Supporting Information for

Modeling the Self-Assembly of Protein Complexes through a Rigid-Body Rotational Reaction-Diffusion Algorithm

Margaret E. Johnson*

**Corresponding Author*

The Johns Hopkins University, TC Jenkins Department of Biophysics, 3400 N Charles St, Baltimore MD 21218

1. **Simulation Methods**
    a. **FPR**
    b. **BD**
    c. **Gillespie**
2. **Optimal hexagonal tiling**

1. Simulation Methods:

1A. FPR simulations for rotational models

Clathrin Simulation Parameters: For the clathrin trimer simulations, for an individual trimer, we used D_t=3nm^2/μs and D_R=0.05rad^2/μs. For K_D=100μM, we used reaction rates of k_a=0.0332nm^3/μs, k_b=1.00022s^{-1}. For K_D=1μM: k_a=3.3971nm^3/μs, k_b=1.0225s^{-1}. For K_D=0.2μM: k_a=18.67nm^3/μs, k_b=1.124s^{-1}. The binding radius σ=1nm.

FPR Method: The Free Propagator Reweighting (FPR) algorithm has been described elsewhere [1-2], so we briefly review here the approach. As indicated in the name, particles or molecules are propagated according to free diffusion, with position updates explicitly defined below. Reaction probabilities are evaluated each time-step Δt for every pair of reactive sites that are separated by r<$R_{\max}$, where $R_{max} = \sigma + 3\sqrt{2dD_{eff}\Delta t}$. Smaller time-steps thus reduce R_{max} and the number of pairs that could react per step. The reaction probabilities (eq 3) must be reweighted by the ratio

$$\omega_{ratio}(r,t|r_0) = p_{\mathrm{irr}}(r,t|r_0)/p^*_{free}(r,t|r_0) \qquad \text{(S1)}$$

of the proper GF (p_{irr}) relative to the free GF (p^*_{free}) used for position updates, where p_{free} is the solution to Eq 2a,c,d, and p^*_{free} is this same distribution after it has been renormalized to maintain excluded volume of a reactive pair (r cannot be $< \sigma$) and rescaled by $S(t|r_0)$ to account for association events. If a reactive site has more than one partner that it could react with in the next step, then each pair is evaluated independently. Time-steps are chosen to minimize the frequency of these non-pairwise interactions, because when new positions are updated, all pairs must enforce excluded volume.

Association events occur if the reweighted reaction probability is greater than a uniform random number (URN). Upon association, particles are placed at a contact separation of r_σ=σ along the vector $\mathbf{r}_{\boldsymbol{\sigma}}$, with each molecule/complex moving by an amount proportional to its diffusion constant. If protein complex orientations are specified, rigid body rotations are used to 'snap' molecules into user-defined geometries, otherwise the

orientations remain unchanged. When complexes dissociate, the reactive sites are already at contact, which is the appropriate separation needed to recover detailed balance. Orientations are left unchanged after dissociation. For molecules or complexes that do no undergo any reaction in Δt, their positions are updated according to free diffusion, and updates that result in overlap ($r<\sigma$) with any possible reaction partners (pairs that had $r<R_{max}$) are rejected and resampled.

Macroscopic rates: For multi-site molecules, geometry could impact measured macroscopic rates. The relationship between the microscopic parameters and the macroscopic rates (e.g. eq 6) is based on the model of particles reacting at $r_\sigma=\sigma$ over the full surface of a sphere with radius σ. This access to binding is reduced by ½, for example, when particles in solution react with particles (e.g. lipids) embedded within a reflective surface. We therefore accordingly adjust the relationship between macroscopic rates and microscopic rates by a factor of ½. This example is easily re-adjusted, but the same occluded access could occur if a single molecule has multiple sites that are densely packed. Over small time-steps, a shared binding partner may only have partial access to each of the sites, which would reduce the net macroscopic rate of binding. Hence, the geometry of the molecules can alter measured macroscopic rates in these cases.

Propagating Molecules: Molecules and complexes are moved according to free translational and rotational diffusion. For translational diffusion, we use simple Euler updates for each complex *i*

$$x_i(t+\Delta t) = x_i(t) + \sqrt{2D_i\Delta t}R \qquad \text{(S2)}$$

and similarly in y and z, where R is a Gaussian distributed random number with mean zero and standard deviation of one. For rotational diffusion, each molecule (or complex) with rotational diffusion constant D_{Ri} is rotated around the global (motionless) x, y and z axis each by

$$\alpha_i = \sqrt{2D_{Ri}\Delta t}R. \qquad \text{(S3)}$$

The rotation matrix from these three rotations is applied to each vector within a complex connecting a site to the center of mass of the complex, as validated in Figure 2. This

includes the reactive sites (p_1) and the molecule centers (c_1), which may be displaced from the center of a multi-protein complex.

Transport properties of complexes: Once two proteins form a complex, we update the diffusion coefficients of the rigid complex to reflect its larger hydrodynamic radius, *a*. Based on the standard Einstein Stokes relations, where $D_t = k_B T/6\pi\eta a$ and $D_R = k_B T/8\pi\eta a^3$ with η the viscosity, k_B Boltzmann's constant, and *T* temperature, we use for a complex made of N_{pro} proteins:

$$D_{t,complex}{}^{-1} = \sum_{k=1}^{N_{pro}} D_k{}^{-1} \quad \text{(S4)}$$

and

$$D_{R,complex}{}^{-1/3} = \sum_{k=1}^{N_{pro}} D_{Rk}{}^{-1/3}. \quad \text{(S5)}$$

This defines the complex's hydrodynamic radius as the sum over all the molecular components. An alternative method would be to re-calculate *a* from the geometry of the complex, or to sum over the component masses, *M* and assume, for example, $a \sim M^{1/3}$.

Hexagonal loop closure probability: For reactive sites that are not diffusing relative to one another because they are in the same rigid complex, $D_{tot}=0$, and the reaction probability for association is either 1 if $r_\sigma=\sigma$, or 0 otherwise. When we use $p_{react}=1$ for these reactions, the hexagon closure is irreversible; even when dissociation occurs after loop closure, only two simultaneous dissociation events can release a trimer from being held within the complex, and this has a miniscule probability. To allow for reversible loop closure, we also used a detailed balance expression for transitioning between unbound (u) and bound (b) states, where $p_{u\to b}(\Delta t) = p_{b\to u}(\Delta t)\frac{p_b}{p_u}$. The transition probability $p_{b\to u}(\Delta t)$ is simply the dissociation probability of eq 4. The ratio $\frac{p_b}{p_u}$ describing the equilibrium states should be dependent only on k_a and k_b. Based on the equilibrium for A+A⇌C, we chose $\frac{p_b}{p_u}=\frac{2N_{C,eq}}{N_{A,eq}}$, which is larger than the thermodynamic expression $\frac{p_b}{p_u}=\frac{N_{C,eq}}{N_{A,eq}{}^2}$, but still produced an overrepresentation of un-closed hexagons. This ad-hoc definition

could be improved upon in future work by, for example, models for pairwise interactions defined for specific equilibrium assembly geometries[3].

1B. BD simulations of full rotational model

To calculate the GFs (Figure 5) and the reaction probabilities (Figure 4) for the full model of Figure 1, we used Brownian Dynamics (BD) simulations with the algorithm of Zhou[4]. Simulations are initialized by placing the reactive particles at a separation $r_\sigma = r_0$ and an initial orientation $\Omega_0 = [\theta_{A0}, \theta_{B0}, \psi_0]$. We considered 5 unique values of θ_{A0} and θ_{B0} each, over the range 0 to π, and 5 unique values of ψ_0, also over the range 0 to π due to symmetry, totaling 125 initial relative orientations (colorful curves in Figures 4-5). Position updates follow the free diffusion propagation described above. For BD, reactions can occur only in the region from $\sigma<r_\sigma<\sigma+\epsilon$ with rate given by κ. The reaction probability is given by $p_{react}(r, \Delta t|r_0) = 1 - \exp\left(-\frac{\Delta t}{2}\left(\frac{\kappa(r_0)}{\epsilon} + \frac{\kappa(r)}{\epsilon}\right)\right)$, which is more accurate as $\epsilon \to 0$. We used a small ϵ to ensure very high accuracy, setting $\epsilon = 0.001 D_{eff}/\kappa$ and the time-step in the reaction region to $\Delta t = 0.001\epsilon^2/2D_{eff}$, which was typically on the order of 10^{-16}s. Time-steps lengthened outside this region. If particles reacted, the trajectory was terminated and re-initialized. Time-dependent rates were simulated out to a total of 1ms (Fig 6), with typically 100 repetitions performed for each initial orientation, which were then averaged over. Each GF/survival probability simulation was for a total of 0.01μs, and typically calculated over 500,000 repetitions for each initial orientation, taking about a day on 120 processors. These time-consuming numerical calculations are not feasible for many-body simulations, where GFs would be needed for all initial separations and orientations and for all changes to reaction parameters.

1C. Gillespie Simulations

For Gillespie simulations, each trimer had three sites, A_1, A_2, and A_3 with the same binding properties. For all reactions, we used $k_{off}=1s^{-1}$. For identical site interactions, $A_i+A_i \to C$, $k_{on}=k_{off}/K_D$. To recover the proper equilibrium for non-identical sites ($A_i+A_k \to C$, $i \neq k$) one must define $k_{on}=2k_{off}/K_D$ because they are treated by the algorithm as

distinct species rather than a self-interaction. This ensures the same equilibrium is reached as one expects from simulating just A_1 species with three times the copy numbers. This feature is naturally captured in the FPR simulations, as the same value of k_a produces a macroscopic rate twice as large for distinct versus self-binding partners.

For the Gillespie simulations, when a binding event occurs $A_i+A_k \rightarrow C$: $i \neq k$, two free sites are selected to bind, with the only requirement that they are not within the same molecule. The propensity for a distinct binding reaction is normally given by $h=NA_i*NA_k$. Neither this propensity nor $h=NA_i*(NA_k-1)$ perfectly captures the combinatorial binding possibilities under the requirement that a molecule not bind to itself. We use the latter propensity; it is nearly correct and is simpler than enumerating all the prohibited binding events each step.

2. Optimal hexagonal tiling

For a hexagonal tiling, each vertex represents a clathrin trimer. The first hexagon formed has 6 trimers, and a second connected hexagon adds 4 trimers. If the lattice grows as a compact cluster to maximize bonds/edges formed (rather than a long chain of hexagons) each hexagon requires an additional 3 or 2 trimers. For an ideal tiling this gives $N_{trimers} \approx 10 + 2.5(N_{hex} - 2)$. Similarly, the total number of bonds formed between trimer legs is $N_{bonds} \approx 11 + 3.5(N_{hex} - 2)$. These equations work well to predict the maximum hexagons and bound legs possible given $N_{trimers}$. For our simulations of 100 trimers, one can verify that there are maximally 38 complete hexagons, with 137 bonds, and 26 'sticky ends' left unbound.

SI References:

1. Johnson, M. E.; Hummer, G., Free-Propagator Reweighting Integrator for Single-Particle Dynamics in Reaction-Diffusion Models of Heterogeneous Protein-Protein Interaction Systems. *Phys Rev X* **2014,** *4* (3), 031037.
2. Yogurtcu, O. N.; Johnson, M. E., Theory of bi-molecular association dynamics in 2D for accurate model and experimental parameterization of binding rates. *J. Chem. Phys.* **2015,** *143*, 084117.
3. Zlotnick, A., To build a virus capsid. An equilibrium model of the self assembly of polyhedral protein complexes. *J Mol Biol* **1994,** *241* (1), 59-67.

4. Zhou, H. X., Kinetics of Diffusion-Influenced Reactions Studied by Brownian Dynamics. *J Phys Chem-Us* **1990,** *94* (25), 8794-8800.