



# Unified regression model of binding equilibria in crowded environments

Byoungkoo Lee<sup>1</sup>, Philip R. LeDuc<sup>2</sup> & Russell Schwartz<sup>1</sup>

SUBJECT AREAS:

MATHEMATICS AND  
COMPUTING

BIOLOGICAL MODELS

COMPUTATIONAL BIOLOGY

BIOPHYSICS

<sup>1</sup>Department of Biological Sciences and Lane Center for Computational Biology, Carnegie Mellon University, 654 Mellon Institute, 4400 Fifth Avenue., Pittsburgh, PA 15213, <sup>2</sup>Departments of Mechanical and Biomedical Engineering, and Lane Center for Computational Biology, Carnegie Mellon University, 420 Scaife Hall, 5000 Forbes Avenue, Pittsburgh, PA 15213.

Received  
7 July 2011

Accepted  
5 September 2011

Published  
20 September 2011

**Molecular crowding is a critical feature distinguishing intracellular environments from idealized solution-based environments and is essential to understanding numerous biochemical reactions, from protein folding to signal transduction. Many biochemical reactions are dramatically altered by crowding, yet it is extremely difficult to predict how crowding will quantitatively affect any particular reaction systems. We previously developed a novel stochastic off-lattice model to efficiently simulate binding reactions across wide parameter ranges in various crowded conditions. We now show that a polynomial regression model can incorporate several interrelated parameters influencing chemistry under crowded conditions. The unified model of binding equilibria accurately reproduces the results of particle simulations over a broad range of variation of six physical parameters that collectively yield a complicated, non-linear crowding effect. The work represents an important step toward the long-term goal of computationally tractable predictive models of reaction chemistry in the cellular environment.**

Correspondence and requests for materials should be addressed to R.S. (russells@andrew.cmu.edu)

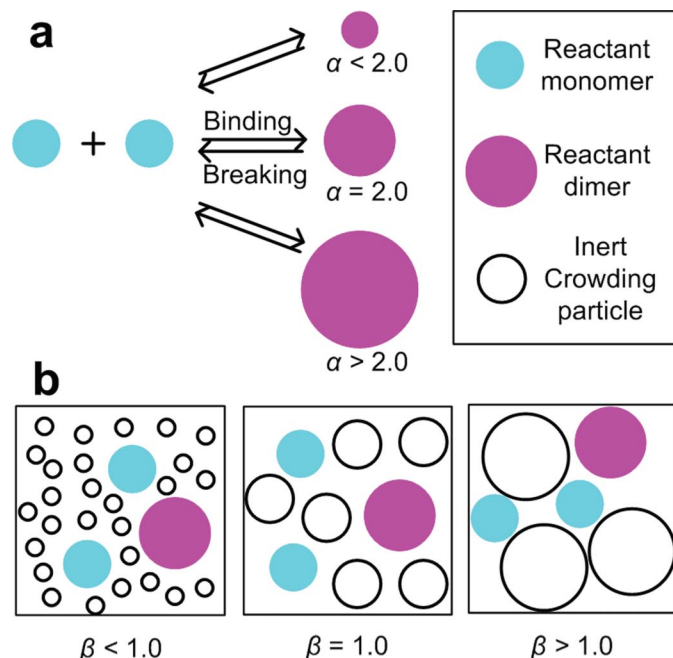
The intracellular environment is a densely concentrated region packed with a mixture of numerous types of macromolecules and subcellular structures<sup>1</sup>. Because of this dense crowding, biochemical reactions *in vivo* behave significantly differently from the same reactions in the well-mixed and dilute conditions of typical *in vitro* models<sup>1</sup>. Molecular crowding, one of critical features of the intracellular environment, affects many fundamental cellular processes, such as protein folding<sup>2</sup>, protein aggregation<sup>3</sup>, enzyme activity<sup>4</sup>, reaction kinetics<sup>5</sup>, and signal transduction<sup>6</sup>. Moreover, the relative size and shape of crowding agents are also crucial parameters in molecular crowding<sup>7,8</sup>. The specific effects of molecular crowding on any particular system, however, are not easily predicted even given detailed knowledge of the specific physical parameters of the given reaction system<sup>9–11</sup>. Therefore, quantitatively characterizing how crowding influences any given reaction system is a very challenging problem using *in vivo* or *in vitro* model systems. Computational modeling and simulation methods provide alternative ways to estimate the effects of individual parameters singly or in combination because these methods allow for us to easily and precisely vary reaction system parameters separately and in combination, a capability difficult to match in any real experimental system. Prior computational models, however, have significant limitations with respect to simulating model reaction systems under crowded conditions. Ordinary differential equation (ODE) models generally assume a well-mixed and diluted solution and thus ignore the crowding effect. Partial differential equation (PDE) models can include spatial constraints but also provide no explicit basis for modeling the crowding effect specifically<sup>12</sup>. Particle-based simulations provide a way to more directly address the molecular crowding effect and explore how many parameter changes might influence it. Lattice-based models provide a computationally efficient means of performing such simulations but require simplified models of particle movement and structures and are therefore prone to exaggerating the crowding effect<sup>13</sup>. Off-lattice models based on Brownian or Langevin dynamics provide for more realistic simulations but at a high computational cost for highly crowded conditions<sup>14,15</sup>, making them unsuitable for modeling crowding for large numbers of particles or long time scales. There is consequently no method that has both the realism to accurately and quantitatively simulate reactions in crowded media and the efficiency to do so for sufficiently large system sizes and time scales to model many biological reactions.

In order to overcome the inherent limitations of the various methods available, we propose a novel strategy intended to yield both efficiency and accuracy in modeling chemistry in crowded media. We accomplish this goal by using relatively computationally costly particle simulations on simple test systems to determine how multiple parameters act singly and in combination to influence the crowding effect and then fit an easily-applied regression model to outputs of these simulations for use in quickly making crowding corrections to more efficient simulation methods. In previous work, developed a coarse-grained two-dimensional stochastic off-lattice particle simulation (2DSOLM)<sup>16</sup> based on Green's function reaction dynamics<sup>17</sup> for simulating binding kinetics in various molecular



crowding conditions. We subsequently showed that one can simplify the problem of modeling of the crowding effect by identifying a subset of “separable” parameters, whose effects one could learn independently and then merge into a model of their collective effects<sup>18</sup>. These four parameters in 2DSOLM are the total concentration ( $C$ ); the probability of binding upon a collision between two reactant monomers ( $B$ ); the mean time for dissociation event ( $M$ ), defined as the inverse of the rate constant; and the diffusion coefficient for reactants and inert particles ( $D$ ). The work showed that such an approach can simplify the problem of quantitatively modeling reactions in crowded media by allowing one to independently account for the influence of several separable parameters. It did not, however, resolve the inherent problem that the crowding effect depends on non-linear interactions of multiple physical parameters.

Here, we generalize the prior approach to account for both separable and inseparable parameters by creating a collective model accounting for the prior parameters and two additional parameters – the area ratio of dimer to monomer ( $\alpha$ ,  $\pi r_{dimer}^2 = \alpha \pi r_{monomer}^2$ ) and the area ratio of inert crowding agent to monomer ( $\beta$ ,  $\pi r_{inert\ particle}^2 = \beta \pi r_{monomer}^2$ ) – whose effects are inseparable from one another and from that of the total concentration ( $C$ ). Figure 1 illustrates these two cross-dependent interaction parameters. The parameter  $\alpha$  captures the change in excluded volume induced by binding between two reactant monomers. Fig. 1a illustrates three different scenarios after the reaction:  $\alpha < 2$ ,  $\alpha = 2$ , and  $\alpha > 2$ . The first case represents a decrease in the area occupancy of two reactants, as might occur if one molecule docks into a solvent-inaccessible binding pocket of another molecule or the two bind tightly along two rough faces whose solvent-accessible areas were significantly larger than their van der Waals areas. The second case represents an unchanged area occupancy of the product relative to the two separate reactants. The third case represents an increase in area occupancy of the product binding, as may occur if the proteins bind so as to establish some new solvent-inaccessible cavities. This parameter would be expected to influence crowding due to an entropic preference for reducing excluded volume



**Figure 1 | Illustrations of the 2DSOLM reaction model for two different cross-dependent interaction parameters.** The parameters  $\alpha$  and  $\beta$  correspond to sizes of reactant dimer and crowding agent relative to reactant monomer. Cyan circles are reactant monomers, magenta circles are reactant dimers, and open circles are inert crowding agents. (a) Illustration of varying  $\alpha$ . (b) Illustration of varying  $\beta$ .

in conditions of high crowding. We would thus expect crowding to tend to drive binding for  $\alpha < 2$  and inhibit binding for  $\alpha > 2$ .

The parameter  $\beta$  represents another cross-dependent property describing the interaction between reactants and inert particles and specifically influencing how effectively inert particles provide steric hindrance to reactants. Three different cases are illustrated in Fig. 1b:  $\beta < 1$ ,  $\beta = 1$ , and  $\beta > 1$  for 2DSOLM. The first case corresponds to individual inert particles smaller than reactant monomers, the second to inert and reactant particles of equal size, and the third to inert particles larger than reactant monomers. Unfolding experiments on ubiquitin<sup>7</sup> have shown that reducing the size of an inert crowding agent can lead to a stronger crowding effect. However, such anecdotal observation provides little basis for quantitatively estimating how the relative size of crowding agents to reactants will affect any given the binding reaction system, especially in a background of potentially multiple other parameter changes. For example, while scaled particle theory in principle allows one to make such inferences for some isolated parameter changes, it is difficult to generalize to more complicated scenarios. In the present work, we develop a unified model accounting for both the previously separable and the inseparable parameters of the simulation, which we refer to as the unified regression model. We show that it is possible to accommodate the three inseparable parameters ( $\alpha, \beta, C$ ) in a single model through a multidimensional polynomial regression model and validate this model over variations in pairs of parameters. In addition, we validate our model of these three inseparable parameters by comparison with estimates from scaled particle theories<sup>19,20</sup>. We then show that it is possible to produce a full unified regression model combining these multidimensional regression models with the previously derived separable components. Finally, we show that the unified model provides an accurate description of random changes in all of the simulation model’s physical parameters over a broad range of biologically relevant values in all parameters. In the process, we show a way to build a model with nearly equivalent predictive power to the costly particle simulations, but capable of easy evaluation for use in correcting computationally efficient simulations of large systems and long time scales. These findings have ramifications for our ability to computational model a diversity of cellular functions that are likely to be heavily influenced by intracellular crowding in two dimensional space, such as protein synthesis, signal transduction, or cytoskeleton assembly and disassembly.

## Results

**Simulation in 2DSOLM.** We investigated the parameter dependence of binding reactions in crowded media by using a two-dimensional model of dimerization of a reactant monomer of radius 2.5nm simulated in a 100nm  $\times$  100nm space with a hard reflective boundary condition. This simple reaction model allowed us to focus on the specific issue of binding equilibrium in a setting tractable for exploring a broad parameter space with sufficiently large particle numbers and numbers of replicates to produce reproducible results. We began with our prior regression model expressing equilibrium constant as a multiplicative function of the effects of the four separable parameters:  $C$  (total concentration),  $B$  (binding probability),  $M$  (inverse dissociation rate), and  $D$  (diffusion coefficient)<sup>18</sup>. We then added the two cross-dependent physical parameters,  $\alpha$  and  $\beta$ , which produce crowding effects inseparable from one another and from  $C$ . To extend the prior regression model to the parameters  $\alpha$  and  $\beta$ , we first established a baseline simulation parameter set with default parameter values of  $B=0.7$ ,  $M=1\text{ns}$ ,  $D=6.95 \times 10^{-11} \text{m}^2 \text{s}^{-1}$ ,  $\alpha=2$ , and  $\beta=1$ , values chosen based on our prior simulation studies<sup>16,18</sup> to produce a reasonably strong crowding effect as well as to approximate the temperature and viscosity conditions of the cytoplasm<sup>21,22</sup>. We fixed one additional simulation parameter,  $d_{th}$ , describing a threshold maximum distance at which two particles can interact with one another, to be  $d_{th}=0.5\text{nm}$  (one fifth of the radius of a reactant monomer). In SOLM simulations,

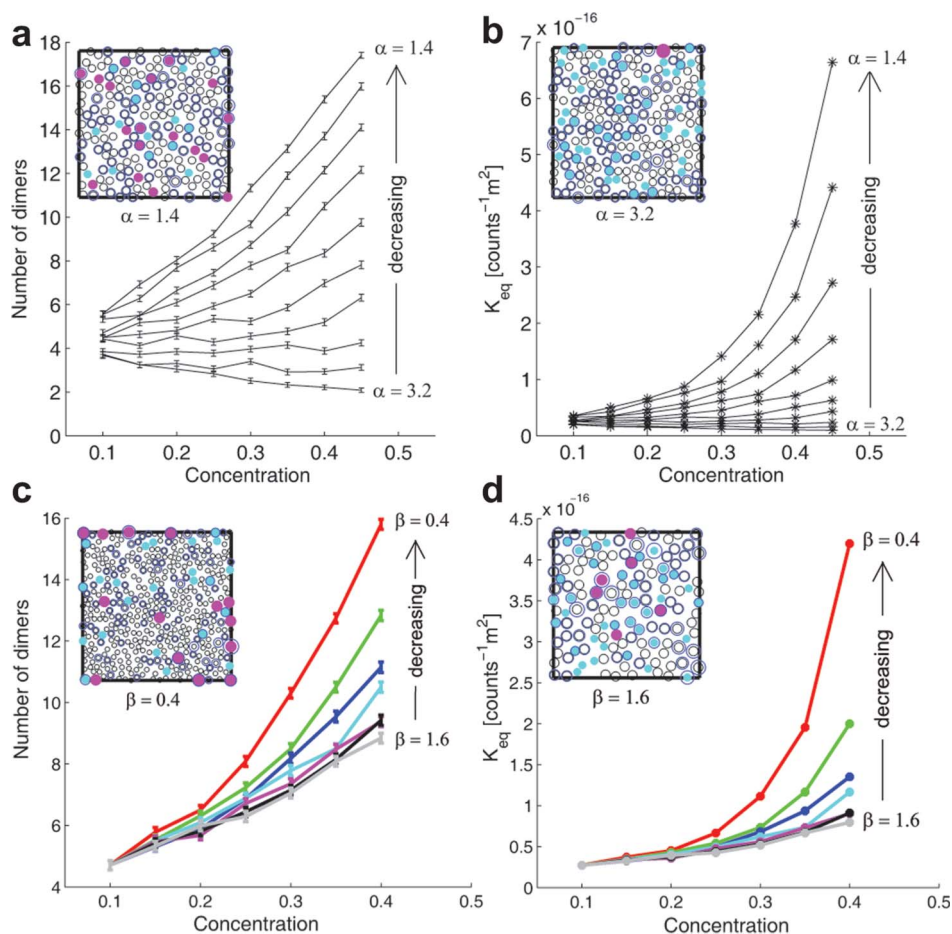


all particles are diffusible and are separated by at least a distance of  $d_{th}$ . When the simulation places two reactant monomers within a distance of  $d_{th}$  of one another, they must either undergo a binding reaction or a collision event, with probability  $B$  of binding or  $1 - B$  of collision<sup>16</sup>. One reactant and one inert particle or two inert particles that are placed within  $d_{th}$  of one another always undergo a collision event<sup>16</sup>. A binding event will result in the formation of a new dimer, while a collision will result in the particle positions being resampled to be beyond  $d_{th}$ . The choice of  $d_{th}$  influences the physical model of the simulator, in part because  $d_{th}$  is effectively a radius at which particles are assumed close enough to exert non-bonded forces on one another and initiate a possible binding interaction and in part because it imposes an upper limit on the maximum possible crowding level in the simulation. It also influences computational efficiency, as small  $d_{th}$  can lead to larger numbers of events when particles are in close proximity. The default  $d_{th}$  value was chosen empirically to permit physiologically reasonable levels of crowding cases (up to 0.45 concentration) while still yielding efficient simulations.

We then separately varied the individual parameter values for  $\alpha$  and  $\beta$  in conjunction with variations in total concentration using the default values for the other parameters to test these cross-dependent parameter effects on the binding chemistry. We simulated ten  $\alpha$  values (1.4, 1.6, 1.8, 2.0, 2.2, 2.4, 2.6, 2.8, 3.0, 3.2) and seven  $\beta$  values (0.4, 0.6, 0.8, 1.0, 1.2, 1.4, 1.6). For different  $\alpha$  value simulations, we simulated eight total concentrations (0.1, 0.15, 0.2, 0.25, 0.3, 0.35, 0.4,

0.45) with dimensionless units of fraction of total simulation area occupied by particles. To produce varying concentrations, we began with a fixed concentration of reactants of 0.1 and then added inert crowding particles to yield each higher value of total concentration (e.g.,  $C=0.1$  corresponds to purely reactant monomers while  $C=0.25$  corresponds to concentration 0.1 of reactant monomers and 0.15 of inert crowding agents). For simulations of varying  $\beta$ , we simulated seven total concentrations (0.1, 0.15, 0.2, 0.25, 0.3, 0.35, 0.4) again using the fixed reactant concentration of 0.1 and varying inert monomer concentrations, similar to the  $\alpha$  simulations. Each set of parameters was run for  $25\mu s$  with 30 repetitions, with progress recorded every  $0.15625\mu s$ . SOLM is a discrete event-driven method and thus has no fixed time step, but rather jumps between discrete changes in system states with randomly sampled waiting times between events<sup>16</sup>. The time interval to collect data on simulation progress only affects visualization of results, not the actual progress of the simulation, and was chosen to provide sufficient resolution to clearly display transient behavior early in the curve and stochastic fluctuations at long time scales. Data was collected at five different simulation times (5, 10, 15, 20,  $25\mu s$ ) to analyze the crowding effect on the test reaction, with values selected based on our prior work<sup>18</sup>. For each condition, we measured reaction progress by the mean number of dimers as a function of time across all simulations.

Figure 2 shows the simulation results for varying values of parameters  $\alpha$  and  $\beta$ . Fig. 2a shows mean dimer counts at



**Figure 2 | Binding equilibrium as a function of concentration for varying  $\alpha$  and  $\beta$ .** Cyan circles represent reactant monomers, magenta circles represent reactant dimers, and open circles represent inert crowding particles. Embedded screen snapshots show an illustrative quasi-equilibrium state for the maximum feasible level of crowding for each curve set. (a) Dimer counts for  $\alpha=1.4$  (top), 1.6, 1.8, 2.0, 2.2, 2.4, 2.6, 2.8, 3.0, 3.2 (bottom). Inset image shows a snapshot for  $C=0.45$  and  $\alpha=1.4$ . (b) Equilibrium constants for  $\alpha=1.4$  (top), 1.6, 1.8, 2.0, 2.2, 2.4, 2.6, 2.8, 3.0, 3.2 (bottom). Inset image shows a snapshot for  $C=0.45$  and  $\alpha=3.2$ . (c) Dimer counts for  $\beta=0.4$  (red, top), 0.6 (green), 0.8 (blue), 1.0 (cyan), 1.2 (magenta), 1.4 (black), 1.6 (gray, bottom). Inset image shows a snapshot for  $C=0.4$  and  $\beta=0.4$ . (d) Equilibrium constants for  $\beta=0.4$  (red, top), 0.6 (green), 0.8 (blue), 1.0 (cyan), 1.2 (magenta), 1.4 (black), 1.6 (gray, bottom). Inset image shows a snapshot for  $C=0.4$  and  $\beta=1.6$ .

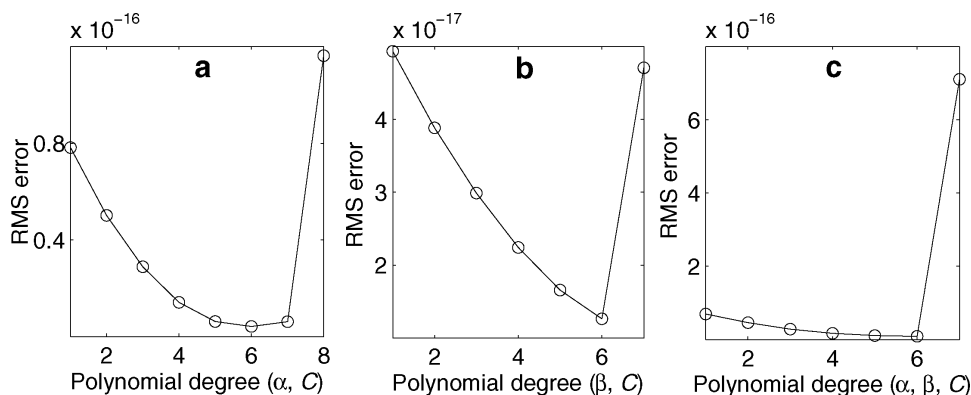


quasi-equilibrium for varying  $\alpha$  values. As the parameter value  $\alpha$  decreases, the number of dimers at equilibrium rapidly increases, especially for the high crowding cases. Interestingly, there is only a slight difference between dimer counts at very low crowding levels but the separation between the equilibrium counts rapidly increases with increasing crowding level. The  $K_{eq}$  curve, calculated from simulation data, more clearly shows how strongly the parameter  $\alpha$  influences the model reaction system, as shown in Fig. 2b. These results show that the parameter  $\alpha$  and the parameter  $C$  are cross-dependent and inseparable. While the size of a dimer does not significantly influence the binding reaction under low crowding conditions, a smaller dimer is much more conducive to binding than a larger dimer under densely crowded conditions. The embedded images in Fig. 2a and 2b show snapshots of the simulator at the quasi-equilibrium state (25  $\mu$ s) for  $\alpha=1.4$  and  $\alpha=3.2$  respectively at under a highly crowded condition ( $C=0.45$ ) in 2DSOLM. Fig. 2c shows mean dimer counts at the quasi-equilibrium state based on simulation data for varying  $\beta$ . As the parameter value  $\beta$  decreases, the number of dimers at the equilibrium state rapidly increases, especially for high crowding cases. Similar to the parameter  $\alpha$  case, variations in  $\beta$  produce little change at low crowding levels, but the difference rapidly increases as the crowding level increases. The  $K_{eq}$  curve, calculated from simulation data, shows how strongly the parameter  $\beta$  influences the model reaction system, as shown in Fig. 2d. These results show that the parameter  $\beta$  and the parameter  $C$  are cross-dependent and inseparable. While the size of an inert crowding agent does not significantly influence the binding reaction under low crowding conditions, binding is much more favorable in the presence of a smaller versus a larger crowding agent under densely crowded conditions. The embedded images in Fig. 2c and 2d show snapshots of the simulator at the quasi-equilibrium state (25  $\mu$ s) for  $\beta=0.4$  and  $\beta=1.6$  respectively at a highly crowded condition ( $C=0.4$ ) in 2DSOLM.

**Unified regression models.** We developed three different polynomial regression models:  $K_{eq}(C,\alpha)$  with default  $\beta$  value ( $=1$ ),  $K_{eq}(C,\beta)$  with default  $\alpha$  value ( $=2$ ), and  $K_{eq}(C,\alpha,\beta)$ . To find the best-fit degree for each model, we applied leave-one-out cross validation to models of varying degree derived from the simulation data. For  $K_{eq}(C,\alpha)$ , a total of 80 different simulation data points was collected, making 11<sup>th</sup> degree the highest possible before the number of regression coefficients exceeds the number of data points. For  $K_{eq}(C,\beta)$ , a total of 49 different simulation data points was collected, making 8<sup>th</sup> degree the highest possible. For  $K_{eq}(C,\alpha,\beta)$ , a total of 122 different simulation data points was collected, making 7<sup>th</sup> degree the maximum possible. Figure 3 shows the root mean square errors of the leave-one-out cross validation for  $K_{eq}(C,\alpha)$ ,  $K_{eq}(C,\beta)$ , and  $K_{eq}(C,\alpha,\beta)$ , respectively. Sixth degree polynomials were selected as the best-fit models after cross validation. The resulting best-fit regression models for  $K_{eq}(C,\alpha)$ ,  $K_{eq}(C,\beta)$ , and  $K_{eq}(C,\alpha,\beta)$  are provided in equations (1–3).

$$\begin{aligned}
 K_{eq}(C,\alpha) &= 10^{-15} \times [(-1.441) + (1.969C + 3.751\alpha) \\
 &+ (-7.5C\alpha + 17.326C^2 - 3.712\alpha^2) \\
 &+ (4.204C\alpha^2 + 24.718C^2\alpha - 163.106C^3 + 1.943\alpha^3) \\
 &+ (-1.029C\alpha^3 - 13.071C^2\alpha^2 - 31.65C^3\alpha + 572.072C^4 - 0.566\alpha^4) \quad (1) \\
 &+ (0.217C\alpha^4 - 0.048C^2\alpha^3 + 44.2C^3\alpha^2 - 167.569C^4\alpha - 533.094C^5 \\
 &+ 0.085\alpha^5) + (-0.044C\alpha^5 + 0.801C^2\alpha^4 - 12.484C^3\alpha^3 \\
 &+ 54.042C^4\alpha^2 - 88.673C^5\alpha + 462.022C^6 - 0.005\alpha^6)] \\
 K_{eq}(C,\beta) &= 10^{-15} \times [(-0.99) + (12.39C + 4.43\beta) \\
 &+ (-22.25C\beta - 95.67C^2 - 10.55\beta^2) \\
 &+ (30.73C\beta^2 + 84.2C^2\beta + 424.36C^3 + 13.56\beta^3) \\
 &+ (-31.3C\beta^3 - 12.66C^2\beta^2 \\
 &- 345.58C^3\beta - 896.2C^4 - 9.3\beta^4) \quad (2) \\
 &+ (18.33C\beta^4 - 20.06C^2\beta^3 + 116.16C^3\beta^2 + 523.21C^4\beta + 852.66C^5 + 3.13\beta^5) \\
 &+ (-4.84C\beta^5 + 16.6C^2\beta^4 - 79.57C^3\beta^3 + 168.14C^4\beta^2 - 777.57C^5\beta \\
 &+ 119.35C^6 - 0.38\beta^6)] \\
 K_{eq}(C,\alpha,\beta) &= 10^{-15} \times [(-2.66) + (19.85C + 4.42\alpha + 2.79\beta) \\
 &+ (-111.51C^2 - 4.14\alpha^2 - 5.83\beta^2 - 14.42C\alpha - 21.44C\beta) \\
 &+ (370.09C^3 + 2.1\alpha^3 + 6.71\beta^3 + 7.21C\alpha^2 + 28.23C\beta^2 \\
 &+ 55.44C^2\alpha + 88.32C^2\beta) + (-566.94C^4 - 0.6\alpha^4 - 3.89\beta^4 \\
 &- 103.33C^3\alpha - 366.94C^3\beta - 1.78C\alpha^3 - 29.21C\beta^3 - 21.42C^2\alpha^2 \\
 &- 11.53C^2\beta^2) + (702.26C^5 + 0.09\alpha^5 + 0.91\beta^5 - 83.81C^4\alpha \\
 &+ 546.94C^4\beta + 0.33C\alpha^4 + 17.81C\beta^4 + 56.4C^3\alpha^2 + 126.84C^3\beta^2 \\
 &+ 0.95C^2\alpha^3 - 23.24C^2\beta^3) + (285.17C^6 - 0.005\alpha^6 - 0.008\beta^6 \\
 &- 137.49C^5\alpha - 777.57C^5\beta - 0.05C\alpha^5 - 4.84C\beta^5 + 50.16C^4\alpha^2 \\
 &+ 156.28C^4\beta^2 + 0.78C^2\alpha^4 + 17.39C^2\beta^4 - 13.57C^3\alpha^3 - 79.57C^3\beta^3)] \quad (3)
 \end{aligned}$$

Surface plots of  $K_{eq}$  values as functions of both parameters  $\alpha$  and  $C$  are shown in Fig. S1 and S2 for simulation data and for the best-fit polynomial regression models of degrees one through eleven. Similarly, the surface plots of  $K_{eq}$  values as functions of both parameters  $\beta$  and  $C$  are shown in Fig. S3 for simulation data and for the best-fit polynomial regression models of degrees one through eight. Finally, the unified regression models were built, which combined with previous regression model<sup>18</sup>, shown in Methods.



**Figure 3** | Least-squares error rates for leave-one-out cross validation test for different polynomial degree. (a)  $K_{eq}(\alpha, C)$ . (b)  $K_{eq}(\beta, C)$ . (c)  $K_{eq}(\alpha, \beta, C)$ .

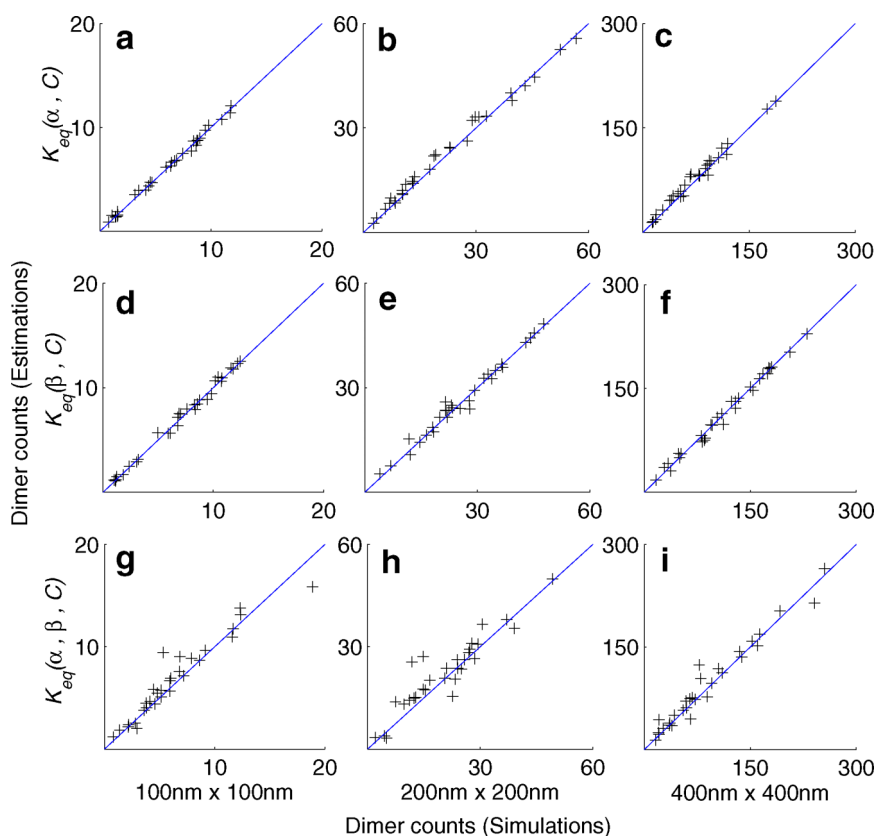


**Evaluation of the unified regression models.** To evaluate three different unified regression models in equations (5–7 in Methods below) with the best-fit sixth degree polynomials (1–3), we randomly selected 30 test cases for each model with fixed reactant concentration 0.1 and other parameters selected uniformly at random from the ranges  $C = 0.1 - 0.4$ ,  $B = 0.1 - 0.9$ ,  $M = 0.6 - 1.4\text{ns}$ ,  $D = 1.95 - 21.95 \times 10^{-11}\text{m}^2\text{s}^{-1}$ ,  $\alpha = 1.4 - 3.2$ , and  $\beta = 0.4 - 1.6$ . For each test case, we compared mean dimer counts from the simulation to the estimated dimer counts predicted by the unified regression models. Simulation values were averages of over 30 repetitions per data point. As an additional test of the performance of the model, we conducted another 30 random trials for each model for a  $200\text{nm} \times 200\text{nm}$  boundary and a  $400\text{nm} \times 400\text{nm}$  boundary in order to explore whether boundary effects significantly impact the fit of the regression model. We similarly compared these simulation results with estimates from the regression model. Simulation values were averages of over 5 repetitions per data point. Figure 4 shows the comparison between the simulated values and estimated values from our regression models using random parameter values from the full parameter space of each regression model. For the  $100\text{nm} \times 100\text{nm}$  cases in Fig. 4a, d, and g, our regression model closely matches the dimer counts derived from the particle simulations at the quasi-equilibrium state. The results suggest that one can use the regression model as a reliable and much faster replacement for the particle simulations for quantitatively predicting equilibrium constants over the parameter range examined. For the  $200\text{nm} \times 200\text{nm}$  cases in Fig. 4b, e, and h, and the  $400\text{nm} \times 400\text{nm}$  cases in Fig. 4c, f, and i, simulations and regression predictions still match very well, although the total particle number increases by four and sixteen

times respectively, suggesting that boundary effects have minimal influence even for these relatively small simulations. The five-parameter regression models ( $K_{eq}(C, B, M, D, \alpha)$  with  $\beta = 1$  in Fig. 4a, b, and c,  $K_{eq}(C, B, M, D, \beta)$  with  $\alpha = 2$  in Fig. 4d, e, and f) accurately predicted the simulation dimer counts for simultaneous variation across the range of parameters previously examined singly or in pairs. The correspondence between the regression models and full simulations was somewhat worse for the full six parameter regression model ( $K_{eq}(C, B, M, D, \alpha, \beta)$  in Fig. 4g, h and i). Overall though, the correlation is strong for the vast majority of the cases, which supports the equivalency of the regression and simulation predictions over a broad range of biologically relevant values for all parameters.

**Comparison with scaled particle theory.** The influence of non-ideal interactions on chemical reactions can be estimated by thermodynamic theories<sup>9,20</sup>. The excluded volume effect that becomes significant under conditions of high molecular crowding is one of critical non-ideal interactions. A correction factor to the equilibrium constant for excluded volume can be approximately calculated, yielding a corrected activity coefficient of reactants and products:  $K_{eq} = \Gamma_{exc} K^o$ , where  $K_{eq}$  is an apparent equilibrium constant,  $K^o$  is the equilibrium constant in the ideal state, and  $\Gamma_{exc}$  is the correction factor for the exclusion effect from interactions among particles in the reaction system<sup>9,20</sup>.

To validate our model, we calculated the correction factor and apparent equilibrium constant for various parameter conditions for the unified regression model in Eq. (3), 2DSOLM simulation, and scaled particle theory in two dimensional space (2DSPT)<sup>19,20</sup>. We first ran additional simulations at an uncrowded 1% concentration case



**Figure 4** | Comparison of dimer counts obtained from regression estimates vs. simulations. A data point on the diagonal line would indicate perfect agreement between the two values while points above the line show overestimates of simulation values and points below the line underestimates. (a)  $K_{eq}(\alpha, C)$  in  $100\text{nm} \times 100\text{nm}$ . (b)  $K_{eq}(\alpha, C)$  in  $200\text{nm} \times 200\text{nm}$ . (c)  $K_{eq}(\alpha, C)$  in  $400\text{nm} \times 400\text{nm}$ . (d)  $K_{eq}(\beta, C)$  in  $100\text{nm} \times 100\text{nm}$ . (e)  $K_{eq}(\beta, C)$  in  $200\text{nm} \times 200\text{nm}$ . (f)  $K_{eq}(\beta, C)$  in  $400\text{nm} \times 400\text{nm}$ . (g)  $K_{eq}(\alpha, \beta, C)$  in  $100\text{nm} \times 100\text{nm}$ . (h)  $K_{eq}(\alpha, \beta, C)$  in  $200\text{nm} \times 200\text{nm}$ . (i)  $K_{eq}(\alpha, \beta, C)$  in  $400\text{nm} \times 400\text{nm}$ . The concentration of reactants is fixed at 0.1 and the concentration of inert crowding agents is randomly selected from 0.0 to 0.3. The other parameters are uniformly randomly selected, with  $B$  from 0.1–0.9,  $M$  from 0.6–1.4ns,  $D$  from  $1.95$ – $21.95 \times 10^{-11}\text{m}^2\text{s}^{-1}$ ,  $\alpha$  from 1.4–3.2, and  $\beta$  from 0.4–1.6.



**Table 1** |  $K^o$  [counts $^{-1}$ m $^2$ ] for two different parameters ( $M$  and  $\alpha$ ) from simulation results of 1% concentration, pure reactants. All the other parameter values are set to default values ( $B=0.7$ ,  $D=6.95\times 10^{-11}$ m $^2$ s $^{-1}$ ,  $\beta=1$ , and  $d_{th}=0.5$ nm). The data was collected at 5 time points (5, 10, 15, 20, 25 $\mu$ s) for 30 independent runs in a 500nm  $\times$  500nm simulation area.

$M$	$\alpha$		
	1.4	2.0	2.6
0.1ns	2.70e-18	3.43e-18	1.97e-18
1ns	2.55e-17	2.46e-17	2.07e-17
10ns	2.57e-16	2.53e-16	2.53e-16

**Table 2** |  $K^o$  [counts $^{-1}$ m $^2$ ] for different  $d_{th}$  parameters from simulation results of 1% concentration, pure reactants. All the other parameter values are set to default values ( $B=0.7$ ,  $M=1$ ns,  $D=6.95\times 10^{-11}$ m $^2$ s $^{-1}$ ,  $\alpha=2$ , and  $\beta=1$ ). The data was collected at 5 time points (5, 10, 15, 20, 25 $\mu$ s) for 30 independent runs in a 500nm  $\times$  500nm simulation area.

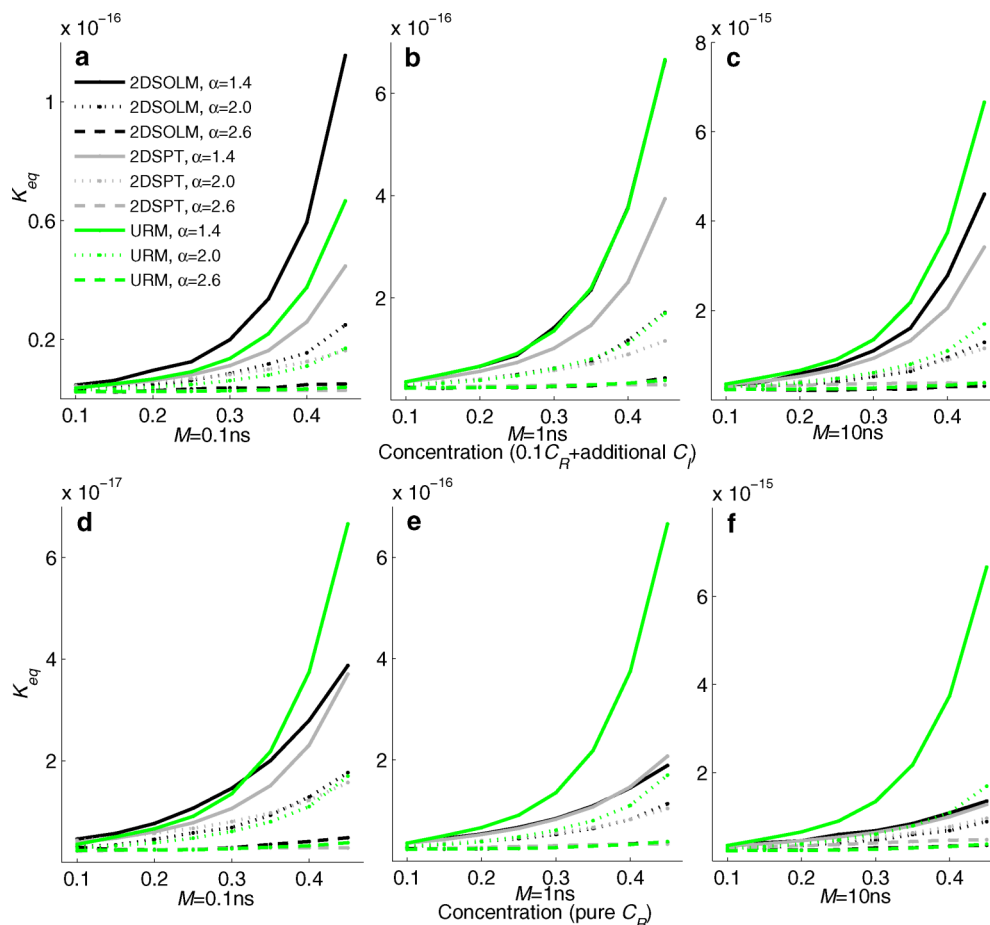
$d_{th}=0.5$ nm	0.25nm	0.125nm
2.46e-17	4.94e-17	9.37e-17

( $C=0.01$ , pure reactants case) with default parameter values of  $B=0.7$ ,  $M=1$ ns,  $D=6.95\times 10^{-11}$ m $^2$ s $^{-1}$ ,  $\alpha=2$ ,  $\beta=1$ , and  $d_{th}=0.5$ nm. We assumed that the interaction among particles at 1% concentration can be reasonably ignored, allowing us to estimate an uncorrected equilibrium constant at the ideal state ( $K^o$ ) from the simulation results. We also ran simulations for three  $\alpha$  values (1.4, 2.0, 2.6) and three  $M$  values (0.1, 1, 10ns). In addition, we ran 1% concentration simulations for three  $d_{th}$  values (0.5, 0.25, 0.125nm). Each of these simulations was used to calculate a  $K^o$  for the appropriate condition. Tables 1 and 2 show the calculated  $K^o$  from these simulation results.

We then estimated apparent equilibrium constants as a function of total concentration for each set of parameters examined above. For tests of variation in  $\alpha$  and  $M$ , we examined two different situations of crowding induced by 0.1 concentration of reactants ( $C_R$ ) plus inert

crowding agents to yield the eight total initial concentrations: 0.1, 0.15, 0.2, 0.25, 0.3, 0.35, 0.4, 0.45. As a secondary validation of the model outside the training conditions, we examined crowding induced purely by reactant particles at the same eight initial concentrations: 0.1, 0.15, 0.2, 0.25, 0.3, 0.35, 0.4, 0.45. In both cases, apparent equilibrium constants for SOLM were calculated from simulations and the apparent equilibrium constants for unified regression model (URM) were estimated from Eq. (7). The apparent equilibrium constants from scaled particle theory (SPT) were calculated using calculated  $K^o$  from SOLM simulation of 1% pure reactant case and estimated correction factors based on the equations in 2DSPT for given parameters (concentration of reactants ( $C_R$ ), concentration of inert crowding agents ( $C_I$ ),  $\alpha$ , and  $\beta$ )<sup>19,20</sup>.

Figure 5 compares apparent equilibrium constants for crowding induced by adding inert crowding agents to fixed 0.1  $C_R$  and for pure reactants for varying  $M$  and  $\alpha$ . URM, SOLM, and SPT all show an apparent equilibrium constant that nonlinearly increases with more



**Figure 5** | Apparent equilibrium constants for SOLM, URM, and SPT for two different parameters ( $M$  and  $\alpha$ ). (a)  $M=0.1$ ns. (b)  $M=1$ ns. (c)  $M=10$ ns for 0.1 concentration of reactants with additional inert crowding agents. (d)  $M=0.1$ ns. (e)  $M=1$ ns. (f)  $M=10$ ns for pure reactant simulations. All other parameter values are set to default values ( $B=0.7$ ,  $D=6.95\times 10^{-11}$ m $^2$ s $^{-1}$ ,  $\beta=1$ , and  $d_{th}=0.5$ nm).



compact dimers and higher total concentrations. In addition, the apparent equilibrium constant for URM, SOLM, and SPT increases as the mean time of dissociation events increases from 0.1 ns to 10 ns. The test case for the exclusion effect mainly from the inert crowding agent, shown in Fig. 5a,b,c, showed that URM more closely tracks the SOLM results than does SPT, consistent with the notion that the regression model is better able to mimic the more complicated interaction patterns found in the full simulations than are the more simplified assumptions of the SPT model. However, tests involving high crowding induced by the reactant particles themselves rather than an inert crowding agent, shown in Fig. 5d,e,f, showed that SPT predicts the apparent equilibrium constant more closely to SOLM simulation results than URM does, especially in cases involving more compact dimers. These latter results demonstrate that the ability of the regression model to accurately capture the behavior of the particle simulation can break down outside the parameter domain to which it was trained.

As a further demonstration of the utility of the regression approach, we added an extension to the model to account for an additional parameter,  $d_{th}$ , a threshold distance used by the simulator for identifying possible reaction events among nearby particles, as well as by the Green's function reaction dynamics (GFRD) algorithm to trigger new sampling of particle positions<sup>16,17</sup>. This parameter can be considered to roughly capture the difference between an interaction radius at which two particles can come close enough to influence and a collision radius at which they would actually sterically interfere with one another. Figure 6 shows the apparent equilibrium constant for fixed 0.1  $C_R$  with added inert crowding agents for three different  $\beta$  values and three different  $d_{th}$  values. The parameter effect of the threshold distance ( $d_{th}$ ) is nonlinear and inseparable from other parameters that influence crowding, particularly total concentration ( $C$ ). Because the URM was not trained to model  $d_{th}$ , we added an additional multiplicative and separable term to the URM by performing a regression fit to a quadratic function in  $C$  and  $d_{th}$ . The resulting term

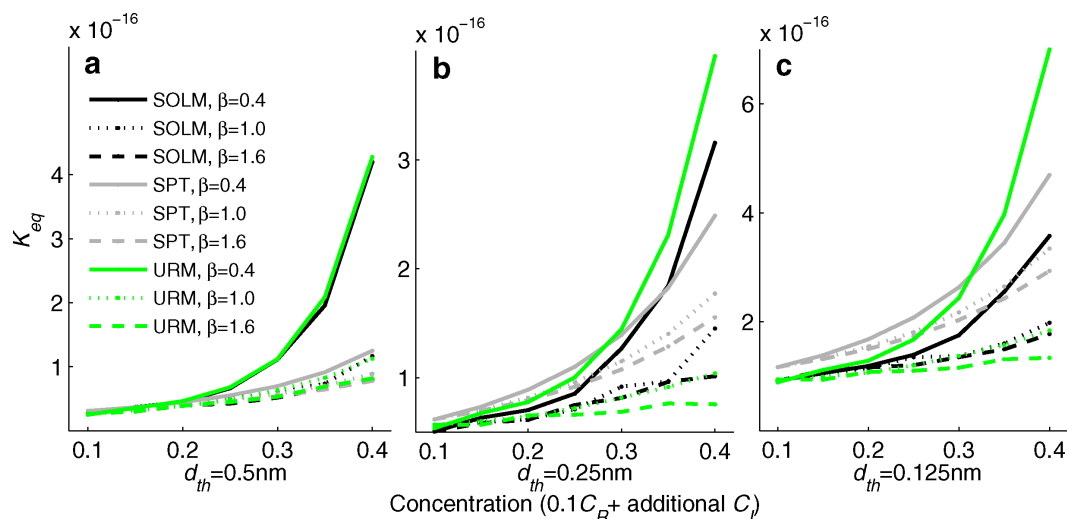
$$(5.72 - 6.1C - 18.03d_{th} + 15.09Cd_{th} - 2.24C^2 + 16.4d_{th}^2)$$

provides the best quadratic fit for the given parameter range using a leave-one-out cross-validation. The figure plots predicted  $K_{eq}$  for SOLM simulations, SPT corrections to low-concentration simulations, and the URM extended with the above  $C$  and  $d_{th}$  cross-term. There is no explicit SPT correction available for this parameter, although the SPT results do reflect the change in the

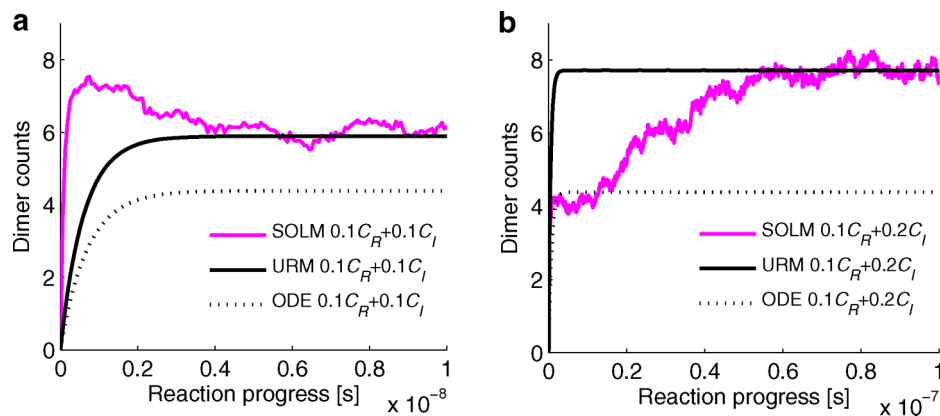
low-concentration equilibrium  $K^0$  induced by changes in  $d_{th}$ . All three models show similar non-linear behavior at the default  $d_{th}$  of 0.5 nm. At  $d_{th} = 0.25$  nm, the extended URM provides a noticeably closer fit to the full particle simulations than does SPT, showing that even a crude and essentially physically naive regression model can reasonably capture the behavior of the full simulations. At  $d_{th} = 0.125$  nm, results are more mixed, with the extended URM showing a better fit for much of the range of  $\beta$  and  $C$ , although eventually breaking down at the most extreme conditions of high concentration and small crowding agents. The results thus demonstrate the ability of regression modeling to capture a physical effect for which we lack a good analytical model, although also the tendency of such regression models to break down as one approaches the limits of the parameter range to which they were trained.

### Incorporation of URM corrections into ordinary differential equation (ODE) simulations.

The major goal of this work is to develop fast but versatile corrections for crowding effects that can be incorporated into more complex simulation models. To that end, we conducted a simple demonstration of how the URM models can be used in building efficient crowding corrections into fast crowding-free simulations. We simulated a homodimerization reaction, comparing SOLM particle simulations of the reaction, standard mass-action ODE simulation without crowding corrections, and a crowding-corrected ODE using our unified regression model (URM). Each SOLM particle simulation used default parameter values ( $B=0.7$ ,  $M=1$  ns,  $D=6.95 \times 10^{-11} \text{ m}^2 \text{ s}^{-1}$ , and  $\alpha=2$ ,  $\beta=1$ ,  $d_{th}=0.5$  nm) and was repeated 30 times with the average of the 30 runs used at each time point for visualization. We repeated each simulation using an ODE model of the system lacking correction for any crowding effects, deriving a dissociation rate  $k_{off}$  from the mean dissociation time  $M$  and an association rate  $k_{on}$  from  $k_{off}$  and the estimated idealized equilibrium constant  $K^0$ . We then repeated each simulation using an ODE with  $k_{off}$  again derived from  $M$  but  $k_{on}$  derived from  $M$  and the URM-predicted equilibrium constant  $K_{eq}(C, B, M, D, \alpha, \beta)$  in Eq. (7). More details of the calculations are provided in Methods. Two sets of simulations were conducted, one using 0.2 total concentration (0.1  $C_R$  + 0.1  $C_I$ ) and one using 0.3 total concentration (0.1  $C_R$  + 0.2  $C_I$ ). All simulations were run for 1  $\mu$ s of simulation time. Figure 7 shows results for  $C=0.2$  (Fig. 7a) and  $C=0.3$  (Fig. 7b), highlighting 0.01  $\mu$ s for  $C=0.2$  and 0.1  $\mu$ s for 0.3, sufficient to display both transient and pseudo-equilibrium behavior in each case. While both ODE models substantially deviate from the



**Figure 6** | Apparent equilibrium constants for SOLM, URM, and SPT for two different parameters ( $\beta$  and  $d_{th}$ ). (a)  $d_{th} = 0.5$  nm. (b)  $d_{th} = 0.25$  nm. (c)  $d_{th} = 0.125$  nm for 0.1 concentration of reactants with additional inert crowding agents. All the other parameter values are set to default values ( $B=0.7$ ,  $M=1$  ns,  $D=6.95 \times 10^{-11} \text{ m}^2 \text{ s}^{-1}$ , and  $\alpha=2$ ).



**Figure 7 | Demonstration of URM as a method for adjusting for crowding effects in simulations.** Each plot shows a comparison of a full SOLM model, an ODE model ignoring crowding effects, and an ODE model using the URM to adjust for crowding in the simulation. Each SOLM simulation uses default parameter values ( $B=0.7$ ,  $M=1\text{ns}$ ,  $D=6.95\times 10^{-11}\text{m}^2\text{s}^{-1}$ , and  $\alpha=2$ ,  $\beta=1$ ,  $d_{th}=0.5\text{nm}$ ). Each SOLM curve shows an average over 30 repetitions, with error bars omitted for clarity. ODEs are simulated using estimated rate parameters from the quasi-ideal equilibrium constant in table 1 and default parameter value of  $M$ . URM simulations are simulated using the default value of  $M$  to derive reverse rates and URM estimates of  $k_{eq}$  to determine forward rates for the given  $M$ . (a) total concentration 0.2 ( $0.1 C_R + 0.1 C_I$ ). (b) total concentration 0.3 ( $0.1 C_R + 0.2 C_I$ )

full particle model in the early transient phase, the URM-corrected model adjusts to yield an equilibrium value consistent with the full SOLM particle model while the uncorrected ODE substantially understates the equilibrium produced by the full particle model. Table 3 shows run times for these sample simulations, as well as larger simulations quadrupling the simulation space and particle counts. As the table shows, both the crowding-naive and URM-corrected ODE models show comparable run times to one another, essentially independent of the scale of the problem. The SOLM particle simulations, however, produce substantially longer run times than those of either ODE model in both conditions, and show approximately a 16-fold increase in run time as the system size is scaled four times. This quadratic increase in problem size for the particle methods is expected due to an approximately linear increase in work per event and linear decrease in mean waiting time between events with increasing problem size<sup>16</sup>.

## Discussion

We have demonstrated that it is possible to establish a unified regression model of binding equilibrium in crowded media that can in principle incorporate any physical parameters one can model in a full particle simulation, even when those parameters exhibit cross-dependencies for which we lack any predictive analytical theory. We have shown that this approach can accurately mimic a computationally costly particle model over a broad range of several such parameters. We have in the process shown the feasibility of a strategy for

quantitative simulation of reaction chemistry in crowded media that will maintain comparable accuracy to detailed particle simulations while enabling the use of much faster simulation methods. Our model is consistent with previous theoretical studies<sup>20</sup> in finding that total concentration ( $C$ ) and size ratio parameters ( $\alpha$ ,  $\beta$ ) act non-linearly to affect the binding equilibrium constant, and further captures both the cross-dependency of three key parameters ( $C$ ,  $\alpha$ ,  $\beta$ ) on the model reaction system and the independent parameter effects of three other parameters ( $B$ ,  $M$ ,  $D$ ). The correction factors and apparent equilibrium constants derived from our regression model generally match well with those derived from scaled particle theory, demonstrating that our model captures the crowding effect for various conditions of  $C$ ,  $\alpha$ , and  $\beta$ , reasonably. However, because the apparent equilibria and corrections of our regression model are based on empirical measurements from particle simulations, the approach will extend in a straightforward manner to a much broader range of parameters. We have further demonstrated how one can use this regression model as a fast correction for crowding effects in efficient ODE simulations. While the results of this corrected ODE model do not capture kinetics of the crowded system as well as the true particle simulation, they nonetheless provide an alternative that accurately captures the effects on system equilibrium of the full particle simulations while exhibiting run times comparable that of a simple ODE.

The principal alternatives for modeling chemistry in crowded conditions are the direct use of detailed particle models, similar to our SOLM but potentially far more involved in the range of parameters and details of the physical model, and scaled particle theory, which uses simpler models for which one can analytically calculate the contribution of the excluded volume event to reaction equilibria. The regression approach is intended to balance the competing benefits of these two alternatives: particle models in allowing one to explore effects of many parameters with potentially complicated cross-dependencies and SPT in allowing one to quickly compute corrections for crowding. Particle models are computationally costly, making them unsuitable for extremely large systems, slow reactions, or very complicated models. SPT's major limitation is that it depends on one's ability to derive analytical models for the specific parameters or phenomena one wishes to consider. SPTs to date make various assumptions that one may wish to relax in practice, and such corrections may prove difficult to analyze. For example, both SPT and SOLM conventionally assume that individual particles are hard spheres in 3D and hard circles in 2D. This assumption has been

**Table 3 | Run time comparison for SOLM particle simulations, ODE simulations using URM corrections for crowding, and uncorrected ODE simulations.** The table reports run times in seconds for  $1\mu\text{s}$  simulations of dimerization reactions using default parameter values for each of the three models and for two system sizes and two crowding levels.

$1\mu\text{s}$		$C=0.2$	$C=0.3$
		$(0.1 C_R + 0.1 C_I)$	$(0.1 C_R + 0.2 C_I)$
$100\text{nm} \times 100\text{nm}$	SOLM	3.13	10.33
	URM	0.36	0.38
	ODE	0.3	0.3
$200\text{nm} \times 200\text{nm}$	SOLM	46.82	163.61
	URM	0.35	0.38
	ODE	0.31	0.29





relaxed in various ways in the SPT literature. Rajagopalan's group extended from the hard-sphere interaction model of SPT to the nonbounded square-well interaction model<sup>23</sup>, and Mittal's group examined both repulsive interaction and attractive interaction of crowding particles<sup>24,25</sup>. In addition, Minton extended from a simple hard sphere model in SPT to two different particle models of a querying protein: Gaussian cloud and equivalent hard sphere to examine the excluded volume effect on more realistic conditions<sup>26</sup>. These are not trivial extensions, however, and are not easily made by users of the technology who are not experts in the underlying theory. By contrast, the results above demonstrate how a comparable extension of the regression approach can be accomplished with no major conceptual changes to account for varying effects of an extra parameter,  $d_{th}$ , that effectively controls a difference between the collision radius and interaction radius of reactant particles. While the essentially naive modeling that goes into such regression fitting has its own limitations and cannot replace good theory when that is available, it does provide a serviceable substitute for theory for systems too complex to allow for analytical solutions. Our regression approach thus provides a key step towards computationally practical simulations of complex reaction systems in crowded media and thus towards realistic models of intracellular biochemistry.

As our demonstration example of URM-corrected ODEs shows, the model exhibits essentially the same thermodynamic behavior as the full particle simulations. However, the corrected ODEs exhibit very different kinetics from the full particle simulations, more comparable to those of the ODE. We attribute this inaccuracy in transient behavior to the fact that the URM at present is fit only to thermodynamic data and thus leaves the rate constants underdetermined. The regression approach can in principle be extended to fit rate constants, not simply equilibrium constants, potentially yielding fast corrections to both kinetics and equilibrium but future work will be needed to explore the effectiveness of such regression corrections on kinetics.

Even for thermodynamics, the equilibrium model is not a perfect representation of the system and additional challenges may be created by extending it to more parameters or unexplored regions of the parameter space. The accuracy of full six parameter model is decreased relative to that of the five-parameter models, which suggests that more accurate accounting of cross-effects may be needed. This observation may derive in part from our decision to omit cross  $\alpha\beta$  terms in the regression equation to reduce the number of regression parameters to a feasible level. It has previously been established that concentration acts nonlinearly on binding chemistry<sup>9,14,20,23</sup>. We can further conclude that allowing for cross-dependent parameters does require a higher order model of  $C$  than the third-order model<sup>18</sup> that proved sufficient when capturing only the effects of separable parameters. Further increasing model complexity may be required to handle a richer model incorporating physical factors neglected in the present model.

It is also important to note that although the motivation behind this work is to improve models of chemistry in the intracellular environment, the model itself is a highly simplified view of binding chemistry in crowded media. The true intracellular environment is far more complicated than our simple reaction model, containing various mixtures of different sizes of proteins and other molecules that are often irregularly shaped and involved in many forms of macromolecular complex or other complicated interaction patterns. Recent simulation work in the area has shown the value of far more complicated physical models, e.g., in the recent work of Kim and Yethiraj using Brownian dynamics to explore effects of crowding on ligand binding to membrane-bound receptors.<sup>27</sup> While our work has not been extended to such detailed models, it nonetheless shows that it is possible to produce a simple unified mathematical model that can reliably reproduce the results of qualitatively similar computationally costly particle simulations for variations in several key

parameters of such a system. Nonetheless, one would need to account for many other parameters to produce a quantitatively predictive model of a realistic intracellular medium or any particular reaction system within it. While no method, including ours, can yet approach that goal, the present work provides a new approach toward more complex simulations at any degree of abstraction. Another obvious extension of the model is to three dimensional systems. We chose to establish the regression approach on two dimensional systems primarily because it provides a simpler and more computationally tractable framework in which to validate and demonstrate the proposed regression strategy. Two dimensional crowding does nonetheless have important applications in modeling association of membrane bound proteins<sup>20</sup>, for example in helping to explain observed patterns of clustering in the syntaxin-1 system<sup>28</sup>. While there are no appreciable conceptual obstacles to adapting the present methods to three dimensions, it remains to be shown that accuracy of the model would be comparable to that in 2D and that the increased computational cost of running particle simulations for training the regression models would not be prohibitive.

In the future, we hope to expand on the general idea of fitting regression models to increasingly detailed particle simulations in order to develop more realistic and complicated reaction models suitable for producing quantitatively predictive simulations of reaction chemistry *in vivo*. A similar approach may prove valuable to a diversity of fields that similarly depend on computationally costly particle models, including various disciplines of biology, physics, chemistry, and engineering.

## Methods

**2DSOLM.** Our previously developed two dimensional stochastic off-lattice model (2DSOLM)<sup>16</sup> was used to conduct all particle simulations described in the paper. A detailed description of the simulation program and its underlying algorithms is provided in our prior work<sup>16</sup>. To effectively simulate some high concentration cases, a modified initialization protocol was used, as described in our prior research<sup>18</sup>. Initially, all reactants are monomers. Initialization of particle positions is performed by establishing a grid of potential particle positions at the maximum possible packing density for whichever of reactant monomers and crowding agents occupies the larger total area and then randomly inserting particles into the corresponding grid positions. This protocol was developed because it makes it possible to initialize in highly crowded conditions where independent uniform placement of particles would usually result in overlapping particles, such as for high concentration cases. The simulation program of 2DSOLM was implemented by C++ programming language and run on a Linux Beowulf cluster. The collected data files were analyzed and plotted using Matlab (R2008a). Figures 2, 3, 4, 5, 6, 7, S1, S2, S3 were generated using Matlab. Simulation snapshots were plotted using Matlab and then simulation movies were made using Adobe Photoshop by concatenating these snapshots. The simulation code for 2DSOLM was released at the web site: <http://www.cs.cmu.edu/~russells/projects/crowding/crowding.html>.

The specific simulations run and their parameters are as described in the Results.

**Calculating thermodynamic equilibrium constants by SPT.** In calculating the thermodynamic correction factor ( $\Gamma_{exc} = K_{eq}/K^0$ ), we first assumed that 1% pure reactant case can be reasonably considered as an ideal state at which interaction among particles in such a diluted condition will be negligible<sup>20</sup>. For example, the calculated correction factor of 1% pure reactant case using simulation data in the default parameter value is 1.02, which is reasonably closed to the ideal state ( $\Gamma_{exc} = 1$ ). The simulation area was set to 500nm  $\times$  500nm for this condition to yield a large enough number of reactants to get accurate measurements for the given parameter conditions at 1% pure reactants case. The correction factor for SPT ( $\Gamma_{exc} = \frac{\gamma_M^2}{\gamma_D} = e^{2 \ln \gamma_M - \ln \gamma_D}$ , where  $\gamma_M$  and  $\gamma_D$  are activity coefficients of reactant monomer and dimer, respectively) was calculated by using the simulation data at 1% concentration to determine  $K^0$  followed by SPT analytical approximation to the activity coefficient under crowded conditions, as described in the prior literature on scaled particle theory in two dimensional space<sup>19,20</sup>. The density of reactant monomers and dimers under any given condition was calculated by the following equations<sup>18</sup>:

$$\rho_{D_{eq}} = \frac{4K_{eq}\rho_R + 1 - \sqrt{8K_{eq}\rho_R + 1}}{8K_{eq}},$$

$$2\rho_{D_{eq}} + \rho_{M_{eq}} = \rho_R,$$

where  $\rho_R$  = density of reactant monomers at initial state [particle counts/simulation area],  $\rho_{M_{eq}}$  = density of reactant monomers at quasi-equilibrium state in the ideal



state, and  $\rho_{D_{eq}}$  = density of reactant dimers at quasi-equilibrium state in the ideal state.  $K$  is then calculated by solving the equation  $K/\Gamma_{exc}(K) - K^0 = 0$  where

$$\Gamma_{exc} = \frac{\prod_i \gamma_i^{\alpha_i}}{\prod_i \gamma_i^{\beta_i}} = \frac{\gamma_M^2}{\gamma_D} = \frac{e^{2 \ln \gamma_M}}{e^{\ln \gamma_D}} = e^{2 \ln \gamma_M - \ln \gamma_D}$$

where activity coefficients are

$$\ln \gamma_M = -\ln(1-S_2) + \frac{2S_1 R_M}{1-S_2} + \frac{S_0 R_M^2}{1-S_2} + \frac{S_1^2 R_M^2}{(1-S_2)^2}$$

$$\ln \gamma_D = -\ln(1-S_2) + \frac{2S_1 R_D}{1-S_2} + \frac{S_0 R_D^2}{1-S_2} + \frac{S_1^2 R_D^2}{(1-S_2)^2}$$

$$S_0 = \pi \sum \rho_i = \pi(\rho_M + \rho_D + \rho_I)$$

$$S_1 = \pi \left( \sum \rho_i R_i \right) = \pi(\rho_M R_M + \rho_D R_D + \rho_I R_I)$$

$$S_2 = \pi \left( \sum \rho_i R_i^2 \right) = \pi(\rho_M R_M^2 + \rho_D R_D^2 + \rho_I R_I^2),$$

where  $\rho$  (density) = number of particles / simulation area,  $R$  = radius of a particle for each particle species:  $M$  (reactant monomer),  $D$  (reactant dimer),  $I$  (inert crowding particle)<sup>19,20</sup>. Finally, the apparent equilibrium constant is calculated by multiplying the correction factor to  $K^0$ , which is calculated from simulation results of 1% pure reactant cases for given parameter conditions.

**Simulation movie files.** We made two movie files to demonstrate the simulation process and show the effects of the two cross-dependent parameters. Each movie contains two different moving images, which are distinguished by different values for a single parameter while all other parameters use the common default values. Video S1 presents  $\alpha=1.4$  vs.  $\alpha=2.6$  for  $C=0.45$  and video S2 presents  $\beta=0.4$  vs.  $\beta=1.6$  for  $C=0.4$ . The first half of each movie shows the system in the initial (pre-equilibration) state, and the second half of the movie shows a quasi-equilibrium state. Movie files illustrating the simulation progress and the effects of each parameter individually were provided for the previous independent parameters<sup>18</sup>. High-resolution versions of the movies can be downloaded from: <http://www.cs.cmu.edu/~russells/projects/crowding/crowding.html>.

**Fitting  $K_{eq}$  to simulation results.** We calculated the equilibrium constant ( $K_{eq}$ ) of the homodimerization reaction by solving for  $K_{eq}$  as a function of the initial monomer concentration and mean steady-state dimer concentration. We then developed a regression model for  $K_{eq}$  as a function of the simulation parameters by extending our prior regression model of four parameters ( $C$ ,  $B$ ,  $M$ , and  $D$ ) up to six parameters ( $C$ ,  $B$ ,  $M$ ,  $D$ ,  $\alpha$ , and  $\beta$ ). First, we used our prior model for the contributions of the parameters  $B$ ,  $M$ , and  $D$ , which contribute linearly and independently to the equilibrium constant<sup>18</sup>, as shown in equation (4):

$$K_{eq}(B,M,D) = \left(\frac{B}{0.7}\right) \left(\frac{M}{1.0ns}\right) \left(\frac{D}{6.95 \times 10^{-11} m^2 s^{-1}}\right) \quad (4)$$

We then built regression models of  $K_{eq}(C,\alpha)$ ,  $K_{eq}(C,\beta)$ , and  $K_{eq}(C,\alpha,\beta)$  to include the parameter sets ( $\alpha$ ), ( $\beta$ ), and ( $\alpha,\beta$ ). Because the three parameters ( $C,\alpha,\beta$ ) are cross-dependent, they cannot each be separately incorporated into the regression model. Unified regression models for these three test cases are provided in equations (5-7):

$$K_{eq}(C,B,M,D,\alpha) = K_{eq}(C,\alpha)K_{eq}(B,M,D) \quad (5)$$

$$K_{eq}(C,B,M,D,\beta) = K_{eq}(C,\beta)K_{eq}(B,M,D) \quad (6)$$

$$K_{eq}(C,B,M,D,\alpha,\beta) = K_{eq}(C,\alpha,\beta)K_{eq}(B,M,D) \quad (7)$$

To find the best-matched regression model for each of the three cases, we performed least-squares polynomial regression for each parameter set, beginning with first-order regression models and increasing the degree until it was not possible to increase further without the number of regression parameters exceeding the number of data points. The general forms for the three regression models of  $K_{eq}(C,\alpha)$ ,  $K_{eq}(C,\beta)$ , and  $K_{eq}(C,\alpha,\beta)$  are as follows:

$$K_{eq}(C,\alpha) = a_0 + (a_1 C + a_2 \alpha) + (a_3 C \alpha + a_4 C^2 + a_5 \alpha^2) + \dots \quad (8)$$

$$K_{eq}(C,\beta) = a_0 + (a_1 C + a_2 \beta) + (a_3 C \beta + a_4 C^2 + a_5 \beta^2) + \dots \quad (9)$$

$$K_{eq}(C,\alpha,\beta) = a_0 + (a_1 C + a_2 \alpha + a_3 \beta) + (a_4 C \alpha + a_5 C \beta + a_6 C^2 + a_7 \alpha^2 + a_8 \beta^2) + \dots \quad (10)$$

Coefficients for the polynomials were calculated by the least-squares polynomial regression. For the polynomial  $K_{eq}(C,\alpha,\beta)$ , we neglected the cross  $\alpha\beta$  terms to reduce the number of coefficients of the model, on the hypothesis that these terms will have a negligible effect beyond the effects for which  $C\alpha$  and  $C\beta$  terms account. We selected the best-matched polynomial across degrees by minimizing the least squares error using a leave-one-out cross validation test for each of the three cases.

**Extension of the URM to account for threshold distance ( $d_{th}$ ).** An extension to the URM to account for effects of  $d_{th}$  was developed by adapting the same protocol from above to learn an additional multiplicative factor to the standard URM. The extension assumed that the previous full URM  $K_{eq}(C,\alpha,\beta)$  could be corrected to account for both the direct influence of  $d_{th}$  and the cross-dependence of  $d_{th}$  and  $C$  with a multiplicative term learned by polynomial regression on  $d_{th}$  and  $C$ . Simulations for three values of  $d_{th}$  (0.5, 0.25, 0.125 nm) and three values of  $\beta$  (0.4, 1.0, 1.6) were fit to a series of polynomial models of the form

$$\Gamma(C,d_{th}) = a_0 + (a_1 C + a_2 d_{th}) + (a_3 C d_{th} + a_4 C^2 + a_5 d_{th}^2) + L$$

Least-squares regression with leave-one-out cross-validation found a second order model to yield the lowest cross-validated error. These fits were then used to create an extended URM,  $K_{eq}(C,\alpha,\beta,d_{th}) = \Gamma(C,d_{th})K_{eq}(C,\alpha,\beta)$ , used to evaluate the ability of the regression approach to incorporate additional parameters for which no equivalent SPT theory is available.

**Use of URM as a correction to ODE simulations.** To demonstrate of the use of the URM as a means of quickly providing crowding corrections for fast space-free simulations, we conducted a set of demonstration simulations using the URM to adjust rates to a fast ODE model. We compared this URM-corrected ODE model to full particle simulations with SOLM and to crowding-naive ODEs. SOLM simulations were conducted using default parameter values ( $B=0.7$ ,  $M=1$  ns,  $D=6.95 \times 10^{-11} m^2 s^{-1}$ , and  $\alpha=2$ ,  $\beta=1$ ,  $d_{th}=0.5$  nm,  $100$  nm  $\times$   $100$  nm space) for  $0.1 C_R + 0.1 C_I$  and  $0.1 C_R + 0.2 C_I$ . Each simulation was repeated 30 times with the resulting values averaged at each time point for visualization. Comparable ODE models were instantiated by assuming the dissociation rate  $k_{off}$  is the inverse of dissociation time  $M$  and that association time  $k_{on}$  is related to the idealized equilibrium constant  $K^0$  by the formula

$$k_{on} = K^0 k_{off}$$

$K^0$  for each simulation was estimated for the idealized no-crowding condition from 1% concentration SOLM simulations, as in table 1. URM-corrected ODEs were derived using an ODE model with dissociation rate  $k_{off}$  again set to  $M^{-1}$  but  $k_{on}$  recalculated at each step of the ODE using the URM from the formula:

$$k_{on} = K_{eq}(C,B,M,D,\alpha,\beta)k_{off}$$

The model is thus corrected so as to yield URM-derived equilibrium values. Simulations were run for 1  $\mu$ s for each model and concentration. As an additional test of run time scaling of the model, we repeated each simulation under identical conditions but using a 200 nm  $\times$  200 nm simulation space, with particle counts scaled proportionately to maintain concentrations at  $C=0.2$  or  $C=0.3$ . Run times were recorded for the ode45 call for both ODE and URM simulations and for the full execution of the SOLM model for particle simulations.

Both ODE models were simulated with Matlab 7.7.0, R2008b using the ode45 embedded Runge-Kutta integration method. SOLM was implemented using C++ and run in GNU bash version 3.2.9 (1), release (x86-64-redhat-linux-gnu). All run time tests were run on a single Intel (R) Core (TM) 2 Quad CPU Q9550 @ 2.83GHz, 4GB RAM workstation using Linux Fedora release 7 (Moonshine).

- Minton, A. P. How can biochemical reactions within cells differ from those in test tubes? *J. Cell Sci.* **119**, 2863–2869 (2006).
- Tokuriki, N. *et al.* Protein folding by the effects of macromolecular crowding. *Protein Sci.* **13**, 125–133 (2004).
- Hatters, D. M., Minton, A. P. & Howlett, G. J. Macromolecular crowding accelerates amyloid formation by human apolipoprotein C-II. *J. Biol. Chem.* **277**, 7824–7830 (2002).
- Jiang, M. & Guo, Z. Effects of macromolecular crowding on the intrinsic catalytic efficiency and structure of enterobactin-specific isochorismate synthase. *J. Am. Chem. Soc.* **129**, 730–731 (2007).
- Grima, R. & Schnell, S. A systematic investigation of the rate laws valid in intracellular environments. *Biophys. Chem.* **124**, 1–10 (2006).
- Eide, J. L. & Chakraborty, A. K. Effects of quenched and annealed macromolecular crowding elements on a simple model for signaling in T lymphocytes. *J. Phys. Chem. B* **110**, 2318–2324 (2006).
- Pincus, D. L. & Thirumalai, D. Crowding effects on the mechanical stability and unfolding pathways of ubiquitin. *J. Phys. Chem. B* **113**, 359–368 (2009).
- Christiansen, A. *et al.* Factors defining effects of macromolecular crowding on protein stability: an in vitro/in silico case study using cytochrome *c*. *Biochemistry* **49**, 6519–6530 (2010).
- Zimmerman, S. B. & Minton, A. P. Macromolecular crowding: biochemical, biophysical, and physiological consequences. *Annu. Rev. Biophys. Biomol. Struct.* **22**, 27–65 (1993).
- Hall, D. & Minton, A. P. Macromolecular crowding: qualitative and semiquantitative successes, quantitative challenges. *Biochim. Biophys. Acta* **1649**, 127–139 (2003).
- LeDuc, P. R. & Schwartz, R. Computational models of molecular self-organization in cellular environments. *Cell Biochem. Biophys.* **48**, 16–31 (2007).
- Mayawala, K., Vlachos, D. G. & Edwards, J. S. Spatial modeling of dimerization reaction dynamics in the plasma membrane: Monte Carlo vs. continuum differential equations. *Biophys. Chem.* **121**, 194–208 (2006).



13. Puskar, K., Ta'asan, S., Schwartz, R. & LeDuc, P. R. Evaluating spatial constraints in cellular assembly processes using a Monte Carlo approach. *Cell Biochem. Biophys.* **45**, 195–201 (2006).
14. Elcock, A. H. Atomic-level observation of macromolecular crowding effects: escape of a protein from the GroEL cage. *Proc. Natl Acad. Sci. USA* **100**, 2340–2344 (2003).
15. Izvekov, S. & Voth, G. A. A multiscale coarse-graining method for biomolecular systems. *J. Phys. Chem. B* **109**, 2469–2473 (2005).
16. Lee, B., LeDuc, P. R. & Schwartz, R. Stochastic off-lattice modeling of molecular self-assembly in crowded environments by Green's function reaction dynamics. *Phys. Rev. E* **78**, 031911 (2008).
17. von Zon, J. S. & ten Wolde, P. R. Green's-function reaction dynamics: a particle-based approach for simulating biochemical networks in time and space. *J. Chem. Phys.* **123**, 234910 (2005).
18. Lee, B., LeDuc, P. R. & Schwartz, R. Parameter effects on binding chemistry in crowded media using a two-dimensional stochastic off-lattice model. *Phys. Rev. E* **80**, 041918 (2009).
19. Lebowitz, J. L., Helfand, E. & Praetgaard, E. Scaled particle theory of fluid mixtures. *J. Chem. Phys.* **43**, 774–779 (1965).
20. Grasberger, B., Minton, A. P., DeLisi, C. & Metzger, H. Interaction between proteins localized in membranes. *Proc. Natl. Acad. Sci. USA* **83**, 6258–6262 (1986).
21. Bicknese, S., Periasamy, N., Shohet, S. B. & Verkman, A. S. Cytoplasmic viscosity near the cell plasma membrane: measurement by evanescent field frequency-domain microfluorimetry. *Biophys. J.* **65**, 1272–1282 (1993).
22. Sengers, J. V. & Watson, J. T. R. Improved international formulations for the viscosity and thermal conductivity of water substance. *J. Phys. Chem. Ref. Data* **15**, 1291–1314 (1986).
23. Hu, Z., Jiang, J. & Rajagopalan, R. Effects of macromolecular crowding on biochemical reaction equilibria: a molecular thermodynamic perspective. *Biophys. J.* **93**, 1464–1473 (2007).
24. Kim, Y. C., Best, R. B. & Mittal, J. Macromolecular crowding effects on protein-protein binding affinity and specificity. *J. Chem. Phys.* **133**, 205101 (2010).
25. Rosen, J., Kim, Y. C. & Mittal, J. Modest protein-crowder attractive interactions can counteract enhancement of protein association by intermolecular excluded volume interactions. *J. Phys. Chem. B.* **115**, 2683–2689 (2011).
26. Minton, A. P. Models for excluded volume interaction between an unfolded protein and rigid macromolecular cosolutes: macromolecular crowding and protein stability revisited. *Biophys. J.* **88**, 971–985 (2005).
27. Kim, J. S. & Yethiraj, A. Crowding effects on association reactions at membranes. *Biophys. J.* **98**, 951–958 (2010).
28. Sieber, J. J. *et al.* Anatomy and dynamics of a supramolecular membrane protein cluster. *Science* **317**, 1072–1076 (2007).

## Acknowledgments

We thank Dr. Robert Murphy and the Center for Bioimage Informatics at Carnegie Mellon University for providing the use of their computer cluster for this research. This work was supported by National Institutes of Health NIAID award #1R01A1076318 (RS), the National Science Foundation (CMMI-0856187, CMMI-1013748) and the Office of Naval Research (N000140910215) (PRL).

## Author contributions

B.L., P.R.L. and R.S. designed the stochastic off-lattice model and the algorithms of the simulation program and participated in planning of simulation experiments. B.L. implemented and ran the simulation code and analyzed the simulation data. B.L., P.R.L. and R.S. wrote and reviewed the manuscript.

## Additional information

**Supplementary information** accompanies this paper at <http://www.nature.com/scientificreports>

**Competing financial interests:** The authors declare no competing financial interests.

**License:** This work is licensed under a Creative Commons Attribution-NonCommercial-ShareAlike 3.0 Unported License. To view a copy of this license, visit <http://creativecommons.org/licenses/by-nc-sa/3.0/>

**How to cite this article:** Lee, B., LeDuc, P.R. & Schwartz, R. Unified regression model of binding equilibria in crowded environments. *Sci. Rep.* **1**, 97; DOI:10.1038/srep00097 (2011).