Dynamics simulations for engineering macromolecular interactions

Avi Robinson-Mosher, Tamar Shinar, Pamela A. Silver, and Jeffrey Way

Citation: Chaos **23**, 025110 (2013); doi: 10.1063/1.4810915
View online: http://dx.doi.org/10.1063/1.4810915
View Table of Contents: http://chaos.aip.org/resource/1/CHAOEH/v23/i2
Published by the American Institute of Physics.

Dynamics simulations for engineering macromolecular interactions

Avi Robinson-Mosher,[1,2,a)] Tamar Shinar,[3,b)] Pamela A. Silver,[1,2,c)] and Jeffrey Way[1,d)]
[1]*Wyss Institute for Biologically Inspired Engineering, 3 Blackfan St., Boston, Massachusetts 02115, USA*
[2]*Department of Systems Biology, Harvard Medical School, 200 Longwood Avenue, Boston, Massachusetts 02115, USA*
[3]*Computer Science and Engineering Department, University of California, Riverside 900 University Ave., Riverside, California 92521, USA*

(Received 24 January 2013; accepted 22 May 2013; published online 12 June 2013)

The predictable engineering of well-behaved transcriptional circuits is a central goal of synthetic biology. The artificial attachment of promoters to transcription factor genes usually results in noisy or chaotic behaviors, and such systems are unlikely to be useful in practical applications. Natural transcriptional regulation relies extensively on protein-protein interactions to insure tightly controlled behavior, but such tight control has been elusive in engineered systems. To help engineer protein-protein interactions, we have developed a molecular dynamics simulation framework that simplifies features of proteins moving by constrained Brownian motion, with the goal of performing long simulations. The behavior of a simulated protein system is determined by summation of forces that include a Brownian force, a drag force, excluded volume constraints, relative position constraints, and binding constraints that relate to experimentally determined on-rates and off-rates for chosen protein elements in a system. Proteins are abstracted as spheres. Binding surfaces are defined radially within a protein. Peptide linkers are abstracted as small protein-like spheres with rigid connections. To address whether our framework could generate useful predictions, we simulated the behavior of an engineered fusion protein consisting of two 20 000 Da proteins attached by flexible glycine/serine-type linkers. The two protein elements remained closely associated, as if constrained by a random walk in three dimensions of the peptide linker, as opposed to showing a distribution of distances expected if movement were dominated by Brownian motion of the protein domains only. We also simulated the behavior of fluorescent proteins tethered by a linker of varying length, compared the predicted Förster resonance energy transfer with previous experimental observations, and obtained a good correspondence. Finally, we simulated the binding behavior of a fusion of two ligands that could simultaneously bind to distinct cell-surface receptors, and explored the landscape of linker lengths and stiffnesses that could enhance receptor binding of one ligand when the other ligand has already bound to its receptor, thus, addressing potential mechanisms for improving targeted signal transduction proteins. These specific results have implications for the design of targeted fusion proteins and artificial transcription factors involving fusion of natural domains. More broadly, the simulation framework described here could be extended to include more detailed system features such as non-spherical protein shapes and electrostatics, without requiring detailed, computationally expensive specifications. This framework should be useful in predicting behavior of engineered protein systems including binding and dissociation reactions.
[http://dx.doi.org/10.1063/1.4810915]

Synthetic biologists frequently take the elements of natural transcription systems and treat them as abstracted black boxes in which the protein elements are used as Nature provides them. Modeling the behavior of such systems typically considers the cell as a bag of genes with certain concentrations of DNA, RNA, and protein, and ignores the complexity of macromolecular movement in three dimensions. This approach limits the possible types of engineering; in particular, it is difficult to design novel three-dimensional protein assemblies that directly and indirectly regulate transcription itself or other biological processes. The goal of the present work is to create a tool that will simulate the movement of synthetic proteins consisting of natural domains connected via engineered linkers in order to make predictions about the interactions of the full system and offer insight into the effects of varying the domain combinations and linker properties.

I. INTRODUCTION

Predictable manipulation of transcriptional networks is a central goal of synthetic biology. Regulation of transcription often involves extensive protein-protein interactions, particularly for processes that are central to the well-being of the organism. For example, in eukaryotes, the decision to transcribe a given gene often follows from elaborate signal transduction pathways involving protein modification, protein-protein

[a)]avi.mosher@wyss.harvard.edu
[b)]shinar@cs.ucr.edu
[c)]pamela_silver@hms.harvard.edu
[d)]jeff.way@wyss.harvard.edu

interactions, and movement between cellular compartments, with positive and negative feedback loops that often result in an all-or-none response (Figure 1(a)). One aspect in the design of such pathways is that a gene is usually present in only one or two copies per cell, so stochasticity of transcription factor binding is a potential issue unless binding/non-binding state of the protein is determined by upstream signaling events.

In bacteria, transcriptional regulation that plays a key role in the physiology of the organism often involves elaborate protein-protein interactions that are spatially and quantitatively tuned to give a desired result. For example, lambda repressor binds to its operators in a highly cooperative manner that involves three distinct protein-protein interactions so that an octameric complex can form in the fully repressed state.[1,2] Another striking example is the Kai clock found in photosynthetic bacteria.[3] In this system, the KaiC protein goes through 24-h cycles of autophosphorylation and dephosphorylation, modulated by the KaiA and KaiB proteins. When these three proteins are placed in a test tube with adenosine triphosphate (ATP), the KaiC phosphorylation state oscillates with a 24-h period until the ATP runs out.[4] In the cell, the phosphorylation state of KaiC determines the transcription pattern of most of the genes in the genome, but this transcription follows from a simple pathway that reads the state of KaiC. Similarly, in higher eukaryotes, the key events in determining transcriptional patterns often take place at the cell surface and in the cytoplasm, and are essentially decided by the time a transcription factor enters the nucleus.

Signaling processes and transcription events often involve movement around flexible junctions within proteins. For example, lambda repressor consists of distinct N- and C-termini attached by a semi-flexible segment; in crystal structures, the N- and C-terminal domains adopt a number of distinct conformations depending on the oligomerization state. The extracellular domain of epidermal growth factor receptor (EGFR) undergoes a major conformational shift upon ligand binding, rotating its N-terminal 310 amino acids almost 180°.[5,6] Antibodies, which control transcriptional events such as the antibody-dependent cell-mediated cytotoxicity response and apoptotic events that lead to B cell elimination, contain a flexible hinge that limits the possible geometries of binding and signal transduction.

At a mathematical level, synthetic biology concerns itself with a number of phenomena that are amenable to computational investigation and modeling. These include engineering novel enzymatic activities, mechanical activities (such as transport or force application), compartmentalization (such as artificial carboxysomes or DNA nanostructures), assembly (such as biopolymer formation and property investigation), and protein interaction networks. (A phosphorylates B which inactivates C.) Differential equations are commonly used to model systems that can be characterized by their concentrations (for instance, protein interaction networks that are uniformly distributed in solution). Here, we concern ourselves with systems where this is not the case. Geometric factors may play a role in the non-equilibrium dynamics of a system (for instance, nucleated assembly or diffusion-limited rates of spread from an initial induction point), particularly with engineered structures that may violate the presumption (generally assumed for natural systems) that the system is in fact functional. Alternatively, the geometric properties of a system may be clearly important but it may not be obvious how to design our engineered structure to achieve it (for the case of DNA nanostructures a tool exists to do this, within certain constraints). Finally, geometry-driven modeling may give the synthetic biologist additional features to engineer, including new interaction surfaces, fusion attachment sites, linker length and stiffness, weak interactions, etc. In the course of evolution, Nature routinely varies these parameters to generate new systems. In contrast, synthetic biologists can

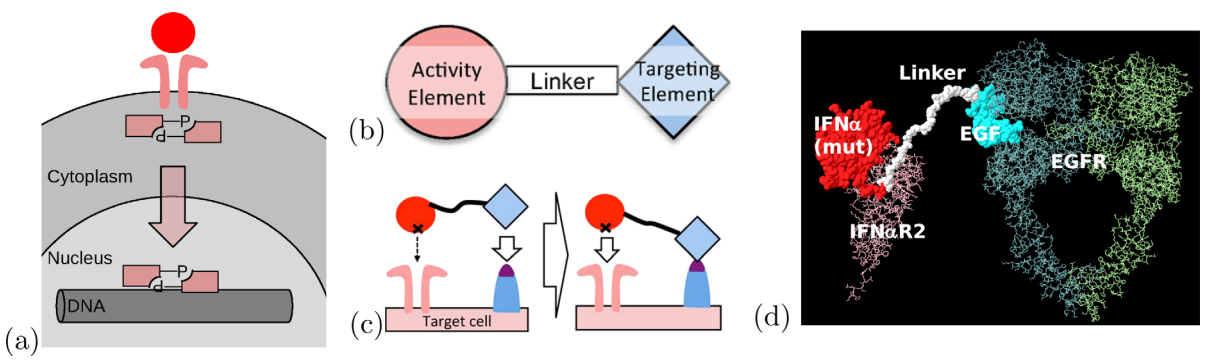

FIG. 1. Synthetic-biological intervention in signal transduction. (a) Mammalian signal transduction generally proceeds by an interaction of a protein ligand with a cell-surface receptor, leading to phosphorylation events on the cytoplasmic face of the cell surface, leading to changes in protein-protein interactions that are directly or indirectly transmitted into the nucleus. Protein-protein interactions occurring at the cell surface, in the cytoplasm, in the nucleus, and on the DNA all represent opportunities for engineering and present common kinetic and thermodynamic problems. (The figure vastly understates the complexity of most mammalian signal transduction.) (b) "Chimeric activators" are engineered signaling proteins with a logical AND function, requiring the presence of two arbitrarily chosen receptors to be present on a cell surface. The Targeting Element binds tightly to one receptor but does not generate the signal of interest. The Activity Element binds to and signals through a second receptor, and contains a mutation (X) that significantly reduces binding. (c) Signaling of the chimeric activator on a target cell is driven through initial binding of the targeting element, which has a much higher affinity for its receptor than the Activity Element. After this initial binding, the Activity Element is present in a high local concentration on the cell surface and can bind to its receptor despite the weakening mutation. (Binding to cells with only the receptor for the Activity Element is minimized due to the mutation.) (d) Spatial considerations in the construction of a specific chimeric activator. IFNα, which kills cells to which it binds, can be targeted with EGF to cancer cells that overproduce EGF receptor 7. To allow simultaneous binding to both receptors, a flexible linker of sufficient length must bridge the domains of the fusion protein. However, the set of conformations adopted in the free and EGFR-bound states (which will determine binding to IFNα receptor and subsequent signaling) is unknown.

rationally engineer linear sequences and adjust concentrations of proteins, but do not design new three-dimensional complexes or systems that change shape over time, in part because we lack the tools to do so.

One class of engineered signal transduction proteins are fusion proteins that bind to multiple receptors on a cell surface. In terms of engineering transcriptional circuits, these proteins are the most distant from DNA, but they embody some of the same protein engineering problems as, for example, engineering transcription factors with hybrid binding specificities. Such fusion proteins are easily studied because they can be simply added to cells, and might serve as therapeutics if properly engineered. Previously we described a class of such fusion proteins, termed "chimeric activators," which are designed to act as AND-gate elements in a protein-based logic system;[7,8] these proteins activate signaling only on cells with two distinct receptors. Chimeric activators consist of two protein ligands that are fused by a flexible linker long enough to allow simultaneous binding to the cell-surface receptors. Typically one ligand is an "Activity Element" that signals, and the other ligand binds to a cell-specific surface protein and provides a targeting function but does not signal (Figures 1(b) and 1(c)).

In constructing such a protein AND-gate, the fusion protein must activate signal transduction on a target cell with both receptors, but it should not activate signal transduction on cells with only one receptor. Simply adding a targeting element to a signaling protein will not prevent the protein from binding to its receptor on non-target cells. Specificity is achieved by introducing a mutation that significantly weakens the signaling element; it binds poorly to its receptor in isolation, but if the fusion protein first binds via its targeting element to the cell surface, the mutated activity element will be present in a high local concentration in the neighborhood of its receptor, and still able to bind. Introduction of the weakening mutation enhances cell type-specificity about 10–20 fold.[7,8] While this is useful, natural signaling proteins that use distinct targeting and signaling receptors show a much larger specificity enhancement. For example, the interleukin-6 (IL-6)-family cytokines leukemia inhibitory factor (LIF) and ciliary neurotrophic factor (CNTF) both signal through a heterodimeric LIFR and gp130 receptor complex. LIF requires only these two receptors, while CNTF requires a third, non-signaling receptor (CNTFR) that serves a targeting function and positions CNTF to bind to LIFR/gp130, but in the absence of CNTFR, CNTF binding to LIFR/gp130 is undetectable.[9] This observation illustrates that when the positioning of a ligand on a cell surface is optimized, subsequent receptor binding can be profoundly improved, and that the spatial aspect of chimeric activator function has not been addressed. The simulation tools described here are designed to do this.

The problem of quantitative and spatial optimization is likely to be important for engineering every aspect of transcriptional regulation. For example, in constructing artificial DNA-binding proteins to bind to arbitrarily chosen sequences, it is straightforward to fuse zinc-finger proteins[10–12] or TALE subunits[13–17] to match a given DNA target, but binding to non-target sequences needs to be minimized by quantitative tuning. In addition, actual binding to DNA may involve bending around the DNA helix, which may require either flexible attachment points to be introduced into the protein or use of multiple subunits that non-covalently assemble onto DNA. The simulation tools we propose here could also be used for designing novel DNA-binding complexes, or artificial proteins acting at any other step in signal transduction.

II. THE APPROACH

This work represents initial efforts in creating a general system for predicting the behavior of flexible, multidomain proteins as they move by constrained Brownian motion and bind to target proteins. Such a system may be useful for understanding natural proteins, and for designing genetically engineered proteins composed of natural protein domains joined at flexible junctions. Examples of natural flexible proteins include antibodies, the phage M13 gene 3 protein that binds to target cells, and fibronectin as well as signaling proteins that can exist in multiple conformational states such as EGFR and Src. Antibodies and gene 3 protein each need to bind in a multivalent manner to target proteins with unpredictable relative conformations; hence the need for flexibility.

The design considerations for engineered fusion proteins with flexible junctions between the domains are poorly understood. For example, if one wishes to direct a signaling protein to a particular cell type in the body, it is possible to fuse antibody V regions that bind to a cell-specific surface protein on the same cell. A secondary binding event of the signaling protein to its receptor could result in enhanced signaling relative to cells lacking the targeting receptor. In principle, this secondary binding event could be optimized by choosing the ideal binding constants of each protein domain for their targets, and the ideal length and flexibility of the linker connecting these proteins. These parameters can be readily manipulated by standard genetic engineering techniques. However, in practice, it is not currently known how to choose these parameters.

To better understand the movement and binding properties of flexible proteins, we are developing a simulation approach that builds on previous work.[18–23] Ultimately, we hope to have a system that simulates the behavior of engineered proteins with multiple binding surfaces, including Brownian motion, binding, and dissociation events. In the present work, we focus on modeling the linker relationship between protein domains which are subject to Brownian forces, varying linker stiffness and binding events (on-events).

The movement of proteins takes place in a highly viscous environment. The mathematical approach we took was to sum the forces acting on each protein element: a distribution of Brownian force calculated from the distribution of Brownian motions, a resistive force resulting from movement in a Newtonian fluid, forces resulting from attachment to other protein elements, attractive forces to lead to binding to specific surfaces, and a hard-sphere non-inertial repulsive force to prevent proteins from being driven into each other. Forces can then be calculated for an engineered protein with arbitrarily chosen domain sizes, linker lengths and flexibilities, and binding properties. In addition, it should be possible to run comparison simulations in which these parameters are varied in a way that corresponds to variant proteins that

could be made by genetic engineering—for instance, varying linker properties or binding parameters.

For protein-sized molecules that interact through specific surfaces, the diffusion-limited on-rate is about $10^6 M^{-1}s^{-1}$. On-rates for protein-protein interactions are generally in this range, but may be somewhat faster or slower depending on electrostatic charges, the Stokes radius of the protein, and whether the amino acid side chains in the binding interface must adopt an unusual conformation before binding. The off-rates of one protein bound to another may vary from $10^9 s^{-1}$ (unmeasurably fast) to less than $10^{-7}s^{-1}$ (i.e., the complex may be stable for several days). In most biological contexts, when an extremely stable complex forms, the off-rate is irrelevant and the binding event is terminated by another process such as proteolysis. In practice, a system that simulates protein binding interaction on a timescale up to a few minutes will be adequate for most purposes.

Minute timescales are difficult to reach for macromolecular systems, however. Molecular dynamics (MD) packages such as NAMD[24] and Anton[25] are limited to time steps on the order of several femtoseconds. All-atom MD simulations involving massive computational effort can reach the microsecond scale.[26] Efforts have been made to extend MD timescales by coarse grained (CG) representations of macromolecules[18,19]—grouping multiple atoms into units which are simulated together, reducing the number of variables in the system and potentially alleviating the restrictions on time step size (up to one to several orders of magnitude larger than all-atom MD). Both MD and CG approaches are impractical for the time scales we intend to investigate, offer levels of detail which are not necessary for decisions about protein engineering strategies, and could provide false precision in our results.

Brownian dynamics (BD) systems[23,27] are used for studying reaction-diffusion systems where interactions occur on a large time scale relative to the dynamics of the bodies, and generally represent individual bodies as hard spheres without an internal structure. Other methods attempting to reach long time scales for coarsely modeling reaction-diffusion systems include the Reaction Before Move[20] BD scheme which extends previous BD methods by analytically determining the probability of two particles colliding in order to avoid missing collisions during large steps. This is related to other event-driven BD schemes such as Green's function reaction dynamics[21,22] which attempt to find the next significant interaction in the system and jump forward to it in time. These methods perform comparatively well for sparse system where regular BD approaches are forced to take many steps where no significant interactions occur, but are less advantageous in dense systems where interactions occur frequently. Our method, described in Secs. III–V, attempts to preserve some of the advantages of these event-driven BD approaches while also modeling flexible connections within molecules.

III. CHOICE OF CONSTRAINED NON-INERTIAL DYNAMICS TO REPRESENT PROTEIN MOVEMENT AND BINDING

The preceding sections have described what we want to be able to do and why from a biological perspective. Here, we address from a mathematical perspective what is necessary to make those goals possible. To model a flexible protein in solution, we make a number of simplifying assumptions beyond those entailed by an all-atom molecular dynamics model. Most significantly we abstract full protein domains as single rigid spheres, ignoring effects due to differences in shape or intra-domain motion. We ignore the details of the solvent, characterizing it only by its viscosity and temperature. We assume that the correlation time of the velocities of the bodies in our system is sufficiently small that the dynamics can be well modeled by a constrained non-inertial dynamics integration scheme as described in Sec. IV and that the effects of temperature can be captured by stochastic Brownian forces on the bodies. The particle Reynolds number is $\mathrm{Re}_p = U_p D_p/\nu$, where U_p is a characteristic particle velocity, D_p is the particle diameter, and ν is the kinematic viscosity of the background fluid. In the problem under consideration, $D_p \approx 3.6\,\mathrm{nm}$, $\nu \approx 3\mathrm{cP}$, and $U_p \approx 1\,\mathrm{m/s}$ (Ref. 28) giving $\mathrm{Re}_p \approx 10^{-3} \ll 1$ (note also that U_p is the instantaneous velocity; the characteristic particle velocity on the time scale we are examining is much smaller). Hence we neglect the inertial terms in the particle dynamics.

We abstract binding interactions between protein domains as constraints within our integration scheme to avoid the stiff force terms that would otherwise be necessary to accurately model such interactions, and which would impose stringent restrictions on the stable integration time step sizes for the system. Similarly, we model excluded volume interactions by constraints which become active when two bodies impinge on one another and become inactive if a constraint attempts to apply a force that would tend to bring the bodies closer together (see Sec. V D).

A variety of integration schemes exist for solving partial differential equations. Broadly there are explicit and implicit stepping methods. All such methods represent the state at a given time t_n and attempt to find an approximation of the state at time t_{n+1}. Explicit stepping methods solve an equation of the form $\mathbf{X}(t_{n+1}) = \mathbf{X}(t_n) + \mathbf{F}(t_n, \mathbf{X}(t_n))$, evaluating the next state as a directly computable equation of the current state. Implicit stepping methods solve an equation of the form $\mathbf{X}(t_{n+1}) = \mathbf{X}(t_n) + \mathbf{F}(t_{n+1}, \mathbf{X}(t_{n+1}))$, where the function evaluated in order to reach the next state uses information about that state; rather than being directly computable, the equation expresses a condition which must be satisfied to some tolerance in order to achieve an admissible approximation for the next state.

Two advantages of an explicit method are first that the evaluation of the next state is comparatively much more computationally efficient than an implicit method, allowing many more steps to be taken, and second that it is often easier to construct energy-conserving explicit integrators. Two disadvantages are that the allowable time step sizes tend to be much smaller,[29] especially for the case of stiff forces (that is, forces which are much larger than other forces in the system), and that it is difficult to directly enforce the satisfaction of constraints such as excluded volumes (attempts to do so in explicit integration schemes include penalty methods,[30] which, if they are to be effective, result in stiff forces,[31] e.g., Lennard-Jones potential). Implicit methods have the advantage of allowing

direct enforcement of constraints as part of the system of equations to be solved, and of allowing for large integration time steps. They have the disadvantage of being much more expensive than explicit stepping for a single step, and also in many cases of dissipating energy. M-SHAKE[32] and LINCS[33] are methods developed for fine-scale molecular dynamics which use Newton's method and Lagrange multipliers to enforce constraints on bond geometry during simulation; they are similar in concept to the implicit approach which we take in this work.

The choice of integration scheme for a problem is dictated by which properties of the system are important to capture. Related to the energy preservation properties of the two schemes, explicit schemes are capable of correctly modeling bond vibrations, at the cost of not permitting large time steps, while implicit schemes damp out high frequency components (in explicit models of interatomic bonds, time steps are constrained largely by bond vibration frequencies; however, even if these vibrations are damped out nonbonded inter-atom collisions come to dominate on a larger but comparable time scale[34]). In our system, we are concerned with domain-scale phenomena over large time scales (up to the order of seconds to minutes), and are not interested in bond-level interactions. If energy conservation is not maintained in an all-atom molecular dynamics simulation, the effective temperature of the system will change. However, in our case, temperature is an exogenous parameter governing Brownian forces rather than a measurable property of the system so this is not a concern. Constraints can be expressed either as stiff penalty functions or as exact equations; since we require large time steps and would like exact enforcement of our conditions, we choose the latter. Use of stiff forces, as opposed to constraints, would require time steps much smaller than those required by accuracy considerations. Due to these considerations, we chose an implicit Newton-step integration scheme (Sec. IV).

BD simulations often model binding as a probabilistic phenomenon.[20,35] We take a hybrid approach between that and the approach similar to Northrup *et al.*[23] where binding occurs with a base probability which is weighted by the goodness of fit between complementary sites on a pair of molecules (Figure 2). The relationship between the properties of the medium, the geometry and resistance properties of the bodies, and the magnitude of the target distance governs the on-rate of the interaction in isolation. Once an interaction is established, it is maintained via a constraint which requires that the bound molecules remain in a fixed relative position and orientation. The constraint has a base off-rate, modulated by the force that the bond must apply to keep the two bodies together (see Sec. V F). Experimentally determined on and off rates are more readily available than measurements of binding energies, which depend on a combination of hydrogen bonding, electrostatic interactions, and shape complementarity. We can empirically relate our system parameters to such on and off rates, and hence use them as input parameters to our system when modeling experiment. We choose to make dissociation based on force rather than purely probabilistic in an attempt to integrate the effects of multiple forces acting on the interacting molecules (e.g., Brownian forces, steric strain as manifested by peptide linker connecting one of the binding partners to another domain).

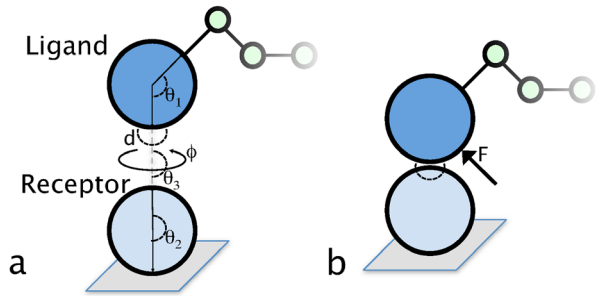

FIG. 2. Simulation of binding on- and off-events. (a) On-event. For a ligand-receptor interaction, the binding sites are defined as points on a spherical surface at defined angles (θ_1, θ_2) relative to other features, such as the attachment to a peptide linker or a cell membrane. When the binding sites approach the ideal binding position and orientation within a threshold distance d, angle ϕ around the bond vectors, and angle θ_3 between the two bond vectors a binding constraint is activated with probability proportional to the goodness of fit between the paired binding sites. (b) Off-event. Binding constraints become inactive and release the bond with a base off-rate modulated by the force F which they experience. In this way, experimental information about on-rates and off-rates can be incorporated into a simulation as distinct parameters.

IV. NEWTON INTEGRATION SCHEME

Our system represents proteins as spheres that differ only in their radii and in their radially defined binding surfaces for other proteins. Peptide linkers are represented as a series of rigidly connected small spheres that are also rigidly connected to the large protein-spheres. In practice, all such spheres are treated by the simulation framework in the same manner—no distinction is made between proteins and peptide elements.

To advance our system in time, we solve the equations for non-inertial Newtonian mechanics. Given m elements, the system satisfies the force balance equation

$$\mathbf{F}(\mathbf{R}) - \zeta\mathbf{v} - \frac{\partial \mathbf{C}^T}{\partial \mathbf{R}}\boldsymbol{\lambda} = \mathbf{0}, \tag{1}$$

where $\mathbf{F} = (\mathbf{F}_1, \ldots, \mathbf{F}_m)^T$ is a vector of forces, ζ is the combined rigid resistance matrix[36] of the elements (in general, the rigid resistance matrix includes hydrodynamic coupling across different molecules; we neglect this), $\mathbf{v} = (\mathbf{v}_1, \ldots, \mathbf{v}_m)^T$ is the vector of element velocities, $\mathbf{R} = (\mathbf{R}_1, \ldots, \mathbf{R}_m)^T$ is the state vector with $\mathbf{R}_i$ a 6-vector representing the position and orientation of element i. The first force term includes non-constraint forces such as linker stiffness. The second force term accounts for hydrodynamic drag on the elements. The last force term on the left hand side represents the forces enforcing the constraints

$$\mathbf{C}(\mathbf{R}) = \mathbf{0}, \tag{2}$$

where $\mathbf{C} = (\mathbf{C}_1, \ldots, \mathbf{C}_l)^T$ is the vector of the l active constraints on the elements. The constraint force is given by the Jacobian of the constraint $\frac{\partial \mathbf{C}}{\partial \mathbf{R}}$ and a set of Lagrange multipliers $\boldsymbol{\lambda} = (\lambda_i, \ldots, \lambda_l)$, the values of which must be determined as part of the dynamics. In our system, the constraint types are for excluded volumes, protein-protein binding (snapping together), and rigid connections within linkers.

The forces and constraints are nonlinear, and hence we use Newton's method to solve Eq. (1) and advance the state from time t^n to time t^{n+1}. Newton's method proceeds by linearizing the equations about the current state $\mathbf{R}^k$, solving the linearized equations for an increment $\delta\mathbf{R}^k$, and updating to the next state $\mathbf{R}^{k+1} = \mathbf{R}^k + \delta\mathbf{R}^k$, until a convergence criteria is satisfied. In particular, in each iteration of Newton's method, we solve the linear system

$$\begin{pmatrix} \frac{1}{\Delta t}\zeta - \frac{\partial \mathbf{F}}{\partial \mathbf{R}}\Big|_{\mathbf{R}^k} & \frac{\partial \mathbf{C}^T}{\partial \mathbf{R}}\Big|_{\mathbf{R}^k} \\ \frac{\partial \mathbf{C}}{\partial \mathbf{R}}\Big|_{\mathbf{R}^k} & \mathbf{0} \end{pmatrix} \begin{pmatrix} \delta\mathbf{R}^k \\ \boldsymbol{\lambda}^k \end{pmatrix} = \begin{pmatrix} \mathbf{F}(\mathbf{R}^k) - \zeta\frac{\mathbf{R}^k - \mathbf{R}^n}{\Delta t} \\ -\mathbf{C}(\mathbf{R}^k) \end{pmatrix}. \tag{3}$$

This linear system of equations is symmetric, indefinite, and can be solved with a standard linear solver. We use PETSc's implementation of MINRES.[37,38] The forces which play a part in $\mathbf{F}$ and $\mathbf{C}$ are described in detail in Sec. V. Note that the drag force depends on the linear and angular velocities, which are expressed as the total change in state over the time step in the drag terms of Eq. (3).

The simulation is initialized by specifying the initial state of each element. The state is then updated from time t^n to the next time as follows:

1. Determine the time step Δt, and set $t^{n+1} = t^n + \Delta t$.
2. Update the state $\mathbf{R}^n$ to $\mathbf{R}^{n+1}$ by solving Eqs. (1) and (2) using Newton's method, where in each Newton step we do the following:
   (a) Compute forces and build the linear system in Eq. (3). If this is the first Newton step, compute and store Brownian forces (which will not change for future Newton steps).
   (b) Solve the linear system Eq. (3) for $\delta\mathbf{R}^k$ and $\boldsymbol{\lambda}^k$.
   (c) Apply a fraction of the resulting position deltas to the bodies in the system ($\mathbf{R}^{k+1} = \mathbf{R}^k + \alpha\delta\mathbf{R}^k$).
   (d) Deactivate any constraints that satisfy deactivation criteria and activate constraints that satisfy activation criteria (Sec. V).

V. FORCE TYPES

For each type of force in the system, we consider the following: activation/deactivation criteria, linearization for the Newton solve, and the effect on the maximum allowable time step.

At present, our system includes six forces: (1) Brownian forces; (2) drag forces; (3) bending forces; (4) excluded volume constraints; (5) relative position constraints defining distances within multidomain proteins; and (6) binding constraints to simulate binding interactions.

A. Brownian forces

Brownian forces as we model them have no position dependence, so they contribute only to the right hand side of the linear system.

For a given spherical element with radius r, the Brownian forces $\mathbf{F}^B$ and torques τ^B acting on it are calculated by sampling from a Gaussian distribution with statistics

$$\begin{aligned} \langle \mathbf{F}^B(t) \rangle &= \mathbf{0}, \\ \langle \mathbf{F}^B(t) \cdot \mathbf{F}^B(t+\Delta t) \rangle &= 2kT\zeta_t\delta(\Delta t) \approx \frac{2k_BT\zeta_t}{\Delta t}, \\ \langle \tau^B(t) \rangle &= \mathbf{0}, \\ \langle \tau^B(t) \cdot \tau^B(t+\Delta t) \rangle &= 2kT\zeta_r\delta(\Delta t) \approx \frac{2k_BT\zeta_r}{\Delta t}, \end{aligned}$$

where k_B is Boltzmann's constant, T is the absolute temperature, δ is the Dirac delta function, and $\zeta_t = 6\pi\mu r$, $\zeta_r = 8\pi\mu r^3$ the translational and rotational drag coefficients, respectively.

B. Drag forces

Each body experience a hydrodynamic drag force acting against its motion, given by

$$\begin{aligned} \mathbf{F}^D &= -\zeta_t\mathbf{V}, \\ \tau^D &= -\zeta_r\boldsymbol{\Omega}, \end{aligned}$$

where $\mathbf{V}$ and $\boldsymbol{\Omega}$ are the linear and angular velocities of the body, respectively. For a general body, the resistance matrix ζ depends on its shape and may couple the 6 degrees of freedom of velocity. For the case of a sphere, the resistance matrix is diagonal. Note also that a more accurate treatment of hydrodynamic drag includes the effects of other bodies in the system. Figure 2 in Parmar *et al.*[39] suggests that inter-protein hydrodynamic interactions may have as much as a 1.5-fold effect on diffusivity across a wide range of concentrations; however, it is not clear how this would manifest in the behavior of individual molecules in close proximity. The effect of shape on protein hydrodynamics is related to the deviation of the protein shape from spherical; this is somewhat challenging, but will need to be addressed in the future. When only Brownian forces and drag forces are present, the displacement statistics of the system are analytically those of Brownian motion.

C. Bending forces

In order to model variable linker stiffness, we impose a bending spring force on the nodes in the linker. This takes the form of a Hookean spring force acting between the center of each such node and the node two steps down the linker from it (i.e., nodes i_n and i_{n+2}) with rest length equal to the maximum allowed distance between those nodes d (see Figure 6(b)). This has the form:

$$\mathbf{F}^{bend} = -k(||x_n - x_{n+2}|| - d)\frac{(x_n - x_{n+2})}{||x_n - x_{n+2}||},$$

where k is the spring constant.

D. Excluded volume constraints

The excluded volume constraints enforce the condition that there is no overlap between two bodies. We add a

constraint $\mathbf{C}_i$ with a corresponding Lagrange multiplier $\boldsymbol{\lambda}_i$ which enforces that the two bodies must have exactly zero overlap as measured in the direction between the body centers before the solve, i.e.,

$$\mathbf{C}_i(\mathbf{R}_1, \mathbf{R}_2) = ||\mathbf{X}_1 - \mathbf{X}_2|| - (r_1 + r_2) = 0,$$

where $\mathbf{X}_i$ is the position of element i and r_i is its radius. Note that this can (temporarily) have the effect of pulling two bodies together.

In each iteration, we first apply the deactivation criteria, and then apply activation criteria. An excluded volume constraint is deactivated if it applied forces that brought the two bodies closer together in the previous Newton iteration. If two bodies intersect by more than a small threshold value (determined by comparing their proximity to the sum of their radii), the excluded volume constraint is activated. A tiny intersection is allowed to avoid spurious sticking effects due to numerical error. Note that if a constraint was on in the previous step and applied only forces that would tend to separate the bodies, we leave it on even if the bodies no longer intersect.

E. Relative position constraints

Relative position constraints require that points embedded within two bodies remain a fixed distance from one another. These constraints are used to model both rigid linkers between bodies (by directly constraining embedded points on the bodies) and flexible linkers (by enforcing rigid relative position constraints among the bodies and a chain of small linker bodies which form the linker, see Figure 3).

We add a constraint $\mathbf{C}_i$ and Lagrange multiplier $\boldsymbol{\lambda}_i$ that attempts to force the embedded points to be at the desired relative position, i.e.,

$$\mathbf{C}_i(\mathbf{R}_1, \mathbf{R}_2) = ||x_1 - x_2|| - d = 0,$$

where $x_i = \mathbf{X}_i + R_i\mathbf{r}_i$ is the embedded position of point i in terms of the position $\mathbf{X}_i$ and rotation R_i of body i, and d is the desired distance between x_1 and x_2.

F. Binding constraints

Binding constraints model pairwise binding interactions between protein domains. We add a constraint $\mathbf{C}_i$ and Lagrange multiplier $\boldsymbol{\lambda}_i$ that attempts to force the binding sites to be equal, i.e.,

$$\mathbf{C}_i(\mathbf{R}_1, \mathbf{R}_2) = x_1 - x_2 = \mathbf{0},$$

where $x_i = \mathbf{X}_i + R_i\mathbf{r}_i$ is the embedded position of binding site i in body i as explained above.

To update the active set of binding constraints, we first check activation criteria and then deactivation criteria. If the constraint is not currently on, we calculate the binding probability density according to

$$p_b(\mathbf{R}_1, \mathbf{R}_2) = e(|\theta_3|, \theta_3^{max})e(|\phi|, \phi^{max})e(d, d^{max})k_{on},$$

where θ_3 captures how close the binding sites in the Activity Element and in the receptor are to aligned, ϕ captures the rotational error around the binding site, d captures the distance error between the sites, and k_{on} is the base on-rate (see Figure 2(a)). The error function $e(a, a^{max})$ is defined as

$$e(a, a^{max}) = \begin{cases} a \le a^{max} & : 1 - \left(\dfrac{a}{a^{max}}\right)^2 \\ a > a^{max} & : 0. \end{cases}$$

We then sample from the exponential distribution $1 - e^{-p_b\Delta t}$ to determine whether to activate the binding constraint. In order to check for constraint deactivation, we take the impulse an active binding force applied over the last time step and project out any component that tends to push the paired bodies apart. We compute a dissociation probability density

$$p_d(\mathbf{R}_1, \mathbf{R}_2, \mathbf{F}) = e^{(\mathbf{F}/\mathbf{F}_0)}k_{off},$$

where $\mathbf{F}_0$ is a force scaling for this bond and k_{off} is the base dissociation rate. We sample from the exponential distribution $1 - e^{-p_d(\mathbf{R}_1,\mathbf{R}_2,\mathbf{F})\Delta t}$ to determine whether to deactivate the constraint.

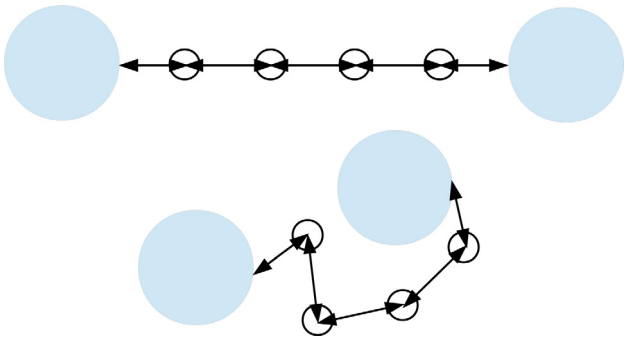

FIG. 3. A flexible linker is created by defining a set of small spherical bodies between domains it links. Each body on a linker segment has a point embedded within it which is constrained to remain a fixed distance from its counterpart on the other body.

G. Effect on time step size

One problem for long simulations is that much computational time can be wasted on short time steps when the relevant elements are far apart and the system does not qualitatively change very much. To address this, we use variable-length time steps, based heuristically on the proximity of proteins that could interact with one another. Specifically, the desired length of each time step is calculated for each pair of reasonably close bodies. The time step is chosen as the time in which the bodies would be expected to close two-thirds of the distance between them under the action of Brownian forces if they move directly towards one another.

H. Techniques for solving the nonlinear system

We use several tricks to accelerate the solution of our system: limiting the size of a Newton step both by a factor which depends on the quality of the step and by the overall size of the step, and preconditioning the linear system. We apply a scale factor α to the calculated step size $\delta\mathbf{R}^k$ such

that $\mathbf{R}^{k+1} = \mathbf{R}^k + \alpha\delta\mathbf{R}^k$, where α is initially set at 0.25 and is bounded between 0.1 and 0.75. We examine the norm of the nonlinear residual after each Newton step; if it has improved by a factor of more than 0.75α (indicating that our linear step guess is good) we increase α; if the residual has improved by less than 0.3α we reduce α; and otherwise we leave α constant. We place a second limitation on the step scaling to ensure that no object point moves more than 0.5 nm in a single Newton step, in order to prevent poor linearizations from destabilizing the simulation. We do this by computing the maximum displacement $\delta\mathbf{R}^k_{max}$ of any linear component of $\delta\mathbf{R}^k$ or any angular component weighted by the associated radius, and clamp α by $0.5/\delta\mathbf{R}^k_{max}$ for this step only.

The computational expense of solving a linear system is directly related to the condition number of that system. Preconditioning the linear system reduces the condition number by making the matrix look closer to an identity matrix; it requires being able to cheaply compute an approximate inverse to the original system. We use $\mathbf{M} = \frac{1}{\Delta t}\zeta$ as our preconditioner; it is block diagonal, cheap to invert, and constant through the simulation up to a factor of Δt and thus only need be inverted once. Instead of Eq. (3) we solve

$$dR^k_M = \mathbf{M}^{1/2}\delta\mathbf{R}^k$$

$$\begin{pmatrix} \mathbf{I} - \mathbf{M}^{-1/2}\left.\frac{\partial \mathbf{F}}{\partial \mathbf{R}}\right|_{\mathbf{R}^k}\mathbf{M}^{-1/2} & \mathbf{M}^{-1/2}\left.\frac{\partial \mathbf{C}^T}{\partial \mathbf{R}}\right|_{\mathbf{R}^k} \\ \left.\frac{\partial \mathbf{C}}{\partial \mathbf{R}}\right|_{\mathbf{R}^k}\mathbf{M}^{-1/2} & \mathbf{0} \end{pmatrix} \begin{pmatrix} \delta\mathbf{R}^k_M \\ \lambda^k \end{pmatrix} = \begin{pmatrix} \mathbf{M}^{1/2}\left(\mathbf{F}(\mathbf{R}^k) - \zeta\frac{\mathbf{R}^k - \mathbf{R}^n}{\Delta t}\right) \\ -\mathbf{C}(\mathbf{R}^k) \end{pmatrix}.$$

After we solve this system, we recover $\delta\mathbf{R}^k$ from $\delta\mathbf{R}^k_M$ as $\delta\mathbf{R}^k = \mathbf{M}^{-1/2}\delta\mathbf{R}^k_M$. In a typical case, this preconditioning improved the condition number of the system by three orders of magnitude.

VI. RESULTS

A. Flexible linker

One purpose of the simulation system described above is to predict the behavior of artificial biological constructions. To address whether the simulation system would generate plausible behavior, and to gain insight into an engineered construction with an eye towards further improvement, we simulated the behavior of a representative targeted fusion protein. The engineered protein consisted of two proteins of molecular weight 20 kD connected by a 35-amino acid glycine-serine linker. The proteins were modeled as spheres of 36 Å diameter (the typical size of a small globular protein[23]), and the linker was modeled as 10 equally spaced rigid segments of total length of 105 Å, joined by 9 nodes of diameter 4 Å (Figure 4(b)). Complete flexibility around the nodes was allowed; this corresponds to a persistence length of about 3 amino acids for the linker. The temperature was set to 310.2 K and the dynamic viscosity was 3.5 cP, corresponding to blood plasma.[40] The simulation was run for 500 000 steps, corresponding to a total duration of 250 μs.

Simulation of the behavior of the linker-tethered fusion protein gives the somewhat surprising result that the large protein domains generally remain quite close, on average closer than the attachment points of the linker connecting the two proteins. Figure 4(b) shows the distribution of distances between the linker attachment points on the surface of the proteins. The modal distance, about 4.25 nm, is slightly larger than that predicted by a random walk of 10 steps, 10.5 Å per step (3.32 nm). However, the average separation between the protein surfaces is 1.60 nm, with a corresponding average distance between the centers of 4.20 nm.

Inspection of the simulation video[45] suggests possible mechanisms. At the Angstrom-nanometer size scale in an aqueous solution, rotational Brownian motion (and thus force) is comparable to translational Brownian motion; for example in a time step of 1 ns, for a spherical protein of 36 Å diameter the mean squared translational displacement will be 0.072 nm^2, while the mean squared angular displacement will be 0.017 rad^2, corresponding to a mean of 0.129 nm motion of a surface point. Thus, the rotational effects on the separation may be expected to be significant.

If translational Brownian motion of the protein elements dominated the simulation, one might expect that the average distance between the centers of the proteins would correspond to the average distance between the center of a sphere of radius approximately 141 Å (linker length plus protein radii) and points within that sphere, which is $3/4R = 105.75$ Å.

B. Förster resonance energy transfer (FRET) efficiency

Evers *et al.*[41] performed FRET studies to determine how different lengths of a flexible glycine-serine linker affect interdomain distances. They used a simple modeling approach treating the linker as a random coil structure in order to examine the behavior of their systems. We applied our modeling approach to the problem, both to attempt to reproduce their results and to see what we can learn about the problem by having an interrogable model system.

The temperature was set to 298 K and the viscosity to 1.31 cP, corresponding to 10% glycerol. The linker segment lengths were 0.45 nm. In order to accommodate these short link lengths, the volume exclusion and hydrodynamic drag radii for the link nodes were set to be different (0.2 nm and 0.3 nm, respectively). Enhanced cyan fluorescent protein

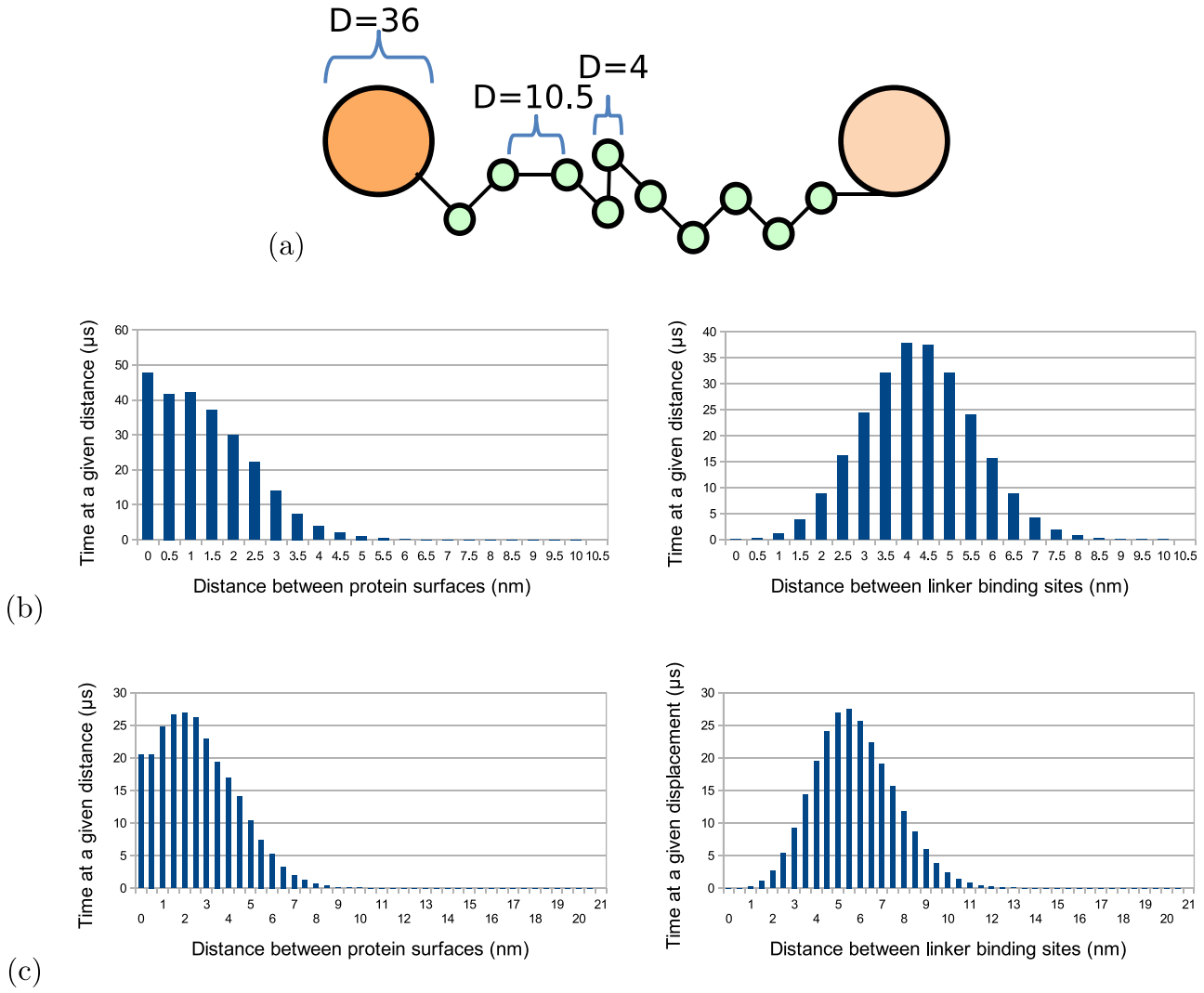

FIG. 4. Average distance between simulated proteins connected by a highly flexible linker. (a) Two spherical proteins of diameter 36 Å were connected by a simulated 35-amino acid glycine-serine linker represented as 10 completely flexible segments of length 10.5 Å attached via 9 intermediate spheres of diameter of 4 Å; 3 amino acids roughly corresponding to a persistence length in a glycine-serine linker. The simulation was performed for 500 000 time steps of 0.5 ns. (b) The left bar chart shows the statistics of the displacement between protein surfaces at each step in the simulation, while the right bar chart shows the statistics for the linker attachment site on the proteins. (c) The simulation conditions are identical to those in (b) except that the linker is composed of 20 flexible segments of length 10.5 Å attached via 19 intermediate spheres for a total length of 210 Å.

(ECFP) was modeled with a hydrodynamic radius of 2 nm and enhanced yellow fluorescent protein (EYFP) with 2.4 nm; we experimented with setting their volume exclusion radii both equal to the hydrodynamic radii and to smaller values, since ECFP and EYFP are somewhat non-spherical.

Rather than choosing a constant value of 2/3 for κ^2 in calculating the Förster radius, our method allows us to calculate it each time step from the emission and absorbance transition dipoles of the ECFP donor and EYFP acceptor as

$$\kappa^2 = (\cos(\theta_T) - 3\cos(\theta_D)\cos(\theta_A))^2,$$

where θ_T is the angle between the emission and absorbance transition dipoles of the donor and the acceptor, θ_D is the angle between the vector joining the two chromophores and the emission transition dipole of the donor, and θ_A is the angle between the vector joining the two chromophores and the absorbance transition dipole of the acceptor. We observed that while the mean value of κ^2 was indeed approximately 2/3, the calculated energy-transfer efficiency differs substantially if the per-frame κ^2 was used to compute it (Figures 5(a) and 5(b)). We also noted that smaller values for the volume exclusion radii of ECFP and EYFP give results closer to accordance with the experimentally measured values, suggesting that protein shape effects are indeed quite important. Each data point corresponds to 40 simulations each of 40 μs, of which the first 1 μs is discarded to allow the simulation to randomize.

C. Chimeric activators

We are interested in engineering linkers which maximize the effectiveness of the chimeric activator approach in increasing the specificity of the Activity Element for its receptor. Accordingly, we performed simulations with a range of linker lengths and stiffnesses and evaluated their effect on receptor binding. Rather than simulate both the chimeric activator and the receptor for the Activity Element, we choose to simulate only the chimeric activator and assume that the

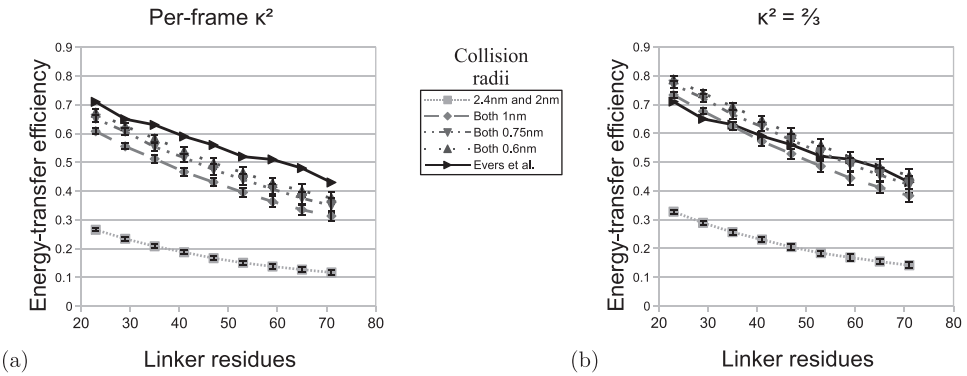

FIG. 5. Calculated energy-transfer efficiency between ECFP and EYFP joined by a flexible glycine-serine linker as described in Evers *et al.*[41] The energy-transfer efficiency was calculated according to the formula for Förster distance which depends on κ^2, which is in turn a function of the relative position and orientation of the fluorophores. We simulated ECFP-linker-EYFP configurations with varying linker lengths and calculated the energy-transfer efficiency with both (a) per-frame calculated κ^2 and (b) constant $\kappa^2 = 2/3$.

receptor will occur in the membrane with a uniform distribution. Given this assumption what we are actually interested in is the distribution of configurations that the Activity Element assumes for a given linker type and problem geometry. We can then write an expression in closed form for $p_b(\mathbf{R}_A)$, the probability of binding in a given time step due to a given configuration of the Activity Element, by integrating over all possible configurations $\mathbf{R}_R$ for the Activity Element receptor

$$p_b(\mathbf{R}_A) = \int_{\mathbf{R}_R} p_b(\mathbf{R}_A, \mathbf{R}_R) p(\mathbf{R}_R),$$

where $\mathbf{R}_A$ is the state of the Activity Element, $p_b(\mathbf{R}_A, \mathbf{R}_R)$ is the probability of binding given the states $\mathbf{R}_A$ and $\mathbf{R}_R$, and $p(\mathbf{R}_R)$ is the probability that the Activity Element receptor is in state $\mathbf{R}_R$, which for these simulations is uniform in position within the $y = 1.8$ nm plane and uniform in rotation around the y-axis. The numbers reported in Figure 6(c) are in arbitrary units which are not scaled by either k_{on} or the concentration of the Activity Element receptor, since the dependence on both is strictly linear.

Each data point in Figure 6(c) corresponds to 40 simulations over 100 μs, of which the first 1 μs is discarded to allow the system to randomize itself. The temperature, viscosity and linker segment lengths of the system are as in Sec. VI A. The maximum error values for binding are set at $d^{max} = 1$nm, $\theta_3 = \pi/4$, and $\phi = \pi/4$, and the target height for the interferon alpha receptor (IFNαR) binding site is set at 3.65 nm above the surface. The error function e for the distance only is cut off such that $e(d, d^{max}) = 0$ when the y-value of the Activity Element binding site is less than 3.65 nm. For all linker configurations, the attachment site of the linker to EGF is at the point offset 1.8 nm along the negative z-axis from its center, while the linker attachment site for interferon alpha is offset $1.8/2^{1/2}$nm along the negative y-axis and $1.8/2^{1/2}$nm along the positive z-axis from its center to better reflect the linker configuration shown in Figure 1(d). The linker stiffness had a significant impact on the overall dynamics, as can be seen in videos with no bending stiffness[45] and the 20 z-N per nm stiffness.[45]

VII. DISCUSSION

Transcriptional control in higher organisms often involves elaborate signal transduction pathways with multiple steps that each offer points for engineering. These steps

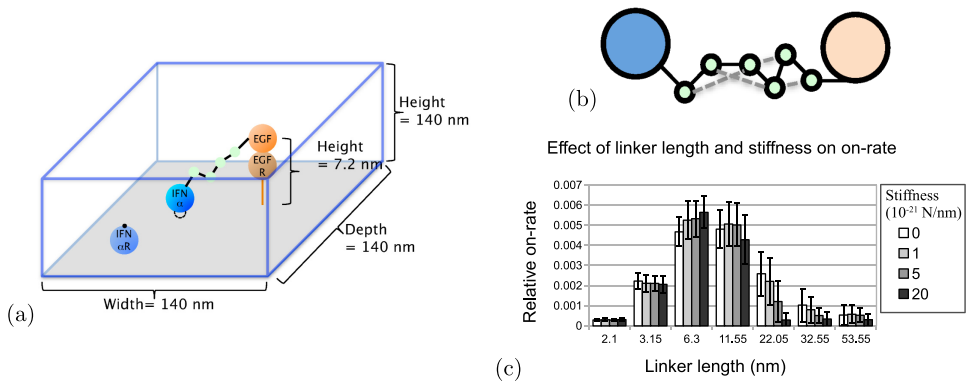

FIG. 6. Effect of modulating the linker length on binding of a tethered ligand to its receptor in simulations of constrained Brownian motion on a cell surface. The rate of binding of IFNα element of an (IFNα)-EGF fusion protein to the IFNαR was estimated, assuming that the EGF was already irreversibly bound to EGFR on the same cell surface. (a) Simulation schematic. The binding reaction takes place in a wrap-around box of dimensions $140 \times 140 \times 140$ nanometers with one copy of each molecule, corresponding to about 75 000 molecules/cell for a typical mammalian cell. The radius of each protein element is 18 Å. The fixed EGF/EGFR element is raised 7.2 nm from the cell surface. Movement of the linker elements was completely unconstrained. (b) Modulation of linker stiffness. The stiffness of the linker was varied by introducing a spring force (gray dotted lines) between alternating elements in the linker, and varying the strength of this force. (c) Relative on-rates in arbitrary units as a function of linker length and stiffness. The on-rates of IFNα for IFNαR in the tethered system of (a) were determined in 100 μm simulations. One nanometer corresponds to about three amino acids.

include protein assembly reactions such as ligand-receptor binding, oligomerization of phosphorylated proteins, and cooperative interactions on DNA to create a transcriptionally active state. To achieve on/off behavior, these reactions are often highly cooperative and involve multiple weak interactions that occur in three dimensions. In the course of evolution, Nature routinely modulates the ability of proteins (as well as RNA and DNA) to bend, rotate, and weakly interact. However, human engineering of these parameters has been limited because of the lack of a mapping function that converts protein-engineerable inputs such as binding strengths; protein shape; the length, flexibility, and attachment sites of linkers, and three-dimensional geometry of binding into an output of system activity.

The goal of the present work is to develop a simulation tool that will predict system outputs such as particular protein assembly events when a user varies such inputs. Our approach builds on previous simulations of natural protein assembly systems in which proteins are abstracted as rigid objects to allow for simulation on long timescales. This framework involves the summation of Brownian forces and viscous drag with forces representing the action of engineerable elements such as linkers and binding positions and strengths.

Based on this framework, we simulated the behavior of two proteins attached by a flexible glycine/serine-type linker. In a simulation of an abstracted pair of proteins and a linker, we found that the average distance between the protein elements was much shorter than the maximal length of the linker, suggesting that the behavior of such a fusion protein might be more influenced by the dynamics of the linker itself rather than Brownian motion of the protein elements (Figure 4). We also tested the ability of such a simulation to reproduce FRET data produced from ECFP and EYFP attached by linkers of various lengths, and found a reasonable correspondence of theory and data (Figure 5). Finally, we simulated the second-step binding of a chimeric activator; a fusion protein consisting of two linker-attached ligands was pre-bound to one cell-surface receptor and the binding rate of the other ligand to its receptor was measured as a function of linker length and stiffness. A maximal binding rate was obtained with an intermediate linker length and higher stiffness (Figure 6).

Previous efforts to simulate the three-dimensional movement and assembly of proteins on long timescales have focused on natural phenomena such as assembly of microtubules and viruses. For example, Hagan and Chandler previously used a similar approach, simulating the assembly of viral capsid proteins into higher-order structures by abstracting the proteins as spheres with radially defined bonding surfaces.[42]

Although our simulation system is at an early stage of development, we were able to perform simulations of flexible protein systems that may offer insight to genetic engineers into how these systems behave. The first simulation was of a fusion protein consisting of two equal-sized 36 Å diameter spherical proteins attached by a linker of 35 amino acids. This corresponds roughly to the chimeric activator protein of Cironi *et al.*,[7] an interferon-alpha–linker (35AA)–EGF fusion protein. One result of the simulation was that while the end-to-end length of the linker is in principle up to 105 Å, the mean distance between the proteins was actually only 16 Å. To achieve simultaneous binding of interferon alpha and EGF to their receptors, the linker attachment sites would need to be separated by at least about 65 Å, which was achieved less than 4% of the time. The effect of increasing the linker length by 2-fold increased the portion of the time at at least 65 Å to 37% a factor of nearly 10-fold (Figure 4(c)).

In this context, it is interesting to note that while the electrostatically neutral $(Gly_4Ser)_n$ linker used by Cironi *et al.*[7] is a standard flexible linker in protein engineering, a naturally occurring linker in the M13 gene 3 attachment protein has the sequence $(Gly_3SerGlu)_n$. The repeating negatively charged glutamic acid may prevent self-interaction of this linker and add some stiffness without overly constraining the conformations adopted by the large domains in the gene 3 protein. From an engineering perspective, this suggests that a linker which did a better job of separating the domains could have a significant effect on the binding efficiency of the system. Note that we modeled an effective persistence length of 10.5 Å; it has been suggested that the physical persistence length may be as short as 4.5 Å,[43] so we may actually be over-estimating the effective separation.

Our specific simulation results illustrate why such simulations may be more useful than simple theoretical predictions. For example, it might be imagined that two protein-sized spheres of diameter D_{big}, connected by a chain of N links of length D_{small} would have an average distance of $D_{big} + N^{\frac{1}{2}} * D_{small}$, corresponding to a random walk in three dimensions. The addition of a "no-knot" constraint would be expected to increase the average distance. However, we actually observed a very short distance between protein spheres in our simulation.

In a second simulation, we measured the predicted FRET signal from a fusion protein consisting of EYFP and ECFP attached by a flexible linker. Evers *et al.*[41] previously determined FRET signals as a function of the linker length between these proteins, and obtained similar results to our predictions, which supports our approach. One factor in the Förster theory of FRET is κ, which is a function of the angles between the dipoles in the fluorescent proteins. On average, κ^2 will be about 2/3 for two dipoles randomly positioned relative to each other, and this value is often used in interpreting experimental FRET data. One aspect of our simulation is that it allowed a direct calculation of κ as well as separation distance at each timepoint in the simulation. We found that when κ was calculated for each frame, the average value was indeed about $\kappa^2 = 2/3$, but that the predicted FRET signal was lower than when a uniform value of $\kappa^2 = 2/3$ was used; this observation illustrates how our simulation approach may reveal subtle effects that are otherwise difficult to identify. The result that the FRET efficiency predicted by us is substantially lower than that observed by Evers *et al.* could be explained by our spherical shape approximation, which may limit close approaches of the dipoles.

In a third simulation, we measured the on-rate of a ligand binding to its cell-surface receptor, in which the ligand was pre-tethered to the cell surface by attachment to a second, prebound ligand via a linker. This situation is a model for the action of targeted fusion proteins that

simultaneously bind to two receptors on a cell surface. Such molecules acting on the cell surface are candidate therapeutics, and even slight improvements in specificity can lead to reduced side effects as perceived by patients. Thus, the existence of a theoretical approach such as ours could aid the design of modified, improved therapeutics through small additive improvements that might be hard to identify in a large parameter space through trial and error. We found that a linker of intermediate length was optimal for maximizing the second binding event, and some effect of linker stiffness was also observed. Other factors could also be taken into account either as variables to be engineer or constraints on the system, such as relative height of the two receptors from the cell membrane, the angle at which the linker emanates from each ligand, and the closest allowable approach of the two membrane to each other.

We envision that our simulation framework could be applied to each aspect of signal transduction, particularly the design of artificial transcription factors. The binding of transcription factors to DNA usually involves non-specific binding to DNA followed by one-dimensional diffusion along the DNA and sometimes transfer between strands that are close in three dimensions.[44] These processes are analogous to the two-dimensional diffusion and binding processes depicted in Figure 6, and could be modeled in the design of new transcription factors. For example, in engineering combinatorial control at the DNA level, it would be useful to have heterodimeric transcriptional activators in which each subunit represents a component of an AND gate. The subunits might have high one-dimensional diffusion rates that would be inevitably reduced upon dimerization due to the more extensive DNA contact; it may be useful to estimate target parameters using our framework in designing such transcription factors. Modeling of tethered configurations should also be useful in design of artificial transcription factors with activation domains, or to engineered flexible cytoplasmic signaling factors, and we expect that our framework could be applied to these problems after further development.

VIII. CONCLUSION

In the present work, we focused on Brownian motion and binding events (on-events) for protein elements approximated as spheres. Ultimately, a refined simulation tool would also represent electrostatic interactions that could affect on-rates or non-specific interactions, off-rates, protein shapes that deviate from spheres, etc. The value of such a system is that it would represent those features that can be altered by amino acid substitution, addition of protein domains, N-linked glycosylation sites, and other tools of the trade of genetic engineers, allowing semi-quantitative *in-silico* prediction and visualization of the properties of engineered systems before spending the time and resources to create them experimentally.

ACKNOWLEDGMENTS

A.R.-M. was supported by National Institutes of Health (NIH) post-doctoral fellowship 1F32CA168274-01 and a Wyss Institute postdoctoral fellowship award. J.C.W. was supported by funds from the Wyss Institute. This work was supported in part by NIH Grant No. 2R01GM036373-29 and DARPA Grant No. W911NF-11-2-0056 to P.A.S. T.S. was supported by the National Institute of General Medical Sciences of the National Institutes of Health under Award No. R01GM104976. The content is solely the responsibility of the authors and does not necessarily represent the official views of the National Institutes of Health.

[1]I. B. Dodd, K. E. Shearwin, A. J. Perkins, T. Burr, A. Hochschild, and J. B. Egan, "Cooperativity in long-range gene regulation by the lambda CI repressor," Genes Dev. **18**, 344–354 (2004).
[2]A. Hochschild and M. Lewis, "The bacteriophage lambda CI protein finds an asymmetric solution," Curr. Opin. Struct. Biol. **19**, 79–86 (2009).
[3]G. Dong and S. S. Golden, "How a cyanobacterium tells time," Curr. Opin. Microbiol. **11**, 541–546 (2008).
[4]M. Nakajima, K. Imai, H. Ito, T. Nishiwaki, Y. Murayama, H. Iwasaki, T. Oyama, and T. Kondo, "Reconstitution of circadian oscillation of cyanobacterial KaiC phosphorylation in vitro," Science **308**, 414–415 (2005).
[5]S. Li, K. R. Schmitz, P. D. Jeffrey, J. J. W. Wiltzius, P. Kussie, and K. M. Ferguson, "Structural basis for inhibition of the epidermal growth factor receptor by cetuximab," Cancer Cell **7**, 301–311 (2005).
[6]T. P. J. Garrett, N. M. McKern, M. Lou, T. C. Elleman, T. E. Adams, G. O. Lovrecz, H.-J. Zhu, F. Walker, M. J. Frenkel, P. A. Hoyne, R. N. Jorissen, E. C. Nice, A. W. Burgess, and C. W. Ward, "Crystal structure of a truncated epidermal growth factor receptor extracellular domain bound to transforming growth factor alpha," Cell **110**, 763–773 (2002).
[7]P. Cironi and I. Swinburne, "Enhancement of cell type specificity by quantitative modulation of a chimeric ligand," J. Biol. Chem. **283**(13), 8469–8476 (2008).
[8]N. D. Taylor, J. C. Way, P. A. Silver, and P. Cironi, "Anti-glycophorin single-chain Fv fusion to low-affinity mutant erythropoietin improves red blood cell-lineage specificity," Protein Eng. **23**, 251–260 (2010).
[9]N. Stahl and G. D. Yancopoulos, "The tripartite CNTF receptor complex: activation and signaling involves components shared with other cytokines," J. Neurobiol. **25**, 1454–1466 (1994).
[10]M. Suzuki and N. Yagi, "DNA recognition code of transcription factors in the helix-turn-helix, probe helix, hormone receptor, and zinc finger families," Proc. Natl. Acad. Sci. U.S.A. **91**, 12357–12361 (1994).
[11]L. Zhang, S. K. Spratt, Q. Liu, B. Johnstone, H. Qi, E. E. Raschke, A. C. Jamieson, E. J. Rebar, A. P. Wolffe, and C. C. Case, "Synthetic zinc finger transcription factor action at an endogenous chromosomal site. Activation of the human erythropoietin gene," J. Biol. Chem. **275**, 33850–33860 (2000).
[12]A. S. Khalil, T. K. Lu, C. J. Bashor, C. L. Ramirez, N. C. Pyenson, J. K. Joung, and J. J. Collins, "A synthetic biology framework for programming eukaryotic transcription functions," Cell **150**, 647–658 (2012).
[13]A. Garg, J. J. Lohmueller, P. A. Silver, and T. Z. Armel, "Engineering synthetic TAL effectors with orthogonal target sites," Nucl. Acids Res. **40**, 7584–7595 (2012).
[14]J. Boch, H. Scholze, S. Schornack, A. Landgraf, S. Hahn, S. Kay, T. Lahaye, A. Nickstadt, and U. Bonas, "Breaking the code of DNA binding specificity of TAL-type III effectors," Science **326**, 1509–1512 (2009).
[15]M. J. Moscou and A. J. Bogdanove, "A simple cipher governs DNA recognition by TAL effectors," Science **326**, 1501 (2009).
[16]P. Perez-Pinera, D. G. Ousterout, J. M. Brunger, A. M. Farin, K. A. Glass, F. Guilak, G. E. Crawford, A. J. Hartemink, and C. A. Gersbach, "Synergistic and tunable human gene activation by combinations of synthetic transcription factors," Nat. Methods **10**, 239–242 (2013).
[17]B. A. Blount, T. Weenink, S. Vasylechko, and T. Ellis, "Rational diversification of a promoter providing fine-tuned expression and orthogonal regulation for synthetic biology," PLoS ONE **7**, e33279 (2012).
[18]S. J. Marrink, H. J. Risselada, S. Yefimov, D. P. Tieleman, and A. H. de Vries, "The MARTINI force field: Coarse grained model for biomolecular simulations," J. Phys. Chem. B **111**, 7812–7824 (2007).
[19]G. A. Voth, *Coarse-Graining of Condensed Phase and Biomolecular Systems* (CRC Press, 2009).
[20]Z. Frazier and F. Alber, "A computational approach to increase time scales in Brownian dynamics-based reaction-diffusion modeling," J. Comput. Biol. **19**, 606–618 (2012).

[21]J. van Zon and P. ten Wolde, "Simulating biochemical networks at the particle level and in time and space: Green's function reaction dynamics," Phys. Rev. Lett. **94**, 128103 (2005).
[22]J. S. van Zon and P. R. ten Wolde, "Green's-function reaction dynamics: A particle-based approach for simulating biochemical networks in time and space," J. Chem. Phys. **123**, 234910 (2005).
[23]S. H. Northrup and H. P. Erickson, "Kinetics of protein-protein association explained by Brownian dynamics computer simulation," Proc. Natl. Acad. Sci. U.S.A. **89**, 3338–3342 (1992).
[24]J. C. Phillips, R. Braun, W. Wang, J. Gumbart, E. Tajkhorshid, E. Villa, C. Chipot, R. D. Skeel, L. Kale, and K. Schulten, "Scalable molecular dynamics with NAMD," J. Comput. Chem. **26**, 1781–1802 (2005).
[25]D. Shaw, M. M. Deneroff, R. Dror, J. S. Kuskin, R. H. Larson, J. K. Salmon, C. Young, B. Batson, K. J. Bowers, and J. C. Chao, "Anton, a special-purpose machine for molecular dynamics simulation," Commun. ACM **51**, 91–97 (2008).
[26]J. L. Klepeis, K. Lindorff-Larsen, R. O. Dror, and D. E. Shaw, "Long-timescale molecular dynamics simulations of protein structure and function," Curr. Opin. Struct. Biol. **19**, 120–127 (2009).
[27]R. R. Gabdoulline and R. C. Wade, "Protein-protein association: Investigation of factors influencing association rates by Brownian dynamics simulations," J. Mol. Biol. **306**, 1139–1155 (2001).
[28]H. C. Berg, *Random Walks in Biology* (Princeton University Press, 1993).
[29]A. Iserles, *A First Course in the Numerical Analysis of Differential Equations* (Cambridge University Press, 2008).
[30]R. Courant, "Variational methods for the solution of problems of equilibrium and vibrations," Bull. Am. Math. Soc. **49**, 1–23 (1943).
[31]M. Hauth, O. Etzmuß, and W. Straßer, "Analysis of numerical methods for the simulation of deformable models," Visual Comput. **19**, 581–600 (2003).
[32]V. Kräutler, W. F. van Gunsteren, and P. H. Hünenberger, "A fast SHAKE algorithm to solve distance constraint equations for small molecules in molecular dynamics simulations," J. Comput. Chem. **22**, 501–508 (2001).
[33]B. Hess, H. Bekker, H. J. C. Berendsen, and J. G. E. M. Fraaije, "LINCS: A linear constraint solver for molecular simulations," J. Comput. Chem. **18**, 1463–1472 (1997).
[34]J. A. Izaguirre, S. Reich, and R. D. Skeel, "Longer time steps for molecular dynamics," J. Chem. Phys. **110**, 9853 (1999).
[35]J. Lipková, K. C. Zygalakis, S. J. Chapman, and R. Erban, "Analysis of Brownian dynamics simulations of reversible bimolecular reactions," SIAM J. Appl. Math. **71**, 714–730 (2011).
[36]S. Kim and S. J. Karrila, *Microhydrodynamics: Principles And Selected Applications* (Dover Publications, 1991).
[37]S. Balay, W. D. Gropp, L. C. McInnes, and B. F. Smith, "Efficient management of parallelism in object oriented numerical software libraries," *in Modern Software Tools in Scientific Computing*, edited by E. Arge, A. M. Bruaset, and H. P. Langtangen (Birkhäuser Press, 1997), pp. 163–202.
[38]C. C. Paige and M. A. Saunders, "Solution of sparse indefinite systems of linear equations," SIAM J. Numer. Anal. **12**, 617–629 (1975).
[39]A. S. Parmar and M. Muschol, "Hydration and hydrodynamic interactions of lysozyme: Effects of chaotropic versus kosmotropic ions," Biophys. J. **97**, 590–598 (2009).
[40]J. Harkness, "The viscosity of human blood plasma: its measurement in health and disease," Biorheology **8**, 171–193 (1971).
[41]T. H. Evers, E. M. W. M. van Dongen, A. C. Faesen, E. W. Meijer, and M. Merkx, "Quantitative understanding of the energy transfer between fluorescent proteins connected via flexible peptide linkers," Biochemistry **45**, 13183–13192 (2006).
[42]M. F. Hagan and D. Chandler, "Dynamic pathways for viral capsid assembly," Biophys. J. **91**, 42–54 (2006).
[43]T. Ohashi, S. D. Galiacy, G. Briscoe, and H. P. Erickson, "An experimental study of GFP-based FRET, with application to intrinsically unstructured proteins," Protein Sci. **16**, 1429–1438 (2007).
[44]O. G. Berg, R. B. Winter, and P. H. von Hippel, "Diffusion-driven mechanisms of protein translocation on nucleic acids. 1. Models and theory," Biochemistry **20**, 6929–6948 (1981).
[45]See supplementary material at http://dx.doi.org/10.1063/1.4810915 for the video of the flexible linker simulation, flexible-linker.mp4; for the video of the chimeric activator simulation where the linker has zero bending stiffness, flexible-chimeric-video.mp4; and for the video of the chimeric activator simulation where the linker has a bending stiffness constant of 20 z-N per nm, stiff-chimeric-video.mp4.