www.nature.com/scientificreports

SCIENTIFIC REPORTS

OPEN

Programmable Potentials: Approximate N-body potentials from coarse-level logic

Received: 26 April 2016
Accepted: 22 August 2016
Published: 27 September 2016

Gunjan S. Thakur[1], Ryan Mohr[2] & Igor Mezić[2]

This paper gives a systematic method for constructing an *N*-body potential, approximating the true potential, that accurately captures meso-scale behavior of the chemical or biological system using pairwise potentials coming from experimental data or ab initio methods. The meso-scale behavior is translated into logic rules for the dynamics. Each pairwise potential has an associated logic function that is constructed using the logic rules, a class of elementary logic functions, and AND, OR, and NOT gates. The effect of each logic function is to turn its associated potential on and off. The *N*-body potential is constructed as linear combination of the pairwise potentials, where the "coefficients" of the potentials are smoothed versions of the associated logic functions. These potentials allow a potentially low-dimensional description of complex processes while still accurately capturing the relevant physics at the meso-scale. We present the proposed formalism to construct coarse-grained potential models for three examples: an inhibitor molecular system, bond breaking in chemical reactions, and DNA transcription from biology. The method can potentially be used in reverse for design of molecular processes by specifying properties of molecules that can carry them out.

Multi-body interactions are ubiquitous in nature and happen at all scales from atomic (quantum description) to molecular (classical approach) to macro scales. A systematic analysis these interactions may unfold the fundamental principles governing a given system. For example, understanding the biophysics of protein folding gives insight into disease pathologies[1]. This understanding can be leveraged to develop new vaccines and drug therapies. Engineering these new products requires accurate and computationally tractable models.

Systems having multibody interactions, in fundamental physics, are often formulated as a "*N*-body potential" problem. In order to fully understand these systems a large number of experiments are needed. Conducting experiments may be expensive and at times even impossible. Another approach is to analyze the *N*-body potential governing the system dynamics. However, at the quantum level, it may be difficult to determine these potentials from first principles due to the complexity of the system. The computational complexity for ab initio methods can scale exponentially in the number of electrons, limiting the practical size of the system to a few thousand atoms[2–4]. Even if the detailed potential is determined, it may not be immediately useful. Such is the case when the properties or behaviors of interest are at a coarser level than that of the detailed potential and simulating the detailed dynamics is too expensive. Very coarse approaches such as those of master equation[5] lack predictability on molecular spatial and time scales due to the assumptions with which they are derived. A potential that models the system is required if one is to make predictions about the system.

It is profitable to restrict one's efforts to considering approximate potentials that respect known behavior. Such coarse-level descriptions may be determined from experimental observation and may correspond to trajectories in some transformed (reaction) coordinate system. For example, consider a signal transduction mechanism[6–10], hierarchical self-assembly[11–20], Kinesin motor protein translocation on a microtubule[21,22], or hydrogen combustion H_2/O_2[23,24]. These systems transition from one stable configuration to another on the occurrence of some trigger event which may comprise of an external stimulus or the system reaching a special configuration. An external stimulus could be an input of energy that initiates hydrogen combustion, leading to a larger release of energy by the reaction itself. A special configuration could be a signaling molecule binding to an active receptor site. These stable configurations can be considered as fixed points in a transformed (reaction) coordinate system. The fixed points, the events, and their associated transitions are the coarse-level descriptions that are to be captured in the

[1]Harvard University, John A. Paulson School of Engineering and Applied Sciences, Cambridge, MA 02138, USA. [2]University of California Santa Barbara, Department of Mechanical Engineering, Santa Barbara, CA 93106, USA. Correspondence and requests for materials should be addressed to G.S.T. (email: gthakur@g.harvard.edu)

approximate N-body potential. However, it is still a challenge to construct a N-body Hamiltonian potential in a systematic manner that encodes the known coarse-level behaviors into a mathematical formulation and successfully predicts intermediate-scale transition events.

This article introduces a method of encoding coarse-level dynamical behavior into logic functions that are used to "stitch" together pairwise interaction potentials into an N-body potential. In this method, the practitioner uses experimentally observed coarse-level behavior to derive logic tables that capture various rules of interaction in the system. The qualitative logic tables are turned into a collection of quantitative logic functions associated with pairwise interaction potentials. The logic functions are then turned into smooth encoding functions via a replacement procedure which in turn are used to modify the pairwise potentials. The effect of an encoding function multiplying a pairwise potential is to smoothly turn the potential on or off when a precise set of conditions are met. The combination of the modified potentials gives an N-body potential that approximates the true potential governing the system.

The method generates a potential that respects what is currently known about the system; it is not claimed that this method results in the unique potential governing the real system. The method does this by leveraging the existing experimental data and the coarse-level behavior that can be derived from it. If more experimental data becomes available, the same procedure can be used to generate a new potential that better models the system. This is equivalent to a refinement of the logic functions and ultimately a refinement of the generated potential. The resulting potential can have a much smaller dimension than the true potential and still accurately capture the relevant physics.

This article begins with a motivating example which is used as an impetus for our modeling framework. In the Methodology section, we define the major components of the framework — logic functions, permissible logical operations, and the translation to the associated encoding functions — and specify how they combine with the pairwise potentials to define the approximate potential. The procedure is depicted in Fig. 1.

The procedure is applied to three examples of increasing complexity to showcase the modeling framework. The first is a simple model of an inhibitor molecule mechanism. It shows how one would go from known coarse-level behavior to an approximate global potential that captures that behavior by explicitly constructing the logic tables, the associated logic functions, and the smooth encoding functions. The inhibitor molecule mechanism has more complicated logic than the motivating example and more effectively demonstrates the modeling procedure.

The second example shows how to model a simple, kinetically controlled, bond breaking chemical reaction using this framework. It shows that bond breaking events, and more generally chemical reactions requiring activation energy, can be naturally modeled in the framework. The general procedure for modeling a bond breaking event and how to account for the activation energy is shown. Furthermore, the derived potential is used with LAMMPS[25] to numerically simulate the chemical reaction. By changing the relative dissociation energies, the reaction can be biased in a particular direction.

As opposed to our method, many force fields have trouble capturing bond breaking events[26]. An exception is the ReaxFF potential[2] that was developed to model reactions of hydrocarbons. The derivation of ReaxFF is based on using interatomic distances to compute the bond order between two atoms and then using the bond order to obtain the bond energy. Corrections to the bond order are dependent on the valency and the deviation of the uncorrected bond order of an atom with its valency. Corrections to the bond energy, in the form of energy penalties (e.g. for over-/under-coordination) are added to get the system energy. This is contrasted with our method where bond weakening and breaking is due to the encoding function which is derived from coarse-level observed behavior.

The final example is a simple model of DNA transcription. It is shown that after the binding of RNA polymerase to the promoter region we can sequentially add the complementary base nucleotides to the DNA strand that is to be transcribed. DNA transcription is a complex process involving the interaction of many different molecules[27,28]. This example shows that we can model such a complex process with a relatively low-dimensional potential that captures the observed mesoscale behavior. To the authors' knowledge there is no other other current potential accomplishing this task.

Motivating Example

There are a number of examples in biology where chemical reactions occurring within a cell are initiated by some signal or stimulus, followed by an ordered sequence of biochemical reactions. Often the term signal transduction is used to refer to such processes. One such example is the epidermal growth factor (EGF) signaling[9,10]. Motivated by this example, we construct a hypothetical system to demonstrate how the proposed formulation can be used to construct a Hamiltonian potential for it. Assume a system of three species, **A**, **B** and **C**, has an evolution dictated by the chemical equation $\mathbf{A} + \mathbf{B} \rightarrow \mathbf{AB} \overset{\mathbf{C}}{\rightarrow} \mathbf{A} + \mathbf{B}$. The sequence in which these reactions happen define logical "interaction rules" used to design the potential. Specifically, these rules are (1) when molecules **A** and **B** are close, and **C** is far, then **A** and **B** bond; and (2) If **C** approaches the **AB** complex, then **A** and **B** dissociate. This mechanism is visualized in Fig. 2.

Each of the species in this system can be modeled as a rod having two sites of interaction at the end points; atoms {1, 2} on **A**, atoms {3, 4} on **B**, and atoms {5, 6} on **C** (Fig. 2). Let us write the force field energy for this system. In general, it is composed of the bonded energies formed from the stretch, bending, and torsional terms, the non-bonded van der Waals and electrostatic terms, and the coupling terms[26]. We can split the potential as

$$U = \sum_{i}\sum_{j>i}\Phi_{(i,j)} + \text{higher order terms}$$

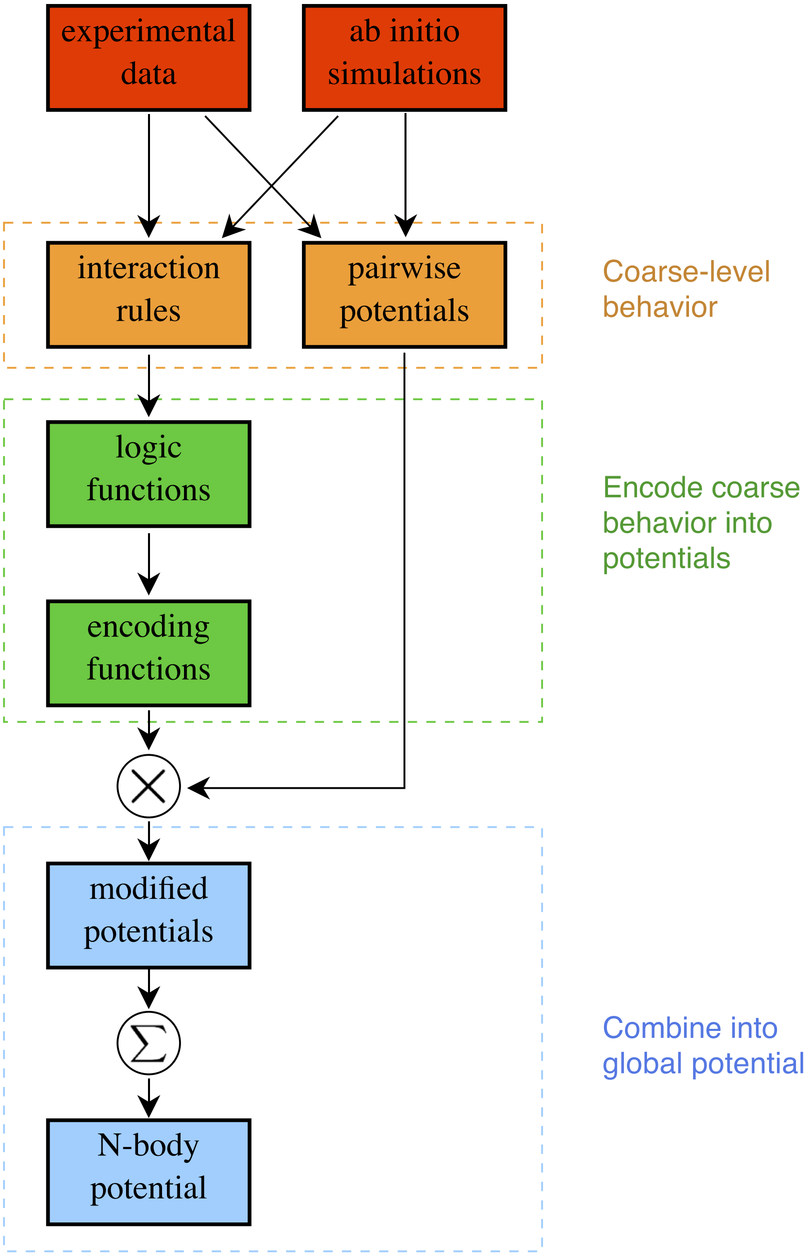


Figure 1. Flow chart of procedure. Using experimentally observed data and quantum calculations (red), we extract coarse-grain behavior (orange) namely: interactions rules and pairwise interaction potentials. This information is used to obtain an N-body potential (blue) for the system by employing the proposed formalism (green).

where $\Phi_{(i,j)}$ denotes a pair-wise interaction potential between two atoms *i* and *j*. These 2-atom terms encompass the stretching, torsional, van der Waals, and electrostatic terms and the higher order terms include the bending energies and all the *k*-atom ($k \geq 3$) coupling terms.

For this system, the bonded energy terms are composed of the stretch energies between the atom-atom pairs (1, 2), (3, 4) and (5, 6), which we can group into a term U_b. Assume that the only non-negligible non-bonded energy terms are the two van der Waals interactions between atoms 2 and 3 and atoms 2 and 5, and the coupling term between atoms 2, 3, and 5. Denoting these three terms by $\Phi_{(2,3)}$, $\Phi_{(2,5)}$, and $\Phi_{(2,3,5)}$, respectively, we get the force field energy of the system as

$$U_{\text{signaling}} = \Phi_{(2,3)} + \Phi_{(2,5)} + \Phi_{(2,3,5)} + U_b$$

The inclusion of the 3-atom potential $\Phi_{(2,3,5)}$ is required in order to capture the transition of the pair (2, 3) being in a stable (bounded) configuration when atom 5 is not present to being in an unstable (free) configuration in the presence of the signaling atom 5.

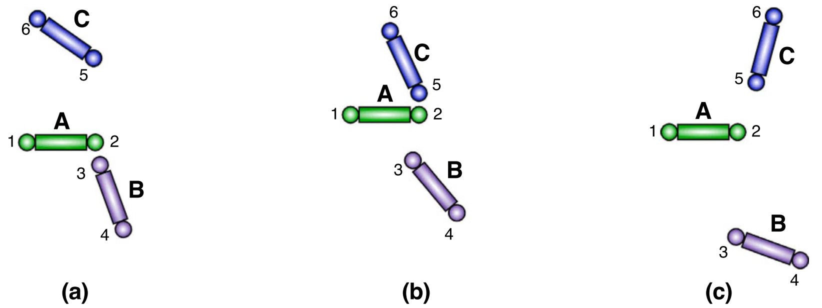


Figure 2. Simple signaling molecule mechanism, A + B → AB $\overset{C}{\rightarrow}$ A + B, modeled by three rods. When **C** is not present, **A** and **B** form a complex. When **C** is present, **A** and **B** dissociate and diffuse apart. Molecule **C** is free to diffuse away from molecule **A**. This behavior is captured with the following rules. When atom 5 is far from atom 2, the potential between atoms 2 and 3 is on. When atom 5 is close to atom 2, the potential between atoms 2 and 3 is turned off allowing molecules **A** and **B** dissociate and diffuse apart. Atom 5 can diffuse away from atom 2.

While in general it may be hard to get the correct forms for the coupling term and other higher-order terms in the expansion, and thus the full potential, we know from the above observations that the effect of the potentials $\Phi_{(2,5)}$ and $\Phi_{(2,3,5)}$ is to basically to turn off $\Phi_{(2,3)}$ when 5 is close to 2. Rewriting the potential as

$$U_{\text{signaling}} = \Phi_{(2,3)}\left(1 + \frac{\Phi_{(2,5)} + \Phi_{(2,3,5)}}{\Phi_{(2,3)}}\right) + U_b,$$

this means that term in parentheses is approximately 1 whenever atoms 2 and 5 are far and approximately 0 whenever atoms 2 and 5 are close. Instead of attempting to find the exactly functional forms of $\Phi_{(2,5)}$ and $\Phi_{(2,3,5)}$, we approximate the potential as

$$U_{\text{signaling}} \approx S_{(2,3)}\Phi_{(2,3)} + U_b, \tag{1}$$

where $S_{(2,3)}$ is an encoding function that acts as a switch turning $\Phi_{(2,3)}$ on and off. In this example, the encoding function is only function the distance between atoms 2 and 5. The encoding function takes values between 0 and 1, it is approximately 0 when atoms 2 and 5 are close, and approximately 1 when atoms 2 and 5 are far; thus it encodes the logic of the coarse-level observed behavior of the system. It is an approximation of the other terms:

$$S_{(2,3)} \approx 1 + \frac{\Phi_{(2,5)} + \Phi_{(2,3,5)}}{\Phi_{(2,3)}}.$$

In the rest of the article, we make this approximation idea (Eq. (1)) precise and derive approximate N-body potentials from simple pairwise interactions that respect observed coarse-level behavior. We give a systematic procedure to construct the encoding functions which allows us to handle systems with more complex interaction rules. We will demonstrate the procedure with three examples. We also use molecular dynamics simulations using the derived potentials that show we can accurately capture the relevant physics.

There are a few items to keep in mind as motivation for the abstract concepts to follow. The basic building blocks for the N-body potential are pairwise interaction potentials (denoted by $\Phi_{(2,3)}$ and $\Phi_{(2,5)}$ for the above example). The explicit form of these potentials can be inferred from the experimental data or ab initio calculations. We approximate the effect of the un-modeled potentials by modifying the relevant pairwise potentials with an encoding function. The encoding function only turns the corresponding potential on and off. The functional form of the potential does not change; it is only scaled between 0 and 1. The logic contained in the encoding functions is obtained from experimental observations or ab initio simulations and the logic only depends on pairwise distances between particles, except the pairwise distance used in the associated potential function; e.g. the logic corresponding to $\Phi_{(2,3)}$ cannot depend on the distance between atoms 2 and 3.

Methodology

It is assumed that there are M interacting entities in a domain $\mathcal{D}$, where $\mathcal{D} \subseteq \mathbb{R}^d$, for $d = 1, 2$, or 3. Each entity is modeled by a finite number of particles with constraint forces between the particles; the totality of these particles over all the entities are labeled from 1 to N. This allows us to treat point particles as well as rigid and and non-rigid bodies. The configuration space is $\mathcal{C} = \mathcal{D}^N$. A particular system configuration, $\vec{\boldsymbol{x}} \in \mathcal{C}$, takes the form $\vec{\boldsymbol{x}} = (\boldsymbol{x}_1, \ldots, \boldsymbol{x}_N)$, where $\boldsymbol{x}_j \in \mathcal{D}$ describes the position of particle j in the domain $\mathcal{D}$.

The dynamics of the system is driven by a potential gradient and external forces. Specifically, the functional form of the dynamics is

$$m_i\ddot{x}_i = -\nabla_{\boldsymbol{x}_i} U(\vec{\boldsymbol{x}}) + F_i(\vec{\boldsymbol{x}}, t), \tag{2}$$

where m_i is the mass of atom i; $\nabla_{\boldsymbol{x}_i}$ is the gradient operator in the configuration space $\mathcal{C}$ with respect to the the position $\boldsymbol{x}_i$ of atom i; $F_i(\vec{\boldsymbol{x}}, t)$ collects the external forces on particle i such as external electric or magnetic fields as well as stochastic effects or boundary constraints; and the approximate potential is defined as

$$U(\vec{\boldsymbol{x}}) = \sum_{\boldsymbol{p}\in\mathcal{I}} \sum_{j\in\mathfrak{m}(\boldsymbol{p})} S_{\boldsymbol{p},j}(\vec{\boldsymbol{x}})\Phi_{\boldsymbol{p},j}(\vec{\boldsymbol{x}}). \tag{3}$$

The notation in this equation is as follows.

$\mathcal{I}$, set of interacting pairs of atoms. $\mathcal{I}$ defines which pairs of atoms interact. For example, if $\boldsymbol{p} = (2, 3)$ is in $\mathcal{I}$, then there exists a pairwise potential between atoms 2 and 3. Since not every pair of atoms in the system has to interact, $\mathcal{I}$ can be a strict subset of $\{1, \ldots, N\} \times \{1, \ldots, N\}$.

$\mathfrak{m}(\cdot)$, multiplicity function. The multiplicity function $\mathfrak{m}: \mathcal{I} \to \mathbb{N}$ determines how many potentials that atom pair $\boldsymbol{p} = (p_1, p_2)$ interacts through. Often $\mathfrak{m}(\boldsymbol{p})$ will be 1 for every atom pair $\boldsymbol{p}$. However, non-unit values become important when a pair $\boldsymbol{p} \in \mathcal{I}$ interacts through multiple different potentials, each with its own encoding function. For example, a non-unit multiplicity is useful when modeling bond-breaking chemical reactions. Initially, two atoms interact through their bond potential; when this bond is broken, another potential is required to model the electron-electron repulsion between the atoms.

$\Phi_{\boldsymbol{p},j}$, pairwise interaction potential. $\Phi_{\boldsymbol{p},j}: \mathcal{C} \to \mathbb{R}$ is the j^{th} interaction potential for the pair of atoms $\boldsymbol{p} = (p_1, p_2) \in \mathcal{I}$. The index j is runs from 1 to $\mathfrak{m}(\boldsymbol{p})$. For $\vec{\boldsymbol{x}} = (\boldsymbol{x}_1, \ldots, \boldsymbol{x}_N)$, it takes the form

$$\Phi_{\boldsymbol{p},j}(\vec{\boldsymbol{x}}) = \phi_{\boldsymbol{p},j}(\|\pi_{p_1}(\vec{\boldsymbol{x}}) - \pi_{p_2}(\vec{\boldsymbol{x}})\|) \equiv \phi_{\boldsymbol{p},j}(\|\boldsymbol{x}_{p_1} - \boldsymbol{x}_{p_2}\|), \tag{4}$$

where for every $i \in \{1, \ldots, N\}$, the coordinate map $\pi_i: \mathcal{C} \to \mathcal{D}$ extracts the position $\boldsymbol{x}_i$ of atom i from the configuration vector $\vec{\boldsymbol{x}} = (\boldsymbol{x}_1, \ldots, \boldsymbol{x}_N)$; the norm $\|\cdot\|$ denotes the normal Euclidean norm; and $\phi_{\boldsymbol{p},j}: \mathbb{R}^+ \to \mathbb{R}$ is the j^{th} 1D pairwise interaction potential through which the pair $\boldsymbol{p}$ interacts. This potential could be, for example, a Lennard-Jones or Morse potential; it can also be different for different interaction pairs. The form given for the potential shows that it is only a function of the distance between $\boldsymbol{x}_{p_1}$ and $\boldsymbol{x}_{p_2}$. When $\mathfrak{m}(\boldsymbol{p}) = 1$, we drop the j index from $\Phi_{\boldsymbol{p},j}$ and write it as $\Phi_{\boldsymbol{p}}$.

$S_{\boldsymbol{p},j}$, encoding function. $S_{\boldsymbol{p},j}: \mathcal{C} \to [0, 1]$ is the encoding function associated with the potential $\Phi_{\boldsymbol{p},j}$. It encodes the coarse-level interaction rules and it is a function of pairwise distances between particles, except for the particle pair $\boldsymbol{p}$ to which it corresponds. That is, for the interaction pair $\boldsymbol{p} = (p_1, p_2) \in \mathcal{I}$, the encoding function $S_{\boldsymbol{p},j}$ is *not* a function of the distance $\|\boldsymbol{x}_{p_1} - \boldsymbol{x}_{p_2}\|$. The effect of $S_{\boldsymbol{p},j}$ is to smoothly turn its associated potential function on and off based on the configuration of the system. Since the encoding functions and potentials are functions of relative distances only, Equation (2) defines a Hamiltonian system[29] when we neglect the forces $F_i(\vec{\boldsymbol{x}})$. When $\mathfrak{m}(\boldsymbol{p}) = 1$, we drop the j index from $S_{\boldsymbol{p},j}$ and write it as $S_{\boldsymbol{p}}$.

A majority of the rest of the paper develops the encoding functions and their properties and shows how one would go from coarse-level interaction rules to encoding functions using a few examples. It is assumed that the coarse-level, interaction rules and the interaction potentials $\phi_j: \mathbb{R}^+ \to \mathbb{R}$ are known. These come from analyzing experimental data or ab initio simulations and are thus application specific and beyond the scope of this article. Ultimately, the encoding function $S_{\boldsymbol{p},j}$ will be a smoothed version — which is made precise later — of a logic function $L_{\boldsymbol{p},j}: \mathcal{C} \to \{0, 1\}$, which assigns 0 or 1 to each configuration $\vec{\boldsymbol{x}}$. The logic function $L_{\boldsymbol{p},j}$ will be the constructed from a finite number of logical operations applied to elementary logic functions from a Boolean algebra. More precisely, the logic function will be an element of the Boolean sub-algebra generated by elementary logic functions. Thus to define the logic functions, it is required to know the specific definitions of the logical operators AND, OR, and NOT (symbolically denoted, $\wedge, \vee, \neg$) and what Boolean functions are used as the elementary logic functions.

A function $b: \mathcal{C} \to \{0, 1\}$, which assigns either 0 or 1 to each configuration vector $\vec{\boldsymbol{x}}$ is called a Boolean function on $\mathcal{C}$ and the set of such all such functions on $\mathcal{C}$ is denoted as $\mathcal{B}$. It is easy to see that the functions that are identically 1 and 0 on $\mathcal{C}$ are Boolean functions. On $\mathcal{B}$, define for all $f, g \in \mathcal{B}$ the two binary logical operations AND ($\wedge$) and OR ($\vee$) and the unary logical operation NOT ($\neg$) by

$$\begin{aligned} &(f \wedge g)(\vec{\boldsymbol{x}}) = f(\vec{\boldsymbol{x}})g(\vec{\boldsymbol{x}}), \quad (f \vee g)(\vec{\boldsymbol{x}}) = f(\vec{\boldsymbol{x}}) + g(\vec{\boldsymbol{x}}) - (f \wedge g)(\vec{\boldsymbol{x}}), \\ &(\neg f)(\vec{\boldsymbol{x}}) = 1 - f(\vec{\boldsymbol{x}}). \end{aligned}$$

These three logical operations will be applied to specific elements of the set of all Boolean functions $\mathcal{B}$ on $\mathcal{C}$ to generate a Boolean sub-algebra. The logic functions $L_{\boldsymbol{p},j}$ will be elements of this sub-algebra.

Proximity functions are used to define the elementary logic functions. A proximity function $P_R: \mathbb{R}^+ \to \{0, 1\}$ has the form $P_R(r) = \chi_{[0,R)}(r)$, for some R satisfying $0 \leq R \leq \infty$. The function $\chi_{[0,R)}: \mathbb{R}^+ \to \mathbb{R}$ is the indicator function for the semi-open interval $[0, R)$ which takes the value 1 if the argument satisfies $0 \leq r < R$ and 0 otherwise. Note that the functions that are identically 1 or 0 are proximity functions. The *elementary logic functions* are defined as compositions of a proximity function with the coordinate functions π_i from above. Specifically, the elementary logic function $\ell_{\boldsymbol{q},R}$ for atom pair $\boldsymbol{q} = (q_1, q_2) \in \mathcal{I}$ and parameter $0 \leq R \leq \infty$ has the form

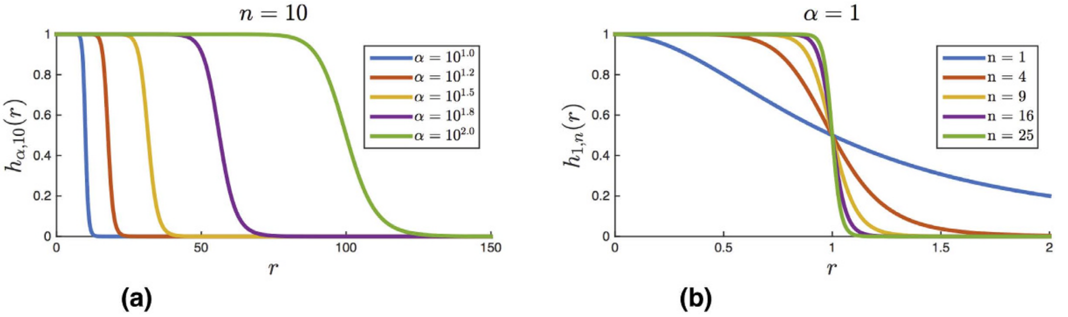


Figure 3. Behavior of the $h_{\alpha,n}$ function from Equation (6). (**a**) α controls the transition point. (**b**) n controls the sharpness of the transition.

$$\ell_{\boldsymbol{q},R}(\vec{\boldsymbol{x}}) = P_R(\|\pi_{q_1}(\vec{\boldsymbol{x}}) - \pi_{q_2}(\vec{\boldsymbol{x}})\|) \equiv \chi_{[0,R)}(\|\boldsymbol{x}_{q_1} - \boldsymbol{x}_{q_2}\|). \tag{5}$$

This function is 1 when $\boldsymbol{x}_{q_1}$ and $\boldsymbol{x}_{q_2}$ are closer than distance R and 0 when not.

A logic function $L_{\boldsymbol{p},j}$ is generated by applying finitely many of the logical operations $\wedge$, $\vee$, and $\neg$ to the elementary logic functions (5) for any finite set of $\boldsymbol{q}$'s, none of which are equal to $\boldsymbol{p}$ — that is, $\ell_{\boldsymbol{p},R}(\vec{\boldsymbol{x}})$ cannot be part of the definition of $L_{\boldsymbol{p},j}$. Each logic function $L_{\boldsymbol{p},j}$ is continuous almost everywhere in $\mathcal{C}$ since each elementary logic function is constant almost everywhere. This follows since it is composed from pairwise products and sums of elementary logic function, which themselves are continuous almost everywhere.

Once the logic function is specified, it must be translated into a smooth encoding function. Ideally, this would be accomplished via a convolution in the (dN-dimensional) configuration space with a smooth, nonnegative summability kernel (see Katznelson[30] for a definition). Analytically, this is intractable, and computationally, this is very expensive. Instead, we individually smooth each of the 1D elementary logic functions $\ell_{\boldsymbol{p},R}(\vec{\boldsymbol{x}})$ in the expression for $L_{\boldsymbol{p},j}(\vec{\boldsymbol{x}})$. This is done by replacing the proximity function of $\ell_{\boldsymbol{q},R}$ with a smoothed version. Again, this could be done via the convolution (now 1-dimenional) of each proximity function with a smooth, 1D summability kernel or, alternatively, by the replacement of each indicator function with a specific functional form. We choose the latter approach and replace each proximity function $\chi_{[0,R)}(r)$ with a function of the form

$$h_{\alpha,n}(r) = \frac{1}{1 + (r/\alpha)^{2n}}, \qquad (0 < \alpha < \infty,\, n \in \mathbb{N}), \tag{6}$$

and we define $h_{0,n}(r) = 0$ and $h_{\infty,n}(r) = 1$. For example, if the logic function has the expression

$$L_{\boldsymbol{p},j}(\vec{\boldsymbol{x}}) = \chi_{[0,R_1)}(\|\boldsymbol{x}_{q_1} - \boldsymbol{x}_{q_2}\|) \wedge (1 - \chi_{[0,R_2)}(\|\boldsymbol{x}_{s_1} - \boldsymbol{x}_{s_2}\|)),$$

then the corresponding encoding function would be

$$S_{\boldsymbol{p},j}(\vec{\boldsymbol{x}}) = h_{\alpha_1,n_1}(\|\boldsymbol{x}_{q_1} - \boldsymbol{x}_{q_2}\|) \wedge (1 - h_{\alpha_2,n_2}(\|\boldsymbol{x}_{s_1} - \boldsymbol{x}_{s_2}\|)),$$

for some choices of parameters α_1, α_2, n_1, and n_2. The parameters α and n control how well $h_{\alpha,n}$ approximates a proximity function (see Fig. 3). In particular, $h_{\alpha,n}(0) = 1$ for any $0 < \alpha < \infty$ and positive n. Furthermore, $\lim_{r\to\infty} h_{\alpha,n}(r) = 0$ and it is strictly monotonically decreasing. On the other hand, for a fixed $0 < \alpha < \infty$, the transition from 1 to 0 becomes sharper as n increases (Fig. 3(b)). To match a specific indicator function $\chi_{[0,R)}$, choose $\alpha = R$. With this choice of α, the function satisfies $h_{R,n}(R) = 1/2$ for all $n \geq 1$;

$$\lim_{n\to\infty} h_{R,n}(r) = \begin{cases} 1 & r < R \\ \dfrac{1}{2} & r = R \\ 0 & r > R. \end{cases}$$

Examples

To show the entire process, starting from coarse, interaction rules and recovering the encoding function, we apply the method to three examples in increasing order of complexity. The first example is a model for an inhibitor molecule system and is used to exhibit the core methodology of the modeling framework. This system can be considered as an extension of the signaling molecule example above (Fig. 2). The second example is a model for a simple bond breaking chemical reaction and makes use of the multiplicity function $\mathfrak{m}$ from the framework. It is shown that the bond dissociation energy is accurately captured in this framework. Numerical simulations show that (i) the system exhibits the same coarse-level behavior that was used to derive the potential and (ii) that biased chemical reactions are easily handled. The final example is a simple model for DNA transcription and is the most complicated of the three. This example shows that the logic, and hence potential, of real systems can be captured in the modeling framework in a straight-forward manner.

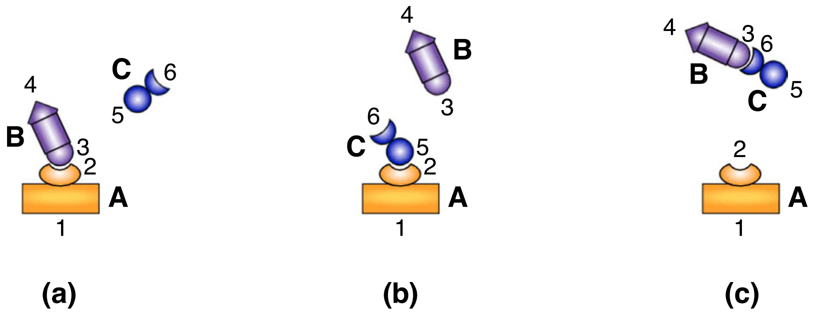


Figure 4. Inhibitor molecule example. When the inhibitor molecule, **C**, is not present (panel **(a)**), a bond between receptor **A**'s site 2 and site 3 on the active molecule **B** can form. When the inhibitor molecule is present, it can either take up the receptor site through a (2, 5) bond (panel **(b)**) or bind to site 3 on **B** with site 6 (panel (**c**)). Either of these cases prevents to active molecule **B** from binding with its receptor site on **A**.

2 and 5 "close"	3 and 6 "close"	$L_{(2,3)}$	$L_{(3,6)}$	$L_{(2,5)}$
0	0	1	1	1
0	1	0	1	0
1	0	0	0	1
1	1	0	0	0

Table 1. Bond logic for the inhibitor molecule mechanism.

Simple inhibitor molecule mechanism. This example can be thought of as a simple model for the action of an inhibitor molecule in a plane. Consider the three interacting molecules in Fig. 4. The configuration space for this example is $\mathcal{C} = (\mathbb{R}^2)^6$, where $\vec{\boldsymbol{x}} \in \mathcal{C}$ is written as $\vec{\boldsymbol{x}} = (\boldsymbol{x}_1, \ldots, \boldsymbol{x}_6)$. The set of interacting atom pairs is $\mathcal{I} = \{(2, 3), (2, 5), (3, 6)\}$. For this example, the multiplicity function $\mathfrak{m}\colon \mathcal{I} \to \mathbb{N}$ is identically 1. Thus we have the pairwise potentials $\Phi_{(2,3)}$, $\Phi_{(2,5)}$, and $\Phi_{(3,6)}$. It is assumed that these potentials are formed using a Morse potential (see (12)). Molecule **C** is an inhibitor molecule and prevents the formation of the **AB** complex. Without **C**, we have $\mathbf{A}+\mathbf{B}\rightarrow\mathbf{AB}$. With **C** present, the there are two possibilities: (i) $\mathbf{A} + \mathbf{B} \overset{\mathbf{C}}{\rightarrow} \mathbf{AC} + \mathbf{B}$ or (ii) $\mathbf{A} + \mathbf{B} \overset{\mathbf{C}}{\rightarrow} \mathbf{A} + \mathbf{BC}$.

This behavior is captured in the logic functions $L_{(2,3)}$, $L_{(2,5)}$ and $L_{(3,6)}$. The logic function $L_{(2,3)}$ is 0, i.e., the potential $\Phi_{(2,3)}$ is turned off, when either atom 5 is close to atom 2 or when atom 6 is close to atom 3. This is different from the motivational example which only turned off the potential if 2 and 5 were close and the bonds between 2 and 5 or 3 and 6 never formed. Additionally, if *AC* has formed (atoms 2 and 5 close), then *BC* cannot form, i.e., $L_{(3,6)}=0$ and $\Phi_{(3,6)}$ is off. Similarly, *BC* has formed (atoms 3 and 6 close), then *AC* cannot form. Table 1 captures this logic. As a general rule when determining the logic, the default state for all the potentials should be set to "on" except when encoding a specific mechanism. In this example, this corresponds to the first row of Table 1 which says that the values of all the logic functions are 1 when all of the atoms are far apart. This means that the associated potentials are turned on. This is exactly the behavior we want since the inhibitor mechanism is inherently a short range phenomena and thus we do not want the mechanism to be active when all the particles are far apart. However, since the atoms are all far apart the long-range behavior of the potential is dominant. For a Lennard-Jones or a Morse potential, the means there is a weak attraction force between the pairs of atoms.

We need to specify what is meant by "close". We assume that "close" is in this case is determined by experiments to mean being within the distances $R_{(2,5)}$ and $R_{(3,6)}$, respectively. Thus, atoms 2 and 5 are close when the elementary logic function $\ell_{(2,5),R_{(2,5)}}$ evaluates to 1 and not close when it evaluates to 0. Using the table, $L_{2,3}(\vec{\boldsymbol{x}})$ corresponding to the interaction potential $\Phi_{(2,3)}(\vec{\boldsymbol{x}})$ can be written as Equation (7).

$$\begin{aligned}
L_{(2,3)}(\vec{\boldsymbol{x}}) &= \neg\Big[\ell_{(2,5),R_{(2,5)}}(\vec{\boldsymbol{x}}) \vee \ell_{(3,6),R_{(3,6)}}(\vec{\boldsymbol{x}})\Big] \\
&= \neg\Big[\chi_{[0,R_{(2,5)})}(\|\boldsymbol{x}_5 - \boldsymbol{x}_2\|) \vee \chi_{[0,R_{(3,6)})}(\|\boldsymbol{x}_6 - \boldsymbol{x}_3\|)\Big] \\
&= \neg\Big[\chi_{[0,R_{(2,5)})}(\|\boldsymbol{x}_5 - \boldsymbol{x}_2\|) + \chi_{[0,R_{(3,6)})}(\|\boldsymbol{x}_6 - \boldsymbol{x}_3\|) \\
&\qquad - \chi_{[0,R_{(2,5)})}(\|\boldsymbol{x}_5 - \boldsymbol{x}_2\|) \cdot \chi_{[0,R_{(3,6)})}(\|\boldsymbol{x}_6 - \boldsymbol{x}_3\|)\Big] \\
&= 1 - \chi_{[0,R_{(2,5)})}(\|\boldsymbol{x}_5 - \boldsymbol{x}_2\|) - \chi_{[0,R_{(3,6)})}(\|\boldsymbol{x}_6 - \boldsymbol{x}_3\|) \\
&\qquad + \chi_{[0,R_{(2,5)})}(\|\boldsymbol{x}_5 - \boldsymbol{x}_2\|) \cdot \chi_{[0,R_{(3,6)})}(\|\boldsymbol{x}_6 - \boldsymbol{x}_3\|).
\end{aligned} \tag{7}$$

The other logic functions are

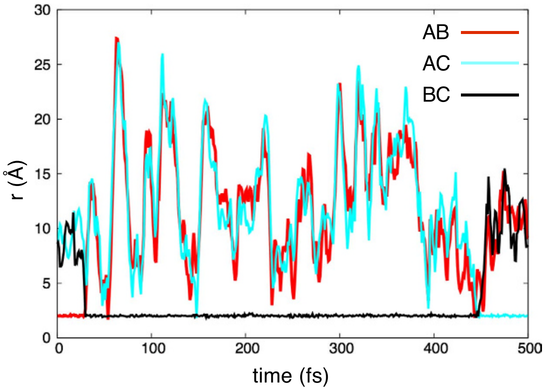


Figure 5. One realization of the inhibitor molecule system (11) simulated in LAMMPS. Initially, the **AB** complex is formed. Around 30 femtoseconds (fs) **C** comes close enough, turns off the **AB** bond and **BC** forms and can diffuse away from **A**. This remains the case until around 450 fs, when **A** approaches **BC**, the **BC** bond turns off and the **AC** bond turns on.

$$L_{(3,6)}(\vec{x}) = \neg\, \chi_{[0,R_{(2,5)})}(\|\boldsymbol{x}_5 - \boldsymbol{x}_2\|) \tag{8}$$

$$L_{(2,5)}(\vec{x}) = \neg\, \chi_{[0,R_{(3,6)})}(\|\boldsymbol{x}_6 - \boldsymbol{x}_3\|). \tag{9}$$

To turn the logic functions into an encoding function, replace each of the proximity functions, $\chi_{[0,R_p)}$, in (7–9) with their smooth versions, h_{R_p,n_p} (Eq. (6)). The encoding function $S_{(2,3)}$ corresponding to $L_{(2,3)}$ is

$$\begin{aligned} S_{(2,3)}(\vec{x}) \;&=\; 1 - h_{R_{(2,5)},n_{(2,5)}}(\|\boldsymbol{x}_5 - \boldsymbol{x}_2\|) - h_{R_{(3,6)},n_{(3,6)}}(\|\boldsymbol{x}_6 - \boldsymbol{x}_3\|) \\ &\quad + h_{R_{(2,5)},n_{(2,5)}}(\|\boldsymbol{x}_5 - \boldsymbol{x}_2\|)\, h_{R_{(3,6)},n_{(3,6)}}(\|\boldsymbol{x}_6 - \boldsymbol{x}_3\|). \end{aligned} \tag{10}$$

The approximate potential for this system is

$$U(\vec{x}) = \sum_{p\in\mathcal{I}} S_p(\vec{x})\Phi_p(\vec{x}) = S_{(2,3)}(\vec{x})\Phi_{(2,3)}(\vec{x}) + S_{(2,5)}(\vec{x})\Phi_{(2,5)}(\vec{x}) + S_{(3,6)}(\vec{x})\Phi_{(3,6)}(\vec{x}). \tag{11}$$

The original configuration space was 12-dimensional. However, (11) is 8-dimensional since it only depends on four atoms (four unique atoms making pairs in $\mathcal{I}$). Thus we were able to reduce the dimension of the configuration space and still capture the relevant physics.

Here, we are only interested in demonstrating the methodology qualitatively so we make the approximation that derivative of each encoding function is 0 almost everywhere (this would be the case if the logic functions were used in place of the encoding functions in (11). With this approximation the force only consists of terms of the form $S_p(\vec{x})\nabla_{x_p}\Phi_p(\vec{x})$. One realization of the inhibitor molecule system (11) simulated in LAMMPS[25] with this approximation is shown in Fig. 5. The potentials Φ_p are formed from Morse potentials

$$\phi_{\text{Morse}}(r) = D(e^{-2a(r-r^{eq})} - 2e^{-a(r-r^{eq})}), \tag{12}$$

where D is the dissociation energy, r^{eq} is the equilibrium distance of the bond, and a is a parameter. Simulations are performed by solving the Langevin equations at constant temperature (i.e. NVE ensemble). The parameters used in the computations are given in Supplementary Table I. Initially, the **AB** complex is formed. Around 30 femtoseconds **C** comes close enough, turns off the **AB** bond and **BC** forms and can diffuse away from **A**. This remains the case until around 450 fs, when **A** approaches **BC**, the **BC** bond turns off and the **AC** bond turns on.

Supplementary Movie 1 shows a simulation of the inhibitor molecule mechanism.

Modeling a bond breaking chemical reaction.

We model a reversible, bond breaking, chemical reaction. In particular, we will model the reaction

$$\mathbf{AB} + \mathbf{C} \rightleftarrows \mathbf{AC} + \mathbf{B}. \tag{13}$$

Modeling such chemical reactions is difficult with traditional force field methods since they cannot describe changes in the electronic structure and, thus, are unable to describe bond-breaking, bond-forming, charge transfer, etc., of the system undergoing a reaction[4,31]. Rather than solving the quantum mechanical equations, we take a coarse-level approach and approximate the bond breaking mechanism with logic functions.

This example makes use of the multiplicity function $\mathfrak{m}(\boldsymbol{i})$ in (2) in order to model the electron-electron repulsion during the transition state. It also shows that the use of a smooth encoding function accurately accounts for the bond dissociation energy. Let $\vec{x} = (\boldsymbol{x}_1, \ldots, \boldsymbol{x}_6) \in (\mathbb{R}^2)^6$ be the configuration vector for this system (see Fig. 6).

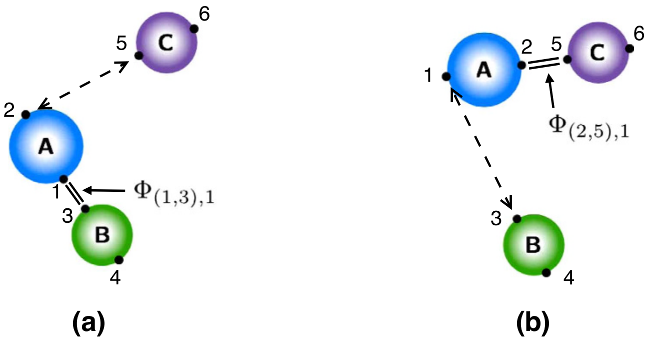


Figure 6. Diagram for chemical reaction example. $\Phi_{(1,3),1}$ and $\Phi_{(2,5),1}$ represent the stable bonds **AB** and **AC**, respectively. The dashed lines represent the repulsion forces induced by the encoding functions. (**a**) the repulsion between **A** and **C** is due to the partial derivatives of $S_{(1,3),1}(\vec{x})$ with respect to both $\boldsymbol{x}_2$ and $\boldsymbol{x}_5$. (**b**) the repulsion between **A** and **B** is due to the partial derivatives of $S_{(2,5),1}(\vec{x})$ with respect to both $\boldsymbol{x}_1$ and $\boldsymbol{x}_3$. See Supplementary Information Sec. III.A for a discussion of the repulsion force induced by the smooth encoding functions.

Potentials	Potential type	Equil. dist.	Interaction range
$\Phi_{(1,3),1}$	**AB** stable mol. bond	$r^{eq}_{\mathbf{AB}}$	—
$\Phi_{(1,3),2}$	**AB** electron repulsion term	—	$<R^{\text{dis}}_{\mathbf{AC,B}}$
$\Phi_{(2,5),1}$	**AC** stable mol. bond	$r^{eq}_{\mathbf{AC}}$	—
$\Phi_{(2,5),2}$	**AC** electron repulsion term	—	$<R^{\text{dis}}_{\mathbf{AB,C}}$

Table 2. Potentials in chemical reaction model.

Table 2 lists all the potentials involved in modeling the system. The potential $\Phi_{(1,3),1}$ is the potential energy of the bond between **A** and **B** when they form the stable molecule **AB**, for example a Lennard-Jones or Morse potential. Similarly, $\Phi_{(2,5),1}$ is the bond potential energy between **A** and **C** when they form the stable molecule **AC**. They have the associated logic functions $L_{(1,3),1}$ and $L_{(2,5),1}$, respectively. The potentials $\Phi_{(1,3),2}$ and $\Phi_{(2,5),2}$ are used to model electron-electron repulsion during the transition state when the reactant bonds have broken and the product bonds have not yet formed. The secondary potentials are usually taken to be the repulsive part of the associated bond potential.

Table 3 lists the logic rules for this system. Consider the forward reaction **AB** + **C** → **AC** + **B**. We model the bond breaking mechanism by turning off the stable bond, $\Phi_{(1,3),1}$, when **C** gets "close enough" to **A**. When **C** moves within the distance $R^{\text{dis}}_{\mathbf{AB,C}}$ to **A**, the **AB** bond ($\Phi_{(1,3),1}$) turns off. Similarly, for the backwards reaction, the **AC** bond ($\Phi_{(2,5),1}$) turns off when **B** gets within a distance $R^{\text{dis}}_{\mathbf{AC,B}}$ of **A**. The logic functions are logical NOT's of the $\boldsymbol{x}_2$–$\boldsymbol{x}_5$ and $\boldsymbol{x}_1$–$\boldsymbol{x}_3$ proximity functions:

$$L_{(1,3),1}\,(\vec{x}) = \neg\ \chi_{[0,R^{dis}_{AB,C})}(\|\boldsymbol{x}_2 - \boldsymbol{x}_5\|)$$
$$L_{(2,5),1}\,(\vec{x}) = \neg\ \chi_{[0,R^{dis}_{AC,B})}(\|\boldsymbol{x}_1 - \boldsymbol{x}_3\|).$$

We assume that $r^{eq}_{\mathbf{AB}} < R^{\text{dis}}_{\mathbf{AC,B}}$ and $r^{eq}_{\mathbf{AC}} < R^{\text{dis}}_{\mathbf{AB,C}}$. With this assumption, the **AC** bond turns off before **B** reaches its equilibrium bond length with **A**.

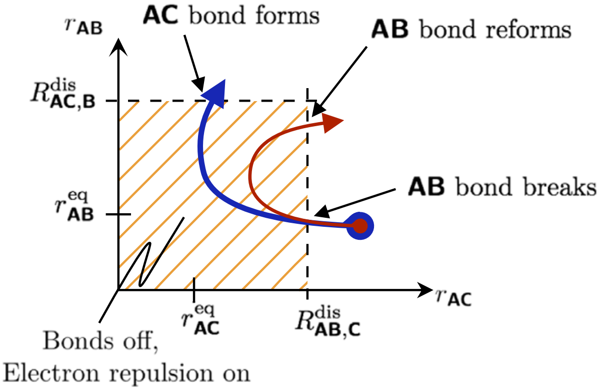
The use of the smooth encoding function in the potential (as opposed to the logic function) allows the transfer of the correct amount of energy from **C** to **AB** in order to break the bond; **C** must transfer an amount of energy equivalent to the bond dissociation energy $D_{\mathbf{AB}}$ of the **AB** bond in order to turn off the $\Phi_{(1,3),1}$ potential. We refer the reader to Sec. III.A in the Supplementary Information for the derivation.

Consider the situation occurring directly after a successful collision of **C** with **AB**. In this case, **A** and **B** are close to their equilibrium distance ($\|\boldsymbol{x}_1 - \boldsymbol{x}_3\| \approx r^{eq}_{\mathbf{AB}}$) and **A** and **C** are closer than the **AB**-bond dissociation distance ($\|\boldsymbol{x}_2 - \boldsymbol{x}_5\| < R^{\text{dis}}_{\mathbf{AB,C}}$). In this state, the bonds are weak and neither **AB** nor **AC** is stable; the system is at its transition state. In this transition state the forces experienced by the molecules due to the bond potentials $\Phi_{(2,3),1}$ and $\Phi_{(2,5),1}$ are small since the encoding functions and their partial derivatives are small, and thus the bond potentials are approximately "off". The dynamics are predominantly dominated by noise and the residual momentum of the molecules.

In this transition state, the electron-electron repulsion should be directly accounted for via a short-range repulsion potential between the molecules; usually this is repulsive part of the associated bond potential. The logic functions are defined such that these repulsion forces are only "on" when the system is in its transition state. This is easily accomplished. Denote the short-range repulsion potential between **A** and **C** by $\Phi_{(2,5),2}$. This force is defined such that $\Phi_{(2,5),2}(\vec{x}) \approx 0$, for $\|\boldsymbol{x}_2 - \boldsymbol{x}_5\| > R^{\text{dis}}_{\mathbf{AB,C}}$. This force is turned on when $\|\boldsymbol{x}_1 - \boldsymbol{x}_3\| < R^{\text{dis}}_{\mathbf{AC,B}}$. The logic function for the **A**-**C** repulsion is

$\|\boldsymbol{x}_1 - \boldsymbol{x}_3\|$	$\|\boldsymbol{x}_2 - \boldsymbol{x}_5\|$	$\Phi_{(1,3),1}$	$\Phi_{(1,3),2}$	$\Phi_{(2,5),1}$	$\Phi_{(2,5),2}$
$<R^{\mathrm{dis}}_{\mathbf{AC,B}}$	—	—	—	OFF	ON
$>R^{\mathrm{dis}}_{\mathbf{AC,B}}$	—	—	—	ON	OFF
—	$<R^{\mathrm{dis}}_{\mathbf{AB,C}}$	OFF	ON	—	—
—	$>R^{\mathrm{dis}}_{\mathbf{AB,C}}$	ON	OFF	—	—

Table 3. Bond-breaking logic rules.



$r_{\mathbf{AB}}$
AC bond forms
AB bond reforms
$R^{\mathrm{dis}}_{\mathbf{AC,B}}$
AB bond breaks
$r^{\mathrm{eq}}_{\mathbf{AB}}$
$r_{\mathbf{AC}}$
$r^{\mathrm{eq}}_{\mathbf{AC}}$
$R^{\mathrm{dis}}_{\mathbf{AB,C}}$
Bonds off,
Electron repulsion on

Figure 7. The two most probable outcomes of a successful AB + C event. Both trajectories start at the same configuration the difference is that **C** has a greater momentum for the blue (thicker) curved arrow. For both the red and blue trajectories, **C** approaches **A**. The **AB** bond breaks when $r_{\mathbf{AC}} = \|\boldsymbol{x}_2 - \boldsymbol{x}_5\| < R^{\mathrm{dis}}_{\mathbf{AB,C}}$. Depending on the momentum and the relative strength of the repulsion terms, either **AB** reforms or **AC** forms.

$$L_{(2,5),2}(\vec{\boldsymbol{x}}) = \chi_{[0,R^{dis}_{AC,B})}(\|\boldsymbol{x}_1 - \boldsymbol{x}_3\|). \tag{14}$$

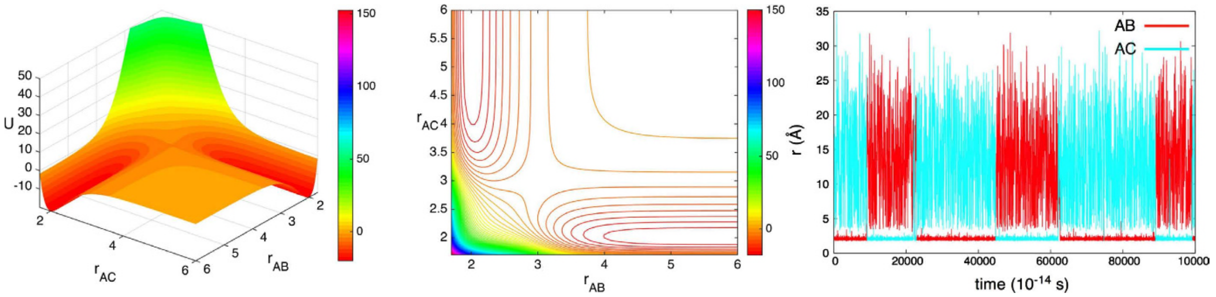
Similarly, repulsions between **A**-**B** and **B**-**C** can be defined with logic functions similar to the above one.

There are three possible outcomes for when the system exits its transition state: (i) either **AC** forms a stable molecule, (ii) **AB** reforms, or (iii) no bonds are formed and all the molecules are free molecules. This depends on the equilibrium distances of the bonds, the dissociation distances, the incoming momentum of **C**, and the repulsion forces. Figure 7 shows the two most probable outcomes for a single **AB** + **C** event.

For our simulations, the **AB** and **AC** bonds ($\Phi_{(1,3),1}$ and $\Phi_{(2,5),1}$, respectively) are given by Morse potentials, (12). In simulations, only the short-range **A**-**B** and **A**-**C** electron-electron repulsions are modeled and are only active during the transition state. The form of these for these repulsions are chosen as the repulsive part of a Morse potential with the same parameters as the full potentials used for the **AB** and **AC** bonds. The logic function for the **A**-**C** repulsion potential, $\Phi_{(2,5),2}$, is given by (14) with obvious modifications for $\Phi_{(1,3),2}$. The associated encoding functions are given by the normal replacement procedure. The full potential used during the numerical experiments is given in (15).

$$\begin{aligned} U(\vec{\boldsymbol{x}}) = & \underbrace{\left(1 - \frac{1}{1 + (\|\boldsymbol{x}_2 - \boldsymbol{x}_5\|/R^{\mathrm{dis}}_{\mathrm{AB,C}})^{2n_{\mathbf{AC}}}}\right)}_{S_{(1,3),1}} \underbrace{D_{\mathbf{AB}}(e^{-2a(\|\boldsymbol{x}_1-\boldsymbol{x}_3\|-r^{eq}_{\mathbf{AB}})} - 2e^{-a(\|\boldsymbol{x}_1-\boldsymbol{x}_3\|-r^{eq}_{\mathbf{AB}})})}_{\Phi_{(1,3),1}} \\ & + \underbrace{\left(1 - \frac{1}{1 + (\|\boldsymbol{x}_1 - \boldsymbol{x}_3\|/R^{\mathrm{dis}}_{\mathrm{AC,B}})^{2n_{\mathbf{AB}}}}\right)}_{S_{(2,5),1}} \times \underbrace{D_{\mathbf{AC}}(e^{-2a(\|\boldsymbol{x}_2-\boldsymbol{x}_5\|-r^{eq}_{\mathbf{AC}})} - 2e^{-a(\|\boldsymbol{x}_2-\boldsymbol{x}_5\|-r^{eq}_{\mathbf{AC}})})}_{\Phi_{(2,5),1}} \\ & + \underbrace{\left(\frac{1}{1 + (\|\boldsymbol{x}_2 - \boldsymbol{x}_5\|/R^{\mathrm{dis}}_{\mathrm{AB,C}})^{2n_{\mathbf{AC}}}}\right)}_{S_{(1,3),2}} \underbrace{D_{\mathbf{AB}}e^{-2a(\|\boldsymbol{x}_1-\boldsymbol{x}_3\|-r^{eq}_{\mathbf{AB}})}}_{\Phi_{(1,3),2}} \\ & + \underbrace{\left(\frac{1}{1 + (\|\boldsymbol{x}_1 - \boldsymbol{x}_3\|/R^{\mathrm{dis}}_{\mathrm{AC,B}})^{2n_{\mathbf{AB}}}}\right)}_{S_{(2,5),2}} \underbrace{D_{\mathbf{AC}}e^{-2a(\|\boldsymbol{x}_2-\boldsymbol{x}_5\|-r^{eq}_{\mathbf{AC}})}}_{\Phi_{(2,5),2}}. \end{aligned} \tag{15}$$

The force derived from (15) is used in LAMMPS[25] to simulate the system for an unbiased and a biased potential (parameters in Supplementary Information Table II). The parameters of the first simulation are chosen so that the **AB** and **AC** are symmetric ($D_{\mathbf{AC}}/D_{\mathbf{AB}} = 1$). In this case, the chemical reaction is unbiased and if averaged over all realizations of the noise, it is expected that the amount of time **AB** is formed is equal to the amount of time



(a) Potential energy surface **(b)** Potential energy level sets **(c)** LAMMPS simulation trace

Figure 8. Simulation of an unbiased (1:1 well-depth), bond breaking chemical reaction, (13). (**a**) The potential energy (15) for the system. The parameters are given under simulation 1 in Supplementary Table II. (**b**) The level set plot of the potential energy. (**c**) A typical trajectory of the simulation. The cyan trace denotes the distance between molecules **A** and **C** ($r_{\mathbf{AC}} = \|\boldsymbol{x}_2 - \boldsymbol{x}_5\|$), whereas the red trace corresponds to the distance between molecules **A** and **B** ($r_{\mathbf{AB}} = \|\boldsymbol{x}_1 - \boldsymbol{x}_3\|$). Initially, **A** and **C** are near their equilibrium length (2) and **B** is far from **A**. We see a successful $\mathbf{AC} + \mathbf{B} \rightarrow \mathbf{AB} + \mathbf{C}$ event happening very soon (red trace is close to the equilibrium distance, then becomes large; cyan trace is large, then becomes small).

AC is formed. Figure 8 shows the potential energy for this simulation (Fig. 8(a)), the corresponding level sets (Fig. 8(b)), and a typical realization of the simulation (Fig. 8(c)). In the energy surface plot and the level set plot, the symmetry of the potential is evident. The realization shown in Fig. 8(c) starts with **AB** near its equilibrium length (2 Å) with **C** far from **A**. The realization shows the approximately equal times that **AB** and **AC** are formed. The deviation is due to this being a particular realization rather than an average over an ensemble of realizations and the finite nature of the simulation.

The parameters of the second simulation are chosen so that the reaction is biased in favor of **AC**. With the chosen parameters ($D_{\mathbf{AC}}/D_{\mathbf{AB}} = 2$), the **AC** bond is twice as stable as **AB**. Figure 9 shows the potential energy for this simulation (Fig. 9(a)), its corresponding level sets (Fig. 9(b)), and a realization of the simulation (Fig. 9(c)). In the energy surface plot and the level set plot, the asymmetry of the potential is evident. The realization shown in Fig. 9(c) starts with **AB** near its equilibrium length (2 Å) with **C** far from **A**. In this particular realization **AC** forms very quickly. Figure 9(c) shows the bias towards the more stable **AC**. The system spends most of its time with a stable **AC** molecule with a relatively small amount of time with a stable **AB** molecule. Thus, biased reactions can be captured in the framework. A movie of a part of the unbiased reaction simulation can be found in Supplementary Movie 2.

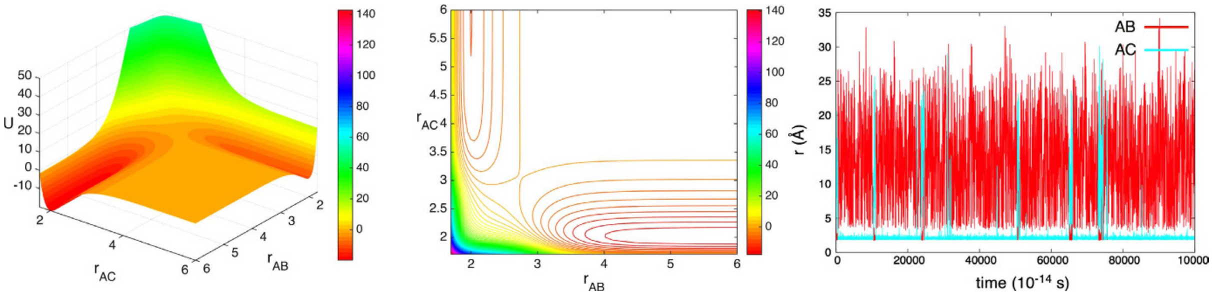
DNA transcription model. The final example is inspired by DNA transcription[27,28]. The model consists of a promoter region (sites 1 and 2) to which RNA polymerase (RNA pol) binds (sites 3 and 4), and a four nucleotide DNA strand, ACTG, to be transcribed (Fig. 10). As a first approximation of the transcription process, the movement of the RNA polymerase down the DNA chain and the unwinding/rewinding of the DNA have not been explicitly modeled.

In the absence of the RNA pol, the free nucleotides cannot bind to their complementary nucleotides in the 4 nucleotide DNA strand (ACTG = (5, 6, 7, 8)). Once RNA pol binds to the promoter, the first first nucleotide (A, atom 5) in the DNA strand can bind to the free version of its complementary nucleotide (U, atom 9). Before this binding happens, the remaining nucleotides in the strand (CTG, atoms 6, 7, 8) cannot bind with their (free) complementary nucleotides (atoms 12, 15, 18). Once A has bound to a free U nucleotide, the next nucleotide in the strand (C, atom 6) can bind with a free G nucleotide (atom 12), while the remaining two nucleotides (TG) still cannot bind with their complementary nucleotides. Once the free G has bound with C, the sugar and phosphate groups (atoms 11 and 13) on T and G can bind to start forming the backbone of the complementary DNA strand. This sequential process continues until each nucleotide in the original DNA strand ACTG has bound with its complementary nucleotide, resulting in the complementary RNA strand UGAC. At this point, the complementary strand and the RNA pol unbind from the original strand and promoter region, respectively.

Supplementary Table III lists the reaction potentials for each of the interacting pairs. The nucleotide base pairs interact via a hydrogen bond ϕ_{H}, whereas the sugar and phosphate groups covalently bond through ϕ_{SP}. The interaction potentials for the system can be easily read from this table (see Supplementary Information Sec. IV.A).

Let us step through the logic in the order the reaction occurs:

1. The bonds between the RNA pol and the promoter region ($\Phi_{(1,3)}$ and $\Phi_{(2,4)}$) are "on" except when the complementary chain has formed and is still attached to the original base strand.
2. The A-U bond ($\Phi_{(5,9)}$) is "on" when the RNA pol has bonded with the promoter and the complementary chain's backbone has not fully formed. This second condition prevents the complementary strand from reattaching to the original DNA strand once it has been formed. It is "off" otherwise.
3. The C-G bond ($\Phi_{(6,12)}$) is "on" when all of following conditions are true: (1) RNA pol has bonded with the promotor, (2) the A-U bond has formed, and (3) the complementary backbone has not formed. It is "off" otherwise.



(a) Potential energy surface **(b)** Potential energy level sets **(c)** LAMMPS simulation trace

Figure 9. Simulation of biased (2:1 well-depth), bond breaking chemical reaction, (13). (**a**) The potential energy (15) for the system. The parameters are given under simulation 2 in Supplementary Table II. (**b**) The level set plot of the potential energy. (**c**) A typical trajectory of the simulation. The cyan trace denotes the distance between molecules **A** and **C** ($r_{\mathbf{AC}} = \|\boldsymbol{x}_2 - \boldsymbol{x}_5\|$), whereas the red trace corresponds to the distance between molecules **A** and **B** ($r_{\mathbf{AB}} = \|\boldsymbol{x}_1 - \boldsymbol{x}_3\|$). Initially, **A** and **B** are near their equilibrium distance (2) and **C** is far from **A**. We see a successful $\mathbf{AB} + \mathbf{C} \rightarrow \mathbf{AC} + \mathbf{B}$ event happening very soon (red trace is close to the equilibrium distance, then becomes large; cyan trace is large, then becomes small). The trace exhibits the bias towards a stable **AC** bond, since the cyan trace is close to equilibrium longer than the red trace.

4. The sugar-phosphate group bond $\Phi_{(11,13)}$ turns "on" when (1) RNA pol has bonded with the promoter and (2) both the A-U and C-G bonds have formed. It remains "on" once the complementary backbone has formed. It is "off" otherwise.
5. The T-A bond ($\Phi_{(7,15)}$) is "on" when all of the following conditions are true: (1) RNA pol has bonded with the promotor, (2) the A-U and C-G bonds have formed, (3) the (11,13) sugar-phosphate bond has formed, and (4) the complementary backbone has not formed. It is "off" otherwise.
6. The sugar-phosphate group bond $\Phi_{(14,16)}$ turns "on" when (1) RNA pol has bonded with the promoter, (2) the A-U, C-G, and T-A bonds have formed, and (3) the (11, 13) sugar-phosphate bond has formed. It remains "on" once the complementary backbone has formed. It is "off" otherwise.
7. The G-C bond ($\Phi_{(8,18)}$) is "on" when all of the following conditions are true: (1) the conditions in (10) are true, (2) the T-A bond has formed, and (3) the (14, 16) sugar phosphate bond has formed. It is "off" otherwise.
8. The sugar-phosphate group bond $\Phi_{(17,19)}$ turns "on" when (1) RNA pol has bonded with the promoter, (2) the A-U, C-G, T-A, and G-C bonds have formed, and (3) the (11, 13) and (14, 16) sugar-phosphate bonds have formed. It remains "on" once the complementary backbone has formed. It is "off" otherwise.

A global potential derived from the above logic rules is given in Eq. (16). The exact form of the logic functions $L_p(\vec{x})$ and the associated smooth encoding functions $S_p(\vec{x})$ comprising the potential are given in the Supplementary Information. The derivation of the potential is not difficult, but lengthy. We refer the reader the Supplementary Information Sec. 4 for the details.

$$\begin{aligned} U(\vec{x}) \;=\; & \underbrace{S_{(1,3)}(\vec{x})\Phi_{(1,3)}(\vec{x}) + S_{(2,4)}(\vec{x})\Phi_{(2,4)}(\vec{x})}_{\text{RNA pol/promoter binding}} \\ & \underbrace{+\, S_{(5,9)}(\vec{x})\Phi_{(5,9)}(\vec{x})}_{\text{complementary } A-U \text{ bond}} + \underbrace{S_{(6,12)}(\vec{x})\Phi_{(6,12)}(\vec{x})}_{\text{complementary } C-G \text{ bond}} \\ & + \underbrace{S_{(7,15)}(\vec{x})\Phi_{(7,15)}(\vec{x})}_{\text{complementary } T-A \text{ bond}} + \underbrace{S_{(8,18)}(\vec{x})\Phi_{(8,18)}(\vec{x})}_{\text{complementary } G-C \text{ bond}} \\ & + \underbrace{S_{(11,13)}(\vec{x})\Phi_{(11,13)}(\vec{x}) + S_{(14,16)}(\vec{x})\Phi_{(14,16)}(\vec{x}) + S_{(17,19)}(\vec{x})\Phi_{(17,19)}(\vec{x}).}_{\text{sugar}-\text{phosphate backbone for complementary RNA strand}} \end{aligned} \tag{16}$$

Figure 11 shows a trace of the pairwise distances between atoms for a typical simulation using this potential in LAMMPS (parameters in Supplementary Information Table IV). We use the same qualitative approximation of the force as was used in the inhibitor molecule example. For simplicity, all the potentials are taken to be Morse potentials. The red trace (TF) corresponds to the distance between the RNA pol and the promoter region. The variables $r_{5,9}$ (cyan), $r_{6,12}$ (gold), $r_{7,15}$ (black), and $r_{8,18}$ (blue) correspond to the sites on the complementary A-U, C-G, T-A, and G-C pairs from the base strand and the free nucleotides. At the start, the RNA pol and the free nucleotides diffuse around in space. Around 900 ps, the RNA pol binds to the promoter region (TF trace ≈ 0). The free nucleotides then bind in the the order of the designed logic. U binds to A ($r_{5,9} \approx 0$) around 1100 ps; G binds with C ($r_{6,12} \approx 0$) between 1800 and 1900 ps; A binds to T ($r_{7,15} \approx 0$) around 2400 ps; and finally C binds to G ($r_{8,18} \approx 0$) around 2700 ps. Once this final free nucleotide has bounded with its complement, the complementary chain has finished forming and unbinds as does the RNA pol. The RNA pol can rebind to the promoter region, but the complementary RNA strand cannot rebind to the original DNA strand. This is exactly the behavior designed

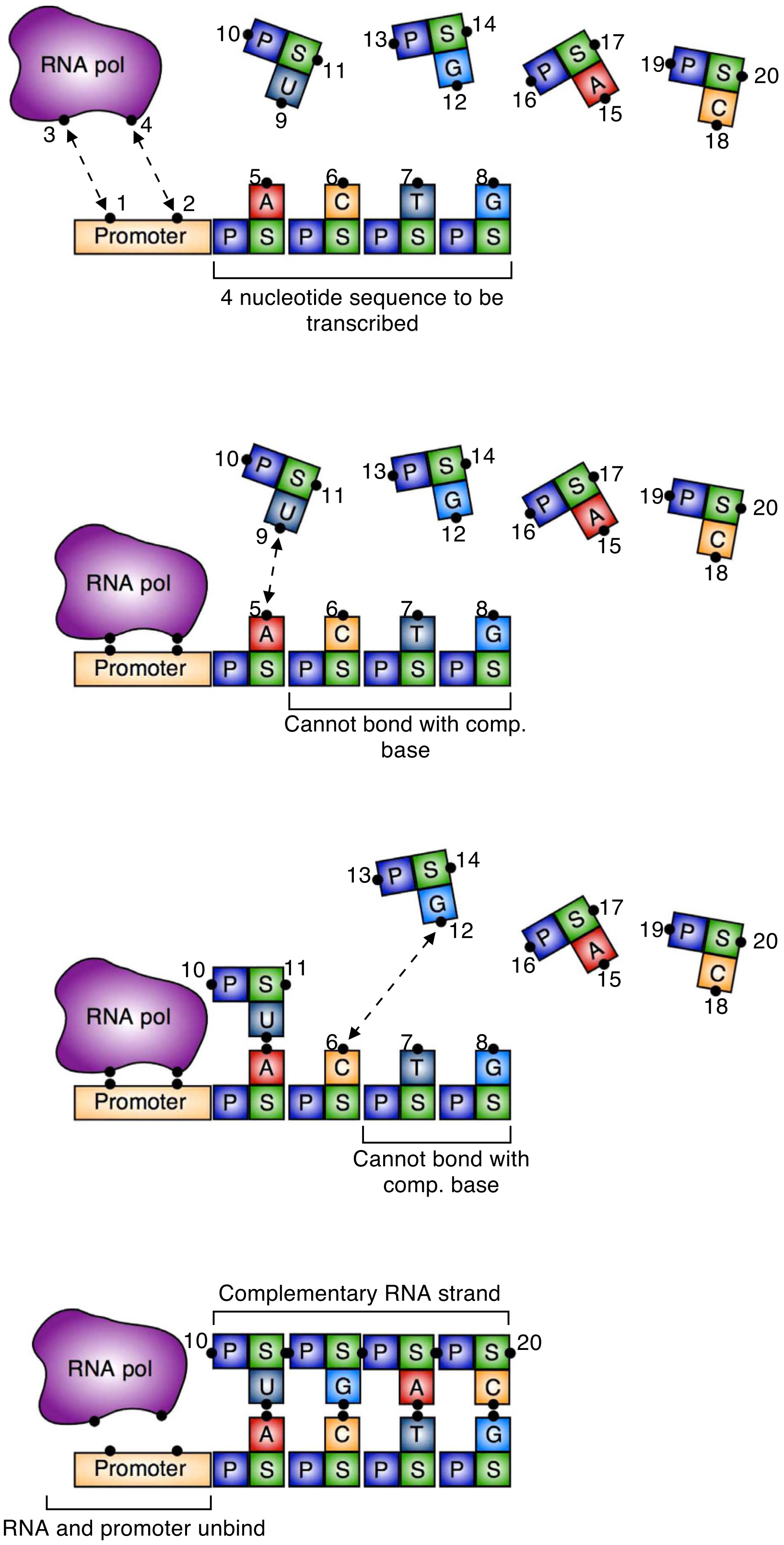


Figure 10. Simple DNA transcription model. Free base nucleotides cannot bind with the DNA strand until RNA pol binds with the promoter. When RNA pol is bound to the promoter, the free nucleotides bind to the DNA strand ACTG sequentially from left to right. Once the complementary strand has formed, the RNA pol unbinds from the promoter and then the complementary strand can diffuse away. Once the RNA pol has diffused far enough away, the bonds between the complementary base pairs turn off and the complementary strand can diffuse away. Dashed arrows between sites denotes an active potential. The P blocks denote a phosphate group and the S blocks denote a sugar group.

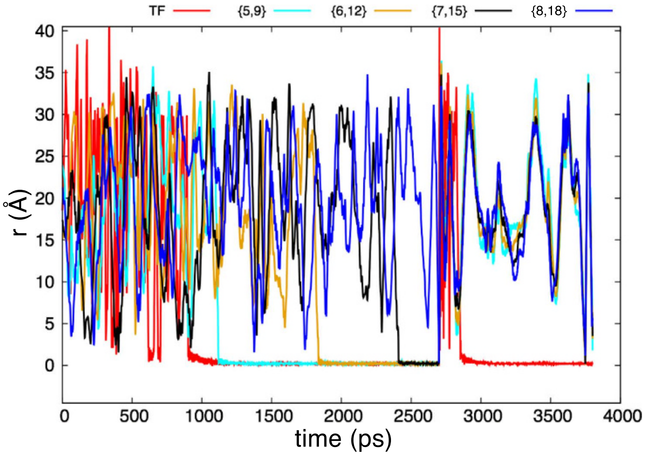


Figure 11. DNA transcription. The red trace (TF) corresponds to the distance between the RNA pol and the promoter region. $r_{5,9}$ (cyan), $r_{6,12}$ (gold), $r_{7,15}$ (black), and $r_{8,18}$ (blue) correspond to the sites on the complementary A-U, C-G, T-A, and G-C pairs from the base strand and the free nucleotides. At the start, the RNA pol and the free nucleotides diffuse around in space. Around 900 ps, the RNA pol binds to the promoter region (TF trace ≈ 0). The free nucleotides then bind in the the order of the designed logic. U binds to A ($r_{5,9} \approx 0$) around 1100 ps; G binds with C ($r_{6,12} \approx 0$) between 1800 and 1900 ps; A binds to T ($r_{7,15} \approx 0$) around 2400 ps; and finally C binds to G ($r_{8,18} \approx 0$) around 2700 ps. Once this final free nucleotide has bounded with its complement, the complementary chain is finished formed and unbinds as does the RNA pol. The RNA pol can rebind to the promoter region, but the complementary strand cannot rebind to the original DNA strand.

into the potential. Supplementary Movie 3 in the Supplementary Information shows one simulation of the DNA transcription.

Conclusions

We have developed and demonstrated a methodology and mathematical framework for obtaining an approximate interaction potential for a system which respects known coarse-level behavior. This methodology develops a semi-empirical model for the system by encoding the known coarse-level physics into logic functions that then modify simple pairwise potentials. Each logic function's only role is to turn its associated pairwise potential on or off. A smooth multi-body interaction potential is obtained by replacing each logic function with a smoothed variant. The reader may wish to think of the resulting approximate potential as a linear combination of pairwise potentials where instead of the coefficients taking scalar values, they are encoding functions capturing the coarse-level logic.

Three relatively simple examples demonstrated our methodology: a simple inhibitor molecule mechanism, a chemical reaction with bond breaking, and a model inspired by DNA transcription. While these examples were simple and inspired by biophysical and chemistry applications, we stress that the methodology is quite general and not restricted to these application domains or only simple problems. Any system that is driven by a potential can utilize this methodology to its benefit.

The result of our procedure is the approximation of a complicated, high-dimensional potential with a lower-dimensional representation that still respects the relevant physics. A significant reduction in the dimensionality of the system is possible; instead of accounting for every interaction between a large number of components, we now only need as many variables as are needed to correctly model the coarse-level logic. In the bond breaking example, the potential capturing the logic was 8-dimensional, whereas the dimension of the configurations space was 12. The same system modeled at the quantum level is much more complicated. Since the bond breaking event is the relevant physics, the reduced order model is accurate enough for this purpose.

With this dimensional reduction, the ability to accurately simulate large, complicated systems within a computational design framework is feasible. The resultant models can be wrapped in an optimization loop as part of exploratory computational experiments, such as for the development of new drug therapies, or as part of an engineering design loop. This in turn allows for the faster and cheaper development of new technologies and products.

We note that the developed framework can be potentially used in reverse: not for approximation to a given physical process with coarse-grained logic given, but for design of molecular processes with logic prescribed by a designer. This is achieved by providing to the designer the specifications of molecules that can carry the logic out.

References

1. Valastyan, J. S. & Lindquist, S. Mechanisms of protein-folding diseases at a glance. *Disease Models & Mechanisms* **7,** 9–14, doi: 10.1242/dmm.013474 (2014).
2. van Duin, A. C. T., Dasgupta, S., Lorant, F. & Goddard III, W. A. ReaxFF: a reactive force field for hydrocarbons. *J. Phys. Chem. A* **105,** 9396–9409, doi: 10.1021/jp004368u (2001).
3. Friesner, R. A. Ab initio quantum chemistry: Methodology and applications. *PNAS* **102,** 6648–6653, doi: 10.1073/pnas.0408036102 (2005).

4. Aktulga, H. M., Pandit, S. A., van Duin, A. C. T. & Grama, A. Y. Reactive Molecular Dynamics: Numerical Methods and Algorithmic Techniques. *SIAM J. Sci. Comput.* **34,** C1–C23, doi: 10.1137/100808599 (2012).
5. Gillespie, D. T. A rigorous derivation of the chemical master equation. *Physica A: Statistical Mechanics and its Applications* **188,** 404–425, doi: 10.1016/0378-4371(92)90283-V (1992).
6. Kiel, C., Yus, E. & Serrano, L. Engineering signal transduction pathways. *Cell* **140,** 33–47, doi: 10.1016/j.cell.2009.12.028 (2010).
7. Laub, M. & Goulian, M. Specificity in two-component signal transduction pathways. *Annual Review of Genetics* **41,** 121–145, doi: 10.1146/annurev.genet.41.042007.170548 (2007).
8. Inui, M., Martello, G. & Piccolo, S. Microrna control of signal transduction. *Nat Rev Mol Cell Biol* **11,** 252–263, doi: 10.1038/nrm2868 (2010).
9. Yarden, Y. & Sliwkowski, M. X. Untangling the erbb signalling network. *Nature Reveiws Molecular Cell Biology* **2,** 127–137, doi: 10.1038/35052073 (2001).
10. Sako, Y., Minoghchi, S. & Yanagida, T. Single-molecule imaging of egfr signalling on the surface of living cells. *Nat Cell Biol* **2,** 168–172, doi: 10.1038/35004044 (2000).
11. Berger, B., Shor, P. W., Tucker-Kellogg, L. & King, J. Local rule-based theory of virus shell assembly. *Proceedings of the National Academy of Sciences* **91,** 7732–7736 (1994).
12. Berger, B., King, J., Schwartz, R. & Shor, P. Local rule mechanism for selecting icosahedral shell geometry. *Discrete Applied Mathematics* **104,** 97–111, doi: 10.1016/S0166-218X(00)00187-6 (2000).
13. Schwartz, R., Shor, P. W., Prevelige, P. E. & Berger, B. Local rules simulation of the kinetics of virus capsid self-assembly. *Biophysical journal* **75,** 2626–2636, doi: 10.1016/S0006-3495(98)77708-2 (1998).
14. Klavins, E., Ghrist, R. & Lipsky, D. A grammatical approach to self-organizing robotic systems. *Automatic Control, IEEE Transactions on* **51,** 949–962, doi: 10.1109/TAC.2006.876950 (2006).
15. Klavins, E. Programmable self-assembly. *Control Systems, IEEE* **27,** 43–56, doi: 10.1109/MCS.2007.384126 (2007).
16. Whitesides, G. M. & Grzybowski, B. Self-assembly at all scales. *Science* **295,** 2418–2421, doi: 10.1126/science.1070821 (2002).
17. Whitesides, G. M. & Boncheva, M. Beyond molecules: Self-assembly of mesoscopic and macroscopic components. *PNAS* **99,** 4769–4774, doi: 10.1073/pnas.082065899 (2002).
18. Vlasov, Y. A., Bo, X.-Z., Sturm, J. C. & Norris, D. J. On-chip natural assembly of silicon photonic bandgap crystals. *Nature* **414,** 289–293, doi: 10.1038/35104529 (2001).
19. Gracias, D. H., Tien, J., Breen, T. L., Hsu, C. & Whitesides, G. M. Forming electrical networks in three dimensions by self-assembly. *Science* **289,** 1170–1172, doi: 10.1126/science.289.5482.1170 (2000).
20. Licata, N. A. & Tkachenko, A. V. Errorproof programmable self-assembly of dna-nanoparticle clusters. *Phys. Rev. E* **74,** 041406, doi: 10.1103/PhysRevE.74.041406 (2006).
21. Valentine, M. T. & Gilbert, S. P. To step or not to step? how biochemistry and mechanics influence processivity in kinesin and eg5. *Current Opinion in Cell Biology* **19,** 75–81, doi: 10.1016/j.ceb.2006.12.011 (2007).
22. Thakur, G. S. *Encoding Information in Coarse Grain Models for Self-Assembling Systems.* Ph.D. thesis, University of California, Santa Barbara (2011).
23. Konnov, A. A. Remaining uncertainties in the kinetic mechanism of hydrogen combustion. *Combustion and Flame* **152,** 507–528, doi: 10.1016/j.combustflame.2007.10.024 (2008).
24. Hong, Z., Davidson, D. F. & Hanson, R. K. An improved H2/O2 mechanism based on recent shock tube/laser absorption measurements. *Combustion and Flame* **158,** 633–644, doi: 10.1016/j.combustflame.2010.10.002 (2011).
25. Plimpton, S. Fast parallel algorithms for short-range molecular dynamics. *J. Comput. Phys.* **117,** 1–19, doi: 10.1006/jcph.1995.1039 (1995).
26. Jensen, F. *Introduction to Computational Chemistry*. 2nd ed. (John Wiley & Sons, 2007).
27. Cramer, P. *et al.* Architecture of RNA Polymerase II and Implications for the Transcription Mechanism. *Science* **288,** 640–649, doi: 10.1126/science.288.5466.640 (2000).
28. Hahn, S. Structure and mechanism of the RNA polymerase II transcription machinery. *Nat Struct Mol Biol* **11,** 394–403, doi: 10.1038/nsmb763 (2004).
29. Arnold, V. *Mathematical methods of classical Mechanics.* 2nd ed., vol. 60 of *Graduate Text in Mathematics* (Springer-Verlag, 1989).
30. Katznelson, Y. An Introduction To Harmonic Analysis. 3rd ed., Cambridge Mathematical Library (Cambridge University Press, 2002).
31. Lin, H. & Truhlar, D. G. QM/MM: what have we learned, where are we, and where do we go from here? *Theoretical Chemistry Accounts* **117,** 185–199, doi: 10.1007/s00214-006-0143-z (2007).

Acknowledgements

Ryan Mohr and Igor Mezić received funding from Army Research Office grant ARO-MURI W911NF-14-1-0359.

Author Contributions

The original conceptualization of the idea is due to G.S.T. and I.M. The precise formulation of the concept is due to G.S.T., I.M. and R.M. The details of mathematical formulation is due to R.M. and numerical simulations were done by G.S.T. The majority of the writing of the manuscript was done by R.M. with G.S.T. contributing. All authors discussed the results and commented on the manuscript at all stages.

Additional Information

Supplementary information accompanies this paper at http://www.nature.com/srep

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

How to cite this article: Thakur, G. S. *et al.* Programmable Potentials: Approximate N-body potentials from coarse-level logic. *Sci. Rep.* **6**, 33415; doi: 10.1038/srep33415 (2016).

This work is licensed under a Creative Commons Attribution 4.0 International License. The images or other third party material in this article are included in the article's Creative Commons license, unless indicated otherwise in the credit line; if the material is not included under the Creative Commons license, users will need to obtain permission from the license holder to reproduce the material. To view a copy of this license, visit http://creativecommons.org/licenses/by/4.0/

© The Author(s) 2016