analytical chemistry

pubs.acs.org/ac Article

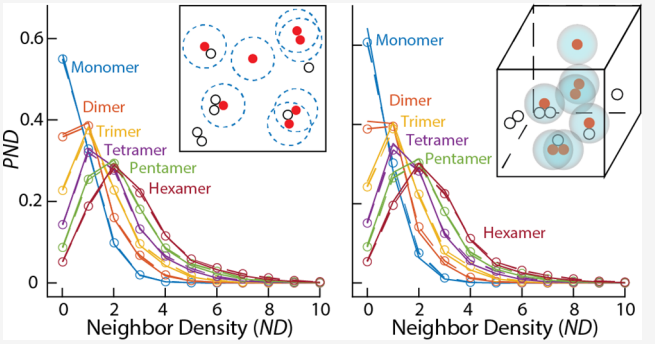
From Monomers to Hexamers: A Theoretical Probability of the Neighbor Density Approach to Dissect Protein Oligomerization in Cells

Huanhuan Chen and Tai-Yen Chen*

Cite This: https://doi.org/10.1021/acs.analchem.3c04728 Read Online

ACCESS | Metrics & More | Article Recommendations | Supporting Information

ABSTRACT: Deciphering the oligomeric state of proteins within cells is pivotal to understanding their role in intricate cellular processes. With the recent advances in single-molecule localization microscopy, previous efforts have harnessed protein location density approaches, coupled with simulations, to extract membrane protein oligomeric states in cells, highlighting the value of such techniques. However, a comprehensive theoretical approach that can be universally applied across different proteins (e.g., membrane and cytosolic proteins) remains elusive. Here, we introduce the theoretical probability of neighbor density (*PND*) as a robust tool to discern protein oligomeric states in cellular environments. Utilizing our approach, the theoretical *PND* was validated against simulated data for both membrane and cytosolic proteins, consistently aligning with experimental baselines for membrane proteins. This congruence was maintained even when adjusting for protein concentrations or exploring proteins of various oligomeric states. The strength of our method lies not only in its precision but also in its adaptability, accommodating diverse cellular protein scenarios without compromising the accuracy. The development and validation of the theoretical *PND* facilitate accurate protein oligomeric state determination and bolster our understanding of protein-mediated cellular functions.



INTRODUCTION

The complex interplay of proteins coming together to form larger structures—known as oligomerization—has far-reaching effects on how cells function and maintain their stability.[1−5] Research has consistently shown that proteins often act in their combined or "oligomeric" forms for key cellular processes, such as regulating genes and controlling enzyme activity.[6−12] The behavior of these protein assemblies is not solely determined by their individual properties but is also shaped by the specific conditions within the cell.[13−16] While in vitro studies provide immensely valuable information, they often fail to fully capture the true essence of protein behaviors in a cellular context. For example, they might miss the effects of changes in cellular components or neglect variables like protein concentration and post-translational modifications. On the contrary, in-cell quantification offers a more complete picture. It accounts for the dynamic cellular environment, including the effects of crowding, varying pH, and other cellular components that directly influence protein states. Furthermore, in-cell studies pave the way for real-time observations, revealing short-lived interactions or transitional states that are difficult to observe in in vitro studies.

Deciphering protein oligomeric states within cells has necessitated a methodological evolution, transitioning meticulously from ensemble techniques to single-molecule methodologies and finally to the precision of single-molecule localization microscopy (SMLM). Ensemble approaches, such as ensemble FRET, BiFC, FCS, and PLA, served as foundational pillars, providing insights into the predominant oligomeric states within cellular populations.[17−20] However, while these methods elucidated average oligomeric states, they could often overlook molecular heterogeneity or rare oligomeric forms. To unmask this concealed diversity, single-molecule techniques like smFRET, single particle tracking, and single-molecule anisotropy were adopted.[14,21−23] These methods illuminated the full spectrum of oligomeric states by capturing individual protein molecules, revealing not just the dominant forms but also transient or less abundant oligomeric states.

By combining single-molecule insights with super-resolution techniques, SMLM enables the identification of individual

Received: October 19, 2023
Revised: December 11, 2023
Accepted: December 11, 2023

oligomeric states in densely populated cellular areas, providing a detailed perspective on protein oligomerization. Among the available methodologies, fluorescence photophysical property-based strategies exploit the unique photophysical behaviors of fluorophores, such as their blinking and bleaching kinetics, to deduce oligomeric forms.[24−29] Protein localization-based methods, on the other hand, capitalize on the difference in local molecule density, calculated from the protein positions, for assemblies with different oligomeric states.[30−32] By extensively simulating the random positions of protein assemblies with different oligomeric states and comparing them with the experimental data, this approach allows for a quantitative assessment of the protein assembly's oligomeric states. Nevertheless, to fully harness the potential of this method, a solid theoretical foundation is essential. An all-encompassing theoretical model would streamline the extraction process with analytical solutions and enhance the accuracy and applicability.

In this study, a theoretical model termed the probability of neighbor density (*PND*) was developed. The model incorporates crucial parameters, including protein concentrations, dye photoactivation efficiency, and the radius of the region of interest, to ensure alignment with actual experimental conditions. Stochastic detections in single-molecule imaging were modeled by using binomial and Poisson distributions. Systematic validation of the theoretical *PND* against simulated data confirmed its efficacy in analyzing both membrane and cytosolic proteins, irrespective of their oligomeric states. The limitations of the theoretical *PND* are also discussed. This research highlights the theoretical *PND* model's potential to reveal both membrane and cytosolic proteins' oligomeric states in intricate cellular contexts, providing a quantitative platform to enhance our comprehension of oligomer-mediated processes in cells.

EXPERIMENTAL SECTION

SMLM, primarily known for spatial mapping, has vast potential when aligned with theoretical models to deduce protein oligomeric states. Here, we present a key parameter, the *PND*, to streamline the process of quantifying protein oligomeric states in a cellular environment. The *PND* represents the likelihood of encountering molecular neighbors within a circular region of interest (ROI) of radius *r*. Specifically, with a given number of detected locations, *N*, within a field of view (FOV, the square with length *L*), we define the neighbor density (*ND*) of an activated subunit as the number of neighbors located within the ROI (Figure 1a). By collecting all *ND* values within the FOV, a probability distribution of observing different *ND* values is derived. For instance, Figure 1a left shows four instances where *ND* = 0 and two instances where *ND* = 1, translating into 66.7 and 33.3% *PND* probabilities, respectively, in Figure 1a, right.

Theoretical *PND* for Protein Assemblies Existing in One Oligomeric State. To derive the theoretical *PND* from experimentally detectable locations (*N*), we start with protein assemblies with a single oligomeric state. Let us postulate that there are A_{O_g} protein assemblies of a defined oligomeric state (O_g) within the FOV. In SMLM, each equivalent is labeled by irreversible photoactivatable fluorophores with a specific photoactivation efficiency (*PE*). Given the stochastic nature of activation, the relationship between the number of detected

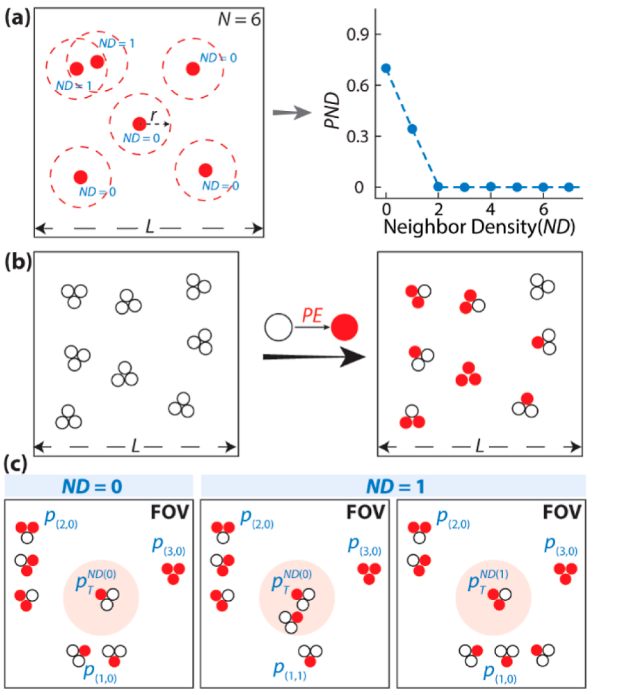


Figure 1. *PND* generation overview. (a) The *ND* at each location is determined by tallying neighbors within an ROI of radius *r*. Aggregating these *ND* values across all locations yields an *ND* histogram. This histogram is then normalized to produce the *PND*. (b) Various activation forms (red) of a trimeric assembly, including singly, doubly, and triply activated forms. (c) Schematic representation of *ND* = 0 and *ND* = 1 of a pure trimeric assembly. *ND* = 0 involves targeting a singly activated protein assembly without any other activated protein assemblies within the ROI (left). *ND* = 1 can arise from two scenarios (right): first, by targeting a singly activated assembly with one additional singly activated assembly in the ROI and second, by targeting a doubly activated assembly without other activated assemblies in the ROI.

locations and the number of protein assemblies can be expressed as $N = A_{O_g} \times O_g \times PE$ or

$$A_{O_g} = \frac{N}{O_g \times PE} \tag{1}$$

For example, if we have eight trimeric protein assemblies (A_{O_g} = 8 and O_g = 3) with the FOV, with a *PE* of 0.5, *N* = 12 equivalents will be detected (Figure 1b), distributing to singly, doubly, and triply activated forms.

Assuming a sufficiently low assembly concentration (i.e., only one single assembly within each ROI) and recognizing that activation is a sequence of independent events, the likelihood of exactly N_a equivalents becoming activated for an assembly with the O_g oligomeric state, $p_{O_g}^{N_a}$, can be described by the binomial distribution

$$p_{O_g}^{N_a} = \binom{O_g}{N_a} \times PE^{N_a} \times (1 - PE)^{O_g - N_a} \tag{2}$$

Using the trimeric model as an example, we can calculate the probability, $p_{O_g}^{N_a}$, for each scenario of the N_a-activated trimers. For a singly activated trimer, p_T^1, where one out of three oligomerization sites is activated, the probability is calculated

https://doi.org/10.1021/acs.analchem.3c04728
Anal. Chem. XXXX, XXX, XXX−XXX

as follows: $p_T^1 = \binom{3}{1} \times 0.5^1 \times (1 - 0.5)^{3-1} = \frac{3}{8}$. In the case of a doubly activated trimer, p_T^2, or fully activated with all three sites, p_T^3, the probability is $p_T^2 = \binom{3}{2} \times 0.5^2 \times (1 - 0.5)^{3-2} = \frac{3}{8}$ and $p_T^3 = \binom{3}{3} \times 0.5^3 = \frac{1}{8}$, respectively.

Here, $\binom{O_g}{N_a}$ is a binomial coefficient, signifying the number of ways to activate N_a subunits from each assembly, disregarding the activation sequence. The terms PE^{N_a} and $(1 - PE)^{O_g - N_a}$ encapsulate the probabilities of activations and nonactivations, respectively. Since $p_{O_g}^{N_a}$ denotes the probability of detecting N_a activated subunits within a single protein assembly, it enables the estimation of the number of assemblies in a specific activated form through

$$A_{O_g}^{N_a} = p_{O_g}^{N_a} \times A_{O_g} \tag{3}$$

Each activated form contributes $N_a \times A_{O_g}^{N_a}$ locations, and they collectively yield the total location number N. The single, double, and triple activations (i.e., N_a = 1, 2, and 3) lead to ND = 0, 1, and 2 conditions, respectively, each with a given ND possibility

$$p_{O_g}^{ND(N_a-1)} = \frac{N_a \times A_{O_g}^{N_a}}{N} \tag{4}$$

Combining the probability of each ND yields the final PND distribution via

$$PND = [p_{O_g}^{ND(0)}, p_{O_g}^{ND(1)}, \cdots, p_{O_g}^{ND(N_a-1)}] \tag{5}$$

Using the above trimeric assemblies ($O_g = T$) as an example, the theoretical PND is calculated as $PND = \left[p_T^{ND(0)} = \frac{1 \times p_T^1 \times 8}{12}, p_T^{ND(1)} = \frac{2 \times p_T^2 \times 8}{12}, p_T^{ND(2)} = \frac{3 \times p_T^3 \times 8}{12}\right]$ or [0.25, 0.50, 0.25].

In real systems, the assumption of a low assembly concentration (or one activated assembly per ROI) might not always hold. As the protein concentration increases, it is possible to have multiple assemblies within a single ROI. To develop a comprehensive theoretical PND, we considered scenarios with multiple protein assemblies in an ROI, distributed randomly across the FOV and each assembly enters the ROI independently.

The ND value is concurrently modulated by two factors: the number of activations and the assemblies present within the ROI. An ND of 0 implies just one singly activated assembly in the ROI (Figure 1c). An ND of 1 could result from two singly activated assemblies or one doubly activated assembly. As the ND values increase, so do the combinations leading to the same ND. Therefore, deriving the theoretical PND necessitates summing the probabilities across these various combinations for each ND. The associated probability, p_{sum}^{ND}, can be quantified as the product of the ND probability of the aimed activated assembly ($p_{O_g}^{ND(N_a-1)}$) and the probability of observing other activated assemblies (p^{roi}) within the ROI

$$p_{sum}^{ND} = p_{O_g}^{ND(N_a-1)} \times p^{roi} \tag{6}$$

The independent nature of each assembly's presence in the ROI, combined with the constant average density across the FOV, strongly aligns with the key prerequisites of the Poisson distribution. As such, the probability of observing k assemblies with N_a equivalents activated in the ROI can be modeled by the Poisson probability, $p_{(N_a,k)}$

$$p_{(N_a,k)} = \frac{\lambda^k \times e^{-\lambda}}{k!} \tag{7}$$

Here, λ is the expected number of assemblies with N_a equivalents activated in the ROI. If neighboring assemblies exhibit a different N_a from the aimed one, given the random distribution of activated assemblies $A_{O_g}^{N_a}$, λ can be deduced from the area ratio between the ROI and FOV and is defined as $\lambda = \frac{area_{ROI}}{area_{FOV}} \times A_{O_g}^{N_a}$. In cases where neighboring assemblies have the same N_a as the aimed one, the number of activated neighboring assemblies becomes $A_{O_g}^{N_a} - 1$. Consequently, λ is recalculated as $\lambda = \frac{area_{ROI}}{area_{FOV}} \times (A_{O_g}^{N_a} - 1)$.

Using trimeric assemblies as an illustrative example (Figure 1c, left), when there are no neighbors (ND = 0), the sole scenario is a targeted singly activated trimer without other activated trimers with the ROI. For this scenario, the k = 0 for N_a varying from 1 to 3, the probability of not observing singly, doubly, and triply activated trimers can be represented as $p^{roi} = p_{(1,0)} \times p_{(2,0)} \times p_{(3,0)}$. The probability of no neighboring activated subunits can be expressed as

$$p_{sum}^{ND(0)} = p_T^{ND(0)} \times p_{(1,0)} \times p_{(2,0)} \times p_{(3,0)} \tag{8}$$

For the ND = 1 case, $p_{sum}^{ND(1)}$ encompasses the probabilities of two scenarios (Figure 1c, right): two singly activated assemblies or a single doubly activated assembly. In the former scenario, the probability is derived from $p_T^{ND(0)}$ and the likelihood of observing one singly activated but neither doubly nor triply activated assemblies $[p^{roi} = p_{(1,1)} \times p_{(2,0)} \times p_{(3,0)}]$. For the latter scenario involving the targeted doubly activated assembly, the probability arises from $p_T^{ND(1)}$ and the probability of not observing any other activated assemblies $[p^{roi} = p_{(1,0)} \times p_{(2,0)} \times p_{(3,0)}]$. These two scenarios collectively result in

$$p_{sum}^{ND(1)} = p_T^{ND(0)} \times p_{(1,1)} \times p_{(2,0)} \times p_{(3,0)} + p_T^{ND(1)} \times p_{(1,0)} \times p_{(2,0)} \times p_{(3,0)} \tag{9}$$

Specifically, $p_{(1,0)}$ is a Poisson probability representing the likelihood of observing zero assemblies with exactly one activated subunit within a given ROI. This contrasts with $p_{(1,1)}$, which denotes the probability of observing exactly one such assembly in the ROI. Meanwhile, $p_T^{ND(0)}$ and $p_T^{ND(1)}$ quantify the probability of a singly activated trimeric assembly and a doubly activated trimeric assembly (indicated by the subscript "T") in the ROI, respectively. These distinct probabilities collectively facilitate comprehensive modeling of the varying activation states and distributions of protein assemblies in cellular environments, enhancing our understanding of their spatial interactions.

By understanding the unique combinations that result in different ND values, it becomes straightforward to determine the theoretical p_{sum}^{ND} for any specified ND. This approach simplifies the derivation of the final theoretically normalized PND, which can be expressed as an array of probabilities

https://doi.org/10.1021/acs.analchem.3c04728
Anal. Chem. XXXX, XXX, XXX−XXX

$$PND = [p_{sum}^{ND(0)}, p_{sum}^{ND(1)}, p_{sum}^{ND(2)}, \cdots]/\sum p_{sum}^{ND} \quad (10)$$

Detailed mathematical information on ND = 0, 1, and 2 for the trimeric protein assembly and corresponding illuminations are summarized in Tables S1−S3 and Figure S1, respectively.

Theoretical *PND* for Protein Assemblies Existing in a Multiple Oligomeric State. Addressing the inherent complexity of protein assemblies, we refine the *PND* model to characterize a protein assembly existing in equilibrium among its monomeric (M), dimeric (D), and trimeric (T) forms. Each form contains subunit populations of S_M, S_D, and S_T, respectively, and satisfy $S_M + S_D + S_T = 1$ (Figure 2a). With a given *PE*, S_M, S_D, and S_T collectively contribute to the total N detected locations. The number of assemblies with specific oligomeric states ($A_{O_g,sub}$) can be related to N via

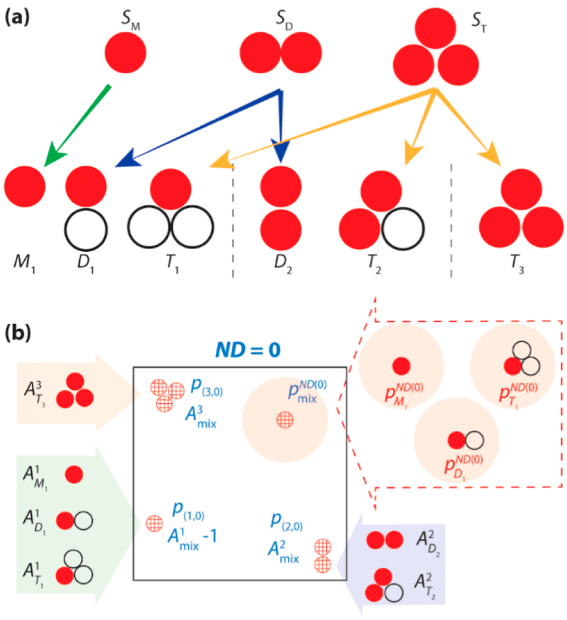


Figure 2. Theoretical *PND* for proteins with varied oligomeric states: (a) activation and reorganization of protein mixtures comprising monomeric (M), dimeric (D), and trimeric (T) states. (b) Schematic example of ND = 0, representing a single activated protein assembly within the ROI. M_1, D_1, and T_1 in the ROI collectively contribute to $p_{mix}^{ND(0)}$. The absence of singly activated (M_1, D_1, and T_1), doublyactivated (D_2 and T_2), and triply activated (T_3) forms in the ROI contribute to p_{mix}^{roi}.

$$A_{O_g,sub} = \frac{N \cdot S_{O_g}}{O_g \cdot PE} \quad (11)$$

The incomplete activation of the fluorescence protein introduces six potential detectable protein assemblies: M_1 for monomeric, D_1 and D_2 for dimeric, and T_1, T_2, and T_3 for the trimeric assemblies, with the subscript indicating specific activation conditions N_a (Figure 2a). To streamline the theoretical *PND* derivation, we reorganize the six assemblies based on their activation levels to form a "pseudo" trimer condition. Specifically, M_1, D_1, and T_1 assemblies are singly activated; D_2 and T_2 are doubly activated assemblies, and T_3 uniquely represents the triply activated assembly.

In scenarios where only a single assembly is present within the ROI, the probability of the assembly having N_a activated equivalents remains the same as that in eq 2. The only modification is to substitute A_{O_g} with $A_{O_g,sub}$ in eq 3 and eq 4 for each unique assembly. This change enables the estimation of the number of assemblies in a specific activated form and the *ND* probability using eqs 12 and 13, respectively.

$$A_{O_g,sub}^{N_a} = p_{O_g}^{N_a} \times A_{O_g,sub} \quad (12)$$

$$p_{O_g,sub}^{ND(N_a-1)} = \frac{N_a \times A_{O_g,sub}^{N_a}}{N} \quad (13)$$

A distinct feature to consider is that various species might contribute to the identical *ND* state. For instance, species M_1, D_1, and T_1 all contribute to the *ND* state of 0. Species D_2 and T_2 contribute to the *ND* state of 1, and only species T_3 contributes to the *ND* of 2. Considering the *ND* of the 0 case, the probability emerges from aggregating contributions from species M_1, D_1, and T_1. Mathematically, this relationship is expressed as

$$p_{mix}^{ND(0)} = p_{M_1}^{ND(0)} + p_{D_1}^{ND(0)} + p_{T_1}^{ND(0)} \quad (14)$$

For the *ND* of 1, the *ND* possibility arises from the combined effect of species D_2 and T_2, represented as

$$p_{mix}^{ND(1)} = p_{D_2}^{ND(1)} + p_{T_2}^{ND(1)} \quad (15)$$

In the case of *ND* of 2, the probability is solely determined by species T_3

$$p_{mix}^{ND(2)} = p_{T_3}^{ND(2)} \quad (16)$$

Combining these probabilities from eqs 14−16 provides the final *PND* distribution as

$$PND = [p_{mix}^{ND(0)}, p_{mix}^{ND(1)}, \cdots, p_{mix}^{ND(N_a-1)}] \quad (17)$$

To extend the derivation to the most general condition, where assemblies exist in various oligomeric states and in high concentrations, two probabilities are pivotal: the *ND* probability of the aimed assembly, denoted as $p_{mix}^{ND(N_a-1)}$, and the probability of observing other activated assemblies, denoted as p_{mix}^{roi}. Using an *ND* of 0 as an example (Figure 2b), the *ND* probability $p_{mix}^{ND(0)}$ will be the sum of M_1, D_1, and T_1, as specified in eq 14. For estimating the likelihood of the absence of other neighboring activated assemblies in the ROI, p_{mix}^{roi}, it is imperative to individually appraise the Poisson distributions, eq 7, for not having singly, doubly, and triply activated conditions. The probability of not observing singly activated assemblies (i.e., no M_1, D_1, and T_1) can be calculated by substituting $A_{O_g}^1$ with $A_{mix}^1 = A_{M_1}^1 + A_{D_1}^1 + A_{T_1}^1 - 1$ in the λ parameter. Analogously, the λ values for not observing doubly (D_2 and T_2) and triply (T_3) activated assemblies can be calculated by using $A_{mix}^2 = A_{D_2}^2 + A_{T_2}^2$ and $A_{mix}^3 = A_{T_3}^3$, respectively. Therefore, the probability of observing zero neighboring activated assemblies, especially when assemblies exhibit diverse oligomeric states and are present in high concentrations, is given by $p_{mix,sum}^{ND(0)}$, which is formulated as

$$p_{mix,sum}^{ND(0)} = p_{mix}^{ND(0)} \times p_{mix}^{roi} = p_{mix}^{ND(0)} \times p_{(1,0)} \times p_{(2,0)} \times p_{(3,0)} \quad (18)$$

By iterating this analytical process, the generalized theoretical *PND* after normalization emerges as

https://doi.org/10.1021/acs.analchem.3c04728
Anal. Chem. XXXX, XXX, XXX−XXX

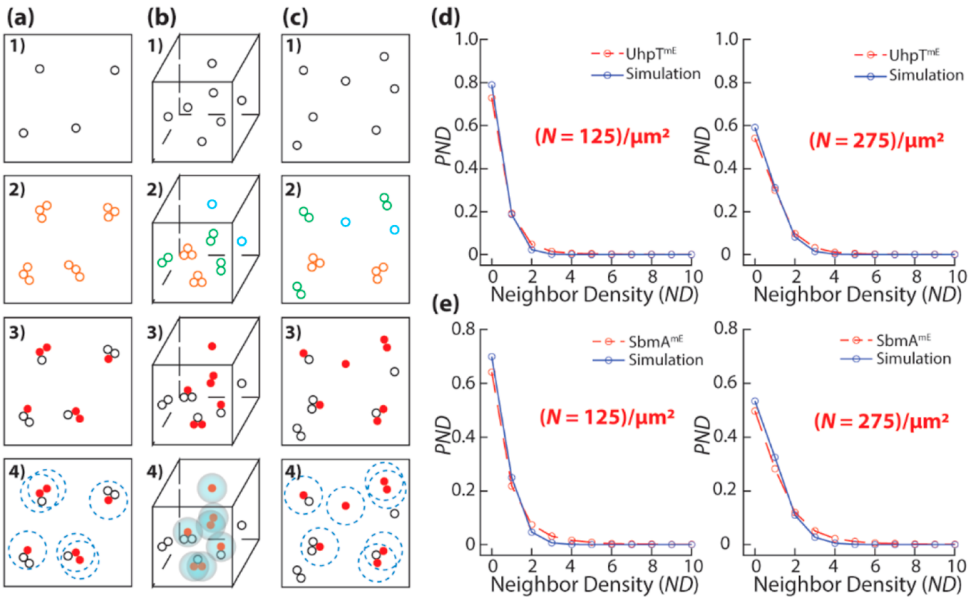


Figure 3. *PND* simulation schemes. For (a) membrane and (b) cytosolic proteins with a single oligomer state or (c) membrane protein with a multiple oligomeric state, the simulation involves four steps: (1) randomly assigned assembly locations, (2) spatial map with a simulated subunit for the monomer (blue), dimer (green), and trimer (orange), (3) randomly selected N locations (red) to account for incomplete fluorophore activation, and (4) employment of a circular ROI for membrane proteins and a spherical ROI for cytosolic proteins to determine the *ND* for the chosen locations, resulting in the final *PND* distributions. (d,e) Experimental (red) and corresponding simulated ground truth (blue) *PND* distributions with different protein concentrations ($N/\mu m^2$) for $UhpT^{mE}$ (d) and $SbmA^{mE}$ (e). All simulations in this study have a typical error of 10^{-4}, which is smaller than the symbol size.

$$PND = [p_{\mathrm{mix,sum}}^{ND(0)}, p_{\mathrm{mix,sum}}^{ND(1)}, p_{\mathrm{mix,sum}}^{ND(2)}, \cdots]/\sum p_{\mathrm{mix,sum}}^{ND} \quad (19)$$

***PND* Simulation for Protein Assemblies.** Through integration with point spread function (PSF) engineering, SMLM localizes proteins within a three-dimensional space.[33−35] In recent advances in single-molecule imaging systems, the double helix has proficient mapping spatial distributions within a z range of 2−3 μm, achieving a lateral and axial resolution of less than ∼50 nm.[36−38] This precision offers invaluable experimental insights, particularly when probing the oligomeric states of both membrane and cytosolic proteins.

Leveraging the detection principles of SMLM, a simulation methodology was formulated to generate *PND* simulations for both membrane and cytosolic protein assemblies with specific oligomeric states. The strategy for simulating the spatial distribution of membrane protein assemblies in confined spaces aligns with established methodologies (Figure 3a).[31] In short, based on the experimentally determined locations, the number of assemblies was first evaluated via eq 1. The spatial distribution of membrane assemblies in a confined square was simulated by randomly positioning monomeric equivalents within the FOV. For multimeric assemblies, a two-dimensional Brownian diffusion model was applied to the original monomeric subunit to predict the positions of the other equivalents. All simulated locations were randomly sampled based on the photoconversion efficiency of the chosen fluorophore. The resultant sampled locations formed the basis for *PND* distribution calculations, with iterations continuing until saturation. Analogous methods were employed for cytosolic proteins (Figure 3b), with the distinction of simulating protein assemblies within a cubic FOV and employing a 3D diffusion model for multimeric assemblies.

To model protein assemblies with multiple oligomeric states, the subpopulation of each oligomer (e.g., S_M for the monomer, S_D for the dimer, and S_T for the trimer) was incorporated (Figure 3c). The number of assemblies for each specific oligomeric state was determined using eq 11. The spatial distribution of assemblies of various oligomeric states was simulated according to previously described methods (Figure 3a,b) and amalgamated to derive the *PND* distributions. This methodology was iteratively applied until a saturation point was achieved. We compared the outcomes from three independent trials. The differences were typically below 10^{-4}, indicating robust simulation results (Figure S2). Simulations were extended to include assemblies as large as hexamers, encompassing more than 92% of the protein population in most cells.[1]

To validate our simulation model, we compared the simulated results to experimental data for two specific membrane proteins: UhpT, which exists predominantly as a monomer, and SbmA, known to be in a monomer−dimer equilibrium with approximately 70% monomer and 30% dimer. The comparison between our simulated data and the experimental distributions revealed a high degree of correlation (Figure 3d,e), suggesting that our simulation routine successfully captured the experimental *PND* distributions for both UhpT-mEos3.2 ($UhpT^{mE}$) and SbmA-mEos3.2 ($SbmA^{mE}$). Moreover, our model is scalable and easily extends to higher oligomeric states. Its flexibility in generating ground truth data for diverse protein assemblies offers a solid foundation for testing and refining theories across a spectrum of oligomeric complexities.

RESULTS AND DISCUSSION

Theoretical *PND* Faithfully Describes Oligomeric States of Membrane and Cytosolic Protein Assemblies.

https://doi.org/10.1021/acs.analchem.3c04728
Anal. Chem. XXXX, XXX, XXX−XXX

The theoretical *PND* distributions were rigorously validated using simulated ground truth data for both membrane and cytosolic proteins across a spectrum of oligomeric states ranging from the monomer to hexamer. Simulations and theoretical derivations were based on specific parameters. Specifically, 120 detectable locations (N = 120), with a photoactivation efficiency of 0.42 (PE = 0.42), were placed into FOVs. These FOVs were defined as either squares for membrane proteins or cubes for cytosolic proteins, both with dimensions of L = 0.5 μm. The *ND*s were collected within the defined ROI: a circular region with a radius of r = 0.02 μm for membrane proteins and a spherical region with a radius of r = 0.05 μm for cytosolic proteins. The selection of r was informed by the spatial resolution limits of the SMLM and the dimensions of the protein assemblies. We ensured that r is larger than typical localization precision at a scale appropriate to encompass entire protein assemblies while avoiding the inclusion of neighboring proteins not part of the assembly. This balanced approach ensured that the radius was optimally set to capture relevant biological interactions and maintain the integrity of our model's predictions.

Figure 4a depicts the *PND* distributions for both membrane and cytosolic protein assemblies existing in a single oligomeric state, spanning the spectrum from monomers to hexamers. The distinct *PND* variations observed among different oligomers emphasize the capability of the *PND* to effectively distinguish between these states. Furthermore, the theoretical *PND* aligns closely with the simulated *PND* for each of the oligomeric state evaluated. This congruence validates the accuracy of the theoretical *PND* model in characterizing the spatial distribution of both membrane (Figure 4a, left) and cytosolic proteins (Figure 4a, right), irrespective of their specific oligomeric states.

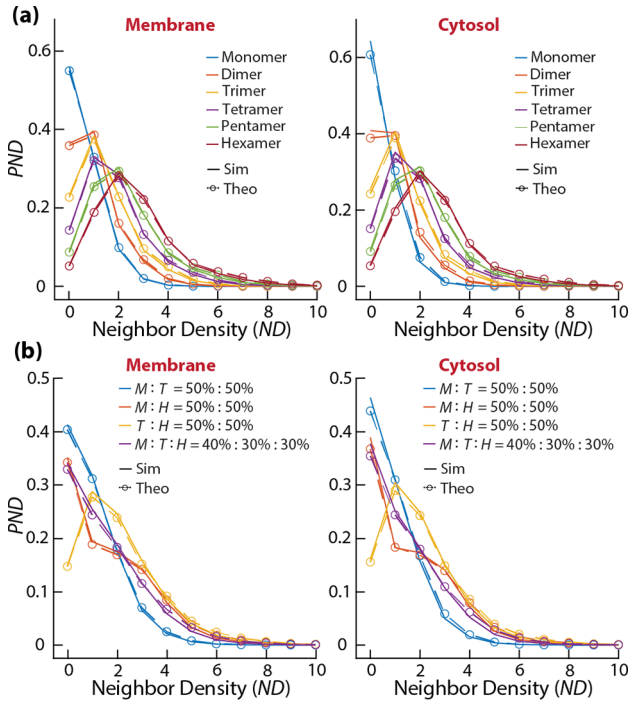


Figure 4. Validation of the theoretical *PND* using simulation. (a) Comparison of simulated (solid line) and theoretical (dashed circle line) *PND* distributions for membrane (left) and cytosolic (right) proteins with single oligomeric states. (b) Analogous comparison for proteins with multiple oligomeric states.

In the evaluation of protein assemblies existing in an equilibrium among varied oligomeric states, we explored four specific combinations, encompassing monomers (M), trimers (T), and hexamers (H). These combinations included equal proportions of monomers and trimers (M:T = 50%:50%), monomers and hexamers (M:H = 50%:50%), trimers and hexamers (T:H = 50%:50%), as well as a configuration with 40% monomers and 30% each of trimers and hexamers (M:T:H = 40%:30%:30%). The congruence observed between the simulated and theoretical *PND* distributions attests to the robustness of the *PND* methodology. This underscores its precision in delineating the oligomeric states of both membrane and cytosolic proteins, as depicted in Figure 4b.

Consistent Performance of the Theoretical *PND* Model across Diverse Experimental Conditions. Having established that the theoretical *PND* faithfully describes oligomeric states of membrane and cytosolic protein assemblies under specific conditions, namely, the size of ROI (r), the photoactivation efficiency (PE), and the concentration of protein assembly concentration (N), we sought to further validate its robustness under varying parameters. This was essential to ascertaining the universal applicability of the theoretical *PND* model across diverse experimental setups.

Using monomeric, trimeric, and hexameric membrane proteins, the effects of the ROI size were first explored (Figure 5a). It is worth noting that the choice of the ROI size often correlates with the location error inherent to imaging approaches. Our analysis revealed that the likelihood of encountering neighboring assemblies grew as the ROI size expanded, leading to a shift toward larger *ND* values. Additionally, a larger ROI attenuated the differences in the *PND* distribution across various oligomeric states.

Since PE is fluorophore specific (e.g., PE of mEos3.2 $PE_{\mathrm{mEos3.2}}$ = 42%, while PE_{mEos2} = 60%),[39] we further investigated the impact of PE on *PND* distributions (Figure 5b). A higher PE naturally heightened the probability of activating multiple subunits, leading to an anticipated surge in the populations at higher *ND* values. With an increase in PE, the *PND* values of different oligomers became more distinguishable.

Within cells, protein concentrations inherently fluctuate, and these changes are intricately linked to the oligomeric state of the protein assemblies. Reflecting this biological reality, we examined a range of concentrations (Figure 5c). The likelihood of detecting neighboring assemblies increases at higher concentrations, resulting in an elevated *ND* value. Additionally, the differences in the *PND* distributions among various oligomers also decrease under these conditions.

To quantitatively assess the influence of r, PE, and N on the model's applicability, we analyzed the discrepancies between the theoretical and simulated *PND* distributions (Figure 5d). An error matric, err, was introduced and defined by

$$\mathrm{err} = \sum |PND_{\mathrm{sim},O_g} - PND_{\mathrm{Theo},O_g}| \tag{20}$$

where $PND_{\mathrm{sim},Og}$ and $PND_{\mathrm{Theo},Og}$ are simulated and theoretical *PND* distributions for proteins with an oligomeric state of O_g, respectively. Across the range of tested conditions (Figure 5a–c), the error remained predominantly below 5%. A slight increase in the error was observed with larger r values, likely due to ROI inconsistencies near the FOV boundaries. When a

https://doi.org/10.1021/acs.analchem.3c04728
Anal. Chem. XXXX, XXX, XXX−XXX

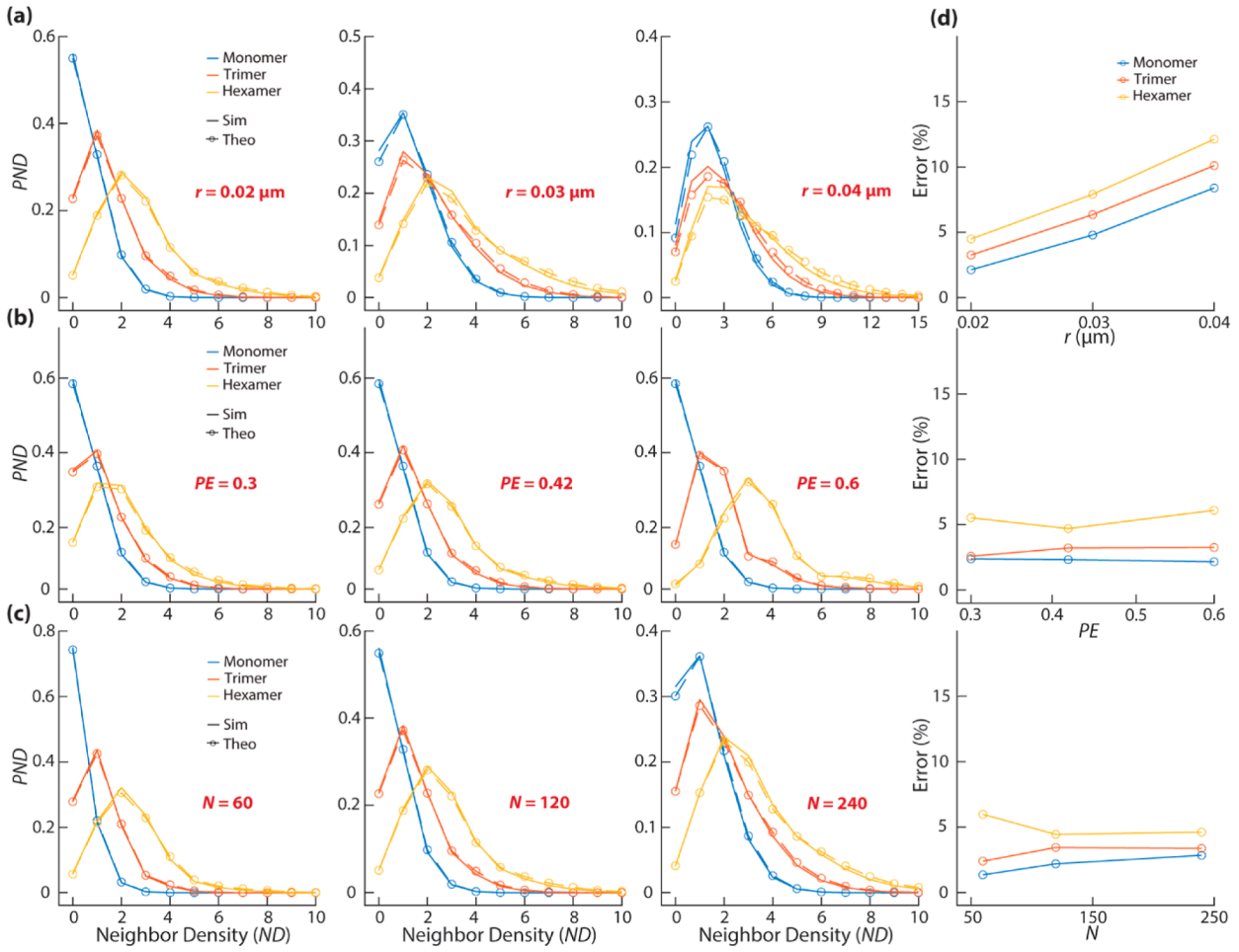


Figure 5. Effects of various experimental parameters on *PND* distributions of membrane proteins. Influence of (a) ROI size (r), (b) photoactivation efficiency (PE), and (c) protein assembly concentration (N) on *PND* distributions for the monomer, trimer, and hexamer. (d) Computed errors associated with variations in r (top), PE (middle), and N (bottom).

detected protein assembly is near the FOV edge, the chances of another protein assembly entering its ROI decrease. This affects *ND* collection and *PND* generation. Such a complication is not considered in our theoretical model, eventually leading to larger discrepancies between simulated and theoretical *PND* values as r grows. It is crucial to note that the localization precision of the imaging method predominantly determines the choice of r. Our *PND* approach might not be ideal for systems with significant location errors such as conventional fluorescence systems with around 200 nm errors. A similar analysis is performed for the cytosolic protein, which also shows consistent alignment between theoretical and simulated *PND* distributions (Figure S3). These results affirm the consistent accuracy and reliability of the theoretical *PND* model, even with variations in the r, PE, and N.

Extracting Subpopulation of the Mixed Oligomeric State Is Feasible for Membrane and Cytosolic Protein in a Certain Protein Concentration Range. Extracting the subpopulations of different oligomeric forms is pivotal for understanding oligomer-mediated processes that drive specific cellular functions. To evaluate the effectiveness of the theoretical *PND* model in extracting the subpopulations of different oligomers, we examined membrane assemblies with four combinations of monomers (M), dimers (D), and trimers (T): M:D = 50%:50%, M:T = 50%:50%, D:T = 50%:50%, and M:D:T = 40%:30%:30%. By allowing the subpopulation of each oligomeric form to float during the fitting step, a residue-based algorithm was employed to identify the theoretical *PND* distributions that are most closely matched to the experimental data. Results from Figure 6a demonstrate that fitting with theoretical *PND* effectively extracted the protein oligomeric subpopulations with a typical error < 6 ± 1% across the four test conditions. Comparable outcomes were observed for cytosolic proteins.

Recognizing that protein concentrations can vary across cellular regions, the theoretical *PND*'s capability was further explored for proteins with different concentrations. A set of *PND* distributions was simulated for N values varying from 60 to 240 in increments of 30. The average of these *PND* distributions served as the ground truth input for the fitting algorithm. Within the algorithm, theoretical *PND* distributions with the same concentration distributions were generated and subsequently averaged to form the fitting metric to obtain the subpopulation of each oligomeric state. As depicted in Figure 6b, the theoretical *PND* again successfully determines protein oligomeric populations, even amidst fluctuations in protein concentrations. Conclusively, the theoretical *PND* emerges as a reliable approach for ascertaining the oligomeric states of both membrane and cytosolic proteins.

One interesting observation is that we observed a consistent overestimation of monomers in all cases, suggesting that our fitting algorithm exhibits increased sensitivity to monomeric

https://doi.org/10.1021/acs.analchem.3c04728
Anal. Chem. XXXX, XXX, XXX−XXX

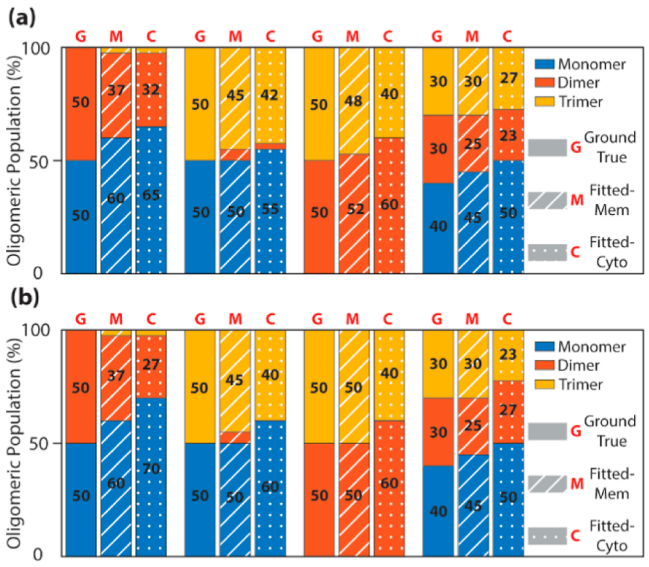


Figure 6. Extraction of oligomer populations through a theoretical *PND*. (a) Comparison between the ground truth (G) and fitted oligomer populations for both membrane (M) and cytosolic (C) proteins exhibiting a single oligomeric state. (b) The same as (a) but for proteins with various concentrations.

populations, essentially setting an upper limit on their estimation. We have quantified this overestimation to be around 10%. Consequently, when the subpopulation of other oligomeric forms is less than 10%, our method might underestimate them. This limitation should be considered when employing the *PND* to deduce protein oligomeric states.

Enhancing the *PND* Model for Single-Molecule Localization Microscopy with Future Prospects and Considerations. The *PND* offers a promising approach for analyzing protein oligomeric states. However, several considerations should be addressed when utilizing this methodology to ensure accurate results. For example, it is essential to recognize the impact of dye blinking and dark states in SMLM for an accurate data interpretation. Our theoretical model, while not detailing specific experimental protocols, assumes the use of SMLM data processed to mitigate these effects. This includes common SMLM practices, such as background noise reduction and blinking correction algorithms, crucial for distinguishing true signals from artifacts. Such processing ensures that the data reliably represent the spatial distribution and density of activated dyes, providing a solid foundation for applying our *PND* model. This understanding is vital for researchers integrating our theoretical framework with practical SMLM applications, aiming for precise analysis of protein oligomeric states in cellular environments.

Our *PND* model is primarily designed for analyzing protein structures ranging from monomers to hexamers, typically not exceeding 10 nm in size. However, it is important to note that the model in its current form may not be suitable for analyzing larger protein aggregates exceeding 10 nm. Addressing these larger structures would require recalibration of both our theoretical approach and specific parameters, such as the radius of the ROI. Investigating protein aggregates and their statistical distributions, potentially following power law or exponential decay models, presents an exciting direction for future research. While our current study focuses on quantifying specific oligomeric states, exploring general aggregation behaviors could offer deeper insights into protein dynamics.

CONCLUSIONS

Understanding the oligomeric states of proteins is crucial to understanding the complex landscape of cellular processes. While recent advancements in SMLM have highlighted the capabilities of protein location density techniques, there remains a need for a versatile method that can accurately assess proteins across varied environments. This study introduced the theoretical *PND* as a significant advancement in the field. Our results demonstrate that the theoretical *PND* consistently matches simulated data for both membrane and cytosolic proteins. Furthermore, this consistency is maintained across different protein concentrations and oligomeric states, highlighting the versatility of the *PND* approach. However, it is crucial to address certain limitations. The assumption of a random protein distribution within the FOV may render the *PND* less suitable for proteins with specific distribution patterns. An optimal concentration range is also essential for reliable results as extremely low or high concentrations can hinder data interpretation. In the broader context, the introduction of the *PND* offers a valuable avenue for accurately determining protein oligomeric states in cells and likely enhances our comprehension of protein-driven cellular processes.

ASSOCIATED CONTENT

Supporting Information

The Supporting Information is available free of charge at https://pubs.acs.org/doi/10.1021/acs.analchem.3c04728.

Theoretical expressions for *ND* = 0, 1, and 2 for pure membrane trimers; schematic representations illustrating potential scenarios of *ND* = 0, 1, and 2 for pure membrane trimers; confirmation of *PND* algorithm saturation; and effects of various experimental parameters on *PND* distributions of cytosolic proteins (PDF)

MatLab scripts for theoretical *PND* generation (ZIP)

AUTHOR INFORMATION

Corresponding Author

Tai-Yen Chen − *Department of Chemistry, University of Houston, Houston, Texas 77204, United States;* orcid.org/0000-0002-2881-3068; Email: tchen37@central.uh.edu

Author

Huanhuan Chen − *Department of Chemistry, University of Houston, Houston, Texas 77204, United States*

Complete contact information is available at: https://pubs.acs.org/10.1021/acs.analchem.3c04728

Notes

The authors declare no competing financial interest.

ACKNOWLEDGMENTS

The authors gratefully acknowledge financial support from NIH (R35GM133505) and the University of Houston.

REFERENCES

(1) Hashimoto, K.; Nishi, H.; Bryant, S.; Panchenko, A. R. *Phys. Biol.* **2011**, *8* (3), 035007.

(2) Wei, G.; Xi, W.; Nussinov, R.; Ma, B. *Chem. Rev.* **2016**, *116* (11), 6516−6551.
(3) Barducci, A.; Bonomi, M.; Prakash, M. K.; Parrinello, M. *Proc. Natl. Acad. Sci. U.S.A.* **2013**, *110* (49), No. E4708-E4713.
(4) Christis, C.; Lubsen, N. H.; Braakman, I. *FEBS J.* **2008**, *275* (19), 4700−4727.
(5) Tsvetkov, P.; Coy, S.; Petrova, B.; Dreishpoon, M.; Verma, A.; Abdusamad, M.; Rossen, J.; Joesch-Cohen, L.; Humeidi, R.; Spangler, R. D.; et al. *Science* **2022**, *375* (6586), 1254−1261.
(6) Bugaj, L. J.; Choksi, A. T.; Mesuda, C. K.; Kane, R. S.; Schaffer, D. V. *Nat. Methods* **2013**, *10* (3), 249−252.
(7) Ortiz-Guerrero, J. M.; Polanco, M. C.; Murillo, F. J.; Padmanabhan, S.; Elías-Arnanz, M. *Proc. Natl. Acad. Sci. U.S.A.* **2011**, *108* (18), 7565−7570.
(8) Lassar, A. B.; Davis, R. L.; Wright, W. E.; Kadesch, T.; Murre, C.; Voronova, A.; Baltimore, D.; Weintraub, H. *Cell* **1991**, *66* (2), 305−315.
(9) Mas, G.; Burmann, B. M.; Sharpe, T.; Claudi, B.; Bumann, D.; Hiller, S. *Sci. Adv.* **2020**, *6* (43), No. eabc5822.
(10) Krüger, C. L.; Zeuner, M.-T.; Cottrell, G. S.; Widera, D.; Heilemann, M. *Sci. Signaling* **2017**, *10* (503), No. eaan1308.
(11) Karathanasis, C.; Medler, J.; Fricke, F.; Smith, S.; Malkusch, S.; Widera, D.; Fulda, S.; Wajant, H.; van Wijk, S. J. L.; Dikic, I.; Heilemann, M. *Sci. Signaling* **2020**, *13* (614), No. eaax5647.
(12) Lin, T.-Y.; Ma, Y.-W.; Tsai, M.-Y. *J. Phys. Chem. B* **2023**, *127* (5), 1074−1088.
(13) Zhang, Y.; Wen, M.-H.; Qin, G.; Cai, C.; Chen, T.-Y. *Metallomics* **2022**, *14* (11), mfac087.
(14) Chen, H.; Chen, T.-Y. *Chem. Biomed. Imaging* **2023**, *1* (1), 49−57.
(15) Nguyen, D.; Yan, G.; Chen, T.-Y.; Do, L. H. *Angew. Chem., Int. Ed.* **2023**, *62*, No. e202300467.
(16) Zulkifli, M.; Spelbring, A. N.; Zhang, Y.; Soma, S.; Chen, S.; Li, L.; Le, T.; Shanbhag, V.; Petris, M. J.; Chen, T.-Y.; Ralle, M.; Barondeau, D. P.; Gohil, V. M. *Proc. Natl. Acad. Sci. U.S.A.* **2023**, *120* (10), No. e2216722120.
(17) Ojha, N.; Rainey, K. H.; Patterson, G. H. *Nat. Commun.* **2020**, *11* (1), 21.
(18) Goncalves, S. A.; Matos, J. E.; Outeiro, T. F. *Trends Biochem. Sci.* **2010**, *35* (11), 643−651.
(19) Kaliszewski, M. J.; Shi, X.; Hou, Y.; Lingerak, R.; Kim, S.; Mallory, P.; Smith, A. W. *Methods* **2018**, *140−141*, 40−51.
(20) Avin, A.; Levy, M.; Porat, Z.; Abramson, J. *Nat. Commun.* **2017**, *8* (1), 1524.
(21) Roy, R.; Hohng, S.; Ha, T. *Nat. Methods* **2008**, *5* (6), 507−516.
(22) Yang, J.; Dear, A. J.; Michaels, T. C. T.; Dobson, C. M.; Knowles, T. P. J.; Wu, S.; Perrett, S. *J. Am. Chem. Soc.* **2018**, *140* (7), 2493−2503.
(23) Chen, T.-Y.; Santiago, A. G.; Jung, W.; Krzemiński, Ł.; Yang, F.; Martell, D. J.; Helmann, J. D.; Chen, P. *Nat. Commun.* **2015**, *6*, 7445.
(24) Fricke, F.; Beaudouin, J.; Eils, R.; Heilemann, M. *Sci. Rep.* **2015**, *5* (1), 14072.
(25) Song, Y.; Ge, B.; Lao, J.; Wang, Z.; Yang, B.; Wang, X.; He, H.; Li, J.; Huang, F. *Biochemistry* **2018**, *57* (5), 852−860.
(26) Nagata, K. O.; Nakada, C.; Kasai, R. S.; Kusumi, A.; Ueda, K. *Proc. Natl. Acad. Sci. U.S.A.* **2013**, *110* (13), 5034−5039.
(27) Ulbrich, M. H.; Isacoff, E. Y. *Nat. Methods* **2007**, *4* (4), 319−321.
(28) Baldering, T. N.; Karathanasis, C.; Harwardt, M.-L. I. E.; Freund, P.; Meurer, M.; Rahm, J. V.; Knop, M.; Dietz, M. S.; Heilemann, M. *iScience* **2021**, *24* (1), 101895.
(29) Walker, G.; Brown, C.; Ge, X.; Kumar, S.; Muzumdar, M. D.; Gupta, K.; Bhattacharyya, M. *Nat. Nanotechnol.* **2023**.
(30) Nan, X.; Collisson, E. A.; Lewis, S.; Huang, J.; Tamgüney, T. M.; Liphardt, J. T.; McCormick, F.; Gray, J. W.; Chu, S. *Proc. Natl. Acad. Sci. U.S.A.* **2013**, *110* (46), 18519−18524.
(31) Xie, X.; Cheng, Y.-S.; Wen, M.-H.; Calindi, A.; Yang, K.; Chiu, C.-W.; Chen, T.-Y. *J. Phys. Chem. B* **2018**, *122* (46), 10496−10504.
(32) Chen, H.; Xie, X.; Chen, T.-Y. *Curr. Opin. Struct. Biol.* **2021**, *66*, 112−118.
(33) Pavani, S. R. P.; Thompson, M. A.; Biteen, J. S.; Lord, S. J.; Liu, N.; Twieg, R. J.; Piestun, R.; Moerner, W. *Proc. Natl. Acad. Sci. U.S.A.* **2009**, *106* (9), 2995−2999.
(34) Quirin, S.; Pavani, S. R. P.; Piestun, R. *Proc. Natl. Acad. Sci. U.S.A.* **2012**, *109* (3), 675−679.
(35) Huang, B.; Wang, W.; Bates, M.; Zhuang, X. *Science* **2008**, *319* (5864), 810−813.
(36) Pavani, S. R. P.; Piestun, R. *Opt. Express* **2008**, *16* (26), 22048−22057.
(37) Gustavsson, A.-K.; Petrov, P. N.; Lee, M. Y.; Shechtman, Y.; Moerner, W. *Nat. Commun.* **2018**, *9* (1), 123.
(38) Gustavsson, A.-K.; Petrov, P. N.; Lee, M. Y.; Shechtman, Y.; Moerner, W. Tilted light sheet microscopy with 3D point spread functions for single-molecule super-resolution imaging in mammalian cells. *Single Molecule Spectroscopy and Superresolution Imaging XI*; International Society for Optics and Photonics, 2018; Vol. *10500*, p 105000M.
(39) Durisic, N.; Laparra-Cuervo, L.; Sandoval-Álvarez, Á.; Borbely, J. S.; Lakadamyali, M. *Nat. Methods* **2014**, *11* (2), 156−162.

https://doi.org/10.1021/acs.analchem.3c04728
Anal. Chem. XXXX, XXX, XXX−XXX