 63x / 1.40 NA oil DIC M27 objective. Immersol 518F was used.

PA-JF549-HaloTag was excited with a DPSS 561 nm laser. 561 nm laser power measured after the objective without oil was 30.1 mW. With the objective removed, the power at the back focal plane was 40.3 mW. In parallel with 561 nm excitation, PA-JF646-SNAP-tag was excited with a 642 nm HeNe laser. HeNe laser power measured after the objective without oil was 30.2 mW. With the objective removed, the power at the back focal plane was 44.0 mW. Under these illumination conditions, the on-time of a single dye molecule was between 1 and 3 frames (100-300 ms). The sample was also illuminated with a 405 nm photo-activation laser between 561 nm and 641 nm excitation and collection. A schematic of a single frame of imaging is provided in Fig. S9. 405 nm laser power was increased

over the duration of the experiment – as localizations became less frequent – from 0.031 mW to 0.626 mW measured after the objective without oil, corresponding to 0.052 mW and 1.04 mW respectively in the back focal plane.

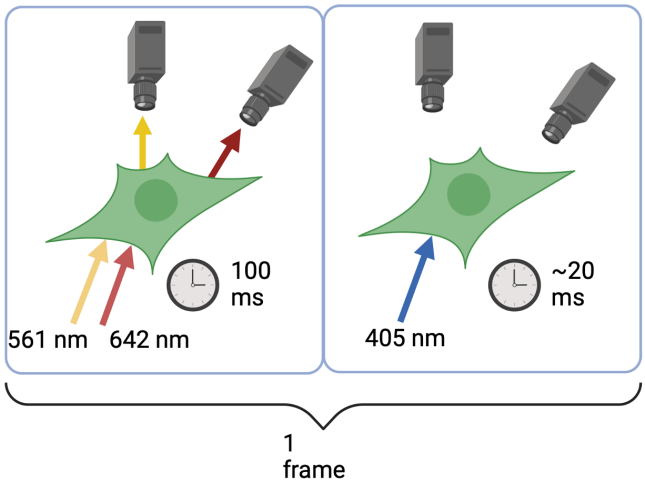

Figure S9. A single frame of SMLM illumination. Each frame has an exposure time of 100 ms for the 561 and 642 nm lasers. For approximately 20 ms during frame transfer and camera data processing, 405 nm laser light stochastically activates more fluorophores. 405 nm light is not present during camera acquisition.

The sample was initially illuminated with 561 nm and 642 nm laser lines (without 405 nm illumination) to photo-bleach prematurely photo-activated fluorophores. After 50-100 frames, the 405 laser was turned on.

Emission from both dyes was collected on two spatially registered cameras. The main beamsplitter for both channels was a quad 405/488/561/642 mirror. The emission was directed to two cameras through a DuoLink unit using a band-pass 490-560 + long-pass 640 dichroic camera splitter that reflected red emission (560-640) to camera 1 and far red emission (640+) to camera 2. The cameras used were Dual PCO.edge 4.2 CLHS sCMOS camera. The frame exposure time was set to 100 ms at 16 bits. The frame size was 49x49 µm.

Supplementary Note 10: Analysis of experimental SMLM data

A. Preprocessing experimental SMLM images

A.1. Generating localizations from SMLM images. SMLM images were processed using Zeiss Zen software to identify localizations. To select localizations, peak mask size was set to 9 pixels and fit a Gaussian mask over potential localizations. Peak intensity-to-noise ratio was set to 6, which selected among potential localizations and kept localizations if the standard deviation of its intensity against the rest of the image was above 6. Overlapping localizations (as defined by the mask of an event) were discarded. Localizations were filtered by localization precision (7 to 40 nm), number of photons (700-8000 photons) and experimental PSF size (90-180 nm). Localization precision was estimated by Zen using a Monte-Carlo Simulation of the experimental PSF including noise. These localization filters were selected based on respective localization histograms. They primarily excluded out-of-focus fluorophores and any dim fluorescent background. Spatially correlated localizations were grouped to aggregate emission of a single fluorophore that persisted over multiple camera frames. Localizations were then corrected for drift as described below.

A.2. Drift correction. Typical SMLM data acquisition lasted from 20 to 40 minutes per image. To characterize drift, we imaged 100 nm TetraSpek Microspheres mounted between a glass slide and a coverslip as described above. We acquired 10,120 frames over approximately 33 minutes in two colors and observed drift on the order of several hundred nanometers to microns (Fig. S10(a)). Drift was corrected using Zen's model-based drift correction algorithm. Zoomed-in images of a bead before and after drift correction are shown in Fig. S10(b). After drift correction was applied, the full width at half maximum of the cluster of localizations was 18 nm for the green channel and 29 nm for the magenta channel.

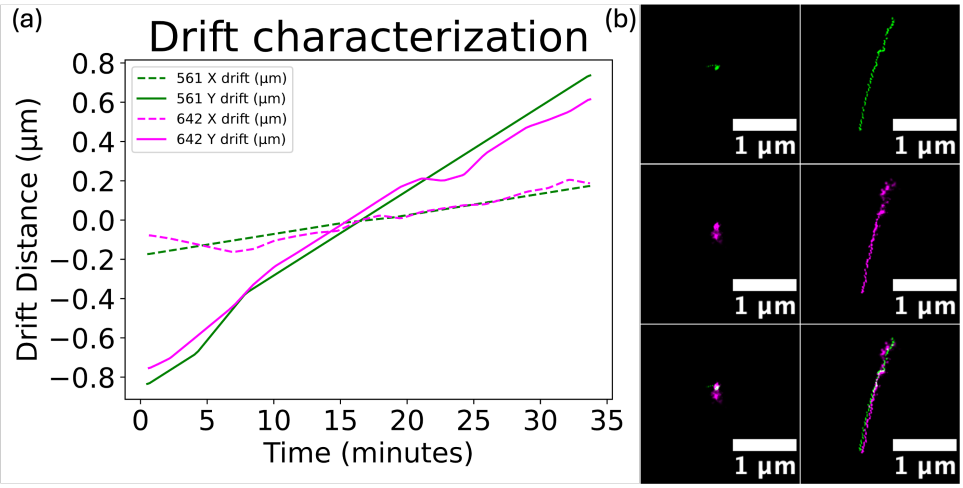

Figure S10. Model based drift correction algorithm was applied to the imaging signal derived from a 100 nm TetrakSpek fluorescent Microsphere. (a) Drift in the x and y direction for both 561 and 642 nm channels. (b) Drift correction as a function of frame number was applied in each channel. Before and after images of the two channels show the drift in the two channels as well as the efficacy of the correction.

A.3. Removing areas of high density and dividing into subregions. Next, we generated a scatter plot based on the positions of drift-corrected localization signals in each cell. In order to remove extraneous localizations outside of the cell, we drew a boundary made of line segments to approximate the shape of the cell. The resulting polygon defined Ω, the effective region of analysis. Any localizations outside of Ω were discarded from further analysis. Next, we eliminated localization signals that were part of tight clusters, using the clustering algorithm DBSCAN (22), with parameters eps $= 75$ nm and min_samples $= 10$.

Then we divided Ω into smaller subregions, by overlaying Ω with $d \times d$ equally sized boxes, where by default $d = 10$. See Fig. S11 for an example. Each subregion R_{ij} was defined as where Ω and one of these boxes overlap, with effective area S_{ij}. Any subregion, where in at least one of two channels the density was $> 35\,\mu\text{m}^{-2}$, was again discarded.

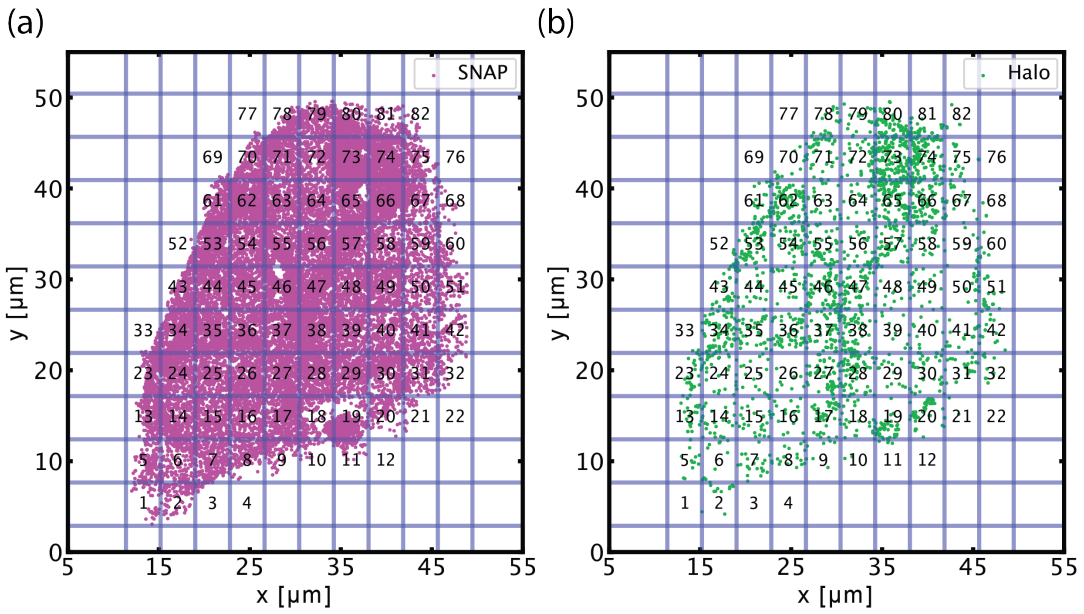

Figure S11. Divide into subregions by overlaying with $d \times d$ boxes ($d = 10$). (a) SNAP localizations, and (b) Halo localizations in a cell expressing Halo-TM-SNAP.

B. Analysis of experimental SMLM dataset

On each subregion R_{ij}, we applied our pipeline (Fig. S4) to the $N_{A,ij} + N_{B,ij}$ localizations to estimate the number of coupled pairs $N_{\text{coupled},ij}$ in R_{ij}. Note that in principle, iMEC can estimate more background pairs than the number of pairs in a subregion (i.e., R_{ij} has less couplings than uniform random distributions of signals), in which case $N_{\text{coupled},ij} = 0$. We calculated the % of coupling as

$$q_{ij} = \frac{N_{\text{coupled},ij}}{\min(N_{A,ij}, N_{B,ij})}.$$

In Fig. 5 of the main text we reported the average $\langle q_{ij} \rangle$ for each cell.

The average over entire cells supported the same qualitative conclusion as from considering subregions with similar densities, that the percentage of couplings in cells expressing Halo-TM-SNAP was significantly higher than cells expressing Halo-β2AR/SNAP-CaaX. In Fig. S12, we plotted $\langle q_{ij} \rangle$, averaged over subregions with similar densities. More specifically, these were subregions whose densities $\min\{\rho_{A,ij}, \rho_{B,ij}\}$ lie in [0,4], [4,8], [8, 12], or [12, 16] μm^{-2} respectively. The percentage $\langle q_{ij} \rangle$ displayed larger variability between different density ranges, but overall, the percentage of coupling is significantly higher for cells expressing Halo-TM-SNAP than those expressing Halo-β2AR/SNAP-CaaX.

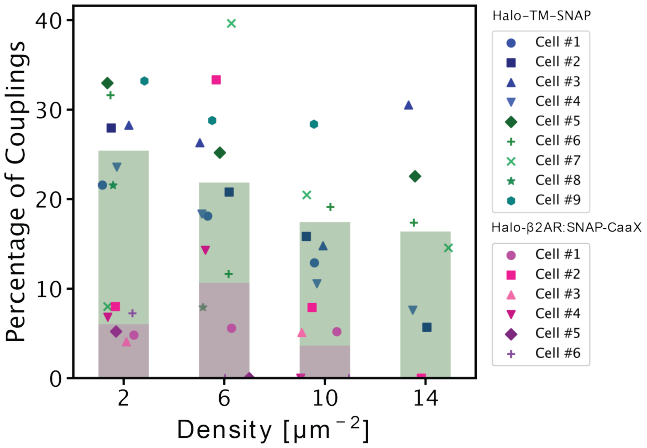

Figure S12. Percentage of couplings in subregions of certain densities (each marker represents a single cell). Across all densities, more couplings were identified in cells expressing Halo-TM-SNAP than those expressing Halo-β2AR/SNAP-CaaX.

C. Comments on choices made in the analysis

On dividing into subregions. Different local regions of the same cell spanned a range of densities. We opted to run the analysis with $d = 10$ as well as $d = 1$. With $d = 10$ (Fig. 5 of the main text), the % of couplings for both cells lines were nearly identical to when $d = 1$ (Fig. S13), with 21% couplings identified for Halo-TM-SNAP and 6% for Halo-β2AR/SNAP-CaaX.

On removing clusters and subregions with high densities. Clusters could have arisen for various reasons, including overcounting or biological factors. Biological factors include cellular recycling into vesicles and endosomal pockets that may trap non-specifically bound dye. These clusters exhibited significantly elevated local densities, typically exceeding 500 μm^2, which is beyond the range in which our algorithm performs well. Thus, clusters were removed by DBSCAN prior to analysis. Similarly, subregions with densities greater than $35\,\mu\text{m}^{-2}$ were excluded from analysis to maintain a certain level of performance.

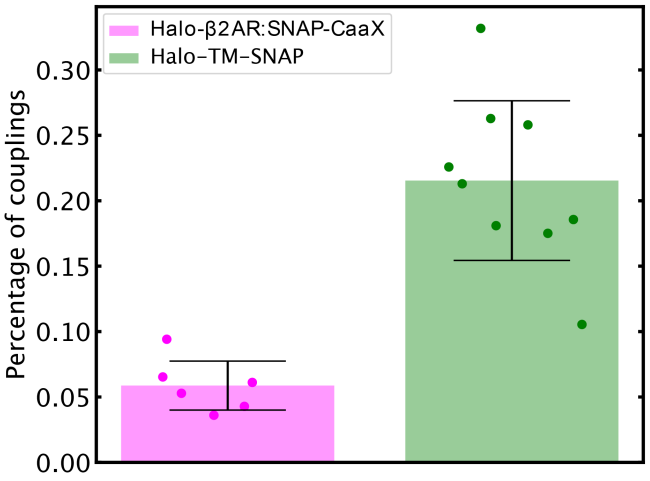

Figure S13. Percentage of couplings without dividing the cell into subregions (6 cells used for Halo-β2AR/SNAP-CaaX, and 9 cells used for Halo-TM-SNAP). Error bars indicate ±1 standard deviation.

Supplementary Note 11: Comments on analysis of experimental images

In an ideal scenario, we would expect the percentage of couplings for Halo-TM-SNAP to be 100%, and 0% for Halo-β2AR/SNAP-CaaX. The experimental data deviated from this scenario, with only ∼20% of couplings identified for cells expressing Halo-TM-SNAP (Figs. 5(d), S12, S13), and in certain subregions, >0% for Halo-β2AR/SNAP-CaaX (Fig. S12). In this section we discuss potential explanations for these observations.

A. Halo-TM-SNAP

The protein Halo-TM-SNAP is a transmembrane protein recombinantly fused to a HaloTag and a SNAP-tag (Fig. 5(a, right) of main text). Thus, ideally, we should have observed 100% couplings, however, not all tags were observed. The missing localizations could be ascribed to several causes. First, the labeling efficiency and the detection rate were not perfect. Second, certain localizations were excluded in the preprocessing step, based on localization precision, point spread function half-width, and what were believed to be instances of overcounting. Third, some dye molecules were activated prior to imaging, which was started after ∼100 frames.

In a positive control such as Halo-TM-SNAP, missing localizations affect the percentage of couplings identified. Consider for example, if only the HaloTag failed to be identified in a protein, the SNAP-tag that was detected might be paired with the HaloTag of a different protein by our algorithm. This situation may become too complex to parse in terms of the weighted bipartite graph in GMO. In Section C below, we showed that if molecules are detected at rate $1-r$ and the true number of couplings is N, then the expected number of couplings identified by our algorithm would be $N(1-r)^2$. For example, a detection rate of $r = 55\%$ would lead to a percentage of couplings ∼20%.

B. Halo-β2AR/SNAP-CaaX

Halo-β2AR and SNAP-CaaX are not expected to interact. Our pipeline identified a positive percentage of couplings for certain subregions of some cells, even though the expected percentage of couplings is 0%. In these subregions, we noticed that the two proteins were not uniformly distributed in space, but there were clusters. However, these clusters were not eliminated by DBSCAN (Fig. S14), which we suspect was due to our choice of parameters (eps=75 nm and min_samples=10). We also observed that when these clusters were present, localizations in both spectral channel were also dense in the same subregion, as demonstrated by Fig. S14.

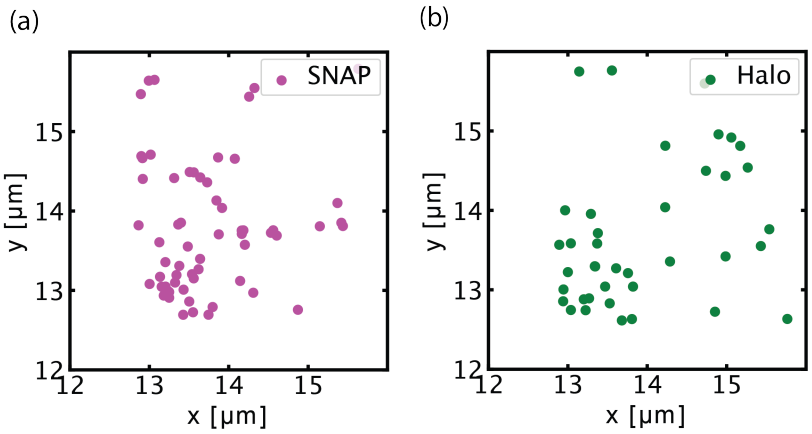

Figure S14. A subregion in a cell expressing Halo-β2AR/SNAP-CaaX. DBSCAN was not able to eliminate some clusters. The two channels also appeared to be correlated in terms of density in this subregion.

C. The estimated number of couplings scales with the square of detection rate

Here we considered the effect of missing detection on the percentage of couplings. In a system with N_{coupled} couplings, if localizations of A and B were dropped at rate r, i.e., a detection rate of $1-r$, then in theory, the detected number of couplings is expected to be $N_{\text{coupled}}(1-r)^2$, which was the conclusion from our numerical experiments.

We conducted two numerical experiments at varying densities. In the first, we modelled a positive control system like Halo-TM-SNAP ($N_A = N_B = N_{\text{coupled}}$); in the second, we assumed equilibrium between the interacting partner and the complex ($N_A = N_B = 2N_{\text{coupled}}$). In each experiment, we simulated SMLM datasets, then randomly deleted localizations of A and B at rate r, and ran our algorithm on the resulting datasets. The parameters used can be found in Table S7. Fig. S15 shows that in both numerical experiments, the numbers of couplings detected by our pipeline matched the theoretical prediction $N_{\text{coupled}}(1-r)^2$ closely. Note that in these numerical experiments, N_A and N_B were kept constant, so the number of couplings was proportional to the percentage of couplings.

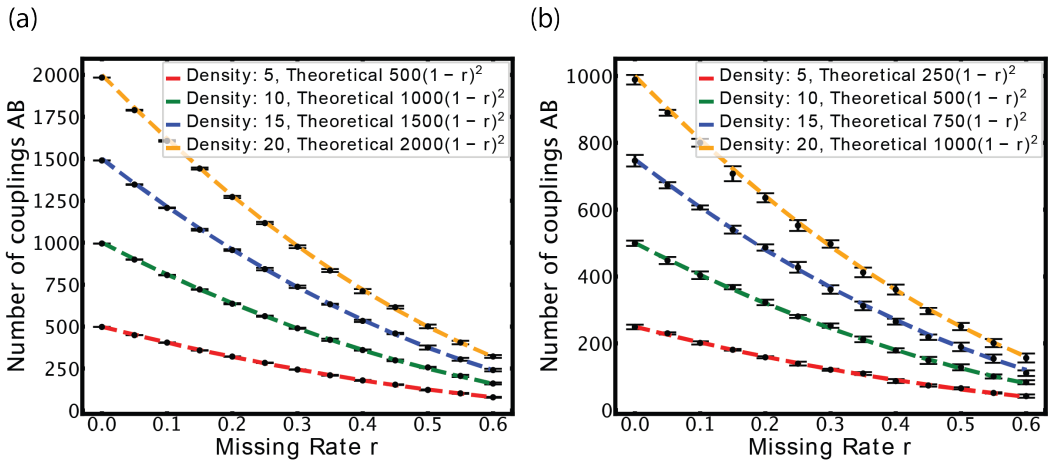

Figure S15. The number of couplings identified scales as $(1-r)^2$ where $1-r$ is the detection rate. (a) Positive control, where all molecules are pairs. (b) Equilibrium, where half the population are paired. Dashed Line: Theoretical Expression; Dots: Algorithm-Predicted Values. Error bars indicate ±3 standard deviations across 10 trials of iMEC.

Table S7. Parameters used to generate Fig. S15

Density $\rho_A = \rho_B = \rho$	See Fig. S15
Area S	$S = 100$ μm^2
Number of AB	ρS or $\rho S/2$
Number of background A	0 or $\rho S/2$
Number of background B	0 or $\rho S/2$
True distance d_{true}	Sampled from Unif(0, 10 nm)
Localization Precision	Sampled from Fig. S6
d_{lower}	0 nm
d_{upper}	20 nm
Search distance factor threshold C	4
Number of sampling points to estimate probability	10^5
Number of Monte Carlo steps to estimate N^*_{bg}	15
Number of Monte Carlo trails in each estimation of N^i_{bg}	10

Supplementary Note 12: Validation at chemical equilibrium

To see whether the algorithm could be used to infer the extent of couplings at the equilibrium, we applied it to the system $A + B \rightleftharpoons AB$ at equilibrium with equilibrium constant K_{eq}. The percentage of couplings is proportional to the concentration of AB, whose analytical expression can be derived:

$$[AB] := \frac{N_{AB}}{S} = \rho + \frac{1}{2K_{\text{eq}}} - \sqrt{\frac{1}{4K_{\text{eq}}^2} + \frac{\rho}{K_{\text{eq}}}}, \tag{S6}$$

where the density ρ refers to the total amount over an area S, i.e., $\rho = (N_A + N_{AB})/S$ and $\rho = (N_B + N_{AB})/S$. We generated simulated dataset for $K_{\text{eq}} = 0.1, 0.5, 1$ and ρ up to 100 molecules μm^{-2}. All parameters used can be found in Table S8.

As shown in Fig. S16, error remained below 3.5% across all densities and equilibrium constants. The error exhibited a slight tendency to grow as density increased. Note that when the number of interacting partners N_{AB} was small (e.g., for $K_{\text{eq}} = 0.1$ and low density), the large error was due to a small denominator in the definition of error.

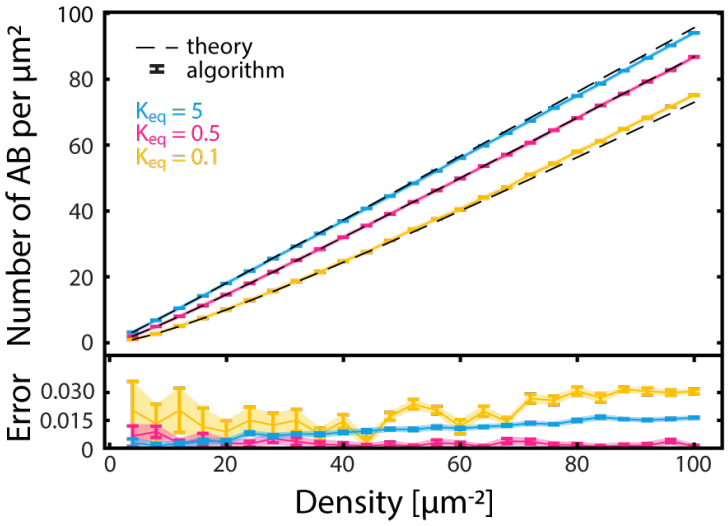

Figure S16. (Top) The density of AB estimated by the algorithm (markers) and the theoretical density (dashed lines) at different K_{eq} and densities. (Bottom) Error rate. Error bars indicate ± 1 standard deviation across 10 trials of iMEC.

Table S8. Parameters used to generate Fig. S16

Density ρ	Varying ρ (see x-axis)
Area S	$S = 100$ μm^2
Number of AB	N_{AB} (from Eq. S6)
Number of background A	$\rho S - N_{AB}$ from
Number of background B	$\rho S - N_{AB}$ from)
True distance d_{true}	Sampled from Unif(0, 10 nm)
Localization Precision σ_A, σ_B	Sampled from Fig. S6
d_{lower}	0 nm
d_{upper}	20 nm
Search distance factor threshold C from SN 3	4
Number of sampling points to estimate probability N from Algorithm 1	10^5
Number of iterations T to estimate N^*_{bg} from Algorithm 2	25
Number of MC trials M in each estimation of N^i_{bg} in Algorithm 2	10

Supplementary Note 13: The algorithm can estimate the level of coupling accurately across different density ratios

Colocalization analysis has been notably affected by density fluctuations. Measuring the degree of coupling becomes challenging when there are variations in the density ratio between A and B (23; 24; 15). A comprehensive assessment requires comparisons across diverse cells with varying densities of molecules A and B. Imagine a scenario where molecules of A (with density ρ_A) colocalize with molecules of B (with density ρ_B) at some percentage of colocalization. Now, if we increase the densities to $2\rho_B$ and $3\rho_B$ while maintaining the same percentage of couplings, an ideal colocalization algorithm should consistently yield the same level of couplings for these different density ratios. However, achieving this ideal outcome has proven challenging. Several approaches on incorporating densities to estimate the level of colocalization (using different definitions of colocalization) have been recently developed (14; 15).

Recall that we defined the percentage of couplings q as the number of coupling molecules divided by the minimum of the number of molecules of A and B. Here we would like to simulate a percentage of couplings γ to either be 3% (low binding) or 20% (high binding), and see whether our pipeline returns a percentage of coupling $\langle q \rangle$ that is close to γ or not. In a region with area S, N_A and N_B molecules of A and B respectively, i.e., at densities $\rho_A = N_A/S$, $\rho_B = N_B/S$, one would expect the number of coupling to be $N_{AB} = S\gamma \min\{\rho_A, \rho_B\}$. Correspondingly, the number of uncoupled A would be $N_A = (\rho_A - \gamma \min\{\rho_A, \rho_B\})S$. A similar calculation for the number of coupled B gives $N_B = (\rho_B - \gamma \min\{\rho_A, \rho_B\})S$. Thus, for different combinations of ρ_A and ρ_B, we generated a dataset with N_A, N_B, and N_{AB} molecules of A, B and AB respectively over a box of area S. Then we applied our algorithm and calculated the percentage of coupling $q(\rho_A, \rho_B)$. All parameters used are shown in Table S9.

A good algorithm would return a percentage of coupling $q(\rho_A, \rho_B)$ that is independent of ρ_A, ρ_B. Indeed, this is what we found; in Fig. S17 we plotted $q(\rho_A, \rho_B)$ for different combinations of (ρ_A, ρ_B), and found that $q(\rho_A, \rho_B)$ is approximately γ, which was set at either 3% or 20%.

Table S9. Simulation Parameters in Fig. S17

Density ρ_A, ρ_B	Varying (see x-axis)
Area S	100 μm^2
Percentage γ	$\gamma = 3\%$ or 20%
Number of AB	$S\gamma \min(\rho_A, \rho_B)$
Number of background A	$(\rho_A - \gamma \min(\rho_A, \rho_B))S$
Number of background B	$(\rho_B - \gamma \min(\rho_A, \rho_B))S$
True distance	Sampled from Unif(0,10 nm)
Localization Precision	Sampled from Fig. S6
d_{lower}	0 nm
d_{upper}	20 nm
Search distance factor threshold C	4
Number of sampling points to estimate probability	10^5
Number of Monte Carlo steps to estimate N^*_{bg}	15
Number of Monte Carlo trails in each estimation of N^i_{bg}	10

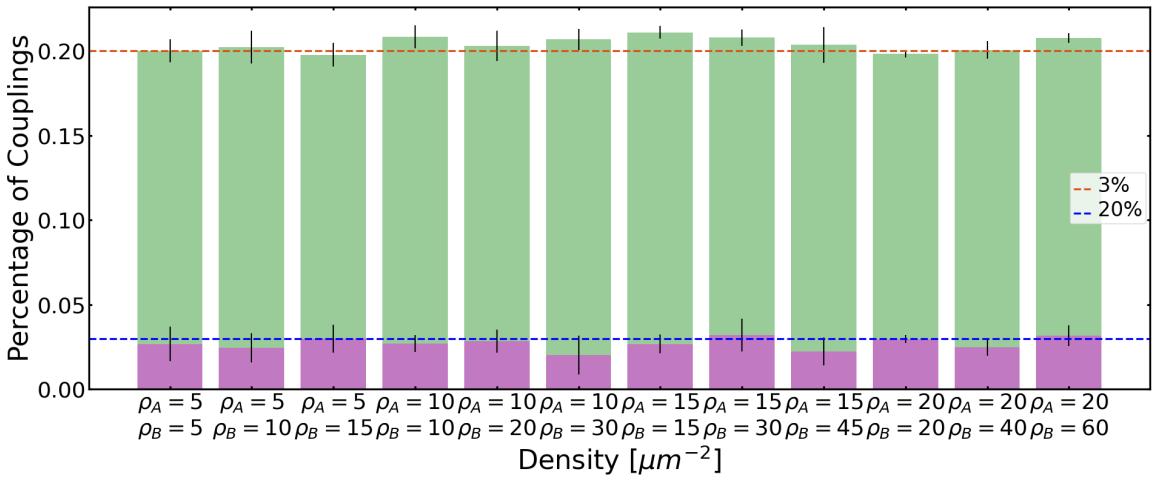

Figure S17. Estimated percentage of couplings q under different combinations of densities for A (ρ_A) and B (ρ_B), where the dataset generated were consistent with percentage of couplings $\gamma = 3\%$ and 20% respectively. Error bars represent one standard deviation across 10 trials of iMEC.

Supplementary Note 14: Inferring rate constants from kinetics data

We ask whether the algorithm can be used to infer kinetic parameters. We used the simulated non-equilibrium dataset in Fig. 6(d) of the main text, and the number of coupled molecules N_{coupled} from our pipeline (GMO+iMEC) to estimate the rate constants k_{on}, k_{off} of the binding reactions $A + B \rightleftharpoons AB$.

A. Analytical solution

In order to infer rate constants from our pipeline, we fitted the number of complexes AB to the solution of the differential equations modelling this system. Consider the reaction $A + B \rightleftharpoons AB$ with rate constants k_{on} and k_{off} for the forward and backward reactions respectively. Under mass-action kinetics, with no AB initially, the concentrations evolve according to the system of differential equations

$$\frac{d[A]}{dt} = \frac{d[B]}{dt} = -\frac{d[AB]}{dt} = -k_1[A][B] + k_2[AB],$$

with initial state $[A](0) = A_0$, $[B](0) = B_0$ and $[AB](0) = 0$. The solution is given by

$$\begin{aligned} [AB](t) &= \frac{A_0 B_0 \left(e^{-2\sqrt{c_2}k_1 t} - 1\right)}{(c_1 - \sqrt{c_2})e^{-2\sqrt{c_2}k_1 t} - c_1 + \sqrt{c_2}}, \\ [A](t) &= A_0 - [AB](t), \\ [B](t) &= B_0 - [AB](t), \end{aligned} \tag{S7}$$

where $c_1 = (A_0 + B_0 + k_2/k_1)/2$ and $c_2 = c_1^2 - A_0 B_0$.

B. Inferring rate constants

The rate constants k_{on} and k_{off} were obtained by fitting the density of AB identified (N_{coupled}/S) at time t, where S is the area of the simulation box, to $[AB](t)$ from Eq. (S7). Simulation and analysis parameters used can be found in Tables S1 and S5.

Compared to the true parameters $k_{\text{on}} = 0.0549$ molecules^{-1}µm^{2}s^{-1} (see Fig. S5 and Supplementary Note 6B) and $k_{\text{off}} = 1$ s^{-1}, the rate constants from fitting were found to be

$$\begin{aligned} \hat{k}_{\text{on}} &= \frac{0.0535 + 0.0492 + 0.0489}{3} = 0.0505 \text{ molecules}^{-1}\mu\text{m}^2\text{s}^{-1}, \\ \hat{k}_{\text{off}} &= \frac{0.947 + 0.808 + 0.831}{3} = 0.862 \text{ s}^{-1}, \end{aligned}$$

where we averaged over the three densities ("cases") in Fig. 6(d) of the main text. Fig. S18 shows the fitted rate constants for each case separately. Note that estimation of the rate constants was more accurate at low density (Case I); this is consistent with our observation that the performance of our algorithm degrades with increasing density (Fig. 4(b) of the main text).

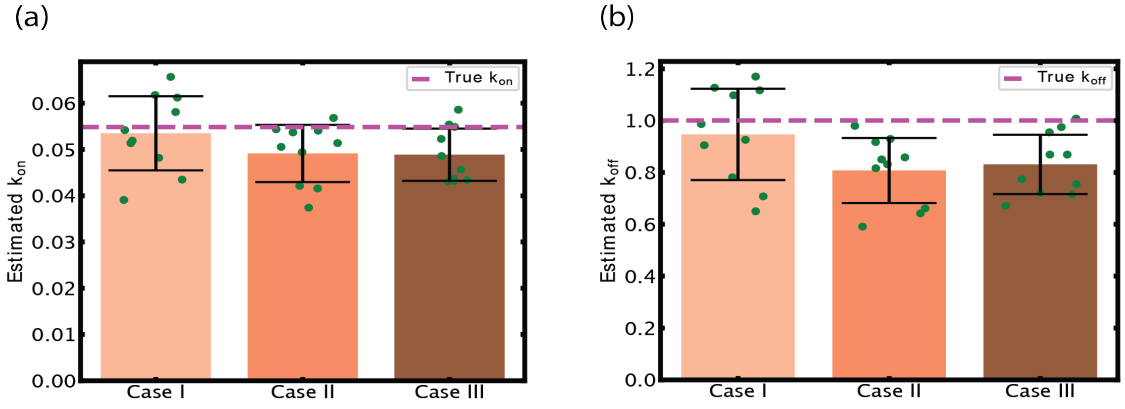

Figure S18. Inference of rate constants k_{on} and k_{off} from stochastic simulation, by fitting to Eq. (S7). (a) Estimated k_{on} and (b) k_{off} for Case I: $d_A = 10, d_B = 20$ μm^{-2}, Case II: $d_A = 20, d_B = 20$ μm^{-2} and Case III: $d_A = 30, d_B = 20$ μm^{-2}. Error bars indicate ± 1 standard deviation over 10 different trials.

Bibliography

1. D Sage, et al., Super-resolution fight club: assessment of 2d and 3d single-molecule localization microscopy software. *Nature Methods* **16**, 387–395 (2019).
2. LS Churchman, H Flyvbjerg, JA Spudich, A non-gaussian distribution quantifies distances measured with fluorescence localization techniques. *Biophysical Journal* **90**, 668–671 (2006).
3. J Park Jr, Moments of the generalized rayleigh distribution. *Quarterly of Applied Mathematics* **19**, 45–49 (1961).
4. A Hagberg, P Swart, D S Chult, Exploring network structure, dynamics, and function using networkx, (Los Alamos National Lab.(LANL), Los Alamos, NM (United States)), Technical report (2008).
5. Z Galil, Efficient algorithms for finding maximum matching in graphs. *ACM Computing Surveys (CSUR)* **18**, 23–38 (1986).
6. LS Churchman, Z Ökten, RS Rock, JF Dawson, JA Spudich, Single molecule high-resolution colocalization of Cy3 and Cy5 attached to macromolecules measures intramolecular distances through time. *Proceedings of the National Academy of Sciences* **102**, 1419–1423 (2005).
7. S Niekamp, et al., Nanometer-accuracy distance measurements between fluorophores at the single-molecule level. *Proceedings of the National Academy of Sciences* **116**, 4275–4284 (2019).
8. NN Batada, LA Shepp, DO Siegmund, Stochastic model of protein–protein interaction: Why signaling proteins need to be colocalized. *Proceedings of the National Academy of Sciences* **101**, 6445–6449 (2004).
9. NN Batada, LA Shepp, DO Siegmund, M Levitt, Spatial regulation and the rate of signal transduction activation. *PLoS Computational Biology* **2**, e44 (2006).
10. M Hoffmann, C Fröhner, F Noé, Readdy 2: Fast and flexible software framework for interacting-particle reaction dynamics. *PLoS computational biology* **15**, e1006830 (2019).
11. SV Costes, et al., Automatic and quantitative measurement of protein-protein colocalization in live cells. *Biophysical Journal* **86**, 3993–4003 (2004).
12. S Malkusch, et al., Coordinate-based colocalization analysis of single-molecule localization microscopy data. *Histochemistry and Cell Biology* **137**, 1–10 (2012).
13. T Lagache, et al., Mapping molecular assemblies with fluorescence microscopy and object-based spatial statistics. *Nature Communications* **9**, 698 (2018).
14. F Levet, et al., A tessellation-based colocalization analysis approach for single-molecule localization microscopy. *Nature Communications* **10**, 2379 (2019).
15. AL Ejdrup, et al., A density-based enrichment measure for assessing colocalization in single-molecule localization microscopy data. *Nature Communications* **13**, 4388 (2022).
16. W Trabesinger, et al., Statistical analysis of single-molecule colocalization assays. *Analytical Chemistry* **73**, 1100–1105 (2001).
17. A Inoue, et al., Illuminating G-protein-coupling selectivity of GPCRs. *Cell* **177**, 1933–1947 (2019).
18. JS Smith, et al., Noncanonical scaffolding of Gαi and β-arrestin by G protein–coupled receptors. *Science* **371**, eaay1833 (2021).
19. Y Ono, et al., Generation of Gαi knock-out HEK293 cells illuminates Gαi-coupling diversity of GPCRs. *Communications Biology* **6**, 112 (2023).
20. R Birke, et al., Sulfonated red and far-red rhodamines to visualize SNAP-and HALO-tagged cell surface proteins. *Organic & Biomolecular Chemistry* **20**, 5967–5980 (2022).

21. H Schihada, R Shekhani, G Schulte, Quantitative assessment of constitutive g protein–coupled receptor activity with bret-based g protein biosensors. *Science Signaling* **14**, eabf1653 (2021).
22. M Ester, HP Kriegel, J Sander, X Xu, , et al., A density-based algorithm for discovering clusters in large spatial databases with noise in *kdd*. Vol. 96, pp. 226–231 (1996).
23. S Bolte, FP Cordelières, A guided tour into subcellular colocalization analysis in light microscopy. *Journal of Microscopy* **224**, 213–232 (2006).
24. C Tameling, et al., Colocalization for super-resolution microscopy via optimal transport. *Nature Computational Science* **1**, 199–211 (2021).