UNIVERSITY OF CALIFORNIA
RIVERSIDE

Simulating Chemical Processes From Brownian Diffusion to Binding Thermodynamics

A Dissertation submitted in partial satisfaction
of the requirements for the degree of

Doctor of Philosophy

in

Chemistry

by

Timothy Cholko

June 2021

Dissertation Committee:
Dr. Chia-en A. Chang, Chairperson
Dr. De-en Jiang
Dr. Leonard Mueller

The Dissertation of Timothy Cholko is approved:

Committee Chairperson

Acknowledgements

I would like to acknowledge my Ph.D. advisor, Chia-en Chang, for encouraging me push my boundaries and learn as much as possible. Over the past five years I was intellectually challenged nearly beyond my limits, and this caused me to grow tremendously.

Graduate school forced me to reconsider my definition of “hard work”. I completed challenges that felt insurmountable. In some cases, this required working far harder and longer than I had on anything in my life before. Going to such lengths and ultimately overcoming the challenges has been one of the most valuable growing experiences I’ve ever had. I thank Dr. Chang for helping me along the way and for creating an environment that facilitated my growth.

I thank all my collaborators from whom I learned a great deal and who contributed to my success as a scientist. I also thank my fellow group members. Some contributed directly to my research, but I thank all who contributed with their support and friendship over the years.

The text of this dissertation, in part, is a reprint of the material as it appears in the following publications:

- Cholko T, C-E Chang. Modeling Effects of Surface Properties and Probe Density for Nanoscale Biosensor Design: A Case Study of DNA Hybridization Near Surfaces **2021** *J. Phys. Chem. B* 125, 1746.

- Cholko T, Chen W, Tang Z. A Molecular Dynamics Investigation of CDK8/CycC and Ligand Binding: Conformational Flexibility and Implication in Drug Discovery **2018** *J Comp. Aid. Mol. Des.* 32, 671.
- Bosken YK, Cholko T, Lou YC, Wu KP. Insights into dynamics of inhibitor and ubiquitin-like protein binding in SARS-CoV-2 papain-like protease **2020** *Front. Molec. Biosci.* 7, 174.

Dedication

I dedicate this dissertation to my parents. They endowed my with the tools to think for myself and taught me to be curious and analytical. Without those abilities I could not have made it through this. Thank you for your constant love and support.

ABSTRACT OF THE DISSERTATION

Modeling Molecular Recognition From Brownian Diffusion to Binding Thermodynamics

by

Timothy Cholko

Doctor of Philosophy, Graduate Program in Chemistry
University of California, Riverside, June 2021
Dr. Chia-en Chang, Chairperson

Molecular recognition is a fundamental part of chemical processes, especially those relevant to biology. It refers to the process by which two molecules diffuse and eventually bind with one another to form a complex. This can be broken down into to broad aspects: kinetics and thermodynamics. Kinetics refers to the motion of molecules and the rates of their reactions with each other. Thermodynamics refers to the transfers of energy that drive the reaction when molecules bind together. The work in this dissertation uses computational methods to study both aspects of molecular recognition in a range of systems, and it includes the application of existing methods and development of new tools for simulation and analysis.

A strong focus is given to protein-ligand systems, in which the ligand is an inhibitory drug designed to shut down function of its target protein. The concept of developing drugs (inhibitors) that bind with and disrupt the activity of their targets is the basis for much of modern medicine, and has had incredible success. The development of more effective inhibitors is a constant challenge. The physical and chemical principles that predict an inhibitor's effectiveness have a complex interplay, and an understanding of these principles is challenging and highly sought after. Two projects described here use molecular dynamics (MD) simulations to elucidate these principles through better understanding inhibitor binding thermodynamics. These techniques are applied to a host of inhibitors for the carcinogenic CDK8 protein and to inhibitors of a protease (PLpro) of the recent SARS-CoV2 virus. Novel inhibitors of PLpro are also developed and validated.

The other work described herein studies molecular binding kinetics in both natural and engineered systems. Brownian dynamics simulation software is described which has been developed by the author and other group members. Its goal is to provide a robust tool with which researchers can study molecular recognition and association in a range of systems under varied conditions. This program, called GeomBD3, has been applied to study association kinetics and mechanisms in enzyme bioconjugates, protein-ligand systems, and nucleic acid biosensors. These studies are included in this dissertation, and we thus demonstrate that Brownian dynamics simulations can aid in rational bio/chemical engineering design efforts and supplement experimental analysis.

Table of Contents

Chapter 1: Introduction

Chapter 7: Conclusion and Future Work

List of Tables

Chapter 2:

Chapter 3:

Chapter 5:

Chapter 6:

List of Figures

Chapter 1:

Chapter 2:

Chapter 3:

Chapter 4:

Chapter 5:

Chapter 6:

Chapter 1: Introduction

1.1 Overview of Select Computational Chemistry Methods

The goal of computer simulations is to generate conformations of a system from which its dynamic properties or state functions can be predicted. Achieving this usually involves some kind of algorithm to propagate the system forward into new conformational states and a potential energy function to evaluate at each state. Various ways of generating the new conformational states and modeling the systems in their environment serve as the primary distinctions between several widely used classes of simulation. Such simulation methods have a range of uses from elucidating molecular recognition processes happening in novel engineered systems, to accurately calculating the free energy change of a protein-drug binding process. This breadth of application places the field of computational chemistry at the confluence of biology, chemistry, pharmacology, and materials science.

One of the simplest approaches taken by researchers to predict the binding affinity of two molecules is known as molecular docking. Typically, docking is used to predict the binding of a small molecule, or ligand, to a larger receptor molecule, such a protein. However, it can also be scaled-up to predict protein-protein binding, for example. Docking is extremely widely used in the pharmaceutical industry as a starting point for drug discovery and optimization.[1, 2] The primary goal is to find the preferred binding site and binding orientation, also called the binding pose, of the drug in its target protein.[3] The ideal pose is determined based on a docking "score", the outcome of scoring function that consists of many additive terms. Usually scoring functions are based on a combination of the drug-protein interaction energy, changes in internal energy of the two, and solvation energy. The robustness and accuracy of docking methods varies greatly, and one should consider the ultimate goal of the process in selecting a method. Often, it is necessary to identify potential “hits”, or candidate drugs, from an initial pool of thousands. In such a case, docking methods such as shape complementarity, where the binding potential is determined solely based on steric considerations can be used. Once promising drug scaffolds have been identified, fragment-based optimization may begin, in which a scaffold is altered one small fragment at a time, and evaluated with detailed docking simulation each time to reach an optimal result. Although docking has become firmly entrenched in industry as an indispensable tool, it is inadequate when high accuracy is required, and is incapable of simulating dynamic processes.

To simulate dynamics, or the sequential changes in a system over time, significantly more complicated methods are needed. One such method is Brownian dynamics (BD) simulations. BD simulations are generally most useful for studying long-range dynamic processes taking place on microsecond to second timescales.[4] This includes molecular recognition, the process by which a molecule diffuses under the influence of random forces from the solvent and intermolecular forces from other solutes to eventually bind with a partner. Molecular recognition is relevant in any system in which the association of two key molecules is critical for function. In particular, these might include protein-drug, enzyme-substrate, or target-probe encounters. In the realm of computer simulation, spatial scales of 10^{-6}-10^{-7} m and temporal scales of milliseconds (10^{-3} s) – both of which are often reached in BD simulations – are considered quite long. Studies at these scales therefore require a reduction in detail compared to most other chemical simulation methods. The reduction can be achieved most easily by choosing a simplified or coarse-grained molecular model and description of the ambient environment. Despite the necessary loss of detail, BD simulations provide valuable insight into critical processes with temporal and spatial resolution that is difficult or impossible to achieve experimentally.

Perhaps the most well-known of the methods that will be discussed here is molecular dynamics (MD) simulation. In an MD simulation, a system is propagated through time by generating a series of subsequent conformations to yield what is known as a trajectory of the system. After obtaining an initial system conformation, often from an experimentally determined structure, subsequent system conformations are generated by assigning velocities to all constituent parts of the system and allowing it to evolve under the influence of some potential energy function. Potential energy functions describing chemical systems are often called force fields, and their development and improvement is constantly ongoing. Standard force fields like AMBER[5] or Charmm[6] attempt to fully describe molecules in terms of their bond lengths, angles, and dihedrals in addition to electrostatic and van der Waals forces arising between non-bonded atoms. The level of detail in an MD simulation can vary dramatically, but typically includes an explicit representation of all atoms in the solute molecules as well as explicit solvent molecules. The applications of MD simulations cover a similarly wide range. Proteins may be the most widely studied macromolecules. Their evolution over time in a simulation can generate millions of unique conformations which can aid in the understanding of their role in a myriad of vital chemical reactions. The same logic applies to study of drug-bound proteins, which has led to MD simulation becoming a workhorse in the field of medicinal chemistry, particularly as a drug discovery and optimization tool. However, applications reach far beyond protein systems to include…. The level of detail in a typical MD simulation relegates them generally to time scales less than a few microseconds for system sizes of 10^5-10^6 atoms. Many methods exist for extending the accessible timescales or conformation sampling of molecular dynamics simulations. Among these are established techniques like accelerated molecular dynamics (aMD),[7] steered MD[8] and metadynamics,[9] which increase the likelihood of system crossing high energy barriers. The technique known as milestoning partitions a process of interest by placing milestones along a reaction coordinate. Many short trajectories are initiated from the milestones and stitched together to elucidate a longer process.[10, 11]

The following chapters detail several projects which employed a handful of different simulation and molecular modeling methodologies. These methods were used to study chemical phenomena ranging from binding kinetics to protein-drug complex free energy calculations. First, a brief description of the theory underlying chemical simulations is given, from statistical mechanics to specific simulation algorithms. Next, common simulation protocols and molecular models are discussed in detail. Finally, there is a discussion of conventional simulation data analysis methods.

1.2 Statistical Mechanics

The theoretical framework from which one can calculate system properties from simulation data is given by statistical mechanics.[12, 13] In the statistical mechanics of a classical system, the macroscopic properties of a system depends on its microstate, that is, the states of all its smallest constituent parts. In simulations, these constituent parts are usually the individual atoms. In particular, the energies of the atoms, which depend on their position *r* and momentum *p*, are required to determine their microstate, and in turn, the macrostate of the system as a whole. Experimentally, the macrostate of a system would be obtained from a measurement taken over a period of time, so the value is an average over the time of the measurement. However, the data produced by simulations are instantaneous snapshots of system. A vast number of these snapshots represents an *ensemble* of system configurations. According to the *ergodic hypothesis*, the average over this ensemble equals the time average, thus thermodynamic properties can be measured from the series of configurations generated in simulations. Statistical mechanics calculates these properties as ensemble averages of a huge number of different macrostates of the system

$$\langle A \rangle = \iint A(p^N, r^N)\rho(p^N, r^N)dp^N dr^N \qquad (1.1)$$

The angle brackets indicate the ensemble average, A is the property being calculated, and ρ is the probability density of the ensemble. Boltzmann statistics is another central concept of statistical mechanics that comes up in computational chemistry frequently. In a simulation, each configuration of the system in the generated ensemble has an energy, and the configurations sampled should follow the Boltzmann distribution if the dynamics are physically realistic.[12] The Boltzmann distribution is

$$p_i = \frac{1}{Q} e^{\frac{-E_i}{k_B T}} \tag{1.2}$$

where p_i is the probability of state i occurring, E_i is the total energy of state i, and k_B is Boltzmann's constant . Essentially, it says that the probability of a state occurring decreases exponentially as the state's energy increases.[14] The Q in the denominator is called the partition function, which is given as

$$Q = \iint e^{\frac{-E_i}{k_B T}} \, dp \, dr \tag{1.3}$$

The partition function relates the macroscopic properties of a system to its microstates, and tells us how the probabilities of the microstates are divided up based on their energies. The partition function depends on the conditions of the system e.g., constant pressure, constant temperature, constant volume, etc. Since simulations only explicitly model an extremely tiny piece of the real world, the surroundings and their effect on the system needs to be taken into account. The concept of considering the simulated system and the way it interacts with its "surroundings" is known as the thermodynamic ensemble. In practice, most simulations are performed in the isothermal-isobaric (NPT) ensemble, in which the number of particles, pressure, and temperature are held constant. This way, the system can exchange heat with the environment by use of a thermostat,[15, 16] and the simulation mimics real-world chemical reactions more closely than other ensembles would allow.[17] In simulations, a property called the free energy F is often of interest. The free energy can be used to predict the favorability of a reaction, such as a drug binding to a protein. Moreover, the relative free energies of binding are an excellent way of ranking the affinities of a set of drug candidates. In statistical mechanics, the free energy expressed as

$$F = -k_B T \ln \int \exp\left(\frac{-E(x)}{k_B T}\right) dx = -k_B T \ln Q \tag{1.4}$$

The relative free energy of two states A and B can therefore be represented as

$$\frac{F_B}{F_A} = -k_B T \ln \frac{Q_B}{Q_A} \tag{1.5}$$

which is a common relation used in relative free energy methods such as Thermodynamic Integration. The free energy can be broken down into enthalpy (potential energy) and entropy. Entropy changes in a system that accompany a reaction are important for understanding the favorability and probability of a given state. The entropy can be defined in terms of the free energy as

$$S = \frac{\langle E \rangle - F}{T} \tag{1.6}$$

or in terms of probability as

$$S = k_B \int p \ln p \, dx \tag{1.7}$$

where p is the probability of a given macrostate of the system. These first equation shows that entropy is proportional to the difference between the total energy E and the free energy F and to the temperature. The second shows that entropy is proportional to the probability of a state, or the number of ways in a which a state can be occupied.

Physically, entropy can be related to an increase or decrease of a system's degrees of freedom.

1.3 Simulation Algorithms

MD simulations allow for observation of system properties or dynamics by generating a vast number of conformations that are sequential in time. To do this, one first provides initial coordinates of the system. Then, an algorithm to propagate it forward in time is employed successively until the desired property or dynamic process can be observed. The algorithm in MD simulations typically starts with Newtons second law

$$F_i = m_i a_i \tag{1.8}$$

Where *F* is the force acting on atom i, *m* is the mass of atom i, and *a* is the acceleration of atom i. The force on an atom can equivalently be related to the gradient of the potential energy of atom i

$$F_i = -\frac{dU}{dr} \tag{1.9}$$

Combining equations 1.8 and 1.9, we have

$$-\frac{dU}{dr} = m_i a_i \tag{1.10}$$

This relationship shows that the acceleration of atom i can be calculated from the potential energy gradient. The velocity *v* and position *r* of atom i have the following relationships with acceleration

$$a = \frac{dv}{dt} \tag{1.11}$$

$$v = \int a\, dt = at + v_0 \tag{1.12}$$

$$v = \frac{dr}{dt} \tag{1.13}$$

$$r_{i+1} = \int v\, dt = \frac{at^2}{2} + v_i t + r_i \tag{1.14}$$

The general algorithm for propagating a system forward in time is then to find the acceleration of the atom from the difference in potential energy of the current configuration and the previous one, update the velocity of the atom, move the atom to its next position $r_{t+\Delta t}$ according to that velocity, and repeat over all atoms in the systems for as many steps as needed.

$$r_{t+\Delta t} = -\frac{1}{2m_i}\frac{dU}{dr}t^2 + v_t t + r_t \tag{1.15}$$

Different methods of employing this same general algorithm have been developed,[18-20] and details of each will not be discussed. However, all have the pitfall that the acceleration is held constant during a finite user-defined simulation timestep. This introduces some fundamental error in all molecular simulations. To avoid non-physical dynamics, one must ensure that the timestep is much smaller than the fastest degree of freedom in the system. For example, a bond stretch involving hydrogen happens on the order of 10^{-14} s. Following a widely accepted rule-of-thumb, the simulation timestep should be no more than 10^{-15} s, or an order of magnitude shorter. Constraining bond stretches involving hydrogen is often permissible and still yields accurate simulations while eliminating the fastest degree of freedom. As a result, common MD simulation timesteps are between 1-4 femtoseconds. With the choice of a timestep Δt, in practice, each subsequent position of an atom is calculated as

$$r_{t+\Delta t} = -\frac{1}{2}a\Delta t^2 + v_t\Delta t + r_t \tag{1.16}$$

Repeating this algorithm many times produces a molecular dynamics trajectory. It is a trajectory because all configurations of the system are generated deterministically and are sequential in time. By this algorithm a given set of initial coordinates and velocities will always result in the same trajectory. In practice, simulations of a system are often repeated using different initial atom velocities and/or slightly different initial coordinates to ensure adequate conformational sampling. A typical modern MD simulation will repeat these steps until the length reaches tens or hundreds of nanoseconds. Such time scales are enough to fully sample a small drug molecule's binding mode in a protein, for example.

Brownian dynamics (BD) simulations are generally aimed at understanding long time- and spatial-scale processes, such as molecular recognition happening across tens or hundreds of nanometers. Such a simulation typically has to be performed in an implicit solvent to be tractable. One therefore needs to use a dynamics algorithm that can incorporate the solvent's effect on molecular transport implicitly. The Langevin equation of motion is commonly used for this purpose. It is partly similar to Newton's equation of motion used in MD simulations, except that additional terms are included to model solvent effects. The Langevin equation is

$$ma = -\nabla U(r) - \gamma v + \sqrt{2\gamma kT}R \qquad (1.17)$$

where *m* is the particle's mass, *a* is the acceleration*, U(r)* is the interparticle potential, *v* is the particle velocity, *k* is the Boltzmann constant, *T* is the temperature, and γ is the viscosity of the solvent, also called the damping coefficient. *R* is zero-mean, stationary Gaussian process imparting random forces on the diffusing particle that causes changes to its direction due to "collisions" with solvent molecules. As γ grows, the dynamics enter the non-inertial, or Brownian regime, where the force on the particle at any step has no correlation with any previous step. This is also called over-damped Langevin dynamics, where collisions with solvent molecules are extremely frequent, causing the solute particle to change direction rapidly.[21]

1.4 Molecular Mechanics

Force fields, the set of functions which describe to potential energy of a system, are at the heart of all chemical simulations. As described previously, they are central to the algorithm for generating new system conformations. As such, their accuracy has immense influence on the ability of simulations to reproduce and predict results of real chemical systems.[22-24] Typically, force fields describe the potential energy of the whole system by calculating interaction energies between pairs or sets of bonded atoms, and between pairs of non-bonded atoms. Such description of molecular systems is often called molecular mechanics. The bonded atom energy terms include bond stretches, bond angles, dihedral angles between three bonded atom pairs, and improper angles between three bonded atom pairs. The bond stretching terms represents the changing energy between a pair of bonded atoms as the bond stretches and contracts

$$E_{bond} = \frac{1}{2}k(x_0 - x)^2 \tag{1.18}$$

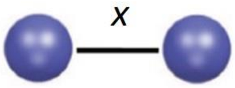


Figure 1.1 A bond distance represented by variable x. Blue balls are atoms.

where k is a force constant and x_0 and x are the current and equilibrium bond lengths. Similarly, bond angle terms represent the energy as the angle between three sequentially bonded atoms changes

$$E_{angle} = \frac{1}{2}k(\theta_0 - \theta)^2 \tag{1.19}$$

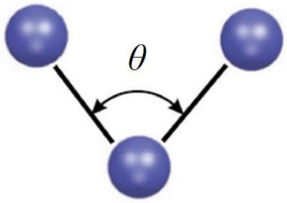


Figure 1.2. A bond angle between three atoms (blue balls) shown by the arrow.

where k is a force constant and θ_0 and θ are the current and equilibrium bond angles. Dihedral angles between a set of three bonded atom pairs are critical for describing the potential energy of differing molecular geometries (Fig XX). A simple example is the difference between a *cis* or *trans* conformation of 2-butene. The dihedral angle energy for any set of three bonded atom pairs is calculated as

$$E_{dihedral} = \frac{1}{2}V_d(1 + cos(n\theta - \theta_0))^2 \tag{1.20}$$

where V_d is the barrier height, n give the periodicity of the function in a 360° rotation, θ is the measured angle and θ_0 is an angle dictating where minima occur.

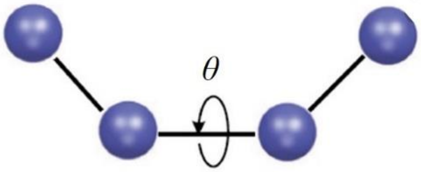


Figure 1.3 Dihedral angle made up of four atoms. As the bond indicated by the arrow rotates, the angle between the plane containing the first three atoms from either end changes.

Another form of dihedral, called an improper dihedral, is used to describe the energy related to planarity (and deviations thereof) of a set of four bonded atoms, where three outer atoms are all bonded to the single central atom but not to each other (Fig XX). A good example is the sp^2 hybridized carbon in a protein backbone. This energy is often calculated using the same function as regular dihedrals above

$$E_{improper} = \frac{1}{2}V_i(1 + cos(n\theta - \theta_0))^2 \tag{1.21}$$

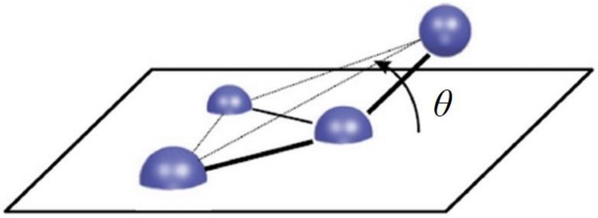


Figure 1.4 An improper angle between four atoms is the angle between two planes containing three of the atoms, as shown by the solid square plane and dashed triangular plane.

In addition to these four potential energy terms representing the interactions between bonded atoms, interactions between non-bonded atoms must also be considered. These are described by two additional terms: one to capture electrostatic interactions, and one to capture very close-range attractive and repulsive forces. Electrostatic interaction energy is described by the Coulomb potential

$$E_{Coulomb} = \frac{q_1 q_2}{4\pi\varepsilon_0 \varepsilon r} \tag{1.22}$$

where q represents the charge on either of the two interacting atoms, r is the distance between the atoms, ε_0 is the electric permittivity of free space, and ε is the dielectric constant of the surrounding medium or solvent. The Lennard-Jones (LJ) potential [ ] is the functional form used to capture the potential energy due to attractive dispersion forces and repulsive orbital overlap between two atoms. It is a simplified model which has the form

$$E_{LJ} = 4\varepsilon\left[\left(\frac{\sigma}{r}\right)^{12} - \left(\frac{\sigma}{r}\right)^{6}\right] \tag{1.23}$$

where r is the distance between two atoms, ε is the depth of the potential well, and σ is the distance at which the potential energy is zero. In a typical implementation, atoms repel strongly as their van der Waals radii being to overlap, attract each other at moderate distances, and have virtually no interaction at distances beyond about 12 Å. Together the bonded and non-bonded potential energy terms make up a full molecular mechanics force field that calculates the potential energy of a system being simulated. Modifications exists, but the preceding general framework is used in almost all well-known MD simulation programs, including Amber[5, 25], CHARMM,[6] GROMOS,[26, 27] NAMD,[28] and many others.

1.5 Solvent and Ion Models

Depending on the properties or processes of interest in the simulation, an accurate model of the solvent is often just as important as the model of the solute. By far the most common solvent in which a simulation takes place is water. Hence, enormous effort has gone toward developing effective and computationally efficient models of water molecules. The most detailed solvent model in simulation is known as explicit solvent, in which the solute is “solvated” or surrounded with atomic models of water molecules. A plethora of models exist, all of which have their own advantages and pitfalls.[29, 30] The simplest and most commonly used models, such as TIP3P, TIP4P, and SPC/E, model water as three to four point charges.[31, 32] The molecule is rigid, vastly reducing the number of degrees of freedom of the system and having little impact for most uses. However, a number other water force fields exist which model the atomic charges differently or allow the O-H bond lengths and H-O-H angle to change.[33, 34] In explicit solvent simulations, the number of water molecules can be enormous, easily exceeding 10^6 in some cases. This drastically increases the time needed to calculate the potential energy of the entire system. Hence, many *implicit* water models have been developed, in which the influence of water is represented only mathematically. This usually comes in the form of random forces acting on the solute which mimic “collisions” with water molecules, and a dielectric screening parameter. Two popular implicit solvent models are the Poisson-Boltzmann (PB) and the Generalized Born (GB) models. The PB method calculates the electrostatic potential in a solvated system by modeling the water as a high dielectric medium.[35] The effects of ions in solution are also built into the PB equation

$$\nabla \varepsilon \nabla \phi = 4\pi\rho - 4\pi \sum_{i} e z_i c_i \exp\left(-e z_i \phi / kT\right) \tag{1.24}$$

where ε is the dielectric, ϕ is the electrostatic potential, ρ is the solute charge density, z is the charge of ion type i, c is the density of ion type i, k is the Boltzmann constant, and T is the absolute temperature. The PB equation can accurately calculated the electrostatics, but it comes at a high computational cost. It is therefore rarely used in actual simulations, and is more commonly used for analysis of simulation data, such as single-point energy calculations, or calculations on a small number of simulation trajectory frames. The GB implicit solvent model is computationally efficient enough to be used during actual dynamic simulations.[36, 37] It is an approximation of the PB equation, which models atoms as hard spheres with a dielectric lower than the surrounding implicit solvent. The degree of electrostatic screening experienced by each atom is embodied in a parameter termed the Born radius.[38] The GB electrostatic interaction is

$$E_{GB} = -\frac{1}{2}\sum_{ij}\frac{q_i q_j}{f_{GB}}\left(1-\frac{e^{-\kappa f_{GB}}}{\varepsilon}\right) \tag{1.25}$$

where q is the magnitude of charge i or j, f_{GB} is a function of the Born radii, κ is the Debye-Huckel screening parameter, and ε is the dielectric constant.

Biological processes and reactions *in vitro* take place in environments with significant dissolved salt concentrations. It is therefore imperative for simulations to be able to model the effects of varying salt concentrations. Much like the aforementioned force fields for water, many force fields for ions exist as well, as researchers have sought to expand and improve the library of salt species covered. Similar to solvent, ions can be represented explicitly or implicitly in a simulation. One of the most commonly used explicit ion models was developed by Aqvist for cations of alkali and alkaline-earth metals.[39] Although reproducing experimental data for ion behavior using simulations was difficult at first, ion force fields were improved many times over the past three decades. Force fields covering anions and cations of almost all metals commonly found in solution *in vivo* and *in vitro* are now available. However, if one wishes to use an implicit solvent in a simulation, any salt concentration must also be treated implicitly. As mentioned above, the PB and GB implicit solvent models can accommodate a built-in salt concentration. Another simpler implicit salt concentration model is the Screened Coulomb model, in which the salt concentration has the effect of screening (weakening) the Coulombic interactions of explicit charges in the system. The Screened Coulomb potential is

$$E_{SC} = \frac{q_1 q_2 e^{-\kappa r}}{4\pi\varepsilon_0 \varepsilon r} \tag{1.26}$$

which is just like the Coulomb potential (eqn. XX) but with an exponential term that depends on a parameter *κ*. *κ* quantifies the screening effect as

$$\kappa = \sqrt{\frac{\sum_i (z_i e)^2 c_i}{\varepsilon\varepsilon_0 kT}} \tag{1.27}$$

where *z* is the formal charge, *e* is the elementary charge, and *c* is the number density of the *i*th ionic species. *κ* is also called the inverse Debye length, which is the distance at which electrostatic forces are effectively screened out.

Implicit solvent and ion models share one major pitfall. Since they model the environment as a continuum of smoothly changing force, rather than a space occupied by discrete molecules, they are inherently incapable of reproducing specific interactions, such as hydrogen bonds between solute and solvent. Therefore, implicit models should only be used when such interactions are not of interest or have a negligible role in the dynamics being studied.

1.6 Preparing Initial Simulation Models

One practical aspect of computational chemistry that has not yet been discussed is how one obtains the initial system coordinates that feed into the simulation software. The algorithm needs some initial set of coordinates from which to generate additional structures. These coordinates typically come from one of a few different sources. The most common is X-ray crystallographic data. This data, obtained by X-ray diffraction experiments, can be downloaded from online databases such as the Protein Data Bank (PDB), which currently contains structural information for over 170,000 biological macromolecules.[40] The coordinate data can be downloaded in the ubiquitous ".pbd" file type, which can typically be made simulation-ready with little to no modification. Another method of obtaining initial system coordinates is to use data obtained form NMR experiments. Similar to the PDB, structural databases like

If no experimentally determined structural data exists, computational chemists can use any of a number of molecular modeling programs[41-43] to construct the molecules manually. In such cases, care must be taken to construct a physically realistic system, which may require days or weeks of literature research on the relevant topic. Manual construction is very often necessary when the system being studied is man-made e.g., chemically-engineered bio- and nanotechnologies, or novel materials.

No matter the source, initial system coordinates must be refined before simulations can begin. Atomic positions are resolved with error of ~2-3 Å in XRD experiments, and H atoms cannot be resolved at all. Moreover, user-created structures will undoubtedly contain sub-optimal molecular geometries. Therefore, the standard MD simulation protocol is to minimize and then equilibrate the system prior to the "production" simulation from which data will be collected. Minimization involves bringing the molecular geometry of the system to a local minimum of the potential energy function used to describe it. This has the effect of quickly removing spatial clashes between atoms that would produce unphysical dynamics. Several methods exist to search for a minimum of the potential energy function, such as the commonly used conjugate-gradient method.[44] This method, similar to many others, takes steps in the direction of the minus gradient of the energy surface to eventually find a minimum as

$$v_k = -g_k + \gamma_k v_{k-1} \tag{1.29}$$

where v_k is the direction to move, g_k is the energy gradient at point x_k and γ_k is a scalar. In the first step of the search v_{k-1} does not exists, so the first move is the direction parallel to the gradient, given by the unit vector

$$s_k = \frac{-g_k}{|g_k|} \tag{1.30}$$

After performing minimization, systems are generally allowed to equilibrate for many pico- or nanoseconds, depending on the size and conditions. The purpose of the equilibration phase is to slowly heat the system and allow it to move out of any far-from-equilibrium states that may have arisen as artifacts of the set-up process or the experimentally obtained structural data. Minimization is carried out at 0 K, so the geometry obtained is still not necessarily a realistic representation of the one that would exist at, for example, 298 K. The atoms of the 0 K geometry may experience unphysical dynamics if suddenly supplied with kinetic energy corresponding to 298 K, so usually systems are "heated" from 0 K to 298 K in 25 or 50 K increments. The equilibration should be monitored and continued until system properties such as temperature, pressure, total energy, or RMSD reach a plateau, indicating that equilibrium has been reached.[12]

1.7 Analysis of Simulation Data

Simulations generate an enormous amount of data. System properties are often calculated during the simulation at a frequency much higher than is needed for analysis. This is done to ensure accuracy of the dynamics, creating many gigabytes or even terabytes of data. Often the raw data can be saved at 1000-fold (or more) lower frequency for analysis. The scope of analyses available for simulation data reflects the vast applicability of simulations across countless fields of research. Here, we will focus only on the most common and popular analysis techniques. Previously, we mentioned that proteins are one of the most commonly simulated macromolecules. A commonly goal of such simulations is to elucidate protein dynamics. A measure called the root-mean-square deviation (RMSD) is a basic way of quantifying structural changes over time. Usually, the deviations of interest are those of the atomic coordinates. They are measured with reference to another set of coordinates, often the initial simulation coordinates or the time average of the coordinates, and are plotted against time. The RMSD is quantified as

$$RMSD = \sqrt{\frac{1}{N}\sum_{i}^{N}\left(r_i - r_{ref}\right)^2} \tag{1.31}$$

where N is the total number of atoms, r_i and r_{ref} and current and reference atom coordinates, respectively. A closely related property is the root-mean-square fluctuation (RMSF), which averages the deviations of a particular part of a system over time. The RMSF is better than the RMSD if the goal is to identify the flexibility of various regions of a system e.g., a protein or enzyme. It is quantified as

$$RMSF = \sqrt{\frac{1}{T}\sum_{i}^{T}\left(r_i - r_{ref}\right)^2} \tag{1.32}$$

where T is the total simulation frames. Often, both RMSD and RMSF are calculated, as the former can show structural changes in time, while the latter shows differences between regions within a system. Another structural measure is called the radius of gyration (R_g), which essentially quantifies the degree of compactness of a molecule. For example, it can be used to track the change in shape of a DNA molecule from elongated to curled as it diffuses on a surface. The R_g is calculated

$$R_g = \sqrt{\frac{\sum_i^n m_i s_i^2}{\sum_i^n m_i}} \tag{1.33}$$

where *m* is the mass and *s* is the distance from the center-of-mass of atom *i* in a molecule of *n* atoms. Often it is useful to know the distribution of parts of a system around a central point. For example, one might wish to quantify the distribution of ions around a charged protein residue, or the distribution of substrate around an enzyme. This type of structural property can be given by the radial distribution function (g(r))

$$g(r) = \frac{n(r)}{\rho 4\pi r^2 dr} \tag{1.34}$$

where *n(r)* is the number of particles in the range *dr*, ρ is the bulk number density of particles, and $4\pi r^2 dr$ is the volume of the space with thickness *dr*.

The process of molecular recognition is commonly simulated. Elucidation of this process lends itself well to the tools of computational chemistry since it is difficult or impossible to observe the entire diffusional process experimentally. The diffusion of a molecule can be influenced by many factors, such as the viscosity of the solvent, temperature, ionic concentration, and intermolecular potentials. A property called the diffusion coefficient (D) can be calculated easily from simulation data as

$$D = \lim_{t \to \infty} \frac{\langle r_t - r_0 \rangle^2}{2nt} \tag{1.35}$$

r_t - r_0 is the displacement of the molecule after time *t* and *n* is the dimensionality of the diffusion. The diffusion coefficient can be used to calculate molecular association and dissociation rates. The diffusion-limited association rate k_{on} of two molecules *a* and *b* is

$$k_{on} = 4\pi(D_a + D_b)(r_a + r_b) \tag{1.36}$$

where *D* is the diffusion coefficient and *r* is the radius of molecule *a* or *b*.[45] If there is an attractive or repulsive potential between the two molecules, k_{on} is

$$\frac{4\pi(D_a + D_b)(r_a + r_b)}{\int_{a+b}^{\infty} \frac{1}{r^2} e^{-\beta E(r)} dr} \tag{1.37}$$

where *β* is inverse temperature divided by Boltzmann's constant and *E(r)* is the intermolecular potential.[21]

Another heavily used method in the field of drug discover and design is calculation of the drug binding free energy. This is a measure of the change in energy between bound and unbound states of a drug-protein complex. The Molecular Mechanics Poisson-Boltzmann Surface Area (MMPB/SA) method[46] is extremely popular for this purpose. It is relatively efficient and offers a useful level of accuracy in most cases, e.g. it can be used to rank the binding affinity of a set of drugs.[47] The calculation itself has three main parts: the change in inter- and intramolecular potential calculated by molecular mechanics functions, the polar part of the solvation energy given by the Poisson-Boltzmann equation, and the non-polar solvation energy which is proportional to the surface area of the drug and protein. Overall, this looks like

$$\Delta G_{MMPBSA} = \Delta G_{MM} + \Delta G_{PB} + \Delta G_{hydrophobic} \tag{1.38}$$

The ΔG_{MM} is the molecular mechanics part, ΔG_{PB} is calculated from Poisson-Boltmann equation (equation 1.24) and $\Delta G_{hydrophobic}$ is the non-polar solvation energy, which can have many different empirically-derived forms. All assume the contribution to binding free energy is proportional the solvent exposed surface area (SASA), such as

$$\Delta G_{hydrophobic} = \gamma SASA + b + G_{disp} \tag{1.39}$$

where γ is the solvent surface tension, b is an empirical fitting term, and G_{dsip} is the energy of attractive solute-solvent dispersion forces. ΔG_{MM} as well can be further broken down into two parts

$$\Delta G_{MM} = \Delta E_{MM} - T\Delta S \tag{1.40}$$

To truly estimate the binding free energy, the entropy change ΔS of the ligand (drug), receptor (protein) and solvent must also be accounted for. Ignoring the entropy change, the binding enthalpy can still be calculated and used to rank the affinity of drug molecules for a protein. Methods that include entropy changes have been devised to calculate the absolute binding free energy (ΔG) of a drug or the relative binding free energy ($\Delta\Delta$G) of two or more drugs. The most popular of these are known as Free Energy Perturbation (FEP)[48, 49] and Thermodynamic Integration (TI).[50] These methods involve slowly morphing one receptor-bound drug molecule into another during the simulation to find the binding free energy difference between the two. To find the absolute binding FE, the drug can also be decoupled from the receptor to find the FE difference between bound and unbound states. In a method such as TI, the FE difference is found by integrating over a number of intermediate states as the system is interpolated between the initial (0) and final state (1)

$$\Delta G = G(1) - G(0) = \int_0^1 \langle \frac{dE(\lambda)}{d\lambda} \rangle_\lambda \, d\lambda \tag{1.41}$$

Here, λ is a function that interpolates between the potential energy function of the two ligands. *G(1)* and *G(0)* are the FE values of the final and initial states, respectively, in which λ is 1 or 0, and *E(λ)* is the potential energy of the system as a function of λ.

References

1. Lill, M., Virtual screening in drug design. In *In Silico Models for Drug Discovery*, Springer: 2013; pp 1-12.

2. F Sousa, S.; MFSA Cerqueira, N.; A Fernandes, P.; Joao Ramos, M., Virtual screening in drug design and development. *Combinatorial chemistry & high throughput screening* **2010,** *13* (5), 442-453.

3. Coleman, R. G.; Carchia, M.; Sterling, T.; Irwin, J. J.; Shoichet, B. K., Ligand pose and orientational sampling in molecular docking. *PloS one* **2013,** *8* (10), e75992.

4. Długosz, M.; Trylska, J., Diffusion in crowded biological environments: applications of Brownian dynamics. *BMC biophysics* **2011,** *4* (1), 1-9.

5. Wang, J.; Wolf, R. M.; Caldwell, J. W.; Kollman, P. A.; Case, D. A., Development and testing of a general amber force field. *Journal of computational chemistry* **2004,** *25* (9), 1157-1174.

6. Vanommeslaeghe, K.; Hatcher, E.; Acharya, C.; Kundu, S.; Zhong, S.; Shim, J.; Darian, E.; Guvench, O.; Lopes, P.; Vorobyov, I., CHARMM general force field: A force field for drug-like molecules compatible with the CHARMM all-atom additive biological force fields. *Journal of computational chemistry* **2010,** *31* (4), 671-690.

7. Hamelberg, D.; Mongan, J.; McCammon, J. A., Accelerated molecular dynamics: a promising and efficient simulation method for biomolecules. *The Journal of chemical physics* **2004,** *120* (24), 11919-11929.

8. Suan Li, M.; Khanh Mai, B., Steered molecular dynamics-a promising tool for drug design. *Current Bioinformatics* **2012,** *7* (4), 342-351.

9. Laio, A.; Gervasio, F. L., Metadynamics: a method to simulate rare events and reconstruct the free energy in biophysics, chemistry and material science. *Reports on Progress in Physics* **2008,** *71* (12), 126601.

10. West, A. M.; Elber, R.; Shalloway, D., Extending molecular dynamics time scales with milestoning: example of complex kinetics in a solvated peptide. *The Journal of chemical physics* **2007,** *126* (14), 04B608.

11. Elber, R., Milestoning: An efficient approach for atomically detailed simulations of kinetics in biophysics. *Annual review of biophysics* **2020,** *49*, 69-85.

12. Leach, A. R., *Molecular modelling: principles and applications*. Pearson education: 2001.

13. Zuckerman, D. M., *Statistical physics of biomolecules: an introduction*. CRC Press: 2010.

14. Hoover, W. G., *Computational statistical mechanics*. Elsevier: 2012.

15. Koopman, E.; Lowe, C., Advantages of a Lowe-Andersen thermostat in molecular dynamics simulations. *The Journal of chemical physics* **2006,** *124* (20), 204103.

16. Hünenberger, P. H., Thermostat algorithms for molecular dynamics simulations. In *Advanced computer simulation*, Springer: 2005; pp 105-149.

17. Allen, M. P.; Tildesley, D. J., *Computer simulation of liquids*. Oxford university press: 2017.

18. Paterlini, M. G.; Ferguson, D. M., Constant temperature simulations using the Langevin equation with velocity Verlet integration. *Chemical Physics* **1998,** *236* (1-3), 243-252.

19. Van Gunsteren, W. F.; Berendsen, H. J., A leap-frog algorithm for stochastic dynamics. *Molecular Simulation* **1988,** *1* (3), 173-185.

20. Allen, M. P., Introduction to molecular dynamics simulation. *Computational soft matter: from synthetic polymers to proteins* **2004,** *23* (1), 1-28.

21. Jackson, M. B., *Molecular and cellular biophysics*. Cambridge University Press: 2006.

22. Showalter, S. A.; Brüschweiler, R., Validation of molecular dynamics simulations of biomolecules using NMR spin relaxation as benchmarks: application to the AMBER99SB force field. *Journal of chemical theory and computation* **2007,** *3* (3), 961-975.

23. Guvench, O.; MacKerell, A. D., Comparison of protein force fields for molecular dynamics simulations. *Molecular modeling of proteins* **2008**, 63-88.

24. Henriques, J.; Cragnell, C.; Skepö, M., Molecular dynamics simulations of intrinsically disordered proteins: force field evaluation and comparison with experiment. *Journal of chemical theory and computation* **2015,** *11* (7), 3420-3431.

25. Maier, J. A.; Martinez, C.; Kasavajhala, K.; Wickstrom, L.; Hauser, K. E.; Simmerling, C., ff14SB: Improving the Accuracy of Protein Side Chain and Backbone Parameters from ff99SB. **2015**.

26. Oostenbrink, C.; Villa, A.; Mark, A. E.; Van Gunsteren, W. F., A biomolecular force field based on the free enthalpy of hydration and solvation: the GROMOS force-field parameter sets 53A5 and 53A6. *Journal of computational chemistry* **2004,** *25* (13), 1656-1676.

27. Soares, T. A.; Hünenberger, P. H.; Kastenholz, M. A.; Kräutler, V.; Lenz, T.; Lins, R. D.; Oostenbrink, C.; van Gunsteren, W. F., An improved nucleic acid parameter set for the GROMOS force field. *Journal of computational chemistry* **2005,** *26* (7), 725-737.

28. Phillips, J. C.; Braun, R.; Wang, W.; Gumbart, J.; Tajkhorshid, E.; Villa, E.; Chipot, C.; Skeel, R. D.; Kale, L.; Schulten, K., Scalable molecular dynamics with NAMD. *Journal of computational chemistry* **2005,** *26* (16), 1781-1802.

29. Smith, M. D.; Rao, J. S.; Segelken, E.; Cruz, L., Force-field induced bias in the structure of Aβ21–30: A comparison of OPLS, AMBER, CHARMM, and GROMOS force fields. *Journal of Chemical Information and Modeling* **2015,** *55* (12), 2587-2595.

30. Zielkiewicz, J., Structural properties of water: Comparison of the SPC, SPCE, TIP4P, and TIP5P models of water. *The Journal of chemical physics* **2005,** *123* (10), 104501.

31. Jorgensen, W. L.; Chandrasekhar, J.; Madura, J. D.; Impey, R. W.; Klein, M. L., Comparison of simple potential functions for simulating liquid water. *The Journal of chemical physics* **1983,** *79* (2), 926-935.

32. Berendsen, H.; Grigera, J.; Straatsma, T., The missing term in effective pair potentials. *Journal of Physical Chemistry* **1987,** *91* (24), 6269-6271.

33. Horn, H. W.; Swope, W. C.; Pitera, J. W.; Madura, J. D.; Dick, T. J.; Hura, G. L.; Head-Gordon, T., Development of an improved four-site water model for biomolecular simulations: TIP4P-Ew. *The Journal of chemical physics* **2004,** *120* (20), 9665-9678.

34. Wu, Y.; Tepper, H. L.; Voth, G. A., Flexible simple point-charge water model with improved liquid-state properties. *The Journal of chemical physics* **2006,** *124* (2), 024503.

35. Lu, B.; Zhou, Y.; Holst, M.; McCammon, J., Recent progress in numerical methods for the Poisson-Boltzmann equation in biophysical applications. *Commun Comput Phys* **2008,** *3* (5), 973-1009.

36. Onufriev, A., Continuum electrostatics solvent modeling with the generalized Born model. *Modeling Solvent Environments* **2010,** *1*.

37. Wojciechowski, M.; Lesyng, B., Generalized Born model: Analysis, refinement, and applications to proteins. *The Journal of Physical Chemistry B* **2004,** *108* (47), 18368-18376.

38. Still, W. C.; Tempczyk, A.; Hawley, R. C.; Hendrickson, T., Semianalytical treatment of solvation for molecular mechanics and dynamics. *Journal of the American Chemical Society* **1990,** *112* (16), 6127-6129.

39. Aqvist, J., Ion-water interaction potentials derived from free energy perturbation simulations. *The Journal of Physical Chemistry* **1990,** *94* (21), 8021-8024.

40. Rose, P. W.; Beran, B.; Bi, C.; Bluhm, W. F.; Dimitropoulos, D.; Goodsell, D. S.; Prlić, A.; Quesada, M.; Quinn, G. B.; Westbrook, J. D., The RCSB Protein Data Bank: redesigned web site and web services. *Nucleic acids research* **2010,** *39* (suppl_1), D392-D401.

41. Humphrey, W.; Dalke, A.; Schulten, K., VMD: Visual molecular dynamics. *Journal of Molecular Graphics* **1996,** *14* (1), 33-38.

42. Hanwell, M. D.; Curtis, D. E.; Lonie, D. C.; Vandermeersch, T.; Zurek, E.; Hutchison, G. R., Avogadro: an advanced semantic chemical editor, visualization, and analysis platform. *Journal of cheminformatics* **2012,** *4* (1), 17.

43. *Molecular Operating Environment (MOE)*, 2018.01; Chemical Computing Group ULC: 1010 Sherbrooke St. West, Suite #910, Montreal, QC, Canada, H3A 2R7, 2018.
44. Powell, M. J. D., Restart procedures for the conjugate gradient method. *Mathematical programming* **1977,** *12* (1), 241-254.

45. Berg, O. G.; von Hippel, P. H., Diffusion-controlled macromolecular interactions. *Annual review of biophysics and biophysical chemistry* **1985,** *14* (1), 131-158.

46. Wang, C.; Greene, D. A.; Xiao, L.; Qi, R.; Luo, R., Recent developments and applications of the MMPBSA method. *Frontiers in molecular biosciences* **2018,** *4*, 87.

47. Hou, T.; Wang, J.; Li, Y.; Wang, W., Assessing the performance of the MM/PBSA and MM/GBSA methods. 1. The accuracy of binding free energy calculations based on molecular dynamics simulations. *Journal of chemical information and modeling* **2011,** *51* (1), 69-82.

48. Christ, C. D.; Mark, A. E.; Van Gunsteren, W. F., Basic ingredients of free energy calculations: a review. *Journal of computational chemistry* **2010,** *31* (8), 1569-1582.

49. Williams-Noonan, B. J.; Yuriev, E.; Chalmers, D. K., Free energy methods in drug design: Prospects of "alchemical perturbation" in medicinal chemistry: miniperspective. *Journal of medicinal chemistry* **2018,** *61* (3), 638-649.

50. Gilson, M. K.; Given, J. A.; Bush, B. L.; McCammon, J. A., The statistical-thermodynamic basis for computation of binding affinities: a critical review. *Biophysical journal* **1997,** *72* (3), 1047-1069.

CHAPTER 2: GeomBD3: Brownian Dynamics Simulation Package for Biological and Engineered Systems

2.1 Abstract

GeomBD3 is a robust, efficient Brownian dynamics simulation package designed to easily handle natural or novel, engineered systems of almost any size. The package described herein allows user to design, execute, and analyze BD simulations. The simulations use all-atom, ridged molecular models that diffuse according to overdamped Langevin dynamics and interact through electrostatic, Lennard-Jones, and ligand desolvation potentials. The program automatically calculates molecular association rates, surface residence times, and association statistics for any number of user defined criteria. Users can also extract molecular association pathways, diffusion coefficients, intermolecular interaction energies, association site probability density maps, and more using the provided supplementary analysis scripts. Here we detail the use of the package from start to finish and apply it to a protein-ligand system and a large nucleic acid biosensor. GeomBD3 provides a versatile tool for researchers from diverse disciplines that can aid in rational design of engineered systems or play an explanatory role as a complement to experiments.

2.2 Introduction

Biological and chemical processes *in vivo* and *in vitro* typically start out with the two eventual binding partners separated by long distances which are traversed *via* Brownian motion of one or both molecules, followed by diffusion steered by intermolecular forces that leads to binding. This can take place on temporal and spatial scales of microseconds and micrometers,[1, 2] making it far beyond the reaches of methods such as molecular dynamics simulation. However, such long-range diffusional processes must be studied in order to elucidate and orientational modifications of systems involving molecular recognition can dramatically affect the kinetics and mechanism of important reactions. Moreover, structural and orientational variables of the system can dramatically affect the reaction, so the ability to rapidly assess these effects is extremely valuable. Researchers often lack a tool to guide rational design of novel biotechnologies since most experimental techniques lack the ability to resolve subtle differences in device functionality that result from nano-scale modifications.

The Brownian dynamics (BD) simulations carried out in our software, GeomBD3, are designed specifically for this purpose. GeomBD3 simulates the molecular recognition process of two molecules over relatively long distances and time scales in a vast array of different systems and system sizes. The software is easily capable of handling molecular models up to hundreds of nanometers in scale, thus allowing simulation of simple protein-ligand systems or very large nanoscale technologies, such as biosensors or surface-tethered multi-enzyme arrays. For example, the software has been used to study enzyme bioconjugates, understand DNA hybridization in biosensors,[3] and replicate SPR experiments under diffusive flux.[4] In many biosensing and catalyst applications, the device surface properties are of high importance, as they can control substrate adsorption and diffusion and thus overall efficiency.[5, 6] Structural modifications to enzymes, such as mutations or conjugation of other macromolecules, can alter catalytic efficiency dramatically.[7] GeomBD3 provides the ability to study different permutations of a system to rapidly assess the effects of such modifications, thus providing a highly useful and unique perspective for researchers.

Here, we present a detailed description of GeomBD3 along with two applications of the software and basic analysis of the results. Users are guided through the general simulation workflow, from set-up, to execution, and finally basic analysis. We then present applications of the program to a simple and well-studied protein-ligand system consisting of acetylcholinesterase and acetylcholine, then to a much larger and more complex nucleic acid biosensor system.

2.3 Methods

2.3.1 Implementation

GeomBD3 source code is written entirely in C++, whereas several pre- and post-simulation scripts are available in both C++ and Python3 to aid in simulation set-up or analysis. Simulation set-up, execution, and analysis are carried out using standard pdb, pqr, and dcd file types, generally in that order. The source code can be divided into two main categories: pre-simulation code for parameterization of the molecular models and calculation of potential energy grids, and simulation code for running, recording, and monitoring the Brownian dynamics. The pre-simulation code consists of *Parameterize.py* for creating pqr files of the molecular model inputs, and the *Gridder* programs, for calculation of potential energy and excluded volume grids. The simulation code is implemented in just one program, *GBD*. All of this code can be found on our group GitHub at http://github.com/chang-group.

2.3.2 Simulation Algorithm

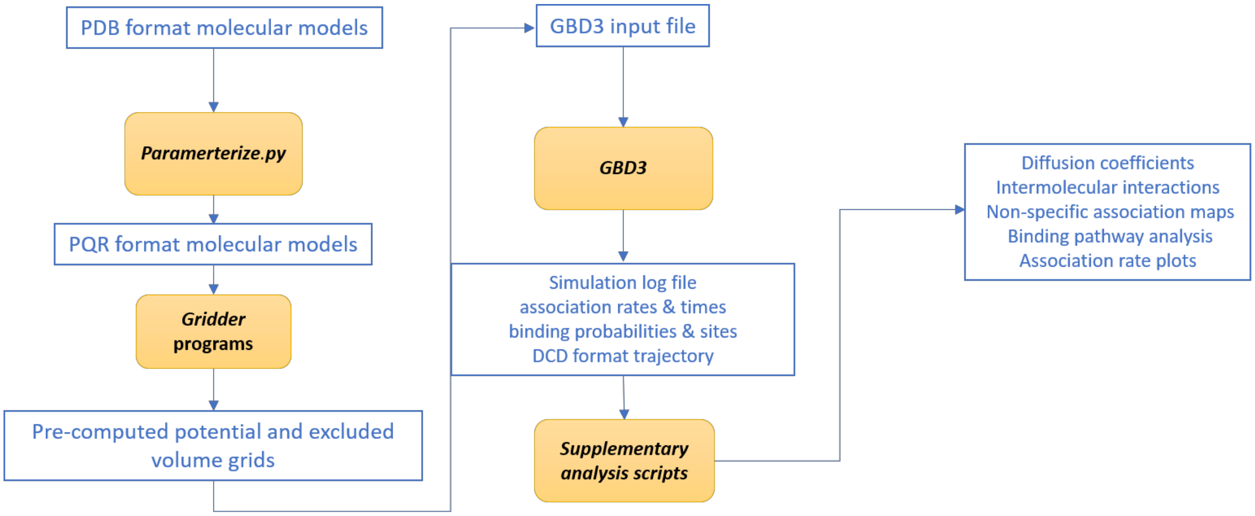

PDB format molecular models
Paramerterize.py
PQR format molecular models
Gridder programs
Pre-computed potential and excluded volume grids
GBD3 input file
GBD3
Simulation log file
association rates & times
binding probabilities & sites
DCD format trajectory
Supplementary analysis scripts
Diffusion coefficients
Intermolecular interactions
Non-specific association maps
Binding pathway analysis
Association rate plots

Figure 2.1. Flow chart of GeomBD3 simulations. Blue text represents program input or output; yellow boxes indicate programs included in the package.

The steps carried out by the software package are summarized in Figure 2.1. The basics steps are as follows. Initially, users should obtain PDB format molecular models of the system, which consists of a ligand and a receptor. The ligand is the molecule which diffuses during the simulation, and the receptor is its intended binding partner. Since simulation time scales linearly with ligand size, the ligand should always be chosen as the smaller of the two; however, it does not need to be a small molecule. For example, the ligand could be a large DNA molecule or protein. Next, the PDBs are converted into PQR type files by using the Parameterize.py script. These pqrs are then fed into the *Gridder* programs that pre-compute as much of the potential energy calculations as possible to save time. Users are not required to generate all three available potential grids, and should choose which are needed for the particular study.

The remaining steps involve calculating the Brownian dynamics trajectories and monitoring output. An input configuration file should be supplied containing the system pqr files, desired grid files, and simulation parameters. Upon running the *GBD* code, first, a number of ligand replicates are positioned according to the desired starting conditions. If ligands lie on the potential energy grids, the intermolecular forces are calculated by the finite difference method. As an example, forces in the *x* dimension F_x are calculated as

$$F_x = \frac{U(gx+1, gy, gz) - U(gx-1, gy, gz)}{2\delta} \tag{2.1}$$

where the potential *U* is a function of an atom's location on the grid specified by *gx, gy,* and *gz*, and δ is the grid spacing. A random force R with the properties

$$R(t) \geq 0 \tag{2.2}$$

$$R(t)R(t') \geq \delta(t - t') \tag{2.3}$$

generated from a Gaussian process is also applied at this time, and ligands are moved to their next positions according to the overdamped Langevin equation

$$r_{t+\Delta t} = r_t - \frac{D}{k_B T}\frac{dU}{dr}\Delta t + \sqrt{2D\Delta t}R \tag{2.4}$$

where r is the ligand position, D is the appropriate diffusion coefficient, $k_B T$ is Boltzmann's constant times the absolute temperature, dU/dr is the potential energy gradient, and Δt is a linearly-scaled timestep. The translational and rotational diffusion coefficients are calculated based on the hydrodynamic radius r_h of either the ligand only or the ligand and receptor; the former case is used when the receptor is attached to a surface i.e., when it is not diffusing.

$$D_{trans} = \frac{k_B T}{6\pi\gamma}\left(\frac{1}{r_{h,lig}} + \frac{1}{r_{h,rec}}\right) \tag{2.5}$$

$$D_{rot} = \frac{k_B T}{8\pi\gamma}\left(\frac{1}{{r_{h,lig}}^3} + \frac{1}{{r_{h,rec}}^3}\right) \tag{2.6}$$

At each new position, each binding criterion is checked and, if satisfied, a binding event is counted and stored. These steps are repeated until certain simulation parameters converge below a defined threshold or the user manually terminates the program. Relevant data pertaining to association rates, binding probabilities, site-specific binding times, ligand-receptor interactions, and simulation convergence are output to a standard log file. Simulation trajectories are output as a .dcd file for viewing and analysis.

2.3.3 Simulation Set-up

GeomBD uses all-atom, rigid molecular models with no internal degrees of freedom.

The initial step in the simulation workflow is to obtain *pqr* format input files containing the molecular structures of the ligand and receptor. These *pqr* files can be generated by the script *Parameterize.py*, which takes in *pdb* files and converts them to *pqr* format. The pdb files should be obtained elsewhere prior to this step. Any file in standard pdb format is acceptable, such as those downloadable from the Protein Data Bank.[8] GeomBD3 also handles non-standard systems easily, as long as PDB molecular models can be created by the user and the parameters for new molecules are added to the .gdbp parameter files. These parameter files are simple to read and edit, making addition of new parameters very straightforward. Instructions for how to prepare a simulation of a novel system are provided in the *USER MANUAL* at http://chemcha-gpu0.ucr.edu/geombd3/. The pqr files outputted by *Parameterize.py* contain the coordinates, partial charges, and vdW radii of each atom, which are used later in energy calculations. The values of partial charges and vdW radii are taken from the AMBER ff14sb[9] and GAFF[10] force fields.

After obtaining properly parameterized *pqr* files of the ligand-receptor system, users should run the series of *Gridder* programs to pre-compute the potential energy at all points in a grid-based representation of the receptor molecule. The following *Gridder* programs are currently available:

Gridder-EX.cc - This creates a grid (the EX grid) defining the excluded volume of the simulation space due to the presence of the receptor. In other words, it defines the space occupied by the receptor. The EX grid should always be generated before running a simulation.

Gridder-ES.cc - This creates a grid (the ES grid) holding the electric field strength of the receptor molecule at each point with the screened Coulomb method.[11] During simulation, the partial charge on each ligand atom can be multiplied by the appropriate grid value to give the electrostatic potential between that atom and the receptor. The default padding value is 40 A.

$$E_{elec} = k_c \frac{q_1 q_2 e^{\frac{-r}{\kappa}}}{\epsilon r} \tag{2.7}$$

$$\kappa = \sqrt{\frac{\sum_i (c_i q_i^2)}{\epsilon \epsilon_0 k_B T}} \tag{2.8}$$

where k_c is Coulomb's constant, q_1 and q_2 are the charges on ligand and receptor atoms, respectively, r is the interatomic distance, and ε is the dielectric of water and ε_0 is vacuum permititivy . κ is the inverse Debye length, where c_i and q_i are the molar concentration and formal charge of ionic species i.

Gridder-LJ.cc - This creates grids (the LJ grids) storing the 12-6 Lennard-Jones potential E_{LJ} between the receptor and each atom type in the ligand. The default padding and spacing values are 12 A and 0.5 A, respectively.

$$E_{LJ} = 4\varepsilon\left[\left(\frac{\sigma}{r}\right)^{12} - \left(\frac{\sigma}{r}\right)^{6}\right] \tag{2.9}$$

where eps is the potential well depth, sigma is the interatomic distance of minimum potential, and r is the interatomic distance.

Gridder-D.cc - This creates a grid holding the desolvation potential of the ligand. This potential is based on the degree of desolvation of each atom and the atom's desolvation parameter, S. The function used in GeomBD3 borrows the functional form and parameters used in AutoDock4.2,[12] which is similar to the concept developed by Stouten, et al.[13]

$$E_{desolv} = \sum_{j=0}^{N} S_i V_j e^{\frac{-r_{ij}^2}{2\sigma^2}} \tag{2.10}$$

$$S_i = solpar + 0.01097q_i \tag{2.11}$$

where Si is the desolvation parameter of ligand atom i, Vj is the fragmental volume of receptor atom j, rij is the distance between them, sigma is 3.5 A and N is the number of atoms in the receptor molecule. In S_i, *solpar* is an empirically-derived part of the desolvation parameter specific to each atom type and q_i is the partial charge on the atom.

During simulation, the potential at the coordinates of each ligand atom is approximated by trilinear interpolation between grid points. Each grid has a user-definable *padding* and *spacing* value. The *padding* variable is the distance the grid will extend in each dimension beyond the minimum/maximum atomic coordinates. The default *padding* values are 40, 12, and 12 Å for the ES, LJ, and D grids, respectively. The *spacing* variable is the distance between adjacent grid points and has a default 0.5 A for all grids.

2.3.4 Configuring and Running the Simulation

Once all the desired pre-computed grids have been made, the simulation configuration file should be set. The file consists of keywords followed by either a number or string specifying the value of the keyword. A user manual with detailed description of each keyword and guidance on selecting appropriate values is provided on our website at http://chemcha-gpu0.ucr.edu/geombd3/. Here we will discuss only a brief overview of the file.

The first handful of keywords deal with simulation output. Next, the pqr files generated previously should be specified after the *ligand* and *receptor* keywords. This is followed by the specification of the grid files to be loaded. The *grid* keyword followed by a specific string denoting the grid type e.g., *es* for electrostatic, should be given first. Then the path to the appropriate grid file should follow.

The next set of keywords specify the simulation parameters. These include *ligandStart, boundary, boundWall,* and *fixedReceptor.* For example, with *ligandStart* and *boundary*, users can select periodic or reflective boundary conditions, and the bounding cell may be either cuboid, hexagonal prismatic, or spherical. In addition, ligands may be started from an xy-aligned plane, a random point, a specified point. Alternatively, simulations can be run under NAM conditions,[14] in which ligands start from a spherical plane and exit at a larger spherical boundary. These options are made available so that simulations can be made to match experimental and natural conditions, or so that more appropriate comparisons can be made between simulated and theoretical results. Association rate constants in most simulations are calculated as

$$k_{on} = \frac{1}{c * t_{avg}} \tag{2.12}$$

where t_{avg} is the average association time and c is the ligand concentration. Under NAM conditions, the association rate constant is calculated as

$$k_{on} = \frac{4\pi D r \beta}{1 - (1-\beta)\frac{k_d(b)}{k_d(q)}} \tag{2.13}$$

where *D* is the relative diffusion coefficient of ligand and receptor, beta is the fraction of ligands that bound before exiting, and *b* and *q* are the radii of the starting and exiting spherical boundaries, respectively. The *biasForce* keyword allows users to apply a biasing force along one dimension to mimic flow of the ligands inside the system. For example, such conditions were used to simulate surface plasmon resonance experimental conditions to investigate the effect of flow on ligand association to HIV protease.[4] The ligand diffusive flux resulting from this force can be calculated as

$$\frac{cF_{bias}}{f} \tag{2.14}$$

where c is the ligand concentration, F_{bias} is the biasing force, and f is the friction force on the diffusing ligand.

The *bindgroups* keyword defines the criteria that must be met in order to achieve binding between ligand and receptor. This should be carefully set based on visualization of the molecular models and knowledge of the system/process being simulated. Each binding criterion consists of an XYZ point, an index number of a ligand atom, and a distance. Taken altogether, this represents the distance between the ligand atom and XYZ point below which the two molecules are considered bound. For example, the input file line

bindgroups 5.1 -9.8 15.0 6 12.0

specifies a condition in which ligand atom number 6 must be less than 12.0 A from the cartesian point (5.1, -9.8, 15.0) to record a binding event.

Simulations can be set to automatically terminate under specific conditions, or they can be manually terminated by the user at any time. Most often, simulations are set to terminate when some value of interest has converged below a certain threshold. For example, if mean binding times t_{avg} are desired, GBD3 can be set to terminate when the SEM of binding times is below a threshold set by the *convergence* input keyword. In such a case, setting *convergence* to 0.01 would cause GBD3 to stop when the SEM/t_{avg} was <

0.01. When a large number of replicates are being run e.g., 5000, the initial data in the log file is necessarily biased toward fast-binding trajectories, and care should be taken to make sure that t_{avg} is actually converged as much as the SEM would suggest. This is easily done by visualization of a graph of t_{avg} versus N_{bind}. An obvious uptrend can be seen, as in the beginning, only trajectories that by chance bound quickly have been sampled. Generally, one needs to sample several times *Nreplicates* binding events before the curve becomes reasonably flat.

2.3.5 Simulation Output

Five files are automatically outputted by GeomBD3. The log file (*.log* extension) contains several useful data. It displays updated information about the number of binding events, the average binding time, the SD and SEM of binding times, convergence of the mean binding time, ligand exit events, and ligand-receptor interaction energies. All information about binding events is displayed separately for each specified binding criterion.

The pqr and dcd are needed for viewing the trajectory. These can be loaded together into any program which supports these file formats e.g., VMD,[15] Pymol, etc., to view the simulation. The pqr contains the initial system configuration, that is, the starting positions of all ligand replicates; it does not contain the receptor coordinates. The dcd file containing the simulation frames should be loaded into the pqr file, and the trajectories can be easily visualized. Users will see trajectories of all ligand replicates displayed at once. Individual trajectories can be viewed by selecting the appropriate representation in the chosen software. Tips for trajectory visualization and analysis are given in the USER MANUAL. The receptor pqr should be loaded separately to view ligand-receptor interactions.

There are two additional data files with the extensions *.t* and *.crd*. The former contains the restart times and binding information, while the later contains the coordinates for each ligand replicate. These files can be used to restart a simulation by restoring the coordinates and trajectory lifetimes of all ligand replicates as well as the statistics for all binding criteria.

2.4 Example Applications

2.4.1 Case 1: Protein-Ligand Binding

One of the most common and straightforward uses of GeomBD3 is to calculate the rate of protein-ligand binding. Here we demonstrate such a simulation applied to the acetylcholinesterase-acetylcholine (AchE-ach) system. The ligand and receptor molecular models were derived from an X-ray crystal structure of AchE from Torpedo Californica (PDB ID: 2ace).[16] Naturally, acetylcholine should be chosen as the ligand and acetylcholinesterase as the receptor. The total charge on ach and AchE were 1 *e* and 8 *e*, respectively. Acetylcholine replicates were started from random points in a a 6.54 × 107 Å3 spherical simulation space with the stationary grid representation of AchE at the center. When ligands passed the boundary of the sphere, they were returned to their previous position and a new forward step was generated, so the sphere essentially acted as an impenetrable wall. The binding site is chosen as point (1.3, 56.7, 74.2), which corresponds to an atom in the gorge leading to the catalytic binding site.[17] Here, we are only interested in calculating the association (binding) rate. The simulation is set to terminate when the SE of the mean binding time has converged below 0.01 of the mean binding time. The timestep is scaled between 30 and 500 Å with a fine setting of 0.025 ps and coarse setting of 0.5 ps. Since AchE is known to be driven by electrostatics,[18] we run simulations under several different conditions: full electrostatics, 0.5 M implicit monovalent salt, and no electrostatics. All simulations contained LJ and ligand desolvation potential grids. Results are summarized in Table 2.1.

$k_{on} \times 10^9$ $M^{-1}s^{-1}$		
0.0 M	**0.5 M**	**no ES grid**
94.0	6.58	2.95

Table 2.1 Association rate constant for AchE-ach with full electrostatics (0.0 M ions), 0.5 M ions, and no electrostatic interactions.

Our results show the dramatic decrease in association rate as electrostatics are screened or completely turned off. We also analyzed the non-specific association of ach to AchE with the program *ProbDX*. This program accepts a PDB-format trajectory and maps ligand coordinates to the nearest grid point, then outputs a DX-format density map which can be visualized (Figure 2.2a). This revealed that the ligand spends much of its time associated to the protein surface and identified multiple association hot spots. All are located on one side of the enzyme owing to its strong dipole.

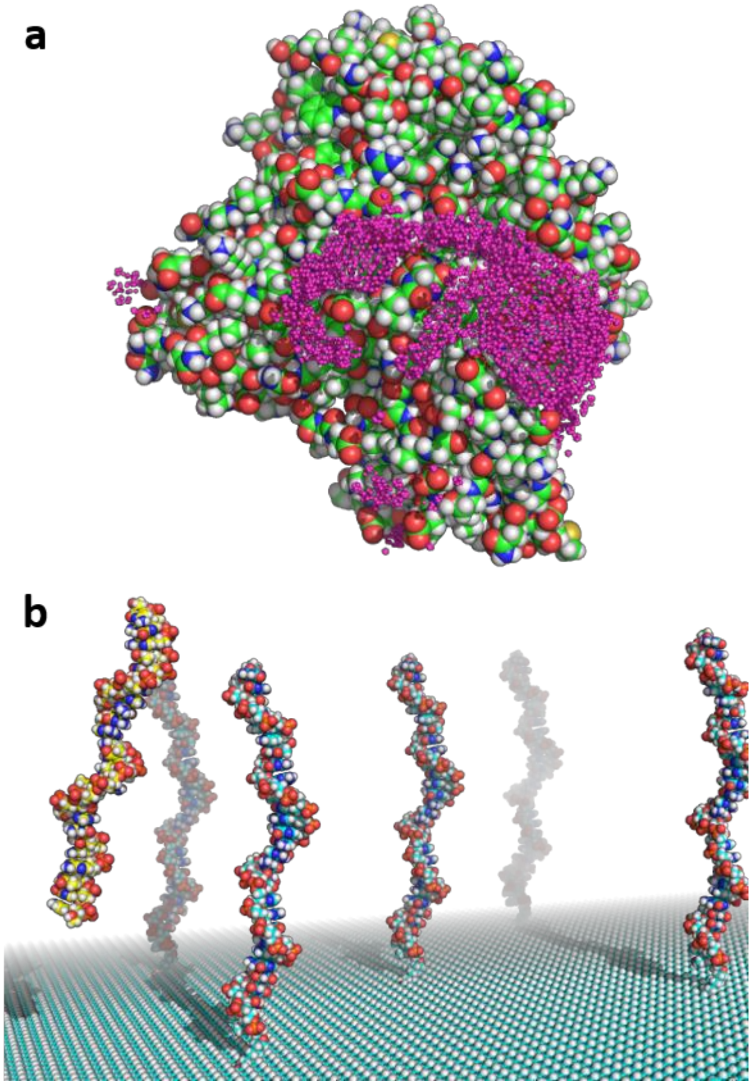


Figure 2.2. (a) Acetylcholinesterase with common ligand association sites indicated by pink dots. Higher dot density indicates higher probability of association. (b) Model DNA biosensor with multiple probes (blue strands) at which the target DNA (yellow strand) can bind.

2.4.2 Case 2: DNA Hybridization in a Biosensor

GeomBD3 is also useful for simulations of large or complex engineered systems that span tens or hundreds of nanometers across. As a second GBD3 application, we present previously published data[3] from a simulation of DNA hybridization (binding) taking place in a biosensor (Figure 2.2b). The sensor consisted of ssDNA probes inserted into a self-assembled monolayer of functionalized alkanethiol molecules on gold. With GeomBD3, it was possible to rapidly study the sensor under different probe surface densities and equipped with different functional groups on the SAM. These simulations used one or more ssDNA probes inserted into a slab of SAM as the receptor, and a ssDNA molecule complementary to the probe strand as the ligand. The simulation space was bounded by a cuboid box with periodic boundary conditions in the *x* and *y* dimensions and a hard wall at the top of the box. The *x* and *y* dimensions of the box were varied to encompass different amounts of SAM, yielding three different probe surface densities. Additionally, hybridization with clusters of five tightly-packed probes was also studied. The probe density was varied while also varying SAM functional groups between -OH, $-COO^-$, and $-CH_3$, yielding 15 different system permutations in total.

Varying the surface properties by changing SAM functional groups can dramatically alter DNA-surface interactions and thus hybridization rates. In some cases, DNA adsorbs and diffuses along the SAM 2-dimensionally, which may slow down or speed up hybridization depending on the strength and nature of the interaction. We analyzed the fraction of hybridization taking place through a surface-mediated mechanism, which we label as 2-dimensional (2D) hybridization.

In the biosensor with a cluster of probes, each probe (labeled 1-5; 5 at center) represented a possible binding site. GeomBD3 accepts multiple binding criteria, and can independently track binding data for each criterion. We were interested in quantifying the proportion of hybridization taking place at each probe in order to understand how tight probe packing may influence accessibility. The results are presented in Table 2.2.

	separation	**τ_{sim} (μs)**	**hybridizing probe (%)**					**γ**	**ΔE_{elec}**	**ΔE_{vdW}**
	(nm)		**1**	**2**	**3**	**4**	**5**		**(kcal/mol)**	**(kcal/mol)**
CH_3	5.0	22.6 ± 0.7	23	23	26	26	2	0.13	-0.35 ± 0.26	-3.71 ± 1.60
	10.0	14.4 ± 0.1	22	20	27	23	7	0.20		
	15.0	15.7 ± 0.4	21	13	24	27	14	0.24		
COO^-	5.0	20.6 ± 0.4	15	30	29	24	2	0.03	0.21 ± 0.33	0.00 ± 0.00
	10.0	15.4 ± 0.2	23	19	28	26	4	0.03		
	15.0	16.1 ± 0.5	15	19	24	22	20	0.04		

Table 2.2 Hybridization time and probe site, fraction 2D hybridization (γ), and DNA-SAM interaction energies on the two different SAMs used for probe cluster simulations.

2.5 Conclusion

GeomBD3 provides researchers with an extremely robust Brownian dynamics software package. The program uses rigid all-atom molecular models of a diffusing ligand and stationary grid-based representation of a receptor. Molecular recognition is simulated by diffusing a molecule according to overdamped Langevin dynamics and evaluating one or more user-defined binding criteria to track binding events. It is capable of quickly simulating binding kinetics for drug-protein systems as well as identifying binding pathways and non-specific adsorption sites as demonstrated by our application to the acetylcholinesterase enzyme. GeomBD3 is also excellent for exploring the effects that specific system properties have on molecular recognition. For example, several different system permutations may be studied and compared rapidly to assess the factors important for optimal design. We demonstrate this in our application of the program to a nucleic acid biosensor with varied probe surface density and equipped with three different SAM surfaces. This also demonstrates the applicability to non-standard engineered systems extending hundreds of nanometers in multiple dimensions. Atomic parameters for novel systems can be easily added to the plain-text parameter files. Thus, the package can easily and efficiently meet the needs of researchers in a wide range of disciplines as both an explanatory and predictive tool.

References

1. Berg, O. G.; von Hippel, P. H., Diffusion-controlled macromolecular interactions. *Annual review of biophysics and biophysical chemistry* **1985,** *14* (1), 131-158.

2. Jackson, M. B., *Molecular and cellular biophysics*. Cambridge University Press: 2006.

3. Cholko, T.; Chang, C.-e. A., Modeling Effects of Surface Properties and Probe Density for Nanoscale Biosensor Design: A Case Study of DNA Hybridization near Surfaces. *The Journal of Physical Chemistry B* **2021,** *125* (7), 1746-1754.

4. Kaushik, S.; Chang, C.-e. A., Molecular mechanics study of flow and surface influence in ligand-protein association. *Frontiers in Molecular Biosciences* **2021,** *8*, 284.

5. Cholko, T.; Kaushik, S.; Chia-en, A. C., Dynamics and molecular interactions of single-stranded DNA in nucleic acid biosensors with varied surface properties. *Physical Chemistry Chemical Physics* **2019,** *21* (29), 16367-16380.

6. Gong, P.; Levicky, R., DNA surface hybridization regimes. *Proceedings of the National Academy of Sciences* **2008,** *105* (14), 5301-5306.

7. Li, Y.-C.; Chao, T.-C.; Kim, H. J.; Cholko, T.; Chen, S.-F.; Li, G.; Snyder, L.; Nakanishi, K.; Chang, C.-e.; Murakami, K., Structure and noncanonical Cdk8 activation mechanism within an Argonaute-containing Mediator kinase module. *Science Advances* **2021,** *7* (3), eabd4484.

8. Berman, H. M.; Westbrook, J.; Feng, Z.; Gilliland, G.; Bhat, T. N.; Weissig, H.; Shindyalov, I. N.; Bourne, P. E., The protein data bank. *Nucleic acids research* **2000,** *28* (1), 235-242.

9. Maier, J. A.; Martinez, C.; Kasavajhala, K.; Wickstrom, L.; Hauser, K. E.; Simmerling, C., ff14SB: improving the accuracy of protein side chain and backbone parameters from ff99SB. *Journal of chemical theory and computation* **2015,** *11* (8), 3696-3713.

10. Wang, J.; Wolf, R. M.; Caldwell, J. W.; Kollman, P. A.; Case, D. A., Development and testing of a general amber force field. *Journal of computational chemistry* **2004,** *25* (9), 1157-1174.

11. Hassan, S. A.; Mehler, E. L., A critical analysis of continuum electrostatics: the screened Coulomb potential–implicit solvent model and the study of the alanine dipeptide and discrimination of misfolded structures of proteins. *Proteins: Structure, Function, and Bioinformatics* **2002,** *47* (1), 45-61.

12. Huey, R.; Morris, G. M.; Olson, A. J.; Goodsell, D. S., A semiempirical free energy force field with charge-based desolvation. *Journal of computational chemistry* **2007,** *28* (6), 1145-1152.

13. Stouten, P. F.; Frömmel, C.; Nakamura, H.; Sander, C., An effective solvation term based on atomic occupancies for use in protein simulations. *Molecular Simulation* **1993,** *10* (2-6), 97-120.

14. Northrup, S. H.; Allison, S. A.; McCammon, J. A., Brownian dynamics simulation of diffusion-influenced bimolecular reactions. *The Journal of Chemical Physics* **1984,** *80* (4), 1517-1524.

15. Humphrey, W.; Dalke, A.; Schulten, K., VMD: Visual molecular dynamics. *Journal of Molecular Graphics* **1996,** *14* (1), 33-38.

16. Raves, M. L.; Harel, M.; Pang, Y.-P.; Silman, I.; Kozikowski, A. P.; Sussman, J. L., Structure of acetylcholinesterase complexed with the nootropic alkaloid,(–)-huperzine A. *Nature structural biology* **1997,** *4* (1), 57-63.

17. Axelsen, P. H.; Harel, M.; Silman, I.; Sussman, J. L., Structure and dynamics of the active site gorge of acetylcholinesterase: Synergistic use of molecular dynamics simulation and X-ray crystallography. *Protein Science* **1994,** *3* (2), 188-197.

18. Ripoll, D. R.; Faerman, C. H.; Axelsen, P. H.; Silman, I.; Sussman, J. L., An electrostatic mechanism for substrate guidance down the aromatic gorge of acetylcholinesterase. *Proceedings of the National Academy of Sciences* **1993,** *90* (11), 5128-5132.

CHAPTER 3. Modeling Effects of Surface Properties and Probe Density for Nanoscale Biosensor Design: A Case Study of DNA Hybridization Near Surfaces

3.1 Abstract

Electrochemical biosensors have extremely robust applications while offering ease preparation, miniaturization, and tunability. By adjusting the arrangement and properties of immobilized probes on the sensor surface to optimize target-probe association, one can design highly-sensitive and efficient sensors. In electrochemical nucleic acid biosensors, a self-assembled monolayer (SAM) is widely used as a tunable surface with inserted DNA or RNA probes to detect target sequences. The effects of inhomogeneous probe distribution across surfaces are difficult to study experimentally due to inadequate resolution. Regions of high probe density may inhibit hybridization with targets, and the magnitude of the effect may vary depending on the hybridization mechanism on a given surface. Another fundamental question concerns diffusion and hybridization of DNA taking place on surfaces and whether it speeds up or hinders molecular recognition. We used all-atom Brownian dynamics simulations to help answer these questions by simulating the hybridization process of single-stranded DNA (ssDNA) targets with a ssDNA probe on polar, non-polar, and anionic SAMs at three different probe surface densities. Moreover, we simulated three tightly-packed probe clusters by modeling clusters with different inter-probe spacing on two different surfaces. Our results indicate that hybridization efficiency depends strongly on finding a balance which allows

attractive forces to steer target DNA toward probes without anchoring it to the surface. Furthermore, we found that the hybridization rate becomes severely hindered when inter-probe spacing is less than or equal to the target DNA length, proving the need for a careful design to both enhance target-probe association and avoid steric hindrance. We developed a general kinetic model to predict hybridization times and found that it works accurately for typical probe densities. These findings elucidate basic features of nanoscale biosensors which can aid in rational design efforts and help explain trends in experimental hybridization rates at different probe densities.

3.2. Introduction

Nanoscale electrochemical sensors have seen extensive use for detecting a range of important biomolecules including viral DNA[1, 2], drug-resistant genes[3], and dangerous damage or mutations in genetic material.[4] Among them, electrochemical nucleic acid biosensors containing a self-assembled monolayer (SAM) of functionalized alkanethiols on gold or other substrates have proven to be excellent designs for detection of biomarkers because of their tunability, ease of miniaturization, and greater accuracy than current techniques.[5-8] In this work, we studied an electrochemical DNA biosensor in which immobilized single-stranded DNA (ssDNA) inserted into a SAM acts as a probe for its complementary sequence. The SAM consists of an alkane chain with a thiol head group that bonds to a gold substrate and a functional tail group on the solvent-exposed end. The tail group largely controls the surface properties of the SAM and can be selected to suit a number of possible applications, including adsorption or repulsion of specific molecules such as proteins or ions[5, 9, 10], modeling biological surfaces[11-13], and immobilization of DNA for controlled assembly[14, 15]. The sensor works by detecting the formation of double-stranded DNA from complementary single-stranded target and probe sequences – a process known as hybridization. Hybridization may occur through a bulk solvent diffusive encounter or through a surface-mediated mechanism.

Current techniques for characterizing the structure of SAMs, which plays a major role in determining the nature of DNA–SAM interactions, lack the resolution to describe nanoscale and sub-nanoscale features.[16-18] Many unanswered questions remain about the organization of the DNA probes on SAMs and its effect on overall sensor function. For example, heterogeneity in surface density of inserted DNA probes has been observed, and these differences in density may have a substantial effect on molecular recognition between the target and probe. Moreover, the nature of hybridization is believed to be much more complicated at a surface than in bulk solution because of the unique thermodynamic environment of probes concentrated on a surface.[19] Interaction between probes and the monolayer or between neighboring individual probes can affect their ability to hybridize with incoming target strands. Indeed, Qiao et al. found evidence that hybridization on surfaces is slowed compared to in solution[20], and that repulsions between neighboring probes can hinder hybridization.[21] Understanding the effect of all these factors on behavior of DNA on various surfaces can help steer rational design of nucleic acid biosensors.

Surface plasmon resonance, atomic force microscopy, and electrochemical measurements[17, 22, 23] have been used to study the behavior of DNA on SAMs, but all have the drawback of detecting only the hybridized double-stranded DNA and provide little to no information about the mechanism of hybridization. Computational modeling and simulation can be tremendously useful in this regard, especially for their ability to elucidate processes on spatial and temporal scales inaccessible to experimental techniques.[24-28] Moreover, a computational approach allows for inexpensive study of many permutations of a system. This can play an explanatory role by systematically varying certain system properties, or can be used to guide design toward optimal function. For example, simulation has been successfully used on its own and in conjunction with experimentation to elucidate SAM interactions with proteins,[29, 30] surface-bound receptors,[31] biosensor dynamics[32] and protein-ligand interactions.[33, 34] One important question concerns the amount of 2-dimensional (2D) hybridization, that is, surface-mediated diffusion of target DNA along the SAM leading to hybridization, versus the amount of 3-dimensional (3D) hybridization, in which hybridization occurs by diffusion of targets through the solvent directly to probes. This property is relevant in many fields; genotyping and gene expression profiling take advantage of DNA microarrays utilizing special surfaces,[35, 36] and catalysis may utilize surfaces with immobilized catalysts to enhance reaction rates.[37] Such processes can be directly observed through simulations to provide essential missing knowledge of sensor function.

We performed all-atom Brownian dynamics simulations to answer open questions about nanoscale features of these biosensors that can aid rational design. In total, sixteen simulations were performed (Table 1). Nine of the simulations modeled a single probe inserted into one of three different SAMs (Figure 3.1a), and each SAM was studied at three different probe densities. Additionally, six other simulations studied effects of high probe density, in which a cluster of five probes on one of two different SAMs was modeled with three different inter-probe spacings (Figure 3.1b). We found that hybridization rate was highly dependent on electrostatic interaction between the target and surface; attractive forces can drastically expedite the search for a probe, but can also strongly inhibit hybridization by hindering target desorption from surfaces. Moreover, we show that the rate of hybridization eventually saturates and becomes severely hindered as inter-probe spacing decreases below a certain threshold. These results indicate that efficiency of hybridization is greatest when surface attraction is weak, allowing an increased target concentration to form near the sensor without targets becoming strongly adsorbed. Furthermore, probe crowding energetically and sterically inhibits the accessibility of probes to incoming targets, resulting in greatly slowed hybridization kinetics on both a polar and hydrophobic surface. Finally, we develop a general model of hybridization kinetics which accurately matched simulations results on all surfaces, but suffered slightly at low probe densities.

	Single-probe simulations	**Multi-probe simulations**
	probe density (nm^{-2})	**probe separation (nm)**
OH-SAM	0.0005	--
	0.001	--
	0.002	--
CH_3-SAM	0.0005	5.0
	0.001	10.0
	0.002	15.0
CH_3-SAM (0.35 M)	0.002	--
COO^--SAM	0.0005	5.0
	0.001	10.0
	0.002	15.0

Table 3.1. Summary of all simulations. In total, 16 simulations were performed. In the single-probe simulations, three different SAMs were used and each was simulated at three different probe surface densities. In the multi-probe simulations, two different SAMs with a cluster of five probes were used, and each was simulated with three different inter-probe separations.

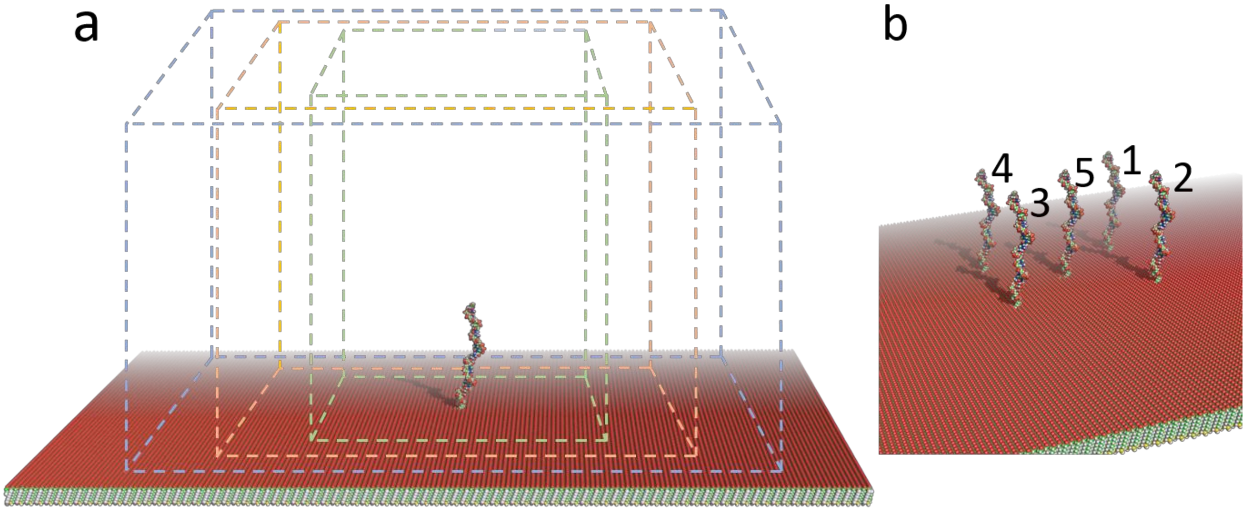


Figure 3.1. The GeomBD3 simulation space. **(a)** The SAM with inserted ssDNA probe is surrounded by a cuboid simulation box with periodic walls at the sides and a hard boundary at the top. Three different sized boxes were used to simulate the different probe surface densities: 0.002 (green), 0.001 (orange) and 0.0005 nm^{-2} (blue). The SAM extends past the walls enough to prohibit targets from hanging off any edges. **(b)** High probe density simulations used the same conditions as stated above, but contained a cluster of five probes space either 5.0, 10.0 or 15.0 nm apart. Simulation box size was the same for all three cluster spacings.

3.3. Methods

3.3.1 Sensor and ssDNA target structure

Simulations were performed with the updated version of the GeomBD2 Brownian dynamics simulation program,[33] available on our group website: http://chemcha-gpu0.ucr.edu/software/. All molecules in the simulations are rigid, containing no internal degrees of freedom. During simulations, one molecule is allowed to diffuse with full translational and rotational degrees of freedom, while another is held stationary. In the present case, the 24-base ssDNA sequence 5'-CGTACTGACTGCTCACGAGGTAGC-3 (hereafter the "target") diffused from a randomly selected position on a plane positioned 42.5 nm above the biosensor surface until it achieved hybridization. The biosensor consisted of a SAM with one or more inserted ssDNA probe(s) (hereafter the "sensor"). All portions of the sensor remained stationary during the simulations. The probe was the 29-base ssDNA sequence 5'- GCTACCTCGTGAGCAGTCAGTACGTTTTT-3', where the first 24 bases were complimentary to the target sequence. The probe extended approximately 10.0 nm above the SAM, which was a 100 × 100 nm square slab (Figure 3.1). The three SAM surfaces used were undecanethiol (hydrophobic), 11-mercapto-1-undecanol (polar), or 11-mercaptoundecanoic acid (anionic) on a gold substrate. We refer to these as the CH_3-, OH-, or COO^--SAM, respectively. On the COO^--SAM, all carboxylic acid tail groups were in the anionic COO^- form to model the surface at pH 7.

The simulation space was constructed as follows. The sensor was placed in the bottom of a rectangular prism extending 50.0 nm in the z-dimension (perpendicular to the surface). A hard boundary was used at the top face of the prism which, if passed, resulted in a target molecule being returned to its previous position and a new random component of force being generated. The side faces of the prism used periodic walls which were placed such that the accessible SAM surface area yielded the desired probe surface density (Figure 3.1a). Target trajectories originated from a random point on a plane spanning the entire simulation space parallel to the SAM positioned 42.5 nm above it so that they began diffusion free from intermolecular potential. All simulations began at this height regardless of box size, so that the only factor being varied was the probe surface density. The length and width of the SAM extended well beyond the periodic walls in all simulations, so no targets could move below the SAM nor come close to the edges.

3.3.2 Simulation protocol

Simulations include a stationary grid representation of the sensor's volume, electric field, and

the van der Waals-like 12-6 Lennard-Jones (LJ) potential. The grid spacing for all three grids was 0.05 nm. A screened Coulomb potential was used to model implicit monovalent salt concentrations of 0.50 and 0.35 M. The electric field and LJ grids extend 4.0 nm and 1.5 nm, respectively, beyond all edges of the sensor. Atomic charges and LJ parameters for DNA were derived from the AMBER ff14SB force field.[38] Atomic charges for the SAM molecules were calculated using AMBER's *antechamber* program[39] with the AM1-BCC semi-empirical charge method, and LJ parameters were taken from AMBER ff14SB. The AMBER force fields have been used successfully in several previous studies of nucleic acids and organic surfaces.[40-42] During simulation, targets diffuse in implicit water solvent according to the over-damped Langevin equation,

$$r_i(t+\ \Delta t) = r_i(t) - \frac{D_i}{k_B T}\frac{dU}{dr}\Delta t + \sqrt{2D_i \Delta t}R \qquad (3.1)$$

where D_i is the translational or rotational diffusion coefficient, k_B is Boltzmann's constant, T is the temperature (298.15 K), $\frac{dU}{dr}$ is the potential energy gradient, Δt is a distance-dependent variable time step between 0.1 and 1.0 ps, and R is a zero-mean, stationary Gaussian process.[43] Translational and rotational diffusion coefficients are calculated from the Stokes-Einstein equation. Additional details on the simulation protocol are given in the supporting information (SI).

3.3.3 Definition of hybridization conditions

When a target satisfied any one of three hybridization conditions, its trajectory was terminated, and its lifetime saved to a log file before a new trajectory was started. The average lifetime, or average hybridization time (τ_{sim}), was recalculated after each new hybridizing trajectory. When the relative standard deviation of the last 100 τ_{sim} values was less than 0.01 a simulation was considered as converged. One hybridization condition required the centers of mass of the target and probe to come within 0.7 nm (Figure 3.2a). Another condition was designed to mimic native base pair formation between target and probe by requiring a base on the target to come within 0.7 nm of its native Watson-Crick base-pairing partner on the probe. The 0.7 nm threshold ensures that this condition is only met if the hydrogen bonding atoms on each base are adjacent (Figure 3.2b). The third and final condition required end-to-end overlap of five or more bases on either end of the target and probe strands, satisfied when the first and fifth overlapping bases were both within 0.7 nm of the probe. (Figure 3.2c). All three conditions were selected to represent formation of a target-probe complex that would likely lead to full hybridization based on results from prior coarse-grained DNA hybridization simulations.[44, 45]

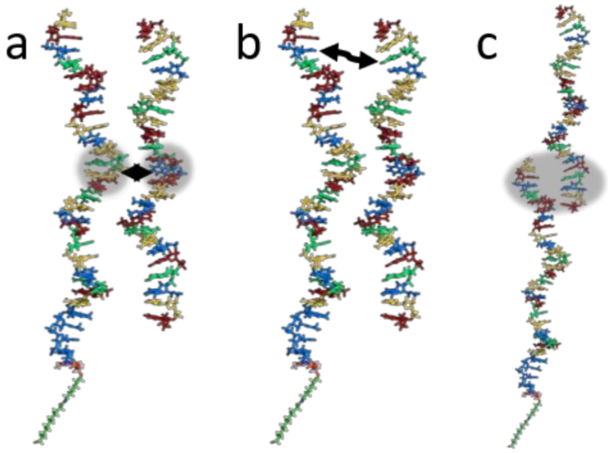


Figure 3.2. ssDNA Probe (left strand, SAM not pictured) and ssDNA target (right strand) in the three possible hybridization conditions: a) the centers of mass are less than 0.7 nm apart, b) any two native Watson-Crick base-pairing partners (indicated by the black arrow), such as the thymine (blue) and adenine (green), are less than 0.7 nm apart, and c) any end-to-end overlap of five or more bases.

3.3.4 Classification of 2D and 3D hybridization

To label trajectories as 2D or 3D hybridizers, we used an in-house program to label each hybridization event based on the pathway of the DNA target just prior to associating with the probe. If targets were adsorbed to the SAM (< 0.4 nm between atomic centers of any SAM atom and any DNA atom) in one or more of the 20 frames (~10 ns) prior to hybridization, they were considered 2D hybridizers. All others were labeled as 3D hybridizers.

3.4. Results and Discussion

Two sets of simulations were performed. The set of single-probe simulations was done on three different SAM surfaces, the OH-SAM, CH_3-SAM, and COO^--SAM. Each surface was simulated at three different probe densities. To achieve a specific probe density, the single probe was placed in the center of a SAM while the length and width of the simulation box were varied to encompass the necessary SAM surface area (Figure 3.1a). A set of multi-probe simulations was also performed, which contained a cluster of five probes on the CH_3-SAM or COO^--SAM (Figure 3.1b). On each SAM, the clustered probes were simulated with three different inter-probe spacings of 5.0, 10.0 or 15.0 nm to elucidate the effect of tight probe packing, which may hinder target-probe recognition. All were performed with implicit 0.50 M monovalent salt concentrations except for one additional test with 0.35 M (Table 1). The additional simulation of the CH_3-SAM with 0.002 nm^{-2} probe density was performed with a 0.35 M salt concentration to understand the effects of ionic screening of electrostatic forces on the dynamics and hybridization rates.

3.4.1 Single-probe simulations with varied surface properties and probe densities

OH-SAM On the polar OH-SAM, target-SAM electrostatic interaction was slightly repulsive. (Table 2). Target adsorption to the SAM was very short-lived, lasting roughly 22 μs, and was maintained primarily by close-range vdW attractions. This is similar to our previous study which showed transient adsorption of DNA on OH-SAMs using molecular dynamics simulations.[32] Moreover, hydrophilic SAMs in general are known to resist non-specific DNA adsorption.[46, 47] Because of the weak adsorption, most of the hybridization happened through a 3D mechanism, and no clear relationship can be seen between the mechanism and probe density (Table 2). One may expect a greater fraction of 3D hybridization as probe density increases, since there is less exposed surface per probe; however, this is not what we observed. We calculated the layer-wise distribution of targets (Figure S1 and Table S1) during the simulation to understand how the SAMs affect diffusion. This analysis shows a substantially increased concentration of targets near the SAM, 2.6-fold higher than would be expected if there were no target-SAM attractions (Figure 3.3a). The increased concentration near the surface coupled with the lack of 2D hybridization indicates that, overall, very weak vdW forces hold targets close to the SAM while allowing target DNA to diffuse freely just above it or 2-dimensionally along its surface. The average hybridization time dropped sharply with increasing probe density, from about 70.3 to 22.7 μs.

CH_3-SAM Behavior on the hydrophobic CH_3-SAM was largely similar to that on the OH-SAM. Here, vdW forces dominate, which depend less on target sequence or contact angle than the electrostatic interactions present on hydrophilic surfaces. Indeed, previous studies have shown that DNA interacts strongly with hydrophobic surfaces through "face-down" adsorption of nucleobases on the surface, and that it diffuses across such surfaces at a rate nearly the same as in bulk water.[32, 48] Although the rigid-body DNA molecules used in our study cannot adopt such conformations to maximize intermolecular contacts, we observed similar behavior. Only on this surface did the target DNA adopt a surface-adsorbed orientation parallel to the surface, which is partly responsible for the stronger electrostatics and vdW interactions. Figure 3.3b shows the distribution of measured contact angles between the DNA and SAM, with the CH_3-SAM showing a relatively high fraction of angles < 20°, indicating the DNA is "laying down" almost parallel to the surface. Analysis of target distribution shows a 4-fold increase in concentration 0-2.0 nm above the CH_3-SAM, but only a 1.5-fold increase in the range 2.0-4.0 nm above, reflecting the dominance of short-ranged vdW forces (Figure 3.3a). Targets showed weak electrostatic attraction and greater vdW attraction on this surface than the OH-SAM, which resulted in a marginally higher fraction of 2D hybridization (Table 2). Notably, this higher 2D fraction corresponds with slightly faster hybridization times for the 0.001 and 0.002 nm^{-2} probe densities on this surface compared to the OH-SAM. (Table 2), indicating that hybridization time was decreased by the high surface concentration of targets. This increased molecular association rate due to reduced search dimensionality has been demonstrated elsewhere both computationally[29] and experimentally[49] for other biomolecules, so long as desorption from the surface is not impeded by attractive forces. Indeed, striking a balance between attractive forces that shuttle targets to a surface and subsequently to the probes, while allowing rapid 2D diffusion and desorption from the surface is the key factor if the goal is to expedite association kinetics. Our simulations of the same CH_3-SAM system at a lower salt concentration of 0.35 M provide strong evidence for this point, which is discussed in a later subsection.

COO^--SAM The COO^--SAM is a negatively charged surface, as the carboxylic acid tail groups are in a deprotonated carboxylate form near pH 7. Since DNA carries a net negative charge of about -1 *e* per nucleotide, electrostatic repulsion between the SAM and incoming targets results in almost entirely 3-dimensional hybridization (Table 2). The repulsion is also evident in the target distribution, which shows that most probes resided in the range 2.0-4.0 nm above the SAM. The range 0-2.0 nm above the SAM had a concentration equal to the bulk (Figure 3.3a). However, the high target concentration in the second layer indicates that there was still a greatly increased concentration near the surface, which effectively channeled targets to probes to the same degree as the other SAMs. Despite the discrepancy in hybridization mechanism compared to the others, the average hybridization time was quite similar to that on the OH- and CH_3-SAMs, and exhibited the same increasing trend as probe density decreased (Table 2). The weak repulsion of 0.21 kcal/mol, or roughly 0.35 kT, which is within the range of thermo-fluctuation, did not significantly affect $\tau_{sim.}$ Hybridization time at the lowest probe density was faster than that on the other SAMs, and notably, had the greatest 3D hybridization fraction of the three densities simulated (Table 2).

surface	**probe density (nm^{-2})**	**γ**	**τ_{sim} (μs)**	**τ_{theo} (μs)**	**ΔE_{elec} (kcal/mol)**	**ΔE_{vdW} (kcal/mol)**
OH-SAM	0.0005	0.14	70.3 ± 1.5	52.7	0.22 ± 0.21	-2.91 ± 2.20
	0.001	0.11	38.1 ± 0.5	35.2		
	0.002	0.13	22.7 ± 0.1	23.4		
CH_3-SAM (0.35 M)	0.002	0.18	40.1 ± 2.5	37.0	-0.96 ± 0.28	-4.10 ± 1.34
CH_3-SAM	0.0005	0.16	70.5 ± 1.9	53.1	-0.35 ± 0.26	-3.71 ± 1.60
	0.001	0.17	36.4 ± 0.9	39.3		
	0.002	0.16	20.4 ± 0.2	28.8		
COO^--SAM	0.0005	0.01	65.5 ± 1.5	57.6	0.21 ± 0.33	0.00 ± 0.00
	0.001	0.01	36.2 ± 1.8	38.3		
	0.002	0.01	20.8 ± 0.4	25.3		

Table 3.2. Fraction of 2D hybridizers (γ), average simulated hybridization time (τ_{sim}), predicted hybridization time (τ_{theo}), and target-SAM electrostatic (ΔE_{elec}) and van der Waals (ΔE_{vdW}) interaction energy for the four single-probe scenarios at different probe densities. τ_{sim} values with ± standard deviation were computed after completion of 500 replicates. ΔE values are ± standard deviation between each analyzed simulation frame.

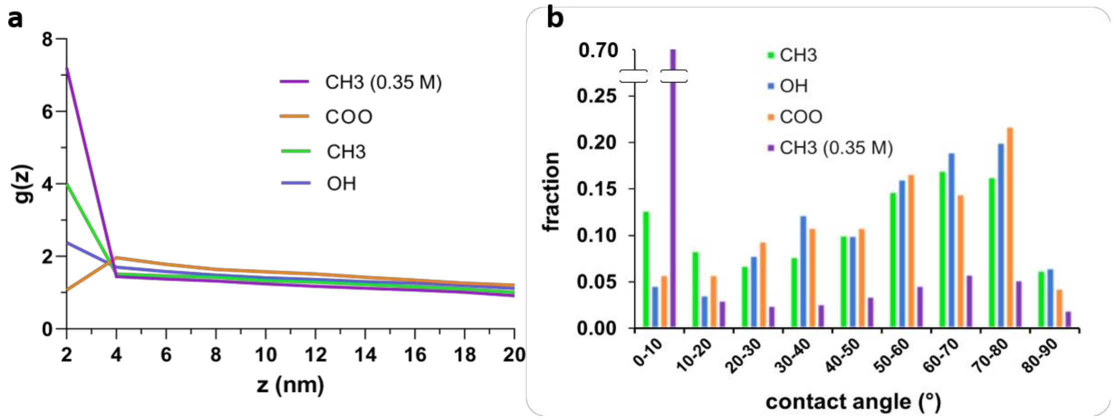


Figure 3.3 (a) Distribution of target ssDNA strands as a function of their distance above the SAM in the four different single-probe scenarios simulated at the 0.001 nm^{-2} probe density. The space was partitioned into equally-sized slices with a height of 2.0 nm. A target was considered to reside in the closest slice to the SAM containing any of its atoms. The number of frames in which the target resided in a slice divided by the expected number of appearances in a slice (assuming totally random diffusion) is calculated for each replicate simulation. The final result is averaged over all replicates. This ratio (g(z)) is plotted on the y-axis. **(b)** Distributions of contact angles measured between surface-adsorbed ssDNA targets and the SAM for each different surface and for the two salt concentrations on the CH_3-SAM. Angles were measured between the vector running from tip-to-tip of the DNA and the plane of the SAM (x-y plane).

3.4.2 Effect of Ionic Strength on DNA Hybridization

Because DNA is a highly charged biomolecule, we examined the effect of ionic concentration on hybridization. We ran a second simulation on the CH_3-SAM at the 0.002 nm^{-2} probe density with the same settings except that the implicit ion concentration was lowered from 0.50 M to 0.35 M. Because the electrostatic interactions are less screened, the target-SAM electrostatic attraction (ΔE_{elec}) was strengthened from -0.35 kcal/mol (0.5M) to -0.96 kcal/mol (0.35M), and ΔE_{vdw} was ~0.4 kcal/mol stronger with 0.35M ionic concentration (Table 2). These stronger attractions nearly doubled the target concentration in the range 0-2.0 nm above the SAM (Figure 3.3a). Interestingly, despite the high surface concentration of targets, the target-SAM interaction is still insufficient to attract the ssDNA to produce more 2D hybridization. As a result, the 2D hybridizing fraction increased negligibly (Table 2). Moreover, the stronger electrostatics induced more of the targets to adopt a side-on adsorption orientation, as opposed to end-on, which can be seen in the strikingly high fraction of DNA making a small contact angle with this surface (Figure 3.3b). These orientations contribute to the increased strength of attraction anchoring targets to the SAM. This, coupled with the increased concentration of targets near SAM, resulted in longer times needed for the targets to overcome the energy barrier associated with leaving the surface and reorienting for hybridization, causing τ_{sim} to nearly double relative to the 0.50 M simulation (Table 2). This highlights the importance of the balance between long-range attractive steering forces and mobility of targets once they reach the surface. Steering target DNA toward probes is obviously helpful for fast kinetics, as shown in the 0.50 M single-probe simulations. However, the same forces responsible for steering may be disruptive. Adsorption of targets to a surface can increase their hybridization rate as long the desorption energy is not prohibitively high. Otherwise, the attractive forces originating from the sensor surface actually hinder the rate dramatically.

3.4.3 High-probe density multi-probe simulations

While a higher density of probes increases the chances of a target-probe encounter, there may be a density beyond which hybridization becomes sterically or energetically hindered by neighboring probes. To study hybridization at tightly-packed probes, we ran multi-probe simulations that included a cluster of five probes on either the CH_3- or COO^--SAM spaced 5.0, 10.0 or 15.0 nm apart in simulation boxes of identical volume (Figure 3.1b). These were performed to test the effect of inter-probe spacing on hybridization. The trend in hybridization time with decreasing separation shows that the hybridization rate has already saturated at inter-probe spacings of 15.0 nm, and hybridization becomes hindered at inter-probe spacings of 5.0 nm or less (Table 3). We saw no change in hybridization time between the 15.0 and 10.0 nm separation cases, but observed a roughly 1.5-fold increase from 10.0 to 5.0 nm, indicating that hybridization is inhibited in such highly crowded probe environments. To understand the source of the inhibition, we counted the total number of trajectories that hybridized with each of the five different probes (labeled 1-5, Figure 3.1b), to elucidate how the neighboring probes affect one another's availability for hybridization. The result shows an unequal distribution of hybridization events. Specifically, hybridization at the central probe is severely limited at the 10.0 and 5.0 nm spacings, and the effect is greater for the smaller spacing (Table 3). In that case, only about 2% of the hybridization happened at the central probe for both the CH_3- and COO^--SAMs, while the outer probes experienced a much more equal share. This effect was less pronounced, but still substantial, at the 10.0 nm spacing. The high negative charge of the DNA molecule makes it difficult for targets to adopt an energetically favorable orientation in the crowded probe environment. Steric limitations may begin to play a role here as well, especially when then target length is approximately equal to the probe spacing, as the target may adsorb to multiple probes at the same time. This was the case in our 5.0 nm-spacing simulation, and was partially responsible for the sharply increased hybridization time observed on both surfaces. Notably, the target and probe are both roughly 10 nm long, indicating that, more generally, tight probe clustering may become an obstacle as the spacing nears the length of the target DNA. The layer-wise distribution of targets in the multi-probe scenarios was virtually the same as in the single-probe scenarios. Overall, both the CH_3- and COO^--SAM sensors showed inhibited hybridization when the probe spacing was equal to or smaller than the target DNA length. We should also note that the cluster of probes here differs from what would exist on a real biosensor surface in an important way: a real surface of the same probe density would present essentially a periodic array of such clusters, in which virtually all the probes would be analogous to the central probe (probe 5, figure 3.1b). Hence, the present simulations contain an unrealistic amount of space surrounding the cluster, and the hybridization times measured for these scenarios should only be compared relative to the other simulated clusters.

			hybridizing probe (%)					
	separation (nm)	**τ_{sim} (μs)**	**1**	**2**	**3**	**4**	**5**	**γ**
CH_3-SAM	5.0	22.6 ± 0.7	23	23	26	26	2	0.13
	10.0	14.4 ± 0.1	22	20	27	23	7	0.20
	15.0	15.7 ± 0.4	21	13	24	27	14	0.24
COO^--SAM	5.0	20.6 ± 0.4	15	30	29	24	2	0.03
	10.0	15.4 ± 0.2	23	19	28	26	4	0.03
	15.0	16.1 ± 0.5	15	19	24	22	20	0.04

Table 3.3 Inter-probe separation, average hybridization time (τ_{sim}), percentage of hybridization at each probe (labeled 1-5, see Fig. 1b), and fraction of 2D hybridization (γ) for multi-probe simulations on the CH_3- and COO^--SAMs. Probes 1-4 are the outer probes of the cluster, and should be energetically and sterically equivalent for incoming targets. Probe 5 is at the center of the cluster, and shows greatly reduced accessibility dependent on the probe separation.

3.4.4 Phenomenological model and comparison with simulation results

A model that relates hybridization time to a sensor's basic properties can aid in the rational design of high-performance biosensors. Real-time hybridization kinetics can be measured experimentally,[49] but it is impossible to determine the mechanism of hybridization, and the details of target interactions with surfaces and probes cannot be discerned. Here we developed a model using kinetic theory of particle association and the simulated percentage of 2D hybridization to predict the average hybridization time, τ_{theo}. We adopt the theoretical framework for predicting the rate of collisions, α, between a diffusing molecule and a spherical target containing a small reactive patch. This sphere can also be imagined as a flat surface containing a small reactive hemispherical bump[50], where in the present case, the flat surface is the SAM and the hemispherical bump is the probe. In that case α is the rate of collisions with the SAM

$$\alpha = \frac{4\pi D}{\int_{a+b}^{\infty} \frac{1}{r^2} e^{\frac{U(r)}{kT}}\, dr} \tag{3.2}$$

where D is the 3-dimensional diffusion coefficient, *a* is the radius of the diffusing molecule, *b* is the radius of the surface or reactive patch of surface, and r is the distance between them. U(r) is the potential energy between the two molecules. For diffusion from the bulk to the SAM, the driving force is the long-range Coulombic potential

$$U(r) = \frac{q_1 q_2}{4\pi\epsilon_0 \epsilon r} \tag{3.3}$$

After integration (1) becomes

$$\alpha = 4\pi D r \left(\frac{U(r)}{e^{U(r)} - 1} \right) \tag{3.4}$$

Multiplying equation 3 by Avogadro's number gives units of $M^{-1}s^{-1}$. The distance r here is the sum of the relevant radii (*a* + *b* in equation 1). Since DNA has a rod-like geometry, the radius of the target DNA, a, is approximated as

$$r_{dna} = \frac{s_1}{\ln \frac{2 s_1}{s_2}} \tag{3.5}$$

Where s_1 and s_2 are the major and minor semi-axes.[51] In studies of molecular association with surface-bound receptors, it has been shown that short-range (vdW) forces which hold the diffusing molecules to the surface can result in molecular association rates as if the entire surface were reactive. This is because an encounter with the secondary search target (probe) is very likely and happens rapidly compared to an encounter with the primary target (SAM surface).[52] Adopting this idea, our model uses the radius of the SAM as b in equation 1, which is taken as the radius of a sphere having equal surface area. Our results show that this model works well for the present system (Table 2). As a comparison, if it is assumed that hybridization times will correspond to the rate of target DNA collision with the probe rather than with any portion of the surface, then the times are overestimated more than 2-fold in most cases (see SI for additional details).

Some fraction of targets will encounter the probe while diffusing 2-dimensionally. This fraction must then overcome an energy barrier associated with desorbing from the surface before hybridizing, and the time of this process is added to the time to diffuse from the bulk to the SAM. The barrier is determined primarily by the strength of short-range forces holding targets to the SAM (Table 2). Borrowing ideas from Kramers' theory for diffusion across an energy barrier, in which there is an exponential dependence on barrier height $E^{\dagger}$ and a dependence on the diffusion coefficient D along the reaction coordinate,[47] we approximate the time to cross the barrier, τ_b,

$$\tau_b = \frac{b^2}{D} e^{\frac{E^{\dagger}}{kT}} \tag{3.6}$$

The term before the exponential gives the time needed to diffuse a distance *b*, which we take as 1.0 nm, since at this distance above the SAM, all short-range vdW forces are broken. We analyzed all hybridizing simulation trajectories and labeled each as either a 2D or 3D hybridization event according to a simple criterion related to the diffusional pathway taken by the target. While the hybridization time from 3D diffusion can be directly approximated using the collision rate α (Equations 1 and 3), the average hybridization time for surface-adsorbed DNA needs to consider the barrier height (Equation 5) to account for the extra time spent desorbing from a surface. As a result, the overall τ_{theo} for all trajectories is

$$\tau_{theo} = \gamma\left[\frac{1}{\alpha C} + \frac{b^2}{D}e^{\frac{E^{\dagger}}{kT}}\right] + (1-\gamma)\frac{1}{\alpha C} \tag{3.7}$$

where C is the target DNA concentration in mol/L, and γ is the fraction of trajectories that hybridized 2-dimensionally. A fully detailed calculation is shown in the SI. In all but one case, the predicted hybridization times τ_{theo} agree quite closely with simulations for probe densities 0.001 and 0.002 nm^{-2}; however, they are underestimated when the probe surface density is low (Figure 3.4). Equation 6 relies on the prior assumption that the time spent on or near the surface before achieving 2D hybridization is negligible. This assumption holds when probe density is high and most targets diffusing on the SAM rapidly undergo 2D hybridization, but it becomes less applicable as probe density decreases. Notably, at normal densities used in biosensors, typically > 0.001 nm^{-2}, the assumption holds reasonably well. On all surfaces, predictions have similar accuracy, and suffer from much larger error at the lowest probe density of 0.0005 nm^{-2}. Results from the 0.35 M CH_3-SAM simulation demonstrate the importance of capturing surface interactions in models of hybridization. While the 0.5 M simulation at the 0.002 nm^{-2} density yielded τ_{sim} = 20.4 μs, the stronger target-SAM attraction at 0.35 M resulted in τ_{sim} doubling to 40.1 μs. The model seems to have difficulty capturing such effects. The green points in Figure 3.4 correspond to the CH_3-SAM, which had the most attractive potential with the target DNA. They show that τ_{sim} was overestimated, estimated accurately, and then underestimated by the model as probe density decreased. The same trend can be seen for the anionic COO-SAM, indicating that some of this error is due to the changing probe density. However, the fact that it is more pronounced on the CH_3-SAM is an indication that the amount of surface interaction has still not been well measured. An improved method for estimating the amount of 2D hybridization could vastly improve such measurements. Additionally, target-probe interactions that form as targets desorb from a surface need to be characterized in detail to allow accurate estimation of the 2D hybridization timescale. Despite this, the model's overall accuracy is still reasonable, which highlights the major influence that surface interactions play in the hybridization process. In summary, this model of hybridization works well under the assumption that the search for a probe is dominated by the 3-dimensional part, which is true under typical conditions. As probe density decreases, the model suffers from larger inaccuracy as 2D search time becomes non-negligible. Additionally, accurately estimating the amount of 2D hybridization and strength of surface interactions is essential for accurate predictions of hybridization rates, as 2D hybridization can take considerably longer than 3D hybridization.

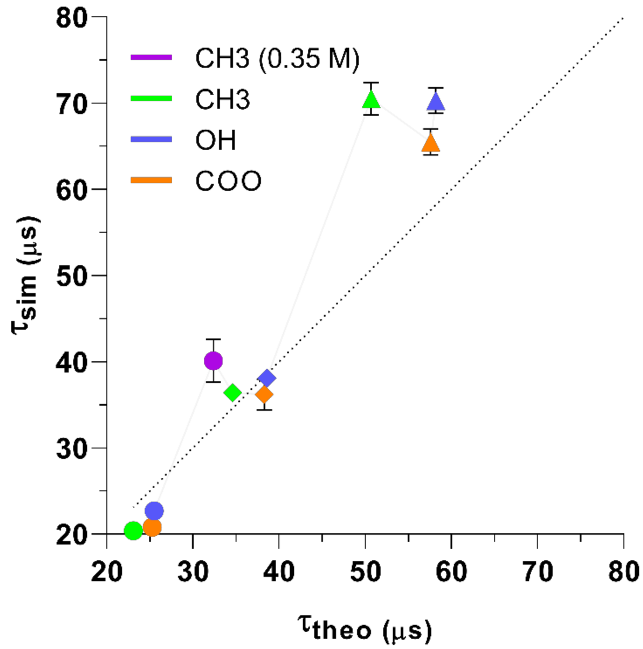


Figure 3.4 Comparison of simulated (τ_{sim}) and predicted (τ_{theo}) hybridization times for the single-probe scenarios at 0.002 nm^{-2} (circles), 0.001 nm^{-2} (squares), and 0.0005 nm^{-2} (triangles) probe densities. Error bars are ± standard deviation of τ_{sim} to show that the value has converged after 500 replicate simulations are completed; some error bars are too small to be displayed.

3.5 Conclusions

High performance biosensors require specific and efficient target-probe association, which depends highly on the surrounding surface properties, probe surface density, and the ionic strength of the solution. In this study, across three surfaces with anionic, polar, or hydrophobic nature, the rate of target DNA hybridization with the probes was largely unchanged at 0.5 M monovalent salt concentration However, upon decreasing the ionic strength of solution to levels where target DNA attraction to the sensor surface was significant, hybridization was markedly slowed as DNA desorption form the surface was impeded, which in turn inhibits hybridization with probes. On the polar and hydrophobic surfaces, a significant fraction of the hybridization took place 2-dimensionally, through a surface mediated mechanism. Simulation of three probe clusters with different inter-probe spacings revealed that hybridization rate eventually reverses as probe density increases since incoming target DNA becomes sterically and energetically hindered from hybridizing with tightly packed probes. Shrinking the probe spacing from 10.0 to 5.0 nm resulted in a doubling of the average hybridization time, revealing that clusters of high probe density in biosensors can severely disrupt hybridization, especially when the spacing is equal to or less than the target DNA length.

We've developed a model DNA of hybridization rate, and found that the key parameter is the fraction of surface-mediated hybridization, since this requires a significant energy barrier to be crossed prior to hybridization. The model accurately predicted hybridization times across all surfaces at typical probe densities, but suffered slightly at low densities where certain assumptions fail to hold. The greatest hybridization efficiency requires striking a balance between attractive forces that enhance target DNA concentration near probes without inhibiting the process by anchoring DNA to the surface. Due to the generality of the model, it can be applied to a variety of other natural or engineered systems to aid in rational design or to help explain experimentally measured reaction rates.

Supporting Information

Explanation of target DNA diffusion coefficient calculation, details for simulation equation of motion, additional details on simulation protocol and models, explanation for analysis of target distribution above the surfaces, full calculation and discussion of modeled hybridization times and comparison to simulated results.

Acknowledgments

This study was supported by the US National Science Foundation (MCB-1932984), and NSF national supercomputer centers (TG-CHE130009). We thank Dr. Joachim Dzubiella for discussion relating to molecular diffusion and associations. We thank Dr. Tao Ye and Qufei Gu for helpful discussions regarding SAM and biosensor structure. Finally, we thank Kevin Maltz for his contributions to the GeomBD source code.

References

1. Faria, H. A. M.; Zucolotto, V. Label-free Electrochemical DNA Biosensor for Zika Virus Identification. *Biosens. Bioelectron.* **2019,** *131*, 149-155.

2. Dong, S.; Zhao, R.; Zhu, J.; Lu, X.; Li, Y.; Qiu, S.; Jia, L.; Jiao, X.; Song, S.; Fan, C. Electrochemical DNA Biosensor Based on a Tetrahedral Nanostructure Probe for the Detection of Avian Influenza A (H7N9) Virus. *ACS Appl. Mater. Interfaces* **2015,** *7*, 8834-8842.

3. Peng, H.-P.; Hu, Y.; Liu, P.; Deng, Y.-N.; Wang, P.; Chen, W.; Liu, A.-L.; Chen, Y.-Z.; Lin, X.-H. Label-free Electrochemical DNA Biosensor for Rapid Detection of Multidrug Resistance Gene Based on Au Nanoparticles/Toluidine Blue–Graphene Oxide Nanocomposites. *Sens. Actuators B: Chem.* **2015,** *207*, 269-276.

4. Hájková, A.; Barek, J.; Vyskočil, V. Electrochemical DNA Biosensor for Detection of DNA Damage Induced by Hydroxyl Radicals. *Bioelectrochemistry* **2017,** *116*, 1-9.

5. Zhou, Y.; Tang, L.; Zeng, G.; Zhang, C.; Zhang, Y.; Xie, X. Current Progress in Biosensors for Heavy Metal Ions Based on DNAzymes/DNA Molecules Functionalized Nanostructures: A Review. *Sens. Actuators B: Chem.* **2016,** *223*, 280-294.

6. Ferapontova, E. E. DNA Electrochemistry and Electrochemical Sensors for Nucleic Acids. *Annu. Rev. Anal. Chem.* **2018,** *11*, 197-218.

7. Drummond, T. G.; Hill, M. G.; Barton, J. K. Electrochemical DNA sensors. *Nat. Biotechnol.* **2003,** *21*, 1192-1199.

8. Kavita, V. DNA Biosensors—A Review. *Journal of Bioeng. Biomedical Sci.* **2017,** *7*, 222-226.

9. Jarczewska, M.; Kierzkowska, E.; Ziółkowski, R.; Górski, Ł.; Malinowska, E. Electrochemical Oligonucleotide-based Biosensor for the Determination of Lead Ion. *Bioelectrochemistry* **2015,** *101*, 35-41.

10. Sun, J.; Gan, Y.; Liang, T.; Zhou, S.; Wang, X.; Wan, H.; Wang, P. Signal Enhancement of Electrochemical DNA Biosensors for the Detection of Trace Heavy Metals. *Curr. Opin. Electrochem.* **2019**, *17*, 23-29.

11. Sassolas, A.; Blum, L. J.; Leca-Bouvier, B. D. Immobilization Strategies to Develop Enzymatic Biosensors. *Biotechnol. Adv.* **2012,** *30*, 489-511.

12. Fan, C.; Plaxco, K. W.; Heeger, A. J. Electrochemical Interrogation of Conformational Changes as a reagentless Method for the Sequence-specific Detection of DNA. *Proc. Natl. Acad. Sci.* **2003,** *100*, 9134-9137.

13. Lin, M.; Song, P.; Zhou, G.; Zuo, X.; Aldalbahi, A.; Lou, X.; Shi, J.; Fan, C. Electrochemical Detection of Nucleic Acids, Proteins, Small Molecules and Cells Using a DNA-Nanostructure-Based Universal Biosensing Platform. *Nat. Protoc.* **2016,** *11*, 1244-1263.

14. Rashid, J. I. A.; Yusof, N. A. The Strategies of DNA Immobilization and Hybridization Detection Mechanism in the Construction of Electrochemical DNA Sensor: A Review. *Sens. Biosens. Res.* **2017,** *16*, 19-31.

15. Wang, S.; Cai, X.; Wang, L.; Li, J.; Li, Q.; Zuo, X.; Shi, J.; Huang, Q.; Fan, C. DNA Orientation-Specific Adhesion and Patterning of Living Mammalian Cells on Self-assembled DNA Monolayers. *Chem. Sci.* **2016,** *7*, 2722-2727.

16. Kelley, S. O.; Mirkin, C. A.; Walt, D. R.; Ismagilov, R. F.; Toner, M.; Sargent, E. H. Advancing the Speed, Sensitivity and Accuracy of Biomolecular Detection Using Multi-Length-scale Engineering. *Nat. Nanotechnol.* **2014,** *9*, 969-980.

17. Josephs, E. A.; Ye, T. Nanoscale Spatial Distribution of Thiolated DNA on Model Nucleic Acid Sensor Surfaces. *ACS Nano* **2013,** *7*, 3653-3660.

18. Luo, X.; Davis, J. J. Electrical Biosensors and the Label Free Detection of Protein Disease Biomarkers. *Chem. Soc. Rev.* **2013,** *42* (13), 5944-5962.

19. Gong, P.; Levicky, R. DNA Surface Hybridization Regimes. *Proc. Natl. Acad. Sci* **2008,** *105*, 5301-5306.

20. Qiao, W.; Chiang, H.-C.; Xie, H.; Levicky, R. Surface vs. Solution Hybridization: Effects of Salt, Temperature, and Probe Type. *Chem. Commun.* **2015,** *51*, 17245-17248.

21. He, Y.; Zhang, J.; Ruffin, S.; Ji, L.; Wang, K.; Levicky, R.; Xia, X. An Electrochemical Study of the Surface Hybridization Process of Morpholino-DNA: Thermodynamics and Kinetics. *Electroanalysis* **2016,** *28*, 1647-1653.

22. Seifpour, A.; Dahl, S. R.; Lin, B.; Jayaraman, A. Molecular Simulation Study of the Assembly of DNA-functionalised Nanoparticles: Effect of DNA Strand Sequence and Composition. *Mol. Simul.* **2013,** *39*, 741-753.

23. Noh, H.; Hung, A. M.; Cha, J. N. Surface-Driven DNA Assembly of Binary Cubic 3D Nanocrystal Superlattices. *Small* **2011,** *7*, 3021-3025.

24. Mereghetti, P.; Kokh, D.; McCammon, J. A.; Wade, R. C. Diffusion and Association Processes in Biological Systems: Theory, Computation and Experiment. *BMC Biophys.* **2011**, *4*, 2.

25. Pastor, R. W.; Zwanzig, R.; Szabo, A. Diffusion Limited First Contact of the Ends of a Polymer: Comparison of Theory with Simulation. *J. Chem. Phys.* **1996,** *105*, 3878-3882.

26. Ando, T.; Skolnick, J. Sliding of Proteins Non-specifically Bound to DNA: Brownian Dynamics Studies with Coarse-Grained Protein and DNA models. *PLoS Comput. Biol.* **2014,** *10*, e1003990.

27. Sprenger, K.; He, Y.; Pfaendtner, J. In *Foundations of Molecular Modeling and Simulation*; Adjiman, C., David A. Kofke, D. A., Snurr, R. Q., Eds.; Springer: Singapore, 2016; pp 21-35.

28. Venable, R. M.; Ingólfsson, H. I.; Lerner, M. G.; Perrin Jr, B. S.; Camley, B. A.; Marrink, S. J.; Brown, F. L.; Pastor, R. W. Lipid and Peptide Diffusion in Bilayers: The Saffman–Delbrück Model and Periodic Boundary Conditions. *J. Phys. Chem. B* **2017,** *121*, 3443-3457.

29. Cholko, T.; Barnum, J.; Chang, C.-E. A. Amyloid-Beta (Aβ42) Peptide Aggregation Rate and Mechanism on Surfaces with Widely Varied Properties: Insights from Brownian Dynamics Simulations. *J. Phys. Chem. B* **2020**, *124*, 5549–5558

30. Liu, J.; Yu, G.; Zhou, J. Ribonuclease A Adsorption onto Charged Self-assembled Monolayers: A Multiscale Simulation Study. *Chem. Eng. Sci.* **2015,** *121*, 331-339.

31. Bell, S.; Terentjev, E. M. Kinetics of Tethered Ligands Binding to a Surface Receptor. *Macromolecules* **2017,** *50*, 8810-8815.

32. Cholko, T.; Kaushik, S.; Chia-en, A. C. Dynamics and Molecular Interactions of Single-stranded DNA in Nucleic Acid Biosensors with Varied Surface Properties. *Phys. Chem. Chem. Phys.* **2019,** *21*, 16367-16380.

33. Roberts, C. C.; Chang, C.-e. A. Analysis of Ligand–Receptor Association and Intermediate Transfer Rates in Multienzyme Nanostructures with All-Atom Brownian Dynamics Simulations. *J. Phys. Chem. B.* **2016,** *120*, 8518-8531.

34. Mishra, G.; Bigman, L. S.; Levy, Y. ssDNA Diffuses Along Replication Protein A via a Reptation Mechanism. *Nucleic Acids Res.* **2020,** *48*, 1701-1714.

35. Graves, D. J. Powerful Tools for Genetic Analysis Come of Age. *Trends Biotechnol.* **1999,** *17*, 127-134.

36. Schena, M.; Heller, R. A.; Theriault, T. P.; Konrad, K.; Lachenmeier, E.; Davis, R. W. Microarrays: Biotechnology's Discovery Platform for Functional Genomics. *Trends Biotechnol.* **1998,** *16*, 301-306.

37. Sheldon, R. A. Enzyme Immobilization: The Quest for Optimum Performance. *Adv. Synth. Catal.* **2007,** *349*, 1289-1307.

38. Maier, J. A.; Martinez, C.; Kasavajhala, K.; Wickstrom, L.; Hauser, K. E.; Simmerling, C. ff14SB: Improving the Accuracy of Protein Side Chain and Backbone Parameters from ff99SB. *J. Chem. Theory Comput.* **2015,** *11*, 3696-3713.

39. Jakalian, A.; Bush, B. L.; Jack, D. B.; Bayly, C. I. Fast, Efficient Generation of High-Quality Atomic Charges. AM1-BCC Model: I. Method. *J. Comput. Chem.* **2000,** *21*, 132-146.

40. Domínguez, C. M.; Ramos, D.; Mendieta-Moreno, J. I.; Fierro, J. L.; Mendieta, J.; Tamayo, J.; Calleja, M. Effect of water-DNA interactions on elastic properties of DNA self-assembled monolayers. *Sci. Rep.* **2017,** *7* (1), 1-8.

41. Elder, R. M.; Jayaraman, A. Structure and Thermodynamics of ssDNA Oligomers Near Hydrophobic and Hydrophilic Surfaces. *Soft Matter* **2013,** *9*, 11521-11533.

42. Park, J. H., and Aluru N. R. Water film thickness-dependent conformation and diffusion of single-strand DNA on poly(ethylene glycol)-silane surface. *Appl. Phys. Lett.* **2010**, *96*, 1-3.

43. Northrup, S. H.; Allison, S. A.; McCammon, J. A. Brownian Dynamics Simulation of Diffusion-Influenced Bimolecular Reactions. *J. Chem. Phys.* **1984,** *80*, 1517-1524.

44. Srinivas, N.; Ouldridge, T. E.; Šulc, P.; Schaeffer, J. M.; Yurke, B.; Louis, A. A.; Doye, J. P.; Winfree, E. On the Biophysics and Kinetics of Toehold-Mediated DNA Strand Displacement. *Nucleic Acids Ress* **2013,** *41*, 10641-10658.

45. Ouldridge, T. E.; Šulc, P.; Romano, F.; Doye, J. P.; Louis, A. A. DNA Hybridization Kinetics: Sippering, Internal Displacement and Sequence Dependence. *Nucleic Acids Ress* **2013,** *41*, 8886-8895.

46. Cha, T.-W.; Boiadjiev, V.; Lozano, J.; Yang, H.; Zhu, X.-Y. Immobilization of Oligonucleotides on Poly (Ethylene Glycol) Brush-Coated Si Surfaces. *Analytical Biochemistry* **2002,** *311*, 27-32.

47. Kastantin, M.; Schwartz, D. K. DNA Hairpin Stabilization on a Hydrophobic Surface. *Small* **2013,** *9*, 933-941.

48. Lin, Y.-C.; Petersson, E. J.; Fakhraai, Z. Surface Effects Mediate Self-Assembly of Amyloid-β Peptides. *ACS Nano* **2014,** *8*, 10178-10186.

49. Xu, S.; Zhan, J.; Man, B.; Jiang, S.; Yue, W.; Gao, S.; Guo, C.; Liu, H.; Li, Z.; Wang, J. Real-time reliable determination of binding kinetics of DNA hybridization using a multi-channel graphene biosensor. *Nat. Commun.* **2017,** *8*, 14902-14911.

50. Jackson, M. B. *Molecular and Cellular Biophysics*. Cambridge University Press: New York, 2006.

51. Berg, O. G.; von Hippel, P. H. Diffusion-Controlled Macromolecular Interactions. *Annu. Rev. Biophys. Biophys. Chem.* **1985,** *14*, 131-158.

52. Adam, G.; Delbrück, M. Reduction of Dimensionality in Biological Diffusion Processes. *Struc. Chem. Mol. Biol.* **1968,** *198*, 198-215.

CHAPTER 4. Simulation-Guided Enzyme Bioconjugate Engineering For Precise Control of Effective Concentration and Enhanced Catalysis

4.1 Abstract

Chemical conjugation of polymers and other macromolecules to enzymes is a widely used strategy to enhance enzyme stability, in vivo circulation, and create immobilized and encapsulated structures. However, rational design of such bioconjugates is limited by a lack of control over the position and quantity of conjugated molecules. Moreover, the catalytic enhancements achieved are often difficult to explain and predict. Here, we demonstrate precise control of DNA conjugation to the phosphotriesterase (PTE) enzyme and show that this technique can be used to predictably control PTE's catalytic rate. Computational modeling and simulation are used to explain and quantify rationally designed DNA's influence on effective substrate concentration around the enzyme. By conjugating DNA at 8 different sites on PTE, our collaborators achieve a 1- to 7- fold decrease in KM of the reaction by increasing the effective substrate concentration and the substrate capture radius. A model which encapsulates these effects is developed and applied to the PTE-DNA bioconjugate and a previously studied horseradish peroxidase (HRP) bioconjugate. It predicts the reduction in KM with remarkable accuracy and is general enough to apply to other similar systems. Using computational insight to guide experiments, we overcome common obstacles in bioconjugate design by showing how catalytic enhancements can be precisely controlled and predicted.

4.2 Introduction

In recent decades, considerable attention has been paid to chemical conjugation of polymers and other nanomaterials on enzymes for enhancement of multipurpose biocatalysts.[1] For example, PEGylation shielding L-asparaginase via covalent attachment reduces its immunogenicity and short half-life along with high maximum saturation velocity (Vmax) and Michaelis constant (Km).[2] The non-degradation and hypersensitivity of PEG in the human body, which raises potential concerns of cumulative chronic toxicity, has been remedied by PEPylation.[3] The obstacles of low overall reusability and cost-effective ratio, usually caused by free biomolecules in solution, have been overcome by immobilization.[4] Additionally, nanoparticles exhibit enzymes' potential applications. Covalent attachment of enzymes on electrodes has emerged as an interest in developing enzymatic fuel cells and biosensors. The immobilized glucose dehydrogenase and bilirubin oxidase on nanoporous AuNPs showed strong oxidative properties compared to free enzymes.[5] Carbon nanotubes covalently attached to high-content glucose oxidase have been used for novel reagentless glucose biosensors.[6] However, controlling the conjugation location precisely remains a challenge. This inability to understand how and where to modify can lead to deleterious effects and loss of a biomolecule's catalytic activity.

Typically, enzymes exist in a non-substrate-rich environment. Thus, the natural metabolic pathway often utilizes a multi-enzyme complexes, which possess structural features to control substrate flux and enhance diffusion toward the active site. This complex promotes the efficient transport and processing of substrates toward the target.[8] Two salient examples are the enzyme superoxide dismutase (SOD) and the bifunctional enzyme thymidylate synthase-dihydrofolate reductase (TS-DHFR). SOD uses charge complementarity to produce substrate-enzyme interactions that enhance enzyme kinetics by directing the substrate to its active site.[9] A positive-charge patch on the surface of TS-DHFR restricts diffusion of a negatively charged reaction intermediate to a pre-defined channel between two active sites, thus promoting substrate channeling and enhancing pathway kinetics.[10] Similarly, kinase proteins are often aided by scaffold proteins that generate high effective substrate concentrations by tethering the kinase and substrate together.[11, 12]

Taking inspiration from nature, DNA modification with a rationally designed sequence has shown promise as a method for controlling substrate flux.[13] The strategy uses sequence-specific DNA, which has a strong binding affinity for the target substrate, allowing tunability of the kinetic enhancement by controlling substrate-DNA interactions. The goal is to enhance the substrate's effective concentration around the enzyme, which, in turn, increases substrate flux through the active site, boosting the catalytic rate.[14]

Phosphotriesterase (PTE), an organophosphate hydrolase, has emerged as an impressive candidate to defend against the organophosphate nerve agent VX.[15, 16] Our collaborators previously found that random cross-linked PTE conjugated by rationally designed DNA produces a roughly 3-fold enhancement of the initial degradation rate of VX by PTE. Even though DNA conjugation unlocks the critical bottleneck of the PTE hydrolysis rate for VX, uncertainty in the amount and position of DNA after conjugation impedes progress. Different mutants yield a 1- to 7- fold change in K_{M} when reacting with paraoxon after the same DNA conjugation.[13] Moreover, a similar effect was seen for the horseradish peroxidase (HRP) enzyme conjugated with DNA (HRP-DNA). HRP contained two attachment positions, one very close to and one very far from the active site, which had widely differing influences on the reaction kinetics.[17] Though the modification at different sites can have strikingly different effects on kinetics, the correlation between the two has not been characterized in detail. This uncertainty currently limits the precise rational design of recombinant proteins. Gentry et al. proposed an enzyme kinetic model in which the improvement is ascribed to an increase in the effective substrate concentration.[14] As a result of the DNA modification, the substrate flux is intensified around the rationally designed DNA. Notably, the DNA fragment also enlarges the capture window to help the active site catch the target substrates.[18] However, very little work has been carried out to optimize and quantify the local concentration effects, and the impact of the site-specific modification on PTE is still unpredictable.

Although engineered bioconjugates have shown significant promise, experimental techniques alone often lack the resolution or precision to accurately explain and predict the associated enhancements. Computational modeling and simulation can be tremendously helpful in this regard since the structure and dynamics of the bioconjugates and their substrates can be studied at sub-nanometer- and picosecond-resolution with methods like molecular dynamics (MD) or Brownian dynamics (BD) simulations. Additionally, computational studies allow for a quick, inexpensive study of countless permutations of a system in order to guide rational design efforts toward the optimal result. Such work has been performed for a range of systems such as nucleic acid biosensors[19], DNA-conjugated enzymes,[20] nanoparticles for drug delivery,[21, 22] and DNA origami nanostructures[23]. Experimental work guided by computational insight therefore presents itself as an extremely powerful method in bioengineering.

Here, we have used a combination of simulation, experiment, and modeling to achieve and accurately predict reductions in K_M after DNA conjugation on PTE. We also compare our model to previous experiments on HRP-DNA. Simulations quantified the effective substrate concentration increase around DNA at two different substrate concentrations. Next, our collaborators created a conjugated system by inducing pAzF mutations at which the DNA was site-clicked to form the complex, PTE-DNA. The PTE-DNA's viability was checked with experimental assays to confirm the correct attachment of DNA and that catalytic enzyme activity was intact. Kinetic assays revealed the reduction in K_M corresponding to each conjugation site, ranging from 1.1- to 7.4-fold. Starting from the substrate concentration data gathered from the simulations, a model created accurately predicts the fold change in K_M after DNA conjugation. Our results show that the K_M decreases because of a combination of increased effective substrate concentration around the DNA and an enlarged substrate capture radius of the conjugated complex.

4.3 Computational Methods

4.3.1 Molecular dynamics simulation protocol

Molecular dynamics (MD) simulations contained the DNA sequence surrounded by substrate molecules in a cuboid simulation box of explicit TIP3P water solvent (Figure 4.1a). Both substrates were simulated at concentrations of 16 mM and 8 mM, requiring 64 or 32 of the molecules, respectively, in a simulation box of approximately 6.5×10^6 Å^3. Systems were minimized using the conjugate gradient method in three steps: first water only, then solute only, and finally the entire system, for 40000, 10000, and 20000 steps, respectively. This was followed by two-stage equilibration, first for water only then for the entire system, for 20-50 ps at 50, 100, 150, 200, 250, 275 and 298 K before production MD runs. Paraoxon and DDVP molecules started their trajectories from the same positions to avoid biasing their relative distribution. Simulations were run for 100 ns in the isothermic-isobaric (NPT) ensemble with three-dimensional periodic boundary conditions and a 2 fs timestep using the *NAMD 2.12* simulation software.[24] The *Amber* force fields Parmbsc1[25] and GAFF[26] were used to model the DNA and substrate, respectively. Parameters for both substrate molecules were assigned by Amber's *antechamber* program using the AM1-BCC semi-empirical charge method.[27] 40 Na^+ ions were added to neutralize the overall system charge. Long-range electrostatics were calculated using the particle mesh Ewald method and a 12 Å cutoff was employed for non-bonded force calculations. The SHAKE algorithm was used to constrain all bonds involving hydrogen and the Langevin thermostat was used to maintain constant temperature.

4.3.2 Brownian dynamics simulation protocol

To study the interaction of paraoxon and DDVP with DNA at 60 μM, it was necessary to use Brownian dynamics (BD) simulations in implicit water in order to sample a statistically significant number of trajectories. Simulations were performed with our in-house BD simulation program, GeomBD3[20], for 750 ns using a 2 fs timestep. The set-up was very similar to the MD simulations, except that, to achieve a 60 μM concentration, only one substrate molecule was present in the 2.75×10^7 Å^3 simulation box at once, starting its trajectory from a randomly selected point around the stationary DNA positioned at the center. All molecules in these simulations were modeled as all-atom rigid bodies (no internal degrees of freedom) and the solvent and ions were both implicit. The simulations use a stationary grid representation of the DNA's electric field computed by a screened Coulomb potential, which extended 100 Å beyond the edges of the DNA. The program ran 500 independent trajectories of both paraoxon and DDVP under identical conditions except for their randomly-chosen starting positions. For both substrates, we repeated another set of 500 trajectories to ensure consistency in the observations (Figure S1).

4.3.3 Radial distribution of paraoxon and DDVP around DNA

To quantify the influence of DNA on the distribution of surrounding paraoxon or DDVP molecules, we calculated the radial distribution of both substrates. The simulation space was partitioned into cylindrical shells centered on DNA (Figure 4.1a). The number of substrate molecules m_i appearing in each cylindrical shell in a given frame i was summed over n total simulation frames and then divided by n to give the average number of molecules in each shell per frame. This quantity divided by the volume V of the shell is [S], the effective substrate concentration in a shell

$$[S] = \frac{\frac{1}{n}\sum_{i=1}^{n} m_i}{\mathrm{V}} \tag{4.1}$$

The bulk concentration $[S]_0$ is the number of substrate molecules being simulated divided by the total simulation volume, which can be compared to [S] to yield the ratio $[S]/[S]_0$, the effective concentration enhancement in each shell.

4.3.4 Calculation of paraoxon and DDVP residence time on DNA

The duration of substrate adsorption onto DNA was quantified at both the 16 mM and 8 mM concentrations from the MD simulation trajectories. We refer to the amount of time over which a molecule remained adsorbed to the DNA as the “residence time” of the molecule. The residence time was calculated by counting the number of consecutive frames in which a molecule was less than a cutoff distance (4.5 Å) from the DNA, measured between the closest atomic centers. A molecule had to below this cutoff for three consecutive frames before being considered as “associated” or beyond this cutoff for three consecutive frames before being considered as “dissociated”.

4.4 Results and Discussion

In our previous work we showed that rationally designed DNA randomly conjugated on PTE resulted in a decrease in K_M without changing k_{cat}.[13] The central hypothesis is that, instead of changing the active site, the rationally designed DNA increased the effective substrate concentration around the enzyme, more effectively shuttling substrate to the active site for catalysis. To prove this concentration increase, we used a combination of MD and BD simulations of two substrates, paraoxon and DDVP, surrounding the DNA sequence and analyzed their distributions and interactions with DNA. Paraoxon is known to have a high affinity for DNA, whereas DDVP, used here as a negative control, experiences much weaker attraction. We ran the simulations at two different concentrations, 16 mM using MD and 60 μM using BD. We then conjugated DNA onto PTE at eight sites covering a wide range of distances from the active site, which resulted in a 1.1- to 7.4-fold K_M decrease. Finally, a predictive model is developed and applied to our PTE-DNA system and a previously studied HRP-DNA system, and shows remarkably high agreement with the experimental K_M enhancement data.

4.4.1 Modeling effective substrate concentrations with MD and BD simulations

Effective substrate concentrations in zones around the DNA were simulated by starting with a bulk concentration of 16 mM in the MD and 60 μM in the BD simulations. We ran two replicates of each simulation to ensure consistency in the results, but we present data from only one set of simulations in the main text. Data from both sets is available in the supporting information (Figure S1). MD simulations are significantly more detailed than BD, and are especially more effective at capturing close-range intermolecular interactions, but come a much greater computational cost. A 16 mM substrate concentration allows for simulation of the maximum number of substrate molecules (64 in total) within a tractable simulation volume, while avoiding concentrations many orders of magnitude higher than experimental conditions. The BD simulations were carried out at 60 μM to match experimental substrate concentrations.

4.4.2 16 mM substrate MD simulation

The plot of $[S]/[S]_0$ versus distance calculated from the MD simulations in Figure 4.1b shows an increase in effective concentration of both paraoxon and DDVP which grows as the distance from DNA decreases. This effective concentration effect was clearly stronger for paraoxon, especially in the first shell. The value of 16.2 for this shell indicates that the effective concentration there was 16.2 times higher than would be expected if the substrate distribution were totally uncorrelated with the position of DNA. As expected, DDVP showed a value of 5.45 in this range, indicating some correlation, but significantly less than paraoxon. Some of the substrate in the closest shell was adsorbed to the DNA, and would therefore be unavailable to the enzyme. This presents a slight complication when trying to classify substrate concentration, since adsorbed molecules don't necessarily contribute to the concentration in a relevant way, and should therefore not be counted. Re-analyzing the distribution, but excluding adsorbed molecules in the count, results in a $[S]/[S]_0$ value of 3.6 for paraoxon and 1.0 for DDVP in the first shell. In more distant shells, both substrates show effective concentrations significantly above the bulk, i.e. $[S]/[S]_0$ values greater than 1, that gradually decrease with distance.

4.4.3 60 µM substrate BD simulation

Distributions of substrate at 60 µM were obtained from Brownian dynamics simulations (Figure 4.1b). In these simulations, only one substrate molecule diffused in the simulation space around the stationary DNA sequence. In this way, we were able to perform 500 replicate trajectories of 800 ns each, at concentrations more than 100-fold lower than in the MD simulations in order to match the experimental conditions. These simulations were performed with *only* electrostatic forces present and no van der Waals forces. By simulating the substrates under the influence of only electrostatics, no adsorption to DNA is observed. This allows substrate to diffuse freely for the entire period, and it is revealed that paraoxon experiences a greater effective concentration increase in all shells.

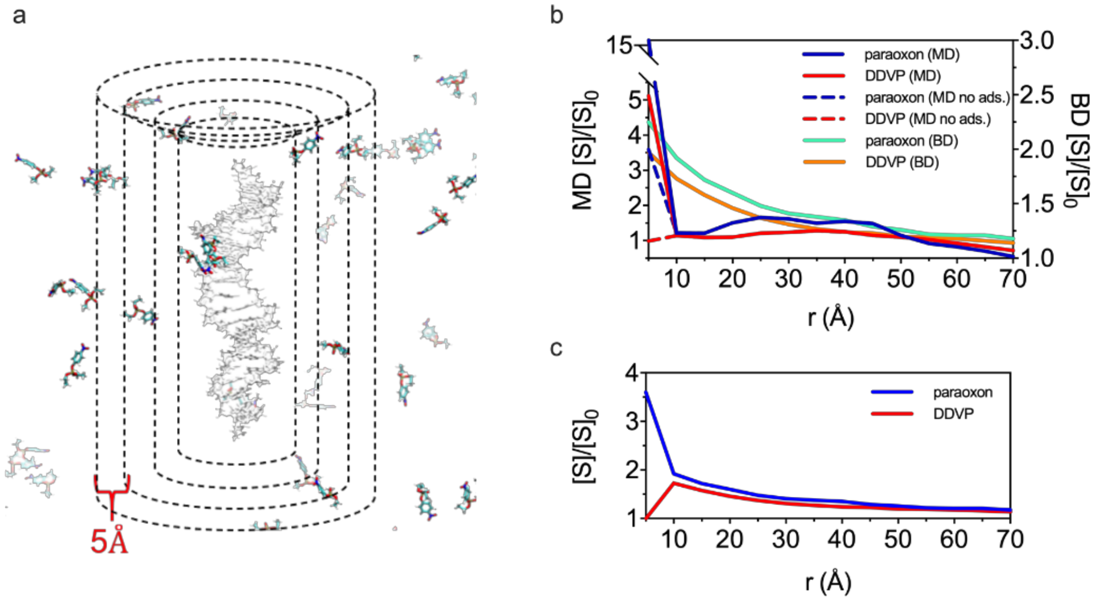


Figure 4.1 (a) An MD simulation snapshot of paraoxon molecules around a segment of dsDNA. Substarte concentration in a 5A-thick cylindrical shells was averaged over simulation time. (b) The effective concentration of paraoxon and DDVP in close proximity to DNA as determined by MD and Brownian dynamics simulations. Left y-axis: MD-determined substrate concentration as a function of distance from the DNA scaffold. Right y-axis: substrate concentration as determined by BD simulations. (c) Combined MD and BD concentration curves.

4.4.4 Interpretation of Simulated Substrate Distributions

Comparing BD simulations to the MD simulations helps discern an effect occurring due to the low number of substrate molecules (64) in the latter. Such a small number results in artificially low concentrations in shells beyond the first, since, when a substrate adsorbs to the DNA, nearby shells become depleted of substrate since there is so surrounding bulk solution from which other molecules can fill in behind it. This results in the dip seen in Figure 4.1a, which causes the gradual uptrend to reverse. Concentration reaches a minimum in the second shell before spiking upward in the closest shell due to the van der Waals (vdW) forces between substrate and DNA that become appreciable at intermolecular distances around 8-10 Å. However, in the BD simulations, this depletion artifact is not present since those simulations included only electrostatic interactions and excluded vdW forces, preventing substrate adsorption. The resulting substrate distributions show a smooth uptrend in all shells, and show that this effect is stronger for paraoxon than for DDVP, in agreement with experimental measurements.[13] This indicates that even with close-range intermolecular vdW forces absent, paraoxon experiences an increase in effective concentration greater than DDVP that intensifies as the distance to DNA decreases. The effects of close-range intermolecular forces were characterized in our MD simulations, and showed stronger attraction between paraoxon and DNA, but caused both substrate concentrations to spike. We therefore use a combination of the MD and BD substrate distribution data to represent the most accurate true distribution. The concentration in the first shell, closest do DNA, is more accurately captured by the MD simulations, while the concentrations in all other shells are more accurately captured in the BD simulations since they do not suffer from the substrate depletion artifact. Combining the data in this way produces Figure 4.1c, with which we proceed to model the overall effect on K_M in later sections.

Additionally, the PTE is conspicuously absent from simulations, which contained only the DNA and substrate. The PTE was excluded from the MD simulations for computational efficiency; including it would have increased the already-large simulation size considerably. BD simulations also excluded PTE in order to make them more comparable to the MD. More importantly, the substrate distribution is certainly driven primarily by DNA due its high concentration of negative charge (-40 *e* overall) compared to PTE, which is just +1 *e* at physiological pH including the two Zn ions in the active site.[29] Therefore, the exclusion of the PTE should introduce no sizeable artifacts.

4.4.5 Adsorption and residence time of substrates on DNA

The effective concentration produced by the conjugated DNA should reflect the increased number of substrate molecules available for catalysis. Therefore, we calculated the average residence time of the two different substrates on the DNA molecule from the MD trajectories to ensure that substrate is available for the chemical reaction. The residence time is the amount of time for which one substrate molecule stayed adsorbed to the DNA. Average residence times for paraoxon and DDVP were 208 ps and 86 ps, respectively. The number of distinct substrate-DNA contacts for each substrate was similar, meaning a roughly equal number of DDVP and paraoxon molecules encountered the DNA, but paraoxon stay adsorbed about 2.4 times longer. This is in agreement with experimental k_d measurements indicating that DDVP dissociates from DNA faster than does paraoxon.[13]

We observed paraoxon adsorbing stably to the minor groove of DNA and to the terminal DNA bases through pi-stacking interactions (Figure 4.2a). At least one paraoxon molecule adsorbed to the major groove of the DNA and subsequently intercalated between two bases (Figure 4.2b), forming a very stable interaction that lasted at least 70 ns, the longest we observed. Intercalation of paraoxon has also been observed experimentally[30], and likely helps maintain a high effective concentration very close to DNA. DDVP interactions with DNA were shorter-lived and electrostatic in nature, mainly taking place near the highly polar backbone and ends of the strand (Figure 4.2c). Overall, the average residence times of both substrates were tens to hundreds of picoseconds, much shorter than the catalytic rate of PTE, which is on the order of 2000 s^{-1}.[28] The simulations show that our rationally-designed DNA sequence successfully brings more substrate molecules near the enzyme and that they are all available for catalysis.

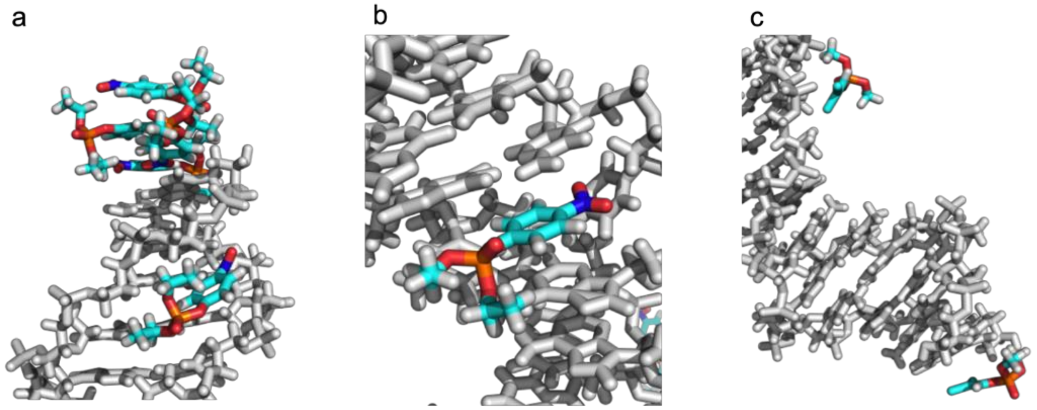


Figure 4.2 MD simulation frames showing various binding modes for paraoxon (a and b) and DDVP (c) with DNA. Paraoxon frequently adsorbed to the major and minor grooves.

4.4.6 Prediction of K_M Enhancement

We have developed a physically intuitive model to predict the fold enhancement of K_M achieved for any conjugation site. To demonstrate the effectiveness of the model, we compare the predictions to experimental data obtained in this work for PTE-DNA as well as experimental data obtained in our previous study on HRP-DNA.[17] K_M is inversely proportional to the enzyme-substate association rate, k_{on}, and can be expressed as

$$K_M = \frac{k_{off} + k_{cat}}{k_{on}} \tag{4.4}$$

where k_{cat} is the rate of the chemical step and k_{off} is the rate of dissociation of substrate from the enzyme.[36] The rate of the chemical step is unchanged by conjugation (Table 1), and it can be safely assumed that k_{off} has little to no change as well. This means the change in K_M is only dependent on the change in k_{on}, and a model that effectively captures this change can predict the change in K_M. A correlation is seen between the effective substrate concentration increase, $[S]/[S]_0$, and the enhancement in K_M ($K_M/K_{M,app}$) in which the two are directly proportional (Figure 4.3)

$$\frac{K_M}{K_{M,app}} \propto \frac{[S]}{[S]_0} \tag{4.5}$$

Although the $[S]/[S]_0$ ratios from the simulation data are proportional to the $K_M/K_{M,app}$, they are clearly not the only factor contributing to the decrease (Figure 4.3). The size of the conjugated DNA may also have a dramatic effect, as it can effectively act as an extension of the enzyme itself, drastically increasing the capture radius of the PTE-DNA complex relative to unconjugated PTE. The standard kinetic theory for collisions per unit time, α, between to molecules experiencing attractive intermolecular forces is

$$\alpha = \frac{4\pi D}{\int_{r_0}^{\infty} \frac{1}{r^2} e^{\frac{-U(r)}{kT}}} = 4\pi D r^* \tag{4.6}$$

where *D* is the relative diffusion coefficient of the molecules and *r* is the sum of the radii of the reactive portions of the two molecules. The new radius *r** is the larger effective radius after accounting for intermolecular attraction[37, 38]. In the present case, we have modeled the effect of DNA conjugation as an increase in r* since the DNA strand acts as a lure for substrate both energetically and sterically. The energetic contribution to the molecular collision rate i.e., the attractive potential between substrate and DNA, is already accounted for in the substrate distribution quantified by the simulations . So, with the $[S]/[S]_0$ ratio as a starting point, we construct a model to predict the fold change in K_M by next accounting for the effective increase in size of the active site after DNA conjugation

$$\frac{r_{AS+DNA}}{r_{AS}} \tag{4.7}$$

where r_{AS} is the radius of the active site and r_{AS+DNA} is the radius of the active site plus the radius of the DNA fragment. The radii of the active site and DNA were measured manually with the *VMD*[39] molecular modeling software. Since both the active site and DNA can be approximated as ellipsoids, the following formula can be used to estimate their effective radii

$$r = \frac{a}{\ln \frac{2a}{b}} \tag{4.8}$$

Where a and b are the major and minor oval semiaxes[38]. However, simply accounting for the linear extension of the active site's radius for capturing substrate is not adequate. As the conjugation site becomes more distant from the active site, the angle between DNA and the active site grows, and the efficiency of substrate capture diminishes. We call this angle the offset angle, denoted θ_{DNA} (Figure S4). The overall effect of the conjugated DNA, γ, is then modeled as

$$\gamma = 1 + \beta \left[\frac{\pi - \theta_{DNA}}{\pi} \left(\frac{r_{AS+DNA}}{r_{AS}} - 1 \right) \right] \tag{4.9}$$

In the parameter γ, the offset angle is allowed to cover the range pi radians, or 180 degrees. When θ_{DNA} is 0, the enlargement of the active site due to DNA conjugation can be fully exploited. When θ_{DNA} is π, the effective extension from DNA conjugation is 0 since the DNA is too far from the active site to increase the flow of substrate there. The factor β is what we refer to as the "blocking factor", and it represents the spatial overlap of the conjugated DNA and active site. If the active site is completely unobstructed, β = 1. If it is 30% blocked, $\beta = 1 - 0.3 = 0.7$, and so on. We note that in all of the modeled $K_M/K_{M,app}$ values for PTE-DNA in Figure 4.3, β has not been adjusted to fit experimental values; it was assumed to be 1 in all cases and thus has no effect on the model output. Altogether, the fold decrease in K_M can then be predicted as

$$\frac{K_M}{K_{M,app}} = \frac{[S]}{[S]_0}\gamma \tag{4.10}$$

4.5. 9 Application of the Model to PTE-DNA and HRP-DNA and Comparison to Experiment

The model matches the experimentally measured $K_M/K_{M,app}$ of the PTE-DNA system quite well (Figure 4.3). Fully detailed example calculations are available in the supporting material. The offset angle effect is evident in the experimental results for Lys294 and Ala364, located 70 Å and 73 Å from the active site, respectively. At these positions, the fold change in K_M was approximately the same as $[S]/[S]_0$, indicating that the DNA was not effective in shuttling substrate to the active site; it only served to increase the effective substrate concentration in the vicinity of the enzyme. This is consistent with our model, which produces $\gamma = 1$ for both Lys294 and Ala364, indicating there is no effective extension of the active site and subsequently $K_M/K_{M,app} = [S]/[S]_0$. Agreement between modeled and experimental $K_M/K_{M,app}$ remains strong for the next 7 conjugation sites between roughly 55 to 8 Å from the active site, where $\gamma > 1$. However, agreement diverges at the closest conjugation site, Asp133. This site, just 6 Å from the active site, demonstrates the importance of β, the blocking factor. Conjugation here yielded only 4.41-fold K_M enhancement, significantly less than that for Lys175, which is 10 Å from the active site (Figure 4.3). We hypothesize that the reason for this sudden reversal is that access to the active site has become severely obstructed by the DNA at

Asp133. This is confirmed by our molecular models which show between a 5-40% blocking of the active site entrance depending on the rotational state of the DNA (Figure S5). Based on our results, a β value for Asp133 of about 0.31 brings the model and experiment into agreement. However, 0.31 implies about 69% reduction in active site accessibility, indicating the blocking effect was underestimated. We discuss the likely causes at the end of this section.

To test the robustness of the model, we applied it to a DNA-conjugated HRP enzyme (HRP-DNA) for which we have previously obtained K_M enhancement data.[17] The DNA used in those experiments was the same fragment used in the current study, and the substrate, tetramethylbenzidine (TMB), has a very similar affinity for the DNA as paraoxon.[17] In the HRP-DNA experiment, conjugation was done at two sites, Lys149 and Lys174, which are 8 and 40 Å, respectively, from the HRP active site. Under the experimental conditions, the DNA is conjugated at one or the other site, but never both, so the measured $K_M/K_{M,app}$ of 2.9 reflects the average of both sites. Assuming the active site is completely unobstructed, the model predicts an 11.68- and 2.67-fold decrease in K_M for Lys149 and Lys174, respectively, yielding an average of 7.17. This is much greater than the experimentally obtained value of 2.9. However, as with Asp133 on PTE, conjugation at Lys149 of HRP appears to severely disrupt the active site. In this case, there can be roughly 5-30% blocking of the active site surface area (Figure S5), which again underestimates the effect. A β value of 0.27 (implying 73% loss of accessibility) for Lys149 is required to reach agreement with experiment, in which case the model predicts $K_M/K_{M,app}$ = 3.15, and the average for the two sites then becomes 2.91.

As we have shown, the factor β can be difficult to estimate, as the blocking of the active site in both cases was worse than would be estimated by simply assuming that accessibility is decreased by the percentage of blocked active site surface area. In reality, the dynamics of the conjugated complex make the effect quite complicated. DNA may lean considerably or even adsorb to the surface of the enzyme, or it could change the conformation of key active site residues. Our results show the effect of conjugating too close to the active site is approximately 2- to 3-fold greater than predicted based simply on the spatial overlap. Fortunately, our results also indicate that this disruption is only an issue at distances less than roughly 8 Å, so β can be entirely ignored the vast majority of the time. The reliance of parameter γ on only the size of the active site, size of the DNA, and conjugation site means it can be generalized to any enzyme-DNA system. Measurement of the enzyme active site is somewhat at the discretion of the user. We measured the maximum distance between residues at the perimeter of the oval-shaped cavity housing the catalytic residues in PTE and HRP (Figure S6b). The reliance of the overall model on $[S]/[S]_0$ means it is substrate-dependent, as the substrate distribution will depend on its unique interactions with the particular DNA sequence used for conjugation. However, the rough substrate distribution can be easily acquired in a matter of hours from molecular simulation techniques such as Brownian dynamics.

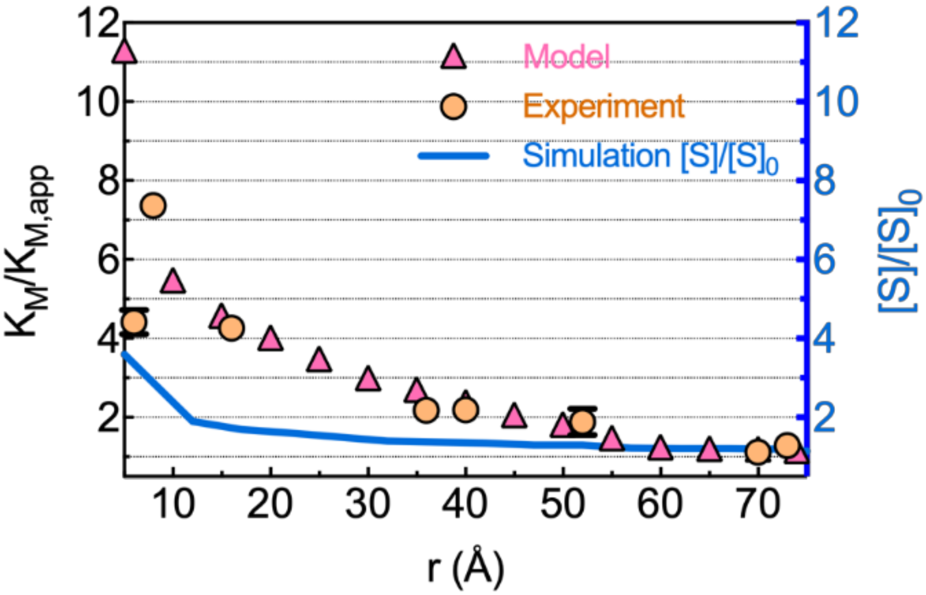


Figure 4.3. Mechanistic modeling captures position-dependent kinetic enhancement PTE-DNA conjugates. Left y-axis: Experimental and modeled $K_{M,app}$ enhancement. Right y-axis Effective susbstrate concentration predicted by combined MD and Brownian dynamics simulations.

4.5 Conclusions

We showed that computationally-guided enzyme bioconjugate design can result in a predictable K_M enhancement. Here, rationally designed DNA conjugated on PTE preferentially attracts paraoxon over DDVP, creating a significantly increased effective concentration around the enzyme, which subsequently decreases K_M of the reaction. Our simulations showed that the concentration increase is due to both electrostatic attraction and short-range vdW forces, and allowed the concentration gradient around DNA to be quantified. Using the site-click method, our collaborators performed precise conjugation of DNA onto PTE at sites covering a wide range of distances from the PTE active site to test the effect on K_M. The results showed that a greater decrease in K_M is seen as DNA is conjugated closer to the active site. This, coupled with simulation data that shows a substrate concentration increase that intensifies closer to DNA, supports a conclusion that K_M enhancement is due in part to an effective substrate concentration increase around the enzyme.

Our best result saw a 7.4-fold decrease with DNA conjugated 10 Å from the active site. Moreover, the K_M is decreased further by the DNA itself increasing the effective size of the active site, resulting in faster shuttling of molecules to the enzyme. These effects are captured in our simple and intuitive model that accurately predicts the fold change in K_M for two different DNA-conjugated enzymes. Both our model and experiments show that the K_M enhancement is maximized by conjugating DNA as close to the active site as possible without obstructing access by the substrate. The model relies only on active site and DNA size as well as the angle between conjugation site and active site, meaning it can be applied to other similar systems with little or no modification of the current parameters.

References

1. Nemzer, L. R.; Schwartz, A.; Epstein, A. J., Enzyme entrapment in reprecipitated polyaniline nano-and microparticles. *Macromolecules* **2010,** *43* (9), 4324-4330.

2. Yadav, D.; Dewangan, H. K., PEGylation: An important Approach for Novel Drug Delivery System. *Journal of Biomaterials Science, Polymer Edition* **2020,** (just-accepted), 1-13.

3. Hou, Y.; Lu, H., Protein PEPylation: A new paradigm of protein–polymer conjugation. *Bioconjugate chemistry* **2019,** *30* (6), 1604-1616.

4. Yata, V. K.; Ranjan, S.; Dasgupta, N.; Lichtfouse, E., Nano-pharmaceuticals: Principles and Applications Vol. Springer: 2020.

5. Siepenkoetter, T.; Salaj-Kosla, U.; Xiao, X.; Conghaile, P. Ó.; Pita, M.; Ludwig, R.; Magner, E., Immobilization of redox enzymes on nanoporous gold electrodes: applications in biofuel cells. *ChemPlusChem* **2017,** *82* (4), 553-560.

6. Zhou, J.; Li, H.; Yang, H.; Cheng, H.; Lai, G., Immobilization of glucose oxidase on a carbon nanotubes/dendrimer-ferrocene modified electrode for reagentless glucose biosensing. *Journal of Nanoscience and Nanotechnology* **2017,** *17* (1), 212-216.

7. Shukla, A. K.; Verma, M.; Acharya, A., Biomolecules Immobilized Nanomaterials and Their Biological Applications. In *Nanomaterial-Based Biomedical Applications in Molecular Imaging, Diagnostics and Therapy*, Springer: 2020; pp 79-101.

8. Park, J. O.; Rubin, S. A.; Xu, Y.-F.; Amador-Noguez, D.; Fan, J.; Shlomi, T.; Rabinowitz, J. D., Metabolite concentrations, fluxes and free energies imply efficient enzyme usage. *Nature chemical biology* **2016,** *12* (7), 482-489.

9. Getzoff, E. D.; Tainer, J. A.; Weiner, P. K.; Kollman, P. A.; Richardson, J. S.; Richardson, D. C., Electrostatic recognition between superoxide and copper, zinc superoxide dismutase. *Nature* **1983,** *306* (5940), 287-290.

10. Elcock, A. H.; Potter, M. J.; Matthews, D. A.; Knighton, D. R.; McCammon, J. A., Electrostatic channeling in the bifunctional enzyme dihydrofolate reductase-thymidylate synthase. *Journal of molecular biology* **1996,** *262* (3), 370-374.

11. Speltz, E. B.; Zalatan, J. G., The relationship between effective molarity and affinity governs rate enhancements in tethered kinase-substrate reactions. *Biochemistry* **2020**.

12. Dyla, M.; Kjaergaard, M., Intrinsically disordered linkers control tethered kinases via effective concentration. *bioRxiv* **2020**.

13. Lang, X.; Hong, X.; Baker, C. A.; Otto, T. C.; Wheeldon, I., Molecular binding scaffolds increase local substrate concentration enhancing the enzymatic hydrolysis of VX nerve agent. *Biotechnology and Bioengineering* **2020**.

14. Gentry, R.; Ye, L.; Nemerson, Y., A microscopic model of enzyme kinetics. *Biophysical journal* **1995,** *69* (2), 356-361.

15. Bigley, A. N.; Xu, C.; Henderson, T. J.; Harvey, S. P.; Raushel, F. M., Enzymatic neutralization of the chemical warfare agent VX: evolution of phosphotriesterase for phosphorothiolate hydrolysis. *Journal of the American Chemical Society* **2013,** *135* (28), 10426-10432.

16. Tsai, P.-C.; Bigley, A.; Li, Y.; Ghanem, E.; Cadieux, C. L.; Kasten, S. A.; Reeves, T. E.; Cerasoli, D. M.; Raushel, F. M., Stereoselective hydrolysis of organophosphate nerve agents by the bacterial phosphotriesterase. *Biochemistry* **2010,** *49* (37), 7978-7987.

17. Gao, Y.; Roberts, C. C.; Toop, A.; Chang, C. e. A.; Wheeldon, I., Mechanisms of enhanced catalysis in enzyme–DNA nanostructures revealed through molecular simulations and experimental analysis. *ChemBioChem* **2016,** *17* (15), 1430-1436.

18. Nakatani, H.; Dunford, H., Meaning of diffusion-controlled association rate constants in enzymology. *Journal of Physical Chemistry* **1979,** *83* (20), 2662-2665.

19. Cholko, T.; Kaushik, S.; Chia-en, A. C., Dynamics and molecular interactions of single-stranded DNA in nucleic acid biosensors with varied surface properties. *Physical Chemistry Chemical Physics* **2019,** *21* (29), 16367-16380.

20. Roberts, C. C.; Chang, C.-e. A., Analysis of Ligand–Receptor Association and Intermediate Transfer Rates in Multienzyme Nanostructures with All-Atom Brownian Dynamics Simulations. *The Journal of Physical Chemistry B* **2016,** *120* (33), 8518-8531.

21. Gupta, R.; Rai, B., In-silico design of nanoparticles for transdermal drug delivery application. *Nanoscale* **2018,** *10* (10), 4940-4951.

22. Ding, H.-m.; Tian, W.-d.; Ma, Y.-q., Designing nanoparticle translocation through membranes by computer simulations. *ACS nano* **2012,** *6* (2), 1230-1238.

23. Roodhuizen, J. A.; Hendrikx, P. J.; Hilbers, P. A.; de Greef, T. F.; Markvoort, A. J., Counterion-dependent mechanisms of DNA origami nanostructure stabilization revealed by atomistic molecular simulation. *ACS nano* **2019,** *13* (9), 10798-10809.

24. Phillips, J. C.; Braun, R.; Wang, W.; Gumbart, J.; Tajkhorshid, E.; Villa, E.; Chipot, C.; Skeel, R. D.; Kale, L.; Schulten, K., Scalable molecular dynamics with NAMD. *Journal of computational chemistry* **2005,** *26* (16), 1781-1802.

25. Ivani, I.; Dans, P. D.; Noy, A.; Pérez, A.; Faustino, I.; Hospital, A.; Walther, J.; Andrio, P.; Goñi, R.; Balaceanu, A., Parmbsc1: a refined force field for DNA simulations. *Nature methods* **2016,** *13* (1), 55.

26. Wang, J.; Wolf, R. M.; Caldwell, J. W.; Kollman, P. A.; Case, D. A., Development and testing of a general amber force field. *Journal of computational chemistry* **2004,** *25* (9), 1157-1174.

27. Jakalian, A.; Jack, D. B.; Bayly, C. I., Fast, efficient generation of high-quality atomic charges. AM1-BCC model: II. Parameterization and validation. *Journal of Computational Chemistry* **2002,** *23* (16), 1623-1641.

28. Roodveldt, C.; Tawfik, D., Directed evolution of phosphotriesterase from Pseudomonas diminuta for heterologous expression in Escherichia coli results in stabilization of the metal-free state. *Protein Engineering Design and Selection* **2005,** *18* (1), 51-58.

29. Breger, J. C.; Ancona, M. G.; Walper, S. A.; Oh, E.; Susumu, K.; Stewart, M. H.; Deschamps, J. R.; Medintz, I. L., Understanding how nanoparticle attachment enhances phosphotriesterase kinetic efficiency. *Acs Nano* **2015,** *9* (8), 8491-8503.

30. Gao, Y.; Or, S.; Toop, A.; Wheeldon, I., DNA nanostructure sequence-dependent binding of organophosphates. *Langmuir* **2017,** *33* (8), 2033-2040.

31. Pickens, C. J.; Johnson, S. N.; Pressnall, M. M.; Leon, M. A.; Berkland, C. J., Practical considerations, challenges, and limitations of bioconjugation via azide–alkyne cycloaddition. *Bioconjugate chemistry* **2017,** *29* (3), 686-701.

32. Chin, J. W.; Santoro, S. W.; Martin, A. B.; King, D. S.; Wang, L.; Schultz, P. G., Addition of p-Azido-l-phenylalanine to the Genetic Code of Escherichia c oli. *Journal of the American Chemical Society* **2002,** *124* (31), 9026-9027.

33. Benning, M. M.; Shim, H.; Raushel, F. M.; Holden, H. M., High resolution X-ray structures of different metal-substituted forms of phosphotriesterase from Pseudomonas diminuta. *Biochemistry* **2001,** *40* (9), 2712-2722.

34. Rochu, D.; Viguie, N.; Renault, F.; Crouzier, D.; Froment, M.-T.; Masson, P., Contribution of the active-site metal cation to the catalytic activity and to the conformational stability of phosphotriesterase: temperature-and pH-dependence. *Biochemical Journal* **2004,** *380* (3), 627-633.

35. Das, S.; Singh, D. K., Purification and characterization of phosphotriesterases from Pseudomonas aeruginosa F10B and Clavibacter michiganense subsp. insidiosum SBL11. *Canadian journal of microbiology* **2006,** *52* (2), 157-168.

36. Canela, E. I.; Navarro, G.; Beltrán, J. L.; Franco, R., The meaning of the Michaelis-Menten constant: Km describes a steady-state. *bioRxiv* **2019**, 608232.

37. Jackson, M. B., *Molecular and cellular biophysics*. Cambridge University Press: 2006.

38. Berg, O. G.; von Hippel, P. H., Diffusion-controlled macromolecular interactions. *Annual review of biophysics and biophysical chemistry* **1985,** *14* (1), 131-158.

39. Humphrey, W.; Dalke, A.; Schulten, K., VMD: Visual molecular dynamics. *Journal of Molecular Graphics* **1996,** *14* (1), 33-38.

CHAPTER 5. A Molecular Dynamics Investigation of CDK8/CycC and Ligand Binding: Conformational Flexibility and Implication in Drug Discovery

5.1 Abstract

Abnormal activity of cyclin-dependent kinase 8 (CDK8) along with its partner protein cyclin C (CycC) is a common feature of many diseases including colorectal cancer. Using molecular dynamics (MD) simulations, this study determined the dynamics of the CDK8-CycC system and we obtained detailed breakdowns of binding energy contributions for four type-I and five type-II CDK8 inhibitors. We revealed system motions and conformational changes that will affect ligand binding, confirmed the essentialness of CycC for inclusion in future computational studies, and provide guidance in development of CDK8 binders. We employed unbiased all-atom MD simulations for 500 ns on twelve CDK8-CycC systems, including apoproteins and protein-ligand complexes, then performed principal component analysis (PCA) and measured the RMSF of key regions to identify protein dynamics. Binding pocket volume analysis identified conformational changes that accompany ligand binding. Next, H-bond analysis, residue-wise interaction calculations, and MM/PBSA were performed to characterize protein-ligand interactions and find the binding energy. We discovered that CycC is vital for maintaining a proper conformation of CDK8 to facilitate ligand binding and that the system exhibits motion that should be carefully considered in future computational work. Surprisingly, we found that motion of the activation loop did not affect ligand binding. Type-I and type-II ligand

binding is driven by van der Waals interactions, but electrostatic energy and entropic penalties affect type-II binding as well. Binding of both ligand types affects protein flexibility. Based on this we provide suggestions for development of tighter-binding CDK8 inhibitors and offer insight that can aid future computational studies.

5.2 Introduction

Cyclin-dependent kinases (CDKs) are among the major regulators of the cell cycle and transcription [1]. The functions of CDKs depend on binding with regulatory proteins called cyclins. CDK8 together with cyclin C (CycC), mediator complex subunit 12 (MED12) and MED13 forms a regulatory kinase module of the mediator complex [2-4], a large protein assembly that couples gene-specific transcriptional regulators to the general RNA polymerase II transcription machinery [5, 6]. A number of studies have shown that CDK8 modulates the transcriptional output from distinct transcription factors involved in oncogenic control [7]. These factors include the Wnt/β-catenin pathway, Notch, p53, and transforming growth factor β [8, 9].

CDK8 has recently attracted considerable attention after it was discovered to have key roles in oncogenesis. The gene expression of CDK8 is related to the activation of β-catenin, a core transcriptional regulator of canonical Wnt signaling in gastric cancers [10-12]. CDK8 is essential in cell proliferation in melanoma and acts as an oncogene in colon cancer in that its expression is amplified in about 60% of colorectal cancer cases [13, 14]. CDK8 gene expression is also related to prognosis in breast and ovarian cancers [15]. Additional cancer-relevant activities of CDK8 include growth factor-induced transcription [16], modulation of transforming growth factor β signaling [17] and phosphorylation of the Notch intracellular domain [18, 19].

The research on selective CDK8 ligands has started only recently but has quickly become highly active. The steroidal natural product cortistatin A was the first-reported high-affinity and selective ligand for CDK8, with IC50 value 12 nM in vitro and complete selectivity against 387 kinases [20]. The existing ligands have two categories based on the major conformations of CDK8 to which they bind. Type I ligands bind to the DMG-in conformation (aspartate-methionine-glycine near the N-terminal region of the activation loop) and occupy the ATP-binding site. The Senexin-type, the newer CCT series, and COT series compounds, which possess 4-aminoquinazoline [21], 3,4,5-trisubstituted pyridine [22] and 6-azabenzothiophene [23] scaffolds, respectively, belong to this category. Type II ligands bind to the DMG-out conformation and occupy mainly the allosteric site (deep pocket) and in some cases the ATP-binding site. The deep pocket is adjacent to the ATP-binding site and is accessible in CDK8 by the rearrangement of the DMG motif from the active (DMG-in) to the inactive state (DMG-out). This pocket is inaccessible in the active conformation (DMG-in), where the Met174 side-chain is reoriented to open up the ATP binding site [24]. Typical type II CDK8 ligands are sorafenib and imatinib analogs that contain an aryl urea core [25]. Research and development of new CDK8 ligands has made significant progress in recent years, and many promising compounds were identified [26-28]. Very recently 4,5-

dihydrothieno[3’,4’:3,4]benzo[1,2-d] isothiazole derivatives were found to have sub-nanomolar in-vitro potency (IC50: 0.46 nM) against CDK8 and high selectivity [29].

Since Scheneider et al. revealed the first crystal structure for human CDK8/CycC complexed with sorafenib (PDBID: 3RGF), in 2011 [30], a total of 25 crystal structures have been made available for this kinase system, thereby providing plenty of structural information for computational approaches to help in understand the atomistic detail of molecular functions and interactions with substrates and ligands. As compared with other CDKs, CDK8 displays additional potential recognition surfaces for interactions, possibly for recognition of MED12, MED13, or the substrates of CDK8. However, all of the crystal structures are lacking 10-20 residues within the activation loop in both the DMG-in and DMG-out conformations, suggesting that the activation loop is highly flexible. However, without structural details of this region, how its motion affects other regions of the protein is unclear and its impact on overall stability and ligand binding is unknown. In addition, although the presence of CycC is crucial in the biological function of CDK8 [30], whether CycC plays a role in ligand binding or protein stability is less clear. Therefore, it is important to gain an understanding of the effect of CycC on the dynamics of CDK8 regions, especially those near the binding sites. If the influence is negligible, CycC could be ignored, thus significantly speeding up calculations. Otherwise CycC must be included in the system to keep calculations accurate and meaningful.

Such information is not available from crystal structures, and one aspect of our study aims to elucidate this relationship. Moreover, we attempt to understand the interaction of ligands with surrounding residues, the stability of the binding modes in crystal structures, and the possibility of alternative binding modes. Computational methods such as molecular dynamics (MD) allow for complementary approaches to understand the details of structural changes during the process of ligand binding [31]. Wu Xu et al., with 50 ns of all-atom MD studies of human CDK8, provided insights into two-point mutations, D173A and D189N, within the activation loop by using hydrogen bond (H-bond) dynamic study of the activation loop residues and the MM/PBSA method [32]. Donatella Callegari and coworkers ranked the residence time of a series of CDK8 type-II inhibitors using metadynamics and the ranking was roughly consistent with the experimental data[33].

In this study, we used all-atom unbiased MD simulations to observe the dynamics of the CDK8-CycC system both with and without bound ligands. The simulations revealed some protein regions that were significantly stabilized by the ligands and showed the effect that other regions may have on ligand binding. We examined and confirmed the importance of CycC to the stability of the CDK8 and found that it facilitates key interactions that stabilize ligand binding modes. Binding of type-I and type-II ligands to CDK8 in DMG-in and DMG-out conformations, respectively, was also simulated. By extensively studying the native bound states of these CDK8/CycC-ligand complexes as well as binding site volume changes, we developed detailed binding energy profiles for each ligand and gained insight that may help improve ligand design.

5.3 Methods

5.3.1 System Description

The subject of this study is the CDK8 protein associated with its partner protein, CycC, along with nine inhibitors of CDK8. CDK8 has two distinct regions known as the N-lobe and the C-lobe. The binding pocket lies between these two lobes and has two regions: the allosteric site, or deep pocket, and the ATP binding site. CDK8 can be in two different conformations, DMG-in or DMG-out, based on the position of a 3-amino-acid sequence called the DMG motif comprising aspartate, methionine and glycine, which are residues 173-175 (Figure 5.1). Type-I ligands bind to the DMG-in conformation and occupy the ATP binding site; type-II ligands bind the DMG-out conformation and occupy the deep pocket and the ATP binding site. CycC associates mainly with the N-lobe and has significant contacts with regions of CDK8 important to the stability of the binding pocket, such as the αC helix.

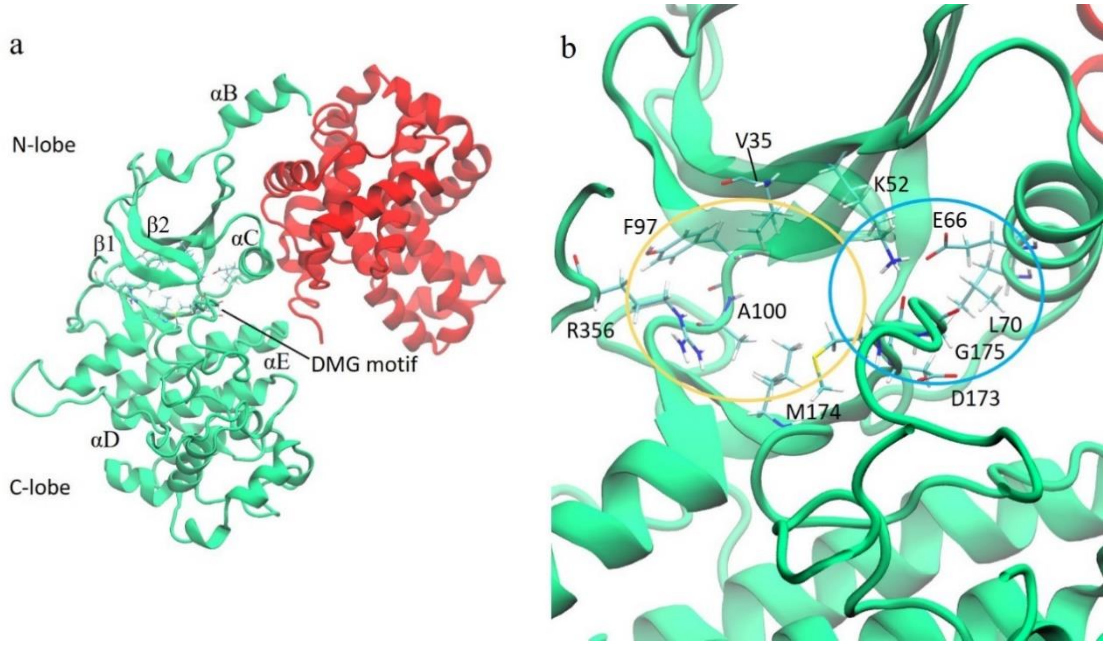


Figure 5.1. (a) CDK8 (green) and Cyclin C (red). (b) a close-up view of the binding pocket of CDK8. Residues that engage in strong interactions with type-I or type-II ligands are labeled with one-letter amino acid codes and shown in licorice. The yellow and blue ovals roughly encircle the ATP and allosteric binding sites, respectively.

We studied DMG-in and DMG-out CDK8/CycC apoproteins, four type-I and five type-II (structure-kinetic relationship series, SKR) CDK8/CycC–ligand complexes. The PDB IDs of the crystal structures used as initial structures and the corresponding MD indices are listed in Table 1. We manually mutated the crystal structure 4F7N and obtained the complex of CDK8/CycC–SKR10, whose crystal structure is not available. For the CDK8/CycC apoproteins in the DMG-in conformation, we used two initial structures, 4G6L and 5CEI, with ligand 50R removed. For the DMG-out apoprotein, we used the initial structure of 4F6W, with ligand SKR1 removed. The molecular structures of the nine ligands are in Figure 5.2. We retained residues 1 to 359 for CDK8 and residues -2 to 257 for CycC for the MD simulations. To build the missing activation loop, we used p38 (PDB ID: 1W82 for the DMG-out conformation and PDBID: 1A9U for the DMG-in conformation) as the reference structure and constructed homology models of the CDK8 activation loop by using SWISS-MODEL [34-36]. Then we aligned residues Asp173 and Arg200 and manually added the homology model of the activation loop to CDK8. We added the other two missing loops αD-αE and αF-αG by using SWISS-MODEL with the native crystal structures as the references.

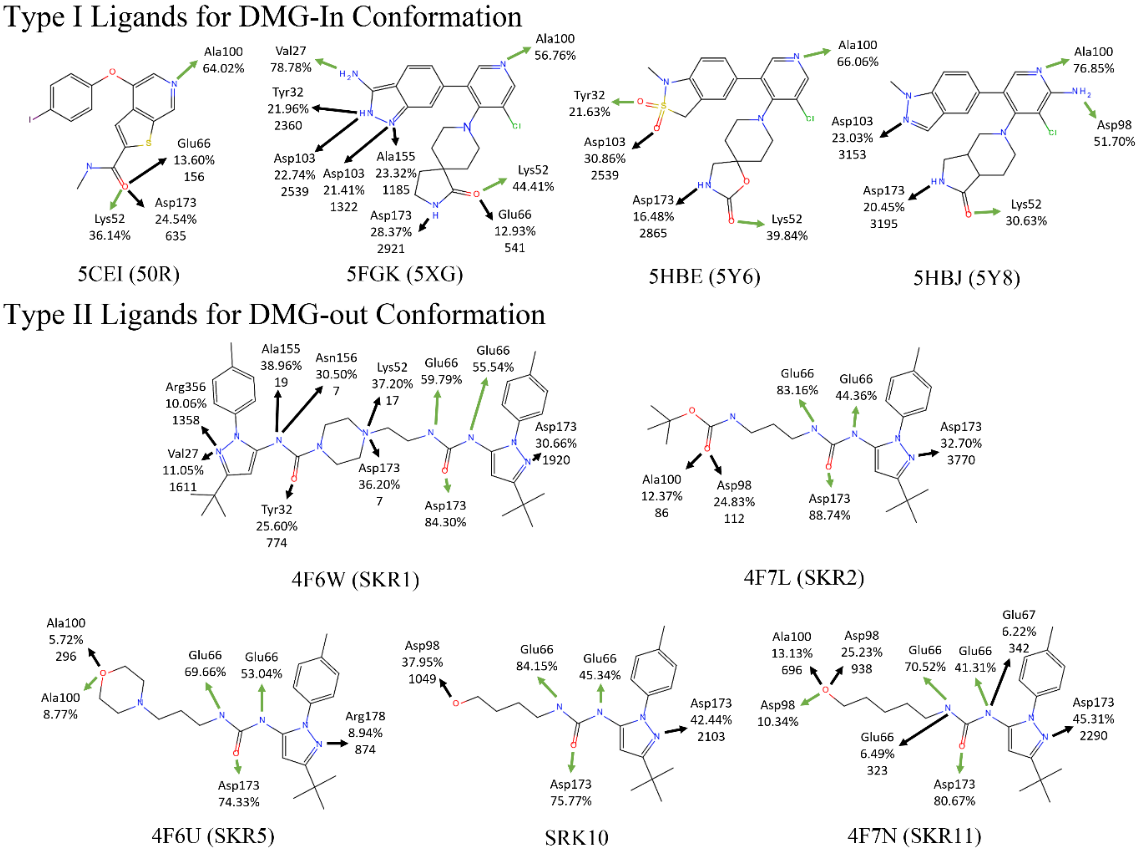


Figure 5.2. Direct H-bonds and water bridges between ligands and CDK8. Green arrows and black arrows indicate direct H-bonds and water bridges between atoms on the ligand with a residue in CDK8, respectively. The occurrence percentages of the H-bonds and water bridges are labeled below the residue name. For water bridges, numbers of observed bridge water molecules are given below the percentage.

MD Index	DMG Conformation	Ligand	PDB ID	Manipulation
1	DMG-In		4G6L [25]	
2	DMG-In		5CEI [24]	Removal of Ligand
3	DMG-In	50R	5CEI [24]	
4	DMG-In	5XG	5FGK [23]	
5	DMG-In	5Y6	5HBE [23]	
6	DMG-In	5Y8	5HBJ [23]	
7	DMG-Out		4F6W [25]	Removal of Ligand
8	DMG-Out	SKR1	4F6W [25]	
9	DMG-Out	SKR2	4F7L [25]	
10	DMG-Out	SKR5	4F6U [25]	
11	DMG-Out	SKR10	4F7N [25]	Manual Ligand Mutation
12	DMG-Out	SKR11	4F7N [25]	

Table 5.1. Initial structures and indices of MD simulations of CDk8/CycC complexes. The references to the crystal structures are provided with PDB IDs.

5.3.2 Unbiased MD simulation

The Amber 14 package with an efficient GPU implementation [37-39] was used for the MD simulations. Amber 99SB and General Amber Force Field (GAFF) [40-42] were used for CDK8/CycC and the 10 ligands, respectively. Single protonation states were used for all histidine residues according to predictions from comparing results for MCCE

[43, 44], ProPKa [45, 46], and DelPhiPKa [47, 48]. Six Cl- ions were placed to maintain a neutral system. Minimization was performed on the hydrogen atoms, side chains and the entire protein complex for 500, 5000, and 5000 steps, respectively, and the system was then solvated with a rectangular TIP3P water box [49] such that the edge of the box was at least 12 Å away from the solutes. The system went through 1000-step water and 5000-step full-system minimization to correct any inconsistencies. Then we equilibrated the water molecules with the solutes fixed for 20 ns at 298K in an isothermic-isobaric (NPT) ensemble. Next, we relaxed the system by slowly heating it during an equilibrium course of 10 ps at 200, 250 and 298 K. We performed the production run in an NPT ensemble with a 2-fs time step and used the Langevin thermostat [50, 51] with a damping constant of 2 ps-1 to maintain a temperature of 298 K. The long-range electrostatic interactions were computed by the particle mesh Ewald method [52]. The SHAKE algorithm [53] was used to constrain water hydrogen atoms during the MD simulations. We performed 500 ns of MD production runs on each complex and the apoprotein by using CPU parallel processing and local GPU machines. We collected the resulting trajectories every 2 ps and re-saved the trajectories for analysis at intervals of 20 ps.

5.3.3 System Dynamics and Flexibility Calculations

Cartesian Principal Component Analysis. To observe major protein motions, we performed classical PCA [54-56] of α-carbon atoms in the 500-ns trajectories saved every 20 ps (25000 frames in total). Using PCA, the complex data set of all α-carbon motions throughout the MD trajectory is reduced to its principal components (PC), the directions which contain the greatest amount of variation (largest motion). The principal components are obtained as the eigenvectors of a covariance matrix consisting of displacements of α-carbons during the trajectory. The first PC mode is the dimension of data with the largest variation, and these were saved and analyzed to reveal the dominant motions. In order to observe motions in different periods of the MD simulations, we divided the aligned 500-ns trajectories of each system into five successive 100-ns trajectories and performed PCA on α-carbon atoms of the entire system in Cartesian coordinates. We calculated the first PC modes of 0-100, 100-200, 200-300, 300-400 and 400-500 ns of MD1, instead of the first PC mode of the entirety of trajectory. In this way, distinct motions of the system occurring in these periods could be captured rather than blending the motion over all 500 ns into one mode. The average positions of the α-carbon atoms were used as a reference to compute the covariance matrix.

Root-mean-square Fluctuation (RMSF) Calculations. The RMSF values of twelve regions of the systems were measured over the 500-ns MD trajectories. We chose regions that were distinct from one another and could plausibly have impacts on ligand binding and or the stability of the CDK8-CycC complex, such as the activation loop, αB helix, and αC helix, among others. A full illustration of the twelve regions and their tabulated RMSF values are shown in Figure 5.4. The RMSF of a region is the average displacement of that region with respect to a reference position taken over the trajectory time, $RMSF = \sqrt{\frac{1}{T}\sum_{t=1}^{T}\left(x_i(t) - x_{i,ref}(t)\right)^2}$ where xi(t) is the position of region i at time t, xi,ref is the reference position of region i, and T is the time interval over which the average is taken. The reference position used here is the average position during the trajectory. The angled brackets mean that the displacement of a region is computed as the average deviation of the atoms that make up that region.

5.3.4 Characterization of Ligand Binding Modes and Binding Energies

Hydrogen bonding analysis. Hydrogen bonds (H-bonds) contribute significantly to the binding interactions of all ligands included in this study. To understand which atoms of the ligands and CDK8 are involved in these interactions, which may provide information that can be used in the design of stronger-binding ligands, we analyzed the trajectory of each protein-ligand complex for H-bonds. In this study, an H-bond (X-H…Y) was considered formed if the distance between H and Y was <2.5 Å and the complimentary angle of X-H…Y was < 30º (Figure S1). We used an in-house script to scan the trajectories for direct H-bonds between ligands and CDK8 as well as mediating water molecules that connect ligands and CDK8. H-bonds between ligands and different atoms on the same residue were merged into one residue–ligand H-bond formation. The occurrence (%) of a H-bond was calculated as the number of the frames containing the H-bond divided by the total frames (25000).

Residue-wise interactions. Further classifying the binding modes, ligand interactions with the 359 CDK8 residues were computed for each of the nine ligands studied. For each residue, we computed the sums of vdW, Coulombic, and Generalized Born (GB) energy terms for the ligand with the residue (EL+R), the ligand alone (EL), and the residue alone

(ER), then computed the interaction energy ΔE = (EL+R) - EL - ER. The vdW term is calculated as a Lennard-Jones potential, EvdW = (A_ij/(r_ij^12)-B_ij/(r_ij^6)), with A = 4εσ12 and B = 4εσ6, where ε is the potential well depth in kcal/mol and σ is the distance at which the potential is zero, and r is the distance between atoms i and j. The Generalized Born energy approximates the solvation energy and was calculated using the Still model [57]. We report only residues that closely interact with the ligands in this analysis.

MM/PBSA. We used the MM/PBSA method [58] to evaluate the intermolecular interactions between a ligand and CDK8/CycC. The method computes the energy (E) of a system from the protein (EP), ligand (EL) and complex (EPL), with the interaction energy computed by Δ<E> = <EPL> - <EP> - <EL>. <E> denotes the computed average energy from a given MD trajectory. The total binding energy term was computed as EMM/PBSA = Ebonded +Eelec + EvdW + GPB + Gnp; where Ebonded is the bonded energy, Eelec and EvdW are electrostatic and vdW energy, GPB is the solvation energy computed by solving the Poisson Boltzmann (PB) equation, and Gnp is the nonpolar energy estimated from the solvent accessible surface area. Because <EPL>, <EP> and <EL> terms were computed using the same bound state trajectory, the bonded term was canceled and is not shown in Table 2.

Binding pocket volume analysis. We used a grid-based in-house program to evaluate the volume of the ATP binding site in order to quantify conformational change of the pocket that accompanies binding. For each conformation, we measured the minimum and maximum of the Cartesian coordinates of the α-carbons of CDK8 binding pocket residues Val27, Gly30, Glu66, Asn156 and Ile171, and divided the space determined by these coordinates into a grid with a spacing of 1 Å along the x, y and z axes. If a grid point is within 1.4 Å (radius of a water molecule) of any atoms of CDK8, it is removed. Otherwise, the grid point is kept in the space. Because the grid spacing is 1 Å, the solvent accessible volume for water of the CDK8 binding pocket is approximated by the number of grid points left over in units of Å3. The same procedure is repeated for each conformation in the trajectories.

5.4 Results and Discussion

Our major areas of analysis were i) CDK8-CycC system dynamics ii) the effect of excluding CycC from MD simulations on dynamics and ligand binding and iii) ligand binding modes and binding site conformational changes with the objective of developing detailed binding energy profiles for four type-I ligands and five type-II ligands (Figure S2). We first present results related the overall system dynamics and regional flexibility. The effects of CycC exclusion are presented next. We repeated MD runs on all twelve systems without the presence of CycC in order to observe the differences in dynamics. These were obtained by measuring RMSF values of twelve major regions of the system and by using PCA on sequential 100-ns portions of the 500-ns MD trajectories of all twelve systems to observe the major global motions. Next, we present binding energy profiles for all nine ligands studied. An in-house script was used to identify all hydrogen bonds (Figure S1) between CDK8 and the ligands and to find their occurrence percentages. Residue-wise interaction analysis was done for all ligand-CDK8 residue pairs to quantify electrostatic, vdW, and desolvation energies important to binding. Finally, MM/PBSA was employed to find the overall binding energies of the ligands. Binding pocket volume analysis for apoproteins and protein ligand complexes in both DMG-in and DMG-out conformations allowed us to assess how the pocket may change to accommodate ligand binding and differences caused by the orientation of the DMG motif.

To further ensure the motions observed in these simulations were not the result of random fluctuations and truly characterized the dynamics of this system, we ran secondary simulations for 200 ns for all twelve of the systems both with and without CycC. These are shorter repeats of the first simulations, run under the exact same conditions but starting with a different random number seed, so that a different trajectory is obtained. The dynamics seen in the secondary runs should recreate that seen in the primary production run and help provide assurance that the observations were not due to randomness and were not strange artifacts of any particular simulation. In our secondary simulations, the same dynamics seen in the first set of simulations was observed in all cases.

5.4.1 CDK8-CycC Dynamics

We examined the first PC modes of the twelve systems over five 100-ns intervals and identified five common global protein motions related to ligand binding and unbinding. These major motions were observed in both DMG conformations, with and without ligands, in all twelve systems. Figure 5.3 shows the five motions: (A) is a breathing motion in which the system bends and unbends about the hinge region connecting the N and C lobes; (C) is rotational motion in which the two lobes rotate back and forth relative to each other; (B) and (D) are the bending and rotational motions between CDK8 and CycC, respectively; (E) consists of the motions of the αB helix, the activation loop, and the loop connecting the αD and αE helixes. All five recurred periodically for all systems over 500 ns but never were all five found within a single 100-ns interval, showing at least 100 ns is required to fully sample system dynamics and indicating that unique motions are unlikely beyond 500 ns. Motions (A) through (D) have considerable impact on ligand binding and unbinding. Because binding sites of proteins are usually enclosed areas, global motions such as these facilitate ligand binding by opening the sites and improving accessibility. The breathing motion was found to be closely related to the binding/unbinding pathways of p38 MAP kinase [61, 62]. Motion (E) may be related to the conversion between DMG-in and DMG-out conformations [63-65]. Projecting MD trajectories onto these PC modes provide a well-defined way to cluster conformations from MD simulations and therefore can be used as a rational way for selecting conformations for molecular docking or other studies that require multiple conformations.

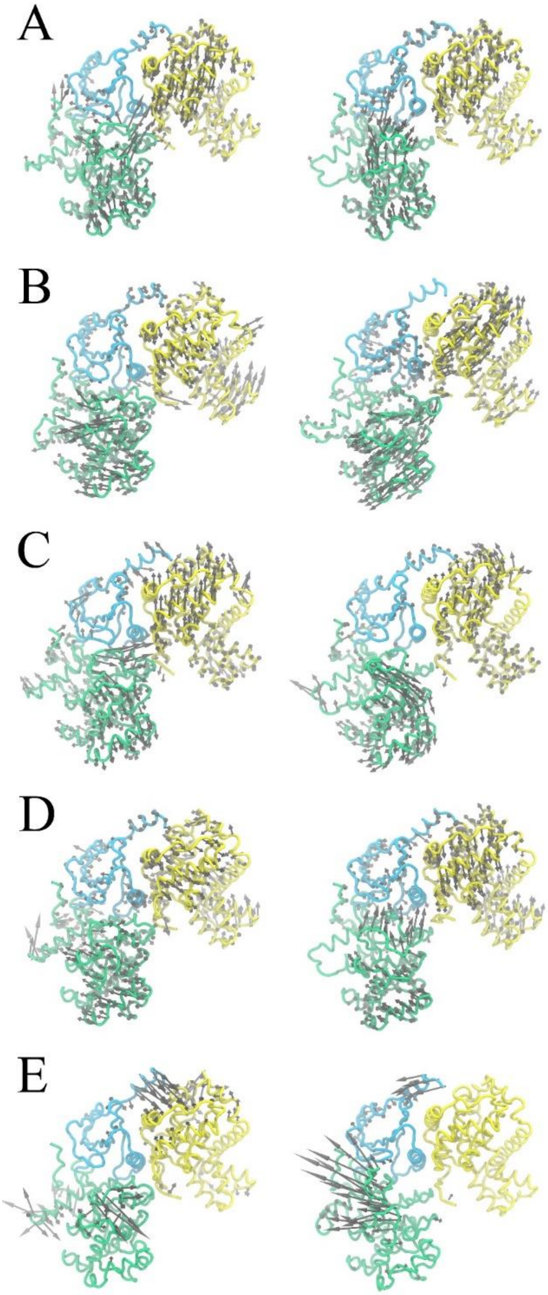
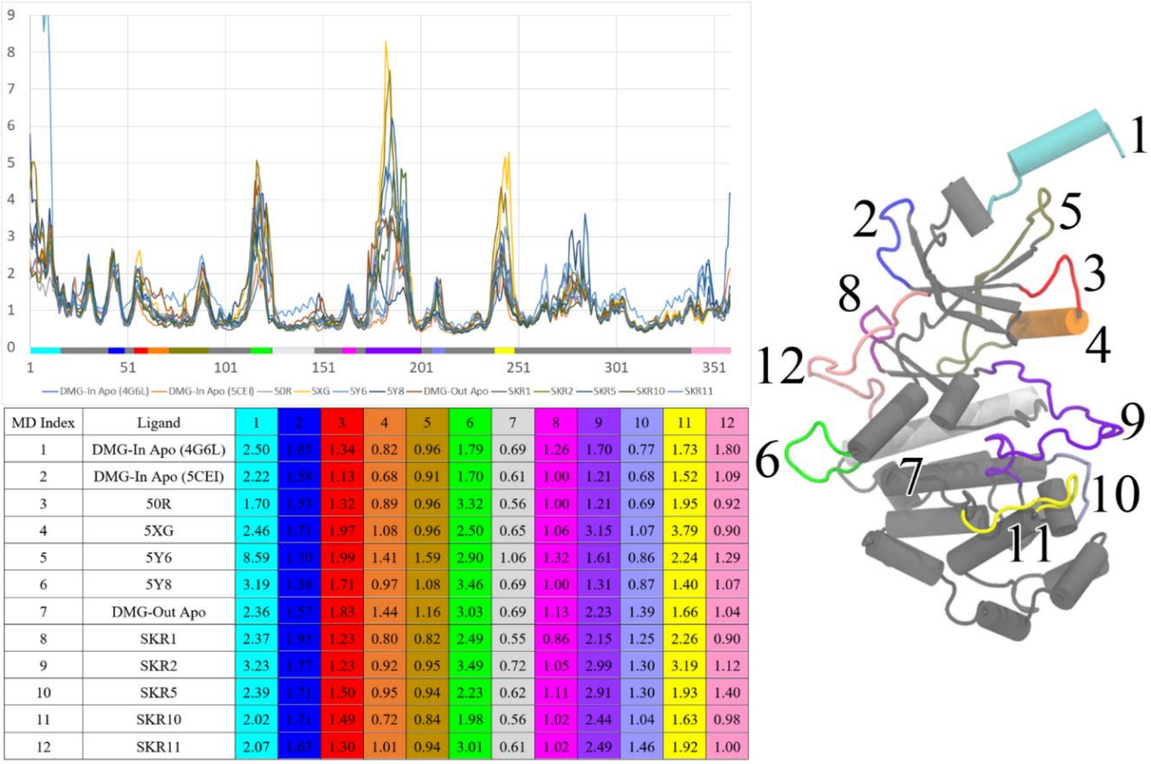

A
B
C
D
E

Figure 5.3. The first PC modes from Cartesian PCA of *apo* CDK8/CycC or CDK8/CycC-Ligand complexes using 100 ns trajectories. The breathing motion between N-lobe (cyan) and C-lobe (green) (A), breathing motion between CDK8 and Cyclin (yellow) (B), rotational motion between N-lobe and C-lobe (C), rotational motion between CDK8 and Cyclin (D), and loop motions (E) are indicated by gray arrows for DMG-In (Left) and DMG-Out (Right) conformations. The way of computing the first PC modes is detailed in Method 2.3 (1). The DMG-In PC modes use MD1 100-200 ns (A), MD3 0-100 ns (B), MD1 200-300 ns (C), MD5 0-100 ns (D), MD5 200-300 ns (E). The DMG-Out PC modes use MD11 100-200 ns (A), MD7 100-200 ns (B), MD11 300-400 ns (C), MD8 100-200 ns (D), MD9 100-200 ns (E).

RMSF measurement of the twelve systems revealed the flexibility of key regions and, most notably, showed that the large motions of the activation loop do not affect binding. The RMSF plots of the twelve systems are shown in Figure 5.4. For clarity, we provide the RMSF values within twelve regions of CDK8. A major difference between RMSFs of DMG-in and DMG-out conformations is in the flexibility of the activation loop (Region 9). Our DMG-in systems consistently showed smaller RMSF values for this region than DMG-out systems (1.2-1.7 vs 2.2-3.0 Å). This finding is supported by DMG-out crystal structures which, due to the higher flexibility, almost always have fewer resolved residues in the activation loop compared to DMG-in structures. Crystal structure 5FGK is missing 17 residues in the activation loop, which indicates a level of flexibility consistent with the abnormally high RMSF we observed of 3.15 Å. Our simulations have revealed that despite the different degrees of flexibility of the activation loop, both type-I and type-II ligands have stable binding conformations, as characterized in section 3.1. Therefore, the natural dynamics of the activation loop may have very limited effects on ligand binding modes in the ATP binding site and allosteric binding site, and it may be unnecessary to consider various loop conformations in future docking-based drug development for CDK8.



MD Index	Ligand	1	2	3	4	5	6	7	8	9	10	11	12
1	DMG-In Apo (4G6L)	2.50	1.85	1.34	0.82	0.96	1.79	0.69	1.26	1.70	0.77	1.73	1.80
2	DMG-In Apo (5CEI)	2.22	1.58	1.13	0.68	0.91	1.70	0.61	1.00	1.21	0.68	1.52	1.09
3	50R	1.70	1.53	1.32	0.89	0.96	3.32	0.56	1.00	1.21	0.69	1.95	0.92
4	5XG	2.46	1.71	1.97	1.08	0.96	2.50	0.65	1.06	3.15	1.07	3.79	0.90
5	5Y6	8.59	1.70	1.99	1.41	1.59	2.90	1.06	1.32	1.61	0.86	2.24	1.29
6	5Y8	3.19	1.58	1.71	0.97	1.08	3.46	0.69	1.00	1.31	0.87	1.40	1.07
7	DMG-Out Apo	2.36	1.57	1.83	1.44	1.16	3.03	0.69	1.13	2.23	1.39	1.66	1.04
8	SKR1	2.37	1.93	1.23	0.80	0.82	2.49	0.55	0.86	2.15	1.25	2.26	0.90
9	SKR2	3.23	1.77	1.23	0.92	0.95	3.49	0.72	1.05	2.99	1.30	3.19	1.12
10	SKR5	2.39	1.71	1.50	0.95	0.94	2.23	0.62	1.11	2.91	1.30	1.93	1.40
11	SKR10	2.02	1.71	1.49	0.72	0.84	1.98	0.56	1.02	2.44	1.04	1.63	0.98
12	SKR11	2.07	1.63	1.30	1.01	0.94	3.01	0.61	1.02	2.49	1.46	1.92	1.00

Figure 5.4. RMSF for all the studied CDK8/CycC systems. The color bar on the x-axis marks the regions of CDK8. The table beneath the plot lists the 12 regions with large motions and highlighted with colors that correspond to the color bar.

Ligand binding has minor effects on the dynamics of the CDK8/CycC complex and influences DMG-in and DMG-out conformations differently. In the apo-form of CDK8, DMG-out CDK8 shows significantly more flexibility in the αC helix and activation loop areas (Figure 5.4). Upon ligand binding, the αC helix of the DMG-out conformation is largely stabilized by the formation of two highly stable H-bonds via the urea linker of type-II ligands with Glu66 on the αC helix and Asp173 on β8, which leads to the activation loop. This stabilizes the binding pose of type-II ligands, in turn stabilizing the αC helix. Type-I ligands form a less stable H-bond with Lys52 (40-50% duration), which forms an H-bond with Asp173, but this interaction is also present in the apo-form CDK8, so ligand binding seems to confer no further stability. The C-terminus (region 12 in Figure 5.4) is also stabilized by ligands. For the DMG-in apo-form of CDK8, the RMSF for this region can be very large depending on initial conformations and sampling, but is greatly reduced upon ligand binding by the π-stacking interaction between the ligand and Arg356. Type-II ligand SKR1 also forms this interaction and stabilizes this region. The recently discovered 4,5-dihydrothieno[3',4':3,4] benzo[1,2-d] isothiazole derivative achieved sub-nanomolar potency despite the fact that the corresponding docking study suggested this ligand had no interaction with Arg356 [29], suggesting this interaction may not be essential. Because our study showed significantly reduced CDK8 flexibility in the C-terminus due to ligand binding, investigation of ligands that avoid this interaction while maintaining the other key interactions is worthwhile and may result in ligands that produce lower entropic penalties. Except for the αC helix and C-terminus, we observed no other important stabilized regions. Ligand binding affects the dynamics of

CDK8 via local, direct interactions and is unable to induce long-range or allosteric effects.

5.4.2 Importance of CycC to CDK8 Stability and Ligand Binding

We performed MD simulations for the systems without CycC, and the results clearly show that CycC stabilizes CDK8 by reducing the fluctuation in the N-terminus, αC helix, and activation loop. The activation of CDKs requires the binding of cyclins and phosphorylation of Thr, Ser, and Tyr on their activation loop [67, 68]. This binding changes the conformation of CDK8 markedly [32,68] and enables ligand binding in the allosteric site [25], which is supported by our observation that in the absence of CycC, the αC helix of CDK8 adopts an αC-out conformation, whereby Glu66 moves away from the DMG motif. By losing the H-bond from Glu66 and interactions from the entire αC helix, the allosteric binding site collapses, thereby disabling the binding of type II ligands.

Figure S3 compares the RMSF of CDK8-50R complexes with and without CycC. Without CycC, the N-terminus of CDK8, αC helix, and activation loop have much larger RMSF values (Figure 5.4). Crystal structures show that the N-terminus of CDK8 rests stably on CycC, however, in our simulations omitting CycC, the N-terminus exhibits extremely large motions, leading to a totally different conformation of this region. Crystal structures show that the αC helix is part of the binding interface of CDK8 and CycC. If CycC is absent in simulation, the αC helix has a wider range of motion and moves toward the space that CycC would normally occupy, becoming too distant form the binding site to form the characteristic H-bond via Glu66 with type-II ligands (Figure 5.5). Although the activation loop is not in direct contact with CycC, the reduced motion of the N- and C-lobes and αC helix by CycC provides stability to this region as well.

Among the three regions stabilized by CycC, the αC helix has the largest impact on ligand binding. The conformation of this helix is characterized as αC-in or αC-out according to the distance between Cα carbon atoms of Glu66 and Asp173 [66]. Structures with a short DMG-αC-helix distance (4−7.2 Å) are classified as αC-in, whereas structures with long distances (9.3−14 Å) are classified as αC-out. Structures with distances in between are classified as αC-out–like structures. Figure S4 shows this distance in MD2 and MD3 and suggests that CDK8 is usually in the αC-out conformation when CycC is absent and αC-in when CycC is present. All type-II ligands included in this study form a very strong H-bond with Glu66 on the αC-helix. This is only possible in the presence of CycC which causes CDK8 to adopt the αC-in conformation. Moreover, although the αC helix is not in direct contact with type-I ligands, it helps stabilize the binding pocket via a key salt bridge between Glu66 and Lys52. Lys52 is one of the most important residues for type-I ligand binding, providing a stable H-bond for in all cases studied here, and in this sense CycC also affects type-I ligands. In the trajectory of CDK8-5Y6 complex without CycC, 5Y6 leaves the native bound state characterized by the crystal structure and found an incorrect binding state conformation due to the absence of CycC and the unstable binding cavity (Figure 5.6). In this conformation, the αC helix moves away from the ligand and the salt bridge between Lys52 and Glu66 is broken. This situation causes the beta sheets β1-2 above the binding site to move upward, providing the ligand with more room to explore the binding site. 5Y6 still keeps a V-shape but rotates by 90 degrees to pick up contacts with Tyr32 and Phe97. The MMPB/SA interaction energy (ΔEMM/PBSA) of this trajectory is -25.5±3.8 kcal/mol and is significantly weaker than its counterpart including CycC (-29.4±4.4 kcal/mol), further indicating the importance of CycC in maintaining proper binding conformations.

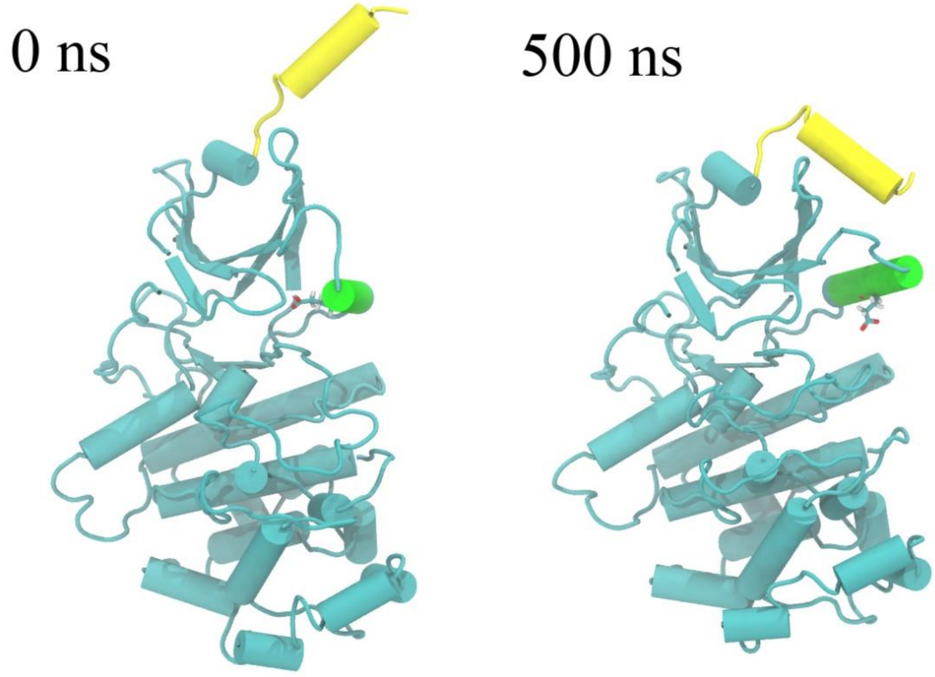


Figure 5.5. Conformation change of the αB and αC helices in MD8 (*apo* CDK8 in DMG-out conformation) in the absence of CycC. The αB helix is in yellow, and the αC helix is in green. GLU 66 is shown in licorice.

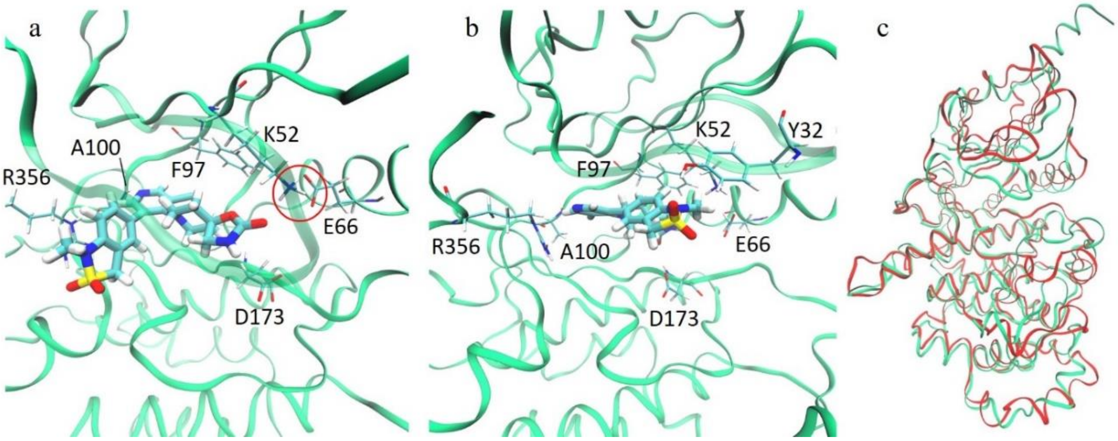


Figure 5.6. Comparison of CDK8-5Y6 complex without CycC at a) 0 ns and b) 200 ns. Two binding modes are observed when CycC is absent, one at 0 ns which is a typical type-I ligand binding mode, and the other at 200 ns which is partially due to loss of the important salt bridge formed between K52 and E66 (red circle in a). (c) is a superposition of the two aligned trajectory frames with 0 ns in green and 200 ns in red showing the large conformational changes that occur in CycC's absence.

5.4.3 Ligand Binding Modes and Binding Pocket Volume Analysis

Our MD simulations revealed binding modes for both type-I and type-II ligands that matched crystal structures and provide a level of detail previously unavailable (Figure 5.7). By calculating the binding energies of the ligands with the MM/PBSA method, we identified the driving forces behind ligand binding to CDK8. The energy breakdowns are shown in Table 2. Type-I ligands formed H-bonds with Lys52, Ala100, and Asp173 in the ATP binding site. Additionally, we found significant vdW interactions with Val27, Val35, Ile79, Tyr99, Leu158, and Arg356. Type-II ligands formed H-bonds with Asp173 and Glu66, which are stabilized by a salt bridge between Glu66 and Lys52 and experience large vdW forces with Leu69, Leu70, Ile79, Phe97, Leu142 and Ala172. Tables 2 and 3 list the strength and durations of these interactions and Figure S7 shows the patterns of H-bond formation and loss for a type-I and type-II ligand. We also computed the total solvent accessible volumes for the ATP and allosteric binding sites for all twelve systems. For type-I ligands, more contacts between ligand and protein lead to better binding affinity through vdW attractions, whereas for type-II ligands, the structural locker formed by Glu66 and Asp173 contributes more significantly. The computed volumes and a few representatives of the volume change over time are in Figure 5.8. The specifics of each ligand binding mode vary and are presented in the following sections.

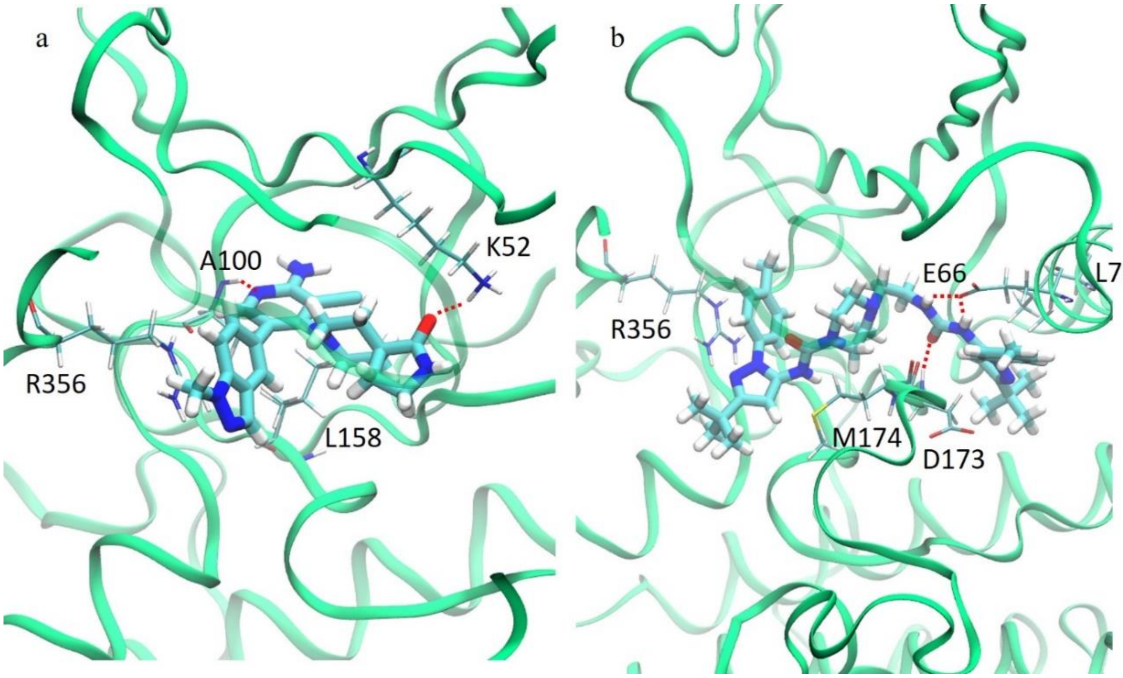


Figure 5.7. Typical binding modes for a) type-I ligands and b) type-II ligands with H-bonds shown by red dotted lines and the strongest-interacting residues labeled and shown in licorice.

5.4.4 Type-I Ligands

H-Bonds. Type-I ligands all share the same direct H-bonds to CDK8 at Lys52 and Ala100, but the H-bond with Ala100 has higher occurrence (56-76% vs. 38-48%) in all cases because of its position in the relatively stationary hinge region and the relative instability of the β-sheet containing Lys52 (Figure 5.2). 50R has a nitrogen on the benzothiophene ring forms a H-bond with Ala100 with an occurrence of 64%, and its amide moiety forms another H-bond with Lys52. The other type-I ligands show a similar binding pattern with the ketone oxygen forming an H-bond with Lys52, and the nitrogen on the pyridine forming an H-bond with Ala100. The 3-aminoindazole moiety of 5XG forms a very highly stable H-bond with Val27, with occurrence 78%, which is a unique feature among the type-I ligands studied here.

Electrostatic, vdW, and other interactions. For all type-I ligands, the overall interaction energy is stronger with Lys52 than Ala100 due to both better vdW and electrostatic interactions despite the greater desolvation penalty. These ligands form vdW interactions (-2.2 to -3.2 kcal/mol) with Leu158, Arg356, Val35, and a few other residues (Table 3). Other important interactions include formation of a cation-π interaction with Arg356 by the aromatic rings of type-I ligands (Figure 5.7). The benzene ring of 50R forms a cation-π interaction with Arg356 and the other three type-I ligands have the same scaffold by which the indazole or its analogue part forms the same cation-π interaction with Arg356. Type-I ligands also form some bridge water interactions with Glu66 (3 to 14%), Asp173 (25 to 38%), and some other residues, but these bridge water molecules are not very stable and are rapidly displaced by bulk water molecules. Bridge waters function as mediating water molecules that hold the interaction between the protein and ligand and may stabilize the binding pose of the ligand, but in this case are unlikely to cause any appreciable decrease in desolvation penalty since they are easily displaced [59, 60]. These interactions are not as strong as direct H-bonds but could still increase ligand binding affinity.

MM/PBSA Binding Energy. All type-I ligands possess a similar scaffold that occupies a nearly identical space in the ATP binding site (Figure S5) and have MM/PBSA interaction energies (ΔEMMPBSA) of about -25 to -32 kcal/mol. This is the net effect of a negative vdW interaction energy term (ΔEvdW) ranging from -40 to -49 kcal/mol and a positive electrostatic plus PB term (ΔEelec+PB) from 15 to 23 kcal/mol. The vdW interaction is the major driving force for binding; the ATP binding site has a small volume in the free state and opens to accommodate the ligands which experience tight contacts once they are established. With this scaffold, these ligands have better binding affinity when the ligand is bulkier, so there may be room to further exploit this property using slightly larger type-I ligands. Figure S6 shows the relationship between binding pocket volume and experimental binding affinity.

5.4.5 Type-II Ligands

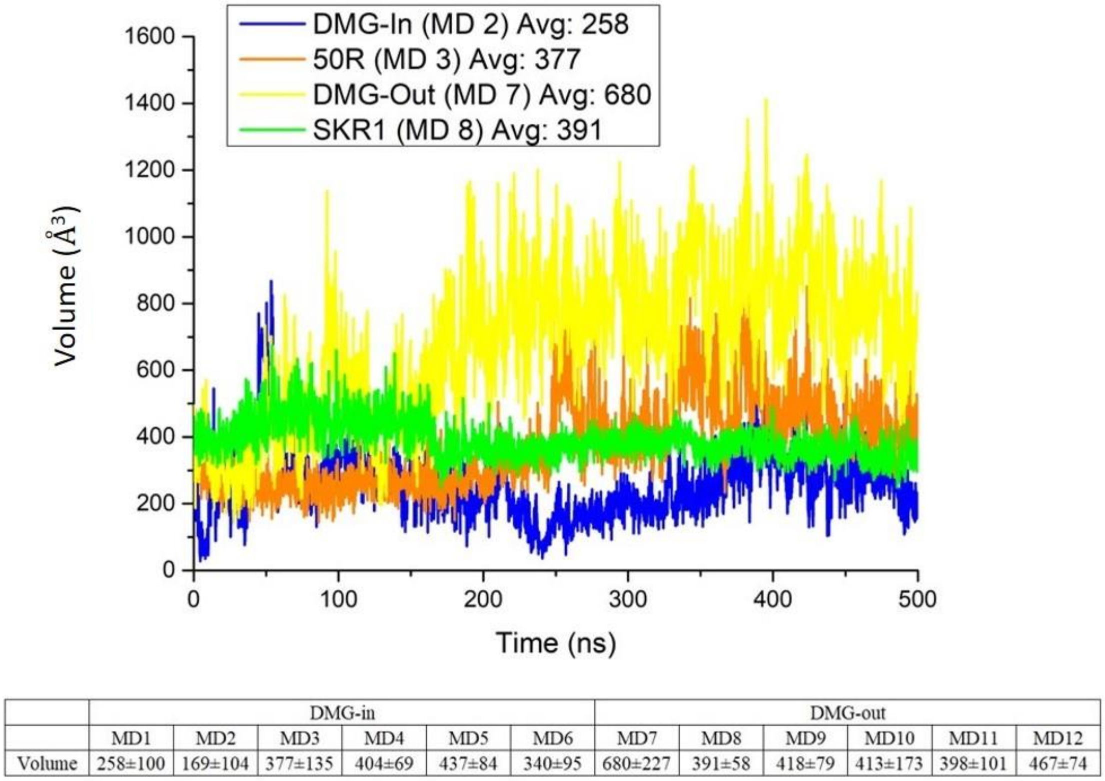
H-Bonds. All Type-II ligands form two strong H-bonds with Glu66 and Asp173 via the urea linker, with occurrences of roughly 90-96% and 76-93%, respectively (Figure 5.2). These two H-bonds function as anchors that stabilize the type-II ligands in the allosteric binding site. The moiety that extends into the ATP binding site for SKR5 and SKR11 has a size and shape which allows these ligands to form an H-bond with Asp98 with occurrences of 8.77 and 10.34%, respectively, that is not seen with the other three type-II ligands. SKR5 has a terminal [3-(morpholine-4-yl)propyl] group that forms another H-bond with Ala100 in the hinge region. Although the occurrence of this H-bond is as low as 17%, it provides SKR5 with detectable residence time [25]

Electrostatic, vdW, and other interactions. The scaffold of all type-II ligands in the allosteric binding site is nearly identical and forms vdW interactions with Leu69 of about -1.0 kcal/mol and with Leu70 of about -2.8 kcal/mol. These ligands have very different structures that extend into the ATP binding site, however, and the vdW interaction in this site and structural flexibility largely account for the variability in binding affinities. SKR1 is the largest ligand and extends into and occupies the entire ATP binding site as well (Figure S5), which results in a very strong vdW interaction (ΔEvdW) with CDK8 at -89 kcal/mol. Because ligand size roughly decreases from SKR2 to SKR11, the vdW interaction (ΔEvdW) decreases from -62 to -50 kcal/mol. Other differences in the ATP site moiety of type-II ligands result in variations of other interactions and may change flexibility. For example, SKR10 cannot form the H-bond with Asp98 that SKR11 can, and its absence causes the analogous part of SKR10 to fold onto the protein surface resulting in a stronger vdW interaction than SKR11.

The benzene ring of SKR1 in the ATP binding site interacts with Arg356 via cation-π stacking in the same way as type-I ligands, but it does not have contacts with the hinge region. SKR1 also forms a few stable bridge water interactions with Asn156 and Asp173. SKR2 binds similarly to SKR1 but occupies less of the ATP binding site and has less vdW interaction in that region, and rather than cation-π stacking, has vdW interaction with Arg356. In addition, the smaller structure extending into the ATP binding site is very flexible during the MD simulations, so SKR2 should have a smaller entropic penalty than SKR1.

The electrostatic plus PB term (ΔEelec+PB) opposes the binding of type-II ligands and decreases with ligand size; however, the electrostatic term (ΔEelec) alone is very similar among the five type-II ligands, except for SKR5. This finding indicates that the greater presence of polar functional groups in the larger type-II ligands doesn't necessarily form favorable interactions with the binding pocket of CDK8 and that the binding of these ligands is driven by non-polar interactions.

MM/PBSA Binding Energies. Type-II ligands share the same scaffold, the minimal compound 7 in [25] that binds to the allosteric binding site of CDK8, but have different structures that extend into the ATP binding site, and MM/PBSA interaction energies of these ligands vary widely due to this difference. MM/PBSA calculations showed that SKR1 has stronger binding energy than SKR2, but experimental data favor SKR2 by 0.5 kcal/mol. Because entropy contributions are not considered in our MM/PBSA calculations, neglected entropic effects may account for this discrepancy. Compared with SKR1, the smaller SKR2 is less confined and retains more freedom, therefore paying less entropic penalty. In addition, RMSF measurements showed that CDK8 bound to SKR2 is more flexible than when bound with SKR1 (Figure 5.4), which suggests that the protein also pays less entropy penalty bound to SKR2. SKR10 and SKR11 have less bulky structures than SKR1, SKR2, and SKR5, and thus less favorable vdW interactions with the ATP binding site, which is largely the reason for their less favorable MM/PBSA energies (Table 2).



	DMG-in						DMG-out					
	MD1	MD2	MD3	MD4	MD5	MD6	MD7	MD8	MD9	MD10	MD11	MD12
Volume	258±100	169±104	377±135	404±69	437±84	340±95	680±227	391±58	418±79	413±173	398±101	467±74

Figure 5.8. Volumes of the ATP binding site in Å^3. Plot: Volume change along the MD time for four selected systems; Table: average volumes with standard deviations of all 12 MD systems.

Type I				
MD Index	3	4	5	6
Ligand	50R	5XG	5Y6	5Y8
PDB ID	5CEI	5FGK	5HBE	5HBJ
$\Delta \mathbf{E}_{\mathbf{vdW}}$	-39.8±2.8	-45.9±2.7	-48.5±2.9	-43.9±3.0
$\Delta \mathbf{E}_{\mathbf{elec}}$	-13.0±4.6	-28.7±6.4	-24.6±7.0	-26.0±5.5
$\Delta \mathbf{G}_{\mathbf{PB}}$	31.2±3.6	47.2±5.7	48.0±6.6	41.4±4.9
$\Delta \mathbf{E}_{\mathbf{elec+PB}}$	18.2±3.5	18.5±4.3	23.4±4.1	15.4±3.3
$\Delta \mathbf{G}_{\mathbf{np}}$	-3.6±0.1	-4.2±0.1	-4.3±0.1	-4.2±0.1
$\Delta \mathbf{E}_{\mathbf{gas}}$	-52.8±5.7	-74.6±6.8	-73.1±8.1	-69.9±6.1
$\Delta \mathbf{G}_{\mathbf{solv}}$	27.6±3.5	43.1±5.6	43.7±6.6	37.3±4.9
$\Delta \mathbf{E}_{\mathbf{MM/PBSA}}$	-25.2±4.1	-31.5±4.1	-29.4±4.4	-32.6±3.9
$\Delta \mathbf{G}_{\mathbf{Expt}}$	-11.4	-12.0	-11.9	-11.3

Type II					
MD Index	8	9	10	11	12
Ligand	SKR1	SKR2	SKR5	SKR10	SKR11
PDB ID	4F6W	4F7L	4F6U	4F7N	4F7N
$\Delta \mathbf{E}_{\mathbf{vdW}}$	-88.5±3.4	-62.4±3.3	-59.9±4.9	-50.6±2.8	-49.5±3.0
$\Delta \mathbf{E}_{\mathbf{elec}}$	-20.4±4.1	-20.0±4.8	-28.3±7.6	-20.9±4.6	-19.0±6.0
$\Delta \mathbf{G}_{\mathbf{PB}}$	66.4±4.7	49.9±5.0	52.6±9.0	40.4±5.0	38.3±5.4
$\Delta \mathbf{E}_{\mathbf{elec+PB}}$	46.0±4.6	22.9±4.3	24.3±5.2	19.6±4.1	19.3±4.2

$\Delta \mathbf{G}_{\mathbf{np}}$	-7.9±0.1	-5.8±0.1	-5.5±0.1	-4.8±0.1	-5.0±0.1
$\Delta \mathbf{E}_{\mathbf{gas}}$	-108.8±4.8	-82.4±5.9	-88.2±10.4	-71.5±4.6	-68.5±6.8
$\Delta \mathbf{G}_{\mathbf{solv}}$	58.5±4.7	37.0±5.0	47.1±9.0	35.6±5.0	33.3±5.4
$\Delta \mathbf{E}_{\mathbf{MM/PBSA}}$	-50.3±4.7	-45.3±4.9	-41.1±5.0	-35.9±3.9	-35.2±4.5
$\Delta \mathbf{G}_{\mathbf{Expt}}$	-10.2	-10.7	-8.4	-8.0	-9.7

Table 5.2. MM/PBSA energy breakdowns for the binding energy of 9 ligands with CDK8/CycC in kcal/mol. ΔE_{vdW} and ΔE_{elec} are the van der Waals and electrostatic energy contributions, respectively, and ΔE_{gas} is the sum of those two terms. ΔG_{PB} and ΔG_{np} are the polar and non-polar solvation energies, respectively, and ΔG_{solv} is the sum of those two terms. $\Delta E_{MM/PBSA}$ is the binding energy predicted by MM/PBSA.

Type I				
MD Index	3	4	5	6
Ligand	50R	5XG	5Y6	5Y8
PDB ID	5CEI	5FGK	5HBE	5HBJ
LEU158	-2.7±0.6	-2.8±0.6	-2.7±0.6	-2.7±0.6
ARG356	-2.5±0.6	-2.7±0.7	-2.7±0.7	-3.2±0.6
VAL35	-2.4±0.5	-2.4±0.6	-2.5±0.7	-2.2±0.6
LYS52	-2.0±1.2	-3.0±1.1	-2.7±0.9	-2.6±1.2
TYR99	-1.7±0.4	-1.8±0.4	-1.6±0.4	-2.0±0.5
ALA100	-1.2±0.7	-1.2±0.7	-1.4±0.6	-1.5±0.5
PHE97	-1.0±0.4	-1.0±0.4	-1.1±0.3	-0.7±0.3
TYR32	-0.6±0.6	-2.3±1.0	-2.8±1.0	-2.0±1.0
VAL27	-1.0±0.5	-1.7±0.7	-0.9±0.5	-0.9±0.5
ILE79	-0.8±0.4	-1.1±0.4	-1.1±0.4	-0.8±0.5
ASP103	+0.1±0.6	+1.2±0.7	+1.9±0.8	+1.4±0.6

Type II					
MD Index	8	9	10	11	12
Ligand	SKR1	SKR2	SKR5	SKR10	SKR11
PDB ID	4F6W	4F7L	4F6U	4F7N	4F7N
GLU66	-5.2±1.1	-5.4±1.2	-4.6±1.4	-5.2±1.1	-4.4±1.4
MET174	-3.3±0.8	-2.8±1.0	-1.1±0.8	-1.4±0.8	-0.7±0.8

ARG356	-2.8±0.9	-0.1±0.5	0.0±0.4	0.0±0.3	0.0±0.3
LEU70	-2.7±0.8	-2.7±0.8	-2.8±0.8	-2.7±0.7	-2.9±0.8
LEU158	-2.4±0.8	+0.9±0.6	-0.6±0.6	0.0±0.4	0.0±0.6
ASP173	-2.0±1.0	-2.3±1.0	-1.8±1.0	-2.2±1.0	-1.7±1.0
PHE97	-1.9±0.7	-2.2±0.8	-2.3±0.7	-1.4±0.8	-1.7±0.8
VAL35	-1.9±0.7	-1.7±0.8	-1.0±0.6	-0.1±0.6	-0.3±0.7
ALA172	-1.7±0.8	-1.7±0.9	-2.0±0.7	-1.7±0.7	-1.6±0.7
ILE79	-1.7±0.8	-2.1±0.9	-2.2±0.8	-1.5±0.8	-1.8±0.9
LEU69	-1.1±0.7	-1.0±0.7	-1.0±0.6	-1.0±0.7	-1.0±0.6
ARG178	0.0±0.1	0.0±0.3	-1.2±1.0	0.0±0.2	0.0±0.4
LEU142	-0.9±0.5	-1.0±0.6	-0.9±0.6	-1.0±-0.6	-0.9±0.7

Table 5.3. The major residues that have interaction energies stronger than -1.0 kcal/mol with type I and type II ligands. All values are in kcal/mol. The standard deviation is marked by ±.

5.5 Conclusion

In this work, we performed MD simulations for nine CDK8/CycC-ligand complexes and three CDK8/CycC apoproteins which included both DMG-in and -out conformations. Our analysis of system dynamics and flexibility shows that the highly flexible activation loop has little effect on ligand binding. Further, ligand binding stabilizes the α-C helix and C-terminus of CDK8 through direct interactions with residues in these regions but does not affect the large-scale dynamics. PCA analysis on sequential 100-ns portions of the MD trajectories revealed the range of protein global motions which are relevant to binding, such as a bending motion about the hinge region, and our simulations provide well-sampled conformations for use in future docking or MD studies.

By repeating simulations with CycC excluded, we were able to discern its stabilizing effect on the system. We found that CycC is critical to maintain the structure of CDK8 and provide proper interactions for ligand binding, namely the stabilization of Glu66 on the αC helix, which forms a critical H-bond with type-II ligands and makes an important salt bridge with Lys52, which H-bonds with type-I ligands.

Analysis of four type-I and five type-II ligand binding modes along with volume measurements of the binding pocket elucidated the protein-ligand interactions. Residues Lys52 and Ala100 form very strong H-bonds with all type-I ligands, and Asp173 and Arg356 provide highly favorable vdW interactions. Additionally, the binding pocket has a smaller volume with type-I ligands, and vdW interactions with the surrounding resides are a major driving force of binding. These ligands can reduce protein flexibility, so entropic penalties need to be taken into consideration. H-bonds may be used to optimize the enthalpic attractions, and, assuming the rigidity of the scaffold can be retained, slightly larger compounds can increase the vdW interaction to optimize the binding affinity.

Type-II ligands bind in both the allosteric and ATP binding site. They all form H-bonds with Glu66 and Asp173; the main variability in type-II binding affinities is due to the varying structures that extend into the ATP binding site. We found that larger structures extending into the ATP site result in favorable vdW interactions and H-bonds. Optimization of type-II binding affinities depends on proper design of the group extending into ATP binding site that achieves a balance between rigidity and size to keep the entropic penalty upon binding minimal while providing enough bulk to stay firmly in the binding pocket and achieve favorable vdW interactions and H-bonds.

Acknowledgements

This study was supported by the US National Institutes of Health (GM-109045) and National Science Foundation national supercomputer centers (TG-CHE130009).

References:

1. Malumbres M (2014) Cyclin-dependent kinases. Genome Biol 15(6):122.

2. Galbraith MD, Donner AJ, Espinosa JM (2010) CDK8: a positive regulator of transcription. Transcription 1: 4−12.

3. Tsutsui T, Fukasawa R, Tanaka A, Hirose Y, Ohkuma Y (2011) Identification of target genes for the CDK subunits of the Mediator complex. Genes Cells 16:1208−1218.

4. Allen BL, Taatjes DJ (2015) The Mediator complex: a central integrator of transcription. Nat Rev Mol Cell Biol 16:155−166.
5. Rickert P, Seghezzi W, Shanahan F, Cho H, Lees E (1996) Cyclin C/CDK8 is a novel CTD kinase associated with RNA polymerase II. Oncogene 12:2631−2640.

6. Xu W, Ji JY (2011) Dysregulation of CDK8 and Cyclin C in tumorigenesis. J Genet Genomics 38(10):439–452.

7. Conaway RC, Sato S, Tomomori-Sato C, Yao T, Conaway, JW (2005) The Mammalian Mediator Complex and Its Role in Transcriptional Regulation. Trends Biochem Sci 30: 250−255.

8. Nemet J, Jelicic B, Rubelj I, Sopta M (2014) The Two Faces of Cdk8, a Positive/Negative Regulator of Transcription. Biochimie 97:22−27.

9. Li N, Fassl A, Chick J, Inuzuka, H, Li X, Mansour MR, Liu L, Wang H, King B, Shaik S, et al (2014) Cyclin C Is a Haploinsufficient Tumour Suppressor. Nat Cell Biol 16: 1080−1091.

10. Morris EJ, Ji JY, Yang F, Di Stefano L, Herr A, Moon NS, Kwon EJ, Haigis KM, Naar AM, Dyson NJ (2008) E2F1 represses beta-catenin transcription and is antagonized by both pRB and CDK8. Nature 455:552−556.

11. Firestein, R, Shima K, Nosho K, Irahara N, Baba Y, Bojarski E, Giovannucci EL, Hahn WC, Fuchs CS, Ogino S (2010) CDK8 expression in 470 colorectal cancers in relation to beta-catenin activation, other molecular alterations and patient survival. Int J Cancer 126:2863−2873.

12. Kim MY, Han SI, Lim SC (2011) Roles of cyclin-dependent kinase 8 and beta-catenin in the oncogenesis and progression of gastric adenocarcinoma. Int J Oncol 38:1375−1383.

13. Adler AS, McCleland ML, Truong T, Lau S, Modrusan Z, Soukup TM, Roose-Girma M, Blackwood EM, Firestein R (2012) CDK8 maintains tumor dedifferentiation and embryonic stem cell pluripotency. Cancer Res 72:2129−2139.

14. Rosenbluh J, Wang X, Hahn WC (2014) Genomic insights into WNT/β-catenin signaling. Trends Pharmacol Sci 35(2):103-109.

15. Broude EV, Győrffy B, Chumanevich AA, et al (2015) Expression of CDK8 and CDK8-interacting Genes as Potential Biomarkers in Breast Cancer Curr Cancer Drug Targets 15(8):739-749.

16. Raithatha S, Su T-C, Lourenco P, Goto S, Sadowski I (2012) Cdk8 Regulates Stability of the Transcription Factor Phd1 To Control Pseudohyphal Differentiation of Saccharomyces cerevisiae. Mol Cell Biol 32(3):664-674.

17. Alarcón C, Zaromytidou A-I, Xi Q, Gao S, Yu J, Fujisawa S, Barlas A, Miller AN, Manova-Todorova K, Macias MJ, Sapkota G, Pan D, Massagué J (2009) Nuclear CDKs Drive Smad Transcriptional Activation and Turnover in BMP and TGF-β Pathways. Cell 139(4):757−769.

18. Rzymski T, Mikula M, Wiklik K, et al (2015) CDK8 kinase–An emerging target in targeted cancer therapy. Biochim Biophys Acta 1854(10 Pt B):1617–1629.

19. Fryer CJ, White JB, Jones KA (2004) Mastermind Recruits CycC:CDK8 to Phosphorylate the Notch ICD and Coordinate Activation with Turnover. Mol Cell 16(4):509−520.

20. Cee VJ, Chen DY, Lee MR, Nicolaou KC (2009) Cortistatin A is a high-affinity ligand of protein kinases ROCK, CDK8, and CDK11. Angew Chem, Int Ed 48:8952−8957.

21. Porter DC, Farmaki E, Altilia S, Schools GP, West DK, Chen M, et al (2012) Cyclin-dependent kinase 8 mediates chemotherapy-induced tumor-promoting paracrine activities. Proc Natl Acad Sci USA 109:13799–804.

22. Mallinger A, Schiemann K, Rink C, et al (2016) Discovery of Potent, Selective, and Orally Bioavailable Small-Molecule Modulators of the Mediator Complex-Associated Kinases CDK8 and CDK19. J Med Chem 59(3):1078-1101.

23. Koehler, MF, Bergeron P, Blackwood EM, Bowman K, Clark KR, Firestein R, Kiefer JR, Maskos K, McCleland ML, Orren L, Salphati L, Schmidt S, Schneider EV, Wu J, Beresini MH (2016) Development of a Potent, Specific CDK8 Kinase Inhibitor Which Phenocopies CDK8/19 Knockout Cells. ACS Med Chem Lett 7(3):223-8.

24. Kumarasiri M, Teo T, Yu M, Philip S, Basnet S.K, Albrecht H, Sykes MJ, Wang P, Wang S (2017) In Search of Novel CDK8 Inhibitors by Virtual Screening. J Chem Inf Model 57(3):413-416.

25. Schneider EV, Bottcher J, Huber R, Maskos K, Neumann L (2013) Structure–kinetic relationship study of CDK8/CycC specific compounds. Proc Natl Acad Sci USA 110: 8081–8086.

26. Czodrowski P, Mallinger A, Wienke D, Esdar C, Pöschke O, Busch M, Rohdich F, Eccles SA, Ortiz-Ruiz MJ, Schneider R, Raynaud FI (2016) Structure-based optimization of potent, selective, and orally bioavailable CDK8 inhibitors discovered by high-throughput screening. J Med Chem 59(20):9337-9349.

27. Wang T, Yang Z, Zhang Y, Yan W, Wang F, He L, Zhou Y, Chen L (2017) Discovery of novel CDK8 inhibitors using multiple crystal structures in docking-based virtual screening. Eur J Med Chem 129:275-286.

28. Schiemann K, Mallinger A, Wienke D, Esdar C, Poeschke O, Busch M, Rohdich F, Eccles SA, Schneider R, Raynaud FI, Czodrowski P (2016) Discovery of potent and selective CDK8 inhibitors from an HSP90 pharmacophore. Bioorganic Med Chem Lett 26(5):1443-1451.

29. Ono K, Banno H, Okaniwa M, Hirayama T, Iwamura N, Hikichi Y, Murai S, Hasegawa M, Hasegawa Y, Yonemori K, Hata A, Aoyama K, Cary DR (2017) Design and synthesis of selective CDK8/19 dual inhibitors: Discovery of 4,5-dihydrothieno[3',4':3,4] benzo[1,2-d] isothiazole derivatives. Bioorg Med Chem 25(8):2336-2350.

30. Schneider EV, Böttcher J, Blaesse M, Neumann L, Huber R, Maskos K (2011) The structure of CDK8/CycC implicates specificity in the CDK/cyclin family and reveals interaction with a deep pocket binder. J Mol Biol 412(2):251-66.

31. Callegari D, Lodola A, Pala D, Rivara S, Mor M, Rizzi A, Capelli AM (2017) Metadynamics Simulations Distinguish Short-and Long-Residence-Time Inhibitors of Cyclin-Dependent Kinase 8. J Chem Inf Model 57(2):159-169.

32. Xu W, Amire-Brahimi B, Xie X-J, Huang L, Ji J-Y (2014) All-atomic Molecular Dynamic Studies of Human CDK8: Insight into the A-loop, Point Mutations and Binding with Its Partner CycC. Comput Biol Chem 51:1-11.

33. Callegari D, Lodola A, Pala D, Rivara S, Mor M, Rizzi A, Capelli AM (2017) Metadynamics Simulations Distinguish Short-and Long-Residence-Time Inhibitors of Cyclin-Dependent Kinase 8. J Chem Inf Model 57(2):159-169.

34. Biasini, M, Bienert S, Waterhouse A, Arnold K, Studer G, Schmidt T, Kiefer F, Cassarino TG, Bertoni M, Bordoli L, Schwede T (2014) SWISS-MODEL: modelling protein tertiary and quaternary structure using evolutionary information. Nucleic Acids Res 42 (W1):W252-W258.

35. Bordoli L, Kiefer F, Arnold K, Benkert P, Battey J, Schwede T (2009) Protein structure homology modelling using SWISS-MODEL Workspace. Nat Protoc 4:1.

36. Arnold K, Bordoli L, Kopp J, Schwede T (2006) The SWISS-MODEL Workspace: A web-based environment for protein structure homology modelling. Bioinformatics 22: 195-201.

37. Case DA, Babin V, Berryman JT, et al (2014) Amber 14. University of California, San Francisco.

38. Case DA, Cheatham TE, Darden T, Gohlke H, Luo R, et al (2005) The Amber biomolecular simulation programs. J Comput Chem 26:1668-1688.

39. Goetz AW, Williamson MJ, Xu D, Poole D, Le Grand S, et al (2012) Routine microsecond molecular dynamics simulations with AMBER on GPUs. 1. Generalized Born. J Chem Theory Comput 8:1542-1555.

40. Hornak V, Abel R, Okur A, Strockbine B, Roitberg A, et al (2006) Comparison of multiple amber force fields and development of improved protein backbone parameters. Proteins: Struct, Funct, Bioinf 65:712-725.

41. Wang JM, Wolf RM, Caldwell JW, Kollman PA, Case DA Development and testing of a general amber force field. J Comput Chem 25:1157-1174.

42. Ozpinar GA, Peukert W, Clark T, An improved generalized AMBER force field (GAFF) for urea. J Mol Model 16:1427-1440.

43. Georgescu RE, Alexov EG, Gunner MR (2002) Combining conformational flexibility and continuum electrostatics for calculating pKa's in proteins. Biophys J 83:1731-1748.

44. Alexov E, Gunner MR (1997) Incorporating protein conformational flexibility into pH- titration calculations: Results on T4 Lysozyme. Biophys J 74:2075-2093.

45. Sondergaard CR, Olsson MHM, Rostkowski M, Jensen JH (2011) Improved Treatment of Ligands and Coupling Effects in Empirical Calculation and Rationalization of pKa Values. J Chem Theory Comput 7(7):2284-2295.

46. Olsson MHM, Sondergaard CR, Rostkowski M, Jensen JH (2011) PROPKA3: consistent treatment of internal and surface residues in empirical pKa predictions. J Chem Theory Comput 7(2):525-537.

47. Wang L, Li L, Alexov E (2015) pKa predictions for proteins, RNAs and DNAs with the Gaussian dielectric function using DelPhiPKa. Proteins 83(12):2117-2125.
48. Wang L, Zhang M, Alexov E (2015) DelPhiPKa Web Server: Predicting pKa of proteins, RNAs and DNAs. Bioinformatics 32(4):614-615

49. Jorgensen WL, Chandrasekhar J, Madura JD, Impey RW, Klein ML Comparison of simple potential functions for simulating liquid water. J Chem Phys 79:926-935.

50. Doll JD, Dion DR (1976) Generalized Langevin Equation Approach For Atom-Solid-Surface Scattering - Numerical Techniques For Gaussian Generalized Langevin Dynamics. J Chem Phys 65:3762-3766.
51. Adelman SA Generalized Langevin Theory For Many-Body Problems In Chemical-Dynamics - General Formulation And The Equivalent Harmonic Chain Representation. J Chem Phys 71:4471-4486.

52. Essmann U, Perera L, Berkowitz ML, Darden T, Lee H, et al (1995) A Smooth particle mesh Ewald method. J Chem Phys 103:8577-8593.

53. Ryckaert JP, Ciccotti G, Berendsen HJC (1977) Numerical-interaction of cartesian equations of motion of a system with constraints - molecular-dynamics of N-alkanes. J Comput Phys 23:327-341.

54. Pearson K (1901) On Lines and Planes of Closest Fit to Systems of Points in Space. Philos Mag 2(11):559–572.

55. Hotelling H (1933) Analysis of a complex of statistical variables into principal components. J Educ Psychol 24:417–441 and 498–520.

56. Hotelling H (1936) Relations between two sets of variates. Biometrika 28:321–377

57. Still WC, Tempczyk A, Hawley RC, Hendrickson T (1990) Semianalytical treatment of solvation for molecular mechanics and dynamics. J Am Chem Soc 112:6127–6129.

58. Miller III B R, McGee Jr TD, Swails JM, Homeyer N, Gohlke H, Roitberg AE MMPBSA.py: an efficient program for end-state free energy calculations. J Chem Theory Comput 8(9):3314-3321.

59. Sun HY, et al (2015) Revealing the favorable dissociation pathway of type II kinase inhibitors via enhanced sampling simulations and two-end-state calculations. Sci Rep 5: 8457.

60. Yang Y, et al (2011) Molecular Dynamics Simulation and Free Energy Calculation Studies of the Binding Mechanism of Allosteric Inhibitors with p38 alpha MAP Kinase. J Chem Inf Model 51:3235–3246.

61. Frembgen-Kesner T, Elcock AH (2006) Computational sampling of a cryptic drug binding site in a protein receptor: Explicit solvent molecular dynamics and inhibitor docking to p38 MAP kinase. J Mol Biol 359:202–214.

62. Filomia F, et al (2010) Insights into MAPK p38 alpha DFG flip mechanism by accelerated molecular dynamics. Bioorg Med Chem Lett 8:6805–6812.

63. Badrinarayan P, Sastry GN (2011) Sequence, Structure, and Active Site Analyses of p38 MAP Kinase: Exploiting DFG-out Conformation as a Strategy to Design New Type II Leads. J Chem Inf Model 51:115–129.

64. Jeffrey PD, Russo AA, Polyak K, Gibbs E, Hurwitz J, Massagué J, Pavletich NP (1995) Mechanism of CDK activation revealed by the structure of a cyclinA-CDK2 complex. Nature 376(6538):313-320.

65. Fisher, RP, David OM (1994) A novel cyclin associates with M015/CDK7 to form the CDK-activating kinase. Cell 78(4):713-724.

66. van Linden OP, Kooistra AJ, Leurs R, de Esch IJ, de Graaf C (2013) KLIFS: a knowledge-based structural database to navigate kinase–ligand interaction space. J Med Chem 57(2):249-277.

67. Levy Y, Onuchic JN (2006) Water mediation in protein folding and molecular recognition. Annu Rev Biophys Biomol Struct 35:389-415.

68. Papoian GA, Ulander J, Wolynes PG (2003) Role of water mediated interactions in protein− protein recognition landscapes. J Am Chem 125(30):9170-9178.

CHAPTER 6. Insights Into Inhibitor and Ubiquitin-like Protein Binding in SARS-CoV-2 Papain-like Protease

6.1 Abstract

Covid-19 is caused by a novel form of coronavirus for which there are currently no vaccines or anti-viral drugs. This virus, termed SARS-CoV-2 (CoV2), contains Papain-like protease (PLpro) involved in viral replication and immune response evasion. Drugs targeting this protease therefore have great potential for inhibiting the virus, and have proven successful in older coronaviruses. Here, we introduce two effective inhibitors of SARS-CoV-1 (CoV1) and MERS-CoV to assess their potential for inhibiting CoV2 PLpro.. We ran 1 μs molecular dynamics (MD) simulations of CoV2, CoV1 and MERS-CoV ligand-free PLpro to characterize the dynamics of CoV2 PLpro, and made comparisons between the three to elucidate important similarities and differences relevant to drug design and ubiquitin-like protein binding for deubiquitinating and deISGylating activity of CoV2. Next, we simulated the inhibitors bound to CoV1 and CoV2 PLpro in various poses and at different known binding sites to analyze their binding modes. We found that the naphthalene-based ligand shows strong potential as an inhibitor of CoV2 PLpro by binding at the putative naphthalene inhibitor binding site in both computational predictions and experimental assays. Our modeling work suggested strategies to improve naphthalene-based compounds, and our results from molecular docking showed that the newly designed compounds exhibited improved binding affinity. The other ligand, chemotherapy drug 6-mercaptopurine (6MP), showed little to no stable intermolecular

interaction with PLpro and quickly dissociated or remained highly mobile. We demonstrate multiple ways to improve the binding affinity of the naphthalene-based inhibitor scaffold by engaging new residues in the unused space of the binding site. Analysis of CoV2 PLpro also brings insights into recognition of ubiquitin-like proteins that may alter innate immune response.

6.2 Introduction

Covid-19, caused by a novel form of coronavirus, has created a global health crisis due to the lack of vaccines and anti-viral drugs. Over the past two decades, coronaviruses such as the Severe Acute Respiratory Syndrome coronavirus (SARS-CoV-1 or CoV1) and Middle East Respiratory Syndrome coronavirus (MERS-CoV) have caused mass human fatality. In late 2019, the novel form of coronavirus, known as SARS-CoV-2 (CoV2), spread rapidly from Wuhan, China to all continents of the world within months, causing widespread mortality and worldwide panic (CDC, 2020). The only way to curtail the spread of the virus thus far has been through strict, indefinite quarantine of millions of people. Clearly, development of anti-viral drugs capable of inhibiting CoV2 is of paramount importance.

CoV2 contains a Papain-like protease (PLpro) that is vital for viral replication (Harcourt et al., 2004). PLpro is responsible for the proteolytic processing of the product of open reading frame 1a (ORF1a) in the replicase gene of CoV2, a large viral polyprotein containing nonstructural proteins which form the replicase complex (Wertz and Murray, 2019). PLpro exists as a monomer in biological settings and has the USP fold, typical for the ubiquitin-specific proteases (USP) family in humans, which is topologically organized into four domains – UBL, thumb, palm, and fingers (Ye et al., 2009) (Figure 6.1a). The peptide bond cleavage in the active site is catalyzed by a conserved catalytic triad comprised of residues Cys111, His272 and Asp286 (Baez-Santos et al., 2015). In addition, PLpro possesses deubiquitinating and deISGylating capabilities (Sulea et al., 2005) which interfere with critical signaling pathways leading to the expression of type I interferons, resulting in antagonistic effect on host innate immune response (Bekes et al., 2016; Devaraj et al., 2007). Therefore, inhibition of PLpro activity can halt viral replication and disrupt its role in host immune response evasion, making it an excellent anti-viral drug target.

CoV2 PLpro exhibits a high sequence similarity to CoV1 PLpro (Figure S1); in particular, the binding site and active site residues are nearly identical. We have introduced a leading naphthalene-based inhibitor, 3k, and chemotherapy agent 6-mercaptopurine (6MP) (Figure 6.1a), which successfully inhibited CoV1 PLpro and MERS-CoV PLpro, respectively (Baez-Santos et al., 2014; Cheng et al., 2015; Chou et al., 2008), to assess their binding affinity for CoV2 PLpro. The 3k binding site is adjacent to the catalytic triad and sterically inhibits the binding of ubiquitin (Ub) and Interferon-stimulated gene 15 (ISG15) by occupying the space normally reserved for their C-terminal (LXGG cleavage site) at ubiquitin binding subsite 1 (SUb1) (Figure 1a). Compounds capable of binding to this site therefore exhibit high inhibitory capabilities. In this work, we carried out several molecular dynamics (MD) simulations of ligand-free and ligand-bound CoV2, CoV1, and MERS-CoV PLpro (Table 1). Based on detailed examination of the CoV2 3k binding site, we provide guidance and suggestions for optimization of compounds targeting this site. Moreover, by simulating 3k bound to CoV2 and CoV1 PLpro, we show that it exhibits a highly similar binding mode in both proteins, suggesting that 3k and similar compounds should have an inhibitory effect on CoV2 PLpro. After analyzing the binding mode and binding site, we constructed and docked new ligands based on the 3k scaffold which showed improved binding affinity over the current molecule. Additionally, we carried out experimental assays to validate 3k binding to CoV2 PLpro and inhibit enzymatic function. We show that the overall dynamics of ligand-free PLpro in all analyzed systems is highly similar, with comparable flexibility in BL2 loop, zinc-binding region and UBL domain. Our detailed description of 3k binding in the protein provides insight into the essential interactions necessary for successful fragment-based drug design. Additionally, we provide well-sampled dynamics of the available CoV2 PLpro crystal structures for wider use as a guide to potential drug binding sites or in docking and drug screening studies.

Summary of Simulations

Simulation Index	**PDB**	**Protein system**	**Length**
MD1 (1, 2)	6W9C	CoV2 PLpro	1 µs, 500 ns
MD2 (1, 2)	6WRH	CoV2 PLpro	1 µs, 500 ns
MD3 (1, 2)	4OW0	CoV1 PLpro	1 µs, 500 ns
MD4 (1, 2)	4RNA	MERS-CoV PLpro	1 µs, 500 ns
MD5a (1-3)	6W9C	CoV2 PL pro complexed w/ 3k (pose A)	1 µs, 500 ns, 200 ns
MD5b (1-3)	6W9C	CoV2 PL pro complexed w/ 3k (pose B)	3 × 200 ns
MD5c (1-3)	6W9C	CoV2 PL pro complexed w/ 3k (pose C)	3 × 200 ns
MD5d (1-3)	6W9C	CoV2 PL pro complexed w/ 3k (pose D)	3 × 200 ns
MD6 (1-3)	4OW0	CoV1 PLpro complexed w/ 3k	1 µs, 500 ns, 200 ns

MD7a (1-3)	6W9C	CoV2 PLpro complexed w/ 6MP (in putative site)	3 × 200 ns
MD7b (1-3)	6W9C	CoV2 PLpro complexed w/ 6MP (in active site)	3 × 200 ns

Table 6.1. Summary of all simulations performed. All ligand-free proteins were simulated twice under identical conditions except for the initial random number seed, first for 1 μs, followed by a 500 ns secondary run to confirm consistency in the observed dynamics. Similarly, all ligand-bound proteins were simulated three times for at least 200 ns. Where necessary, secondary and tertiary runs are referred to by a dash and number after the main designation e.g., MD1-2 means the second run of simulation MD1. These trajectories are available on our group webpage: http://chemcha-gpu0.ucr.edu/software/ and the COVID-19 Molecular Structure and Therapeutics Hub: https://covid.molssi.org/.

6.3 Results and Discussion

We analyzed 1 μs trajectories of ligand-free CoV2, CoV1 and MERS-CoV PLpro to uncover the overall protein dynamics of the novel coronavirus protease and to make comparisons to older conronavirus PLpro for which inhibitors have been developed. In addition, we simulated ligand-bound trajectories of CoV1 and CoV2 PLpro to assess potential effectiveness of one naphthalene-based and one thiopurine inhibitor – 3k and 6MP, respectively – in the 2019 coronavirus. We showed that 3k formed stable interactions with CoV2 PLpro, suggesting that the compound can bind to the protein, which was verified by experimental assays. Moreover, we designed and docked new ligands based on the 3k-scaffold to CoV2 PLpro, and show the they achieve improved binding affinity. Protein flexibility, entropy, and conformational changes were analyzed in the ligand-free protein simulations to characterize the overall protein dynamics and to assess similarities and differences relevant to inhibitor or Ub binding in CoV2 PLpro. The ligand-bound MD simulations were analyzed for a detailed characterization of ligand binding modes by analyzing residue-wise interactions, binding energy, and ligand-induced conformational changes.

6.3.1 Structure and Dynamics of Ligand-free CoV2 PLpro and Implications for Drug Discovery

Dynamic regions of potential importance to small molecule drug or Ub binding in CoV2 PLpro include portions of the thumb domain (containing SUb2), the fingers region (adjacent to SUb1) and the BL2 loop (directly adjacent to the 3k binding site). Principal component analysis (PCA) shows that the dominant overall motion of CoV2 PLpro occurs due to high flexibility of the fingers domain – especially the zinc-binding region, the BL2 loop, and the UBL domain (Figure S2). The fingers domain is the most mobile region of PLpro, and has been shown to crystallize in different conformations (Baez-Santos et al., 2015). Because this region is highly flexible and challenging for a small molecular inhibitor to bind tightly, it is not considered as an ideal druggable site.

This study focuses on the binding site of naphthalene inhibitors (Figure 6.1a), a druggable site reported in previous studies (Baez-Santos et al., 2014) that is directly adjacent to the PLpro active site to prevent off-target binding to the highly similar active site of human proteins (Kemp, 2016). Flexibility of the BL2 loop, which can result in an open or closed conformation, indicates potential of this binding site to accommodate compounds with new scaffolds or different derivatives of 3k, which may include larger substitutions to strengthen binding with underutilized regions. One such region is the hydrophobic portion lined by residues Met208, Pro247 and Pro248 (Figure 6.1b). Closer to the BL2 loop, Gly266 may be able to provide inhibitor binding specificity through hydrogen bond formation . The portion of the binding site extending just past the BL2 loop in the direction of the UBL domain presents substantial space to engage PLpro residues with larger ligands (Figure 6.1b).

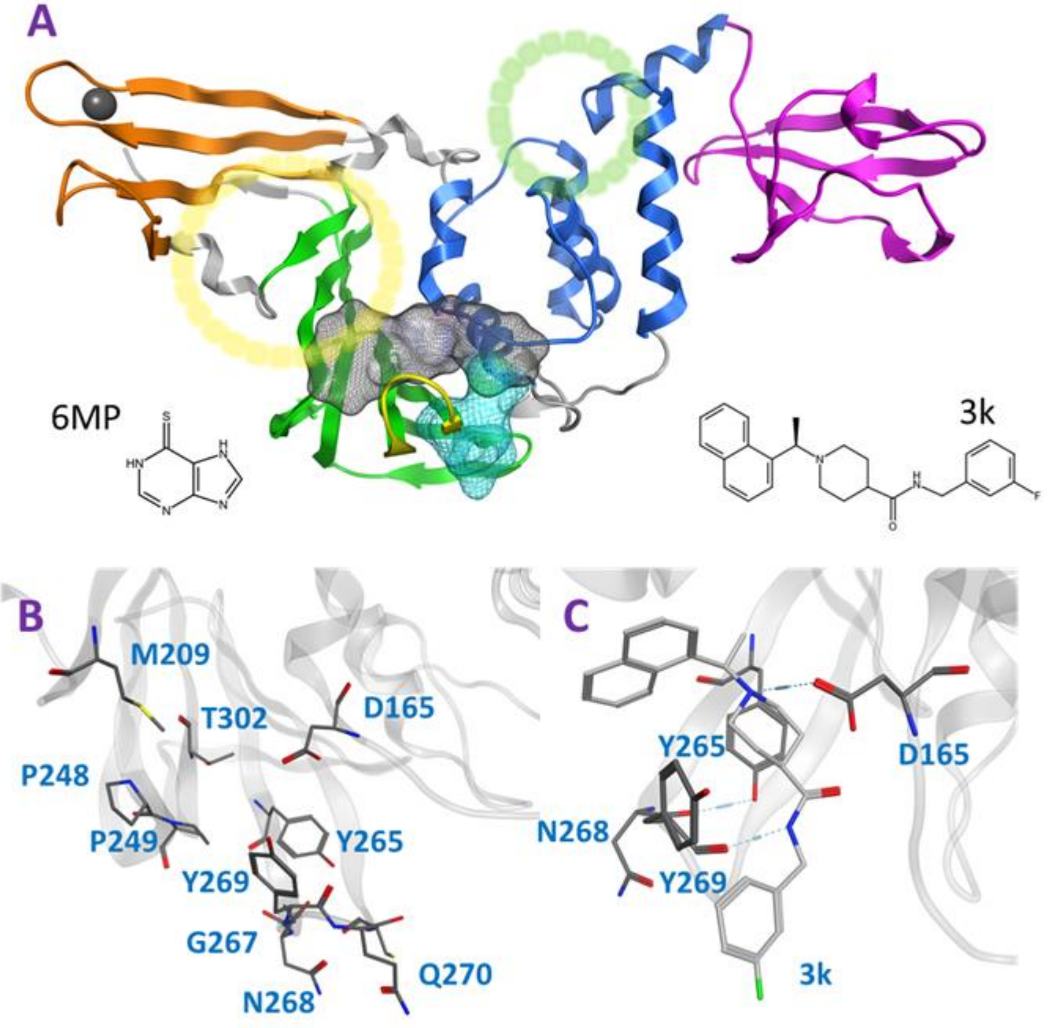


Figure 6.1. Cartoon representation of the entire CoV2 PLpro structure and close-ups of regions important to ligand binding. a) PLpro with the four domains and other major regions indicated as follows: fingers – orange, palm – green (BL2 loop – yellow), thumb – blue, UBL – magenta, SUb1 and SUb2 – yellow and green circles, respectively. The putative 3k binding site is shown as a grey surface and the active site as a teal surface. 6MP was docked to the putative 3k site and active site. b) important binding site residues. c) 3k (light grey) engaging in hydrogen bonds with D164 and Y268, and the important BL2 loop-stabilizing hydrogen bond between Y264 and N267.

One very prominent motion of CoV2 PLpro is partial rotation of the UBL domain and its relative position to “ridge” helix (Asp62 – His73) in the SUb2 region (Figure 6.2). The function of the UBL domain is unknown, and although some studies suggest that it has no effect on function of PLpro (Clasman et al., 2017), we observed one noteworthy interaction involving this domain. Transposition of UBL towards the thumb domain results in hydrophobic interactions between Pro59 of the UBL domain and Pro77 and Thr75; Thr75 then interacts with Phe69 of the “ridge” helix and can alter the latter residues conformation. Mutating this Phe was shown to affect the binding affinity of ISG15 and K48 linked diUb in CoV1 PLpro (Ratia et al., 2014), so the conformational dynamics of this residue may also be important in CoV2. Since CoV1 exhibits this same interaction between UBL residues and this Phe residue, we compared the conformation populations for Phe69/70 (residue numbering differs by 1 between Cov2/CoV1) between the two PLpros. Notably, Cov1 contains Leu at position 75 (rather than Thr75 as in Cov2) and its concerted motion with Phe70 yields four different conformations. The Phe69-Thr75 interaction in CoV2 affects Phe69 to a lesser extent, resulting in just two distinct conformations of the same sidechain (Figure S3). In contrast to the dynamic SUb2 region, SUb1, the binding site for distal Ub, does not show any significant structural fluctuation. Ub-interacting hydrophobic residues Met208 and Pro247 are exposed to the solvent to potentially engage in ligand interactions (Figure S4), which may be an alternative method to disrupt Ub binding at SUb1 in aside from blocking its C-terminal from the LXGG cleavage site.

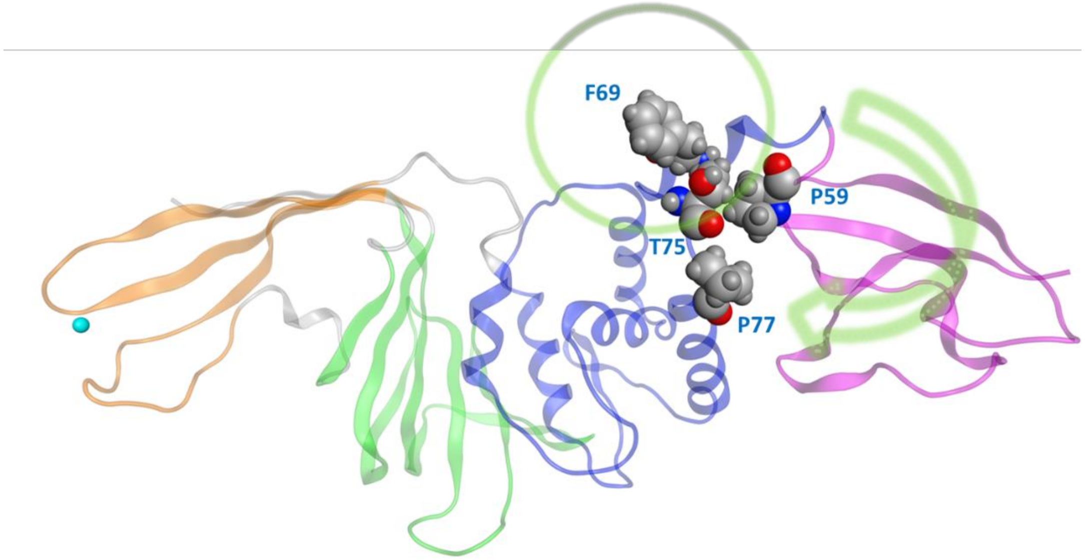


Figure 6.2. The movement of UBL in CoV2 allows for interactions between UBL residue Pro59 and the thumb domain residues Thr75 and Pro77. These interactions subsequently result in different rotameric states for the nearby key Ub-interacting residue Phe69/Phe70 in CoV2/CoV1. Green arrow indicates the major motion of UBL; green circle indicates SUb2.

Overall, the dynamics of the CoV2, CoV1 and MERS-CoV ligand-free PLpro is quite similar. Figure S2 shows the first principal component of overall motion for all three systems, which reveals similarly high mobility in the zinc-binding domain, BL2 loop and UBL domain. CoV2 simulations MD2 and MD3 showed similar flexibility to CoV1 in most of the highly flexible regions during the entire course of the 1 μs MD simulations. Notably, in MD1, the initial crystal structure conformation shows a unique conformation of Asn267 and Tyr268 (Figure 6.3), resulting in larger root-mean-square fluctuation (RMSF) and dihedral entropy values than those computed for the other ligand-free

PLpros (Figure 6.4), as well as additional rotameric states (Figure S5). Around 420 ns into MD1-1 and just 20 ns into MD1-2 the residues change conformation to ones highly similar to those in MD2 (CoV1) and MD3 (CoV2), at which point the RMSFs become nearly identical (Figure S6). This unique conformation of key ligand-binding residue Tyr268 (Chaudhuri et al., 2011) is not preorganized for protein-ligand complex formation, thus it may incur a cost in conformational energy or entropy which can affect inhibitor binding. Figure S7 compares backbone dihedral angle populations of several binding pocket residues between CoV1 and CoV2 PLpro over the simulation time. In terms of dihedral entropy as well, CoV1 and CoV2 are quite similar. The entropy calculations for the backbone torsion show only a few regions with higher conformational sampling in CoV2, mainly in the zinc binding region of the fingers domain and BL2 loop (Figure 6.4). In MERS-CoV, the amino acid composition of the BL2 loop is entirely different from CoV2 with the exception of two flanking Gly residues (Lee et al., 2015). Although the entropy and RMSF show similar flexibility of the loop, its overall conformation relative to the palm domain is more open than in CoV1 and CoV2. BL2 remains in this open conformation, which appears to be stabilized by hydrophobic interactions of Gln270, Glu273, Thr274 and His278 sidechains, for the entirety of MD4 (Figure S8).

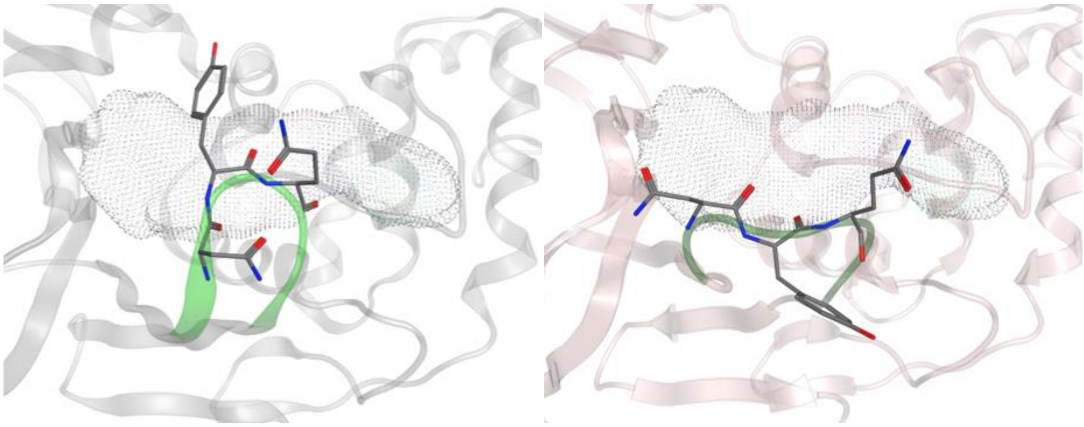


Figure 6.3. Conformation of Asn267 and Tyr268, 3k binding site is indicated with dotted surface. a) common conformation of these residues observed in CoV2 (6WRH) and CoV1 simulations. b) unique conformation observed only in CoV2 (6W9C) simulation.

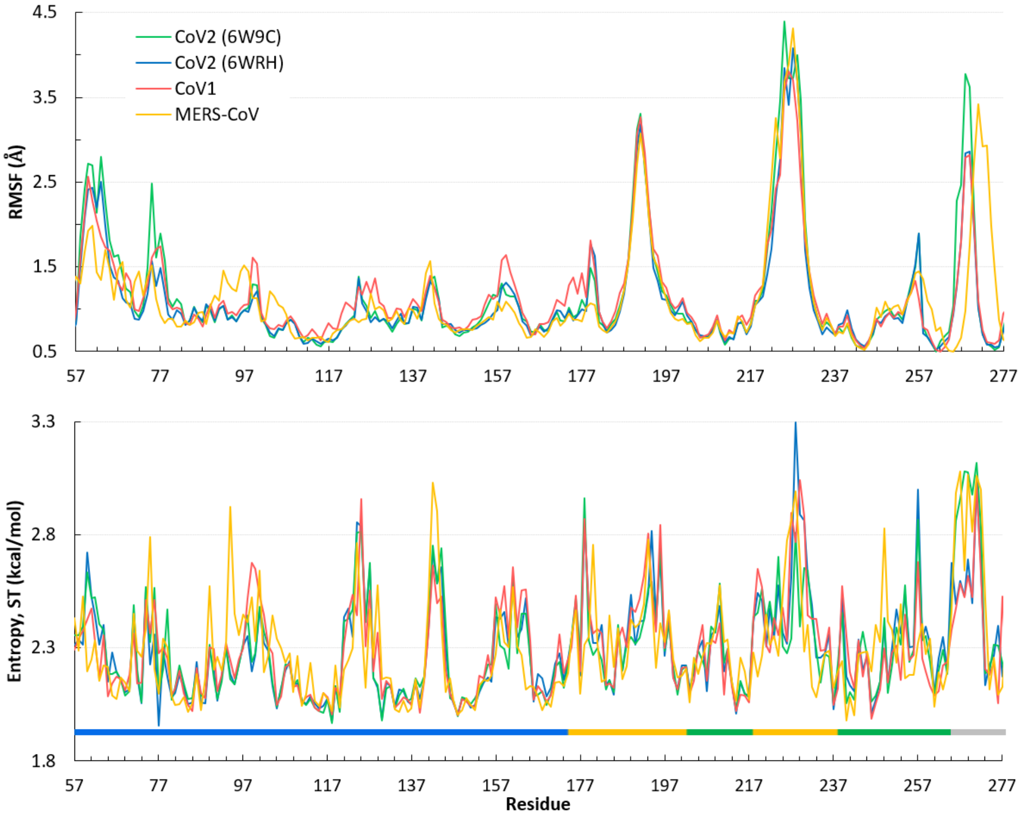


Figure 6.4. Quantifying the overall dynamics and conformational flexibility of PLpro. Top: RMSF of alpha C atoms over the 1 µs trajectory of all four ligand-free PLpros. The spike at the BL2 loop (~ residues 265-275) is larger for CoV2 (6W9C) and MERS-CoV because of their more open conformations during all or part of the simulations. Residues 1-56 (UBL domain) and 300-311 (C-terminal) have been omitted for clarity. Bottom: Dihedral entropy of the psi angle for CoV2, CoV1 and MERS-CoV systems. PLpro regions indicated by color bar: thumb – blue, fingers – yellow, palm – green, BL2 loop – grey. Entropy calculated at 298 K.

6.3.2 Comparison of Ligand-free and Ligand-bound Structures CoV1 and CoV2 PLpro

Revealing detailed protein conformational changes after ligand binding provides insight to the key binding interactions relevant to drug development. We compared ligand-free and ligand-bound systems to identify how binding shifts the populated conformations of surrounding residues.

Simulations show the CoV2 PLpro BL2 loop having significant flexibility in ligand-free proteins. Residues Asn267, Gln269, and most importantly Tyr268, account for most of this motion, which resembles opening and closing of the loop (Figures 6.5a and 6.5b). MD5a-d all show that the BL2 loop in CoV2 PLpro is highly stabilized by ligand binding , as most residues interacting with the ligand are confined to a single conformation (Figure 6.5c). Most notably, the sidechain and backbone rotation of key residue Tyr268 is minimized through a hydrogen bond and strong vdW interactions with the ligand, as detailed in next subsection. The very same ligand-induced stabilization of the BL2 loop is seen for CoV1 PLpro (Figure 6.6). The central portion of this binding pocket, which houses the piperidine, carboxyl and amide moieties of 3k, is narrower and may already be maximized in terms of inhibitor binding potential. Two key hydrogen bonds form here (Figure 6.1c), and it has been shown that substitutions larger than a methyl group or hydrogen at the benzylic-naphthyl or benzyl position, respectively, on naphthalene inhibitors lowered their effectiveness (Baez-Santos et al., 2014).

Residues involved in consistent interactions with the ligand show a significant difference in dihedral entropy. Hydrogen bonds formed with 3k substantially restrict conformational exchange for the associated residues. Asp164 and Tyr268 appear to be a key aspect in the 3k-CoV2 PLpro interactions, which is reflected by the decreased dihedral entropy (Figure S9).

Comparison of MD5a-d to MD6 (Table 1) reveals that 3k binds very similarly in CoV2 and CoV1 PLpro, inducing a closed, ordered conformation of the BL2 loop around the ligand. Moreover, the RMSD values of 3k over 200 ns in MD5a and MD6 of 1.06 and 0.95 Å, respectively, reveal similar stability in the CoV2 and CoV1 putative binding sites. The naphthalene moiety occupies the hydrophobic cleft of the pocket and the fluorophenyl ring protrudes from the opposite end of the pocket while retaining a high degree of mobility relative to the rest of the compound. The high similarity of these binding modes indicates strong potential of naphthalene inhibitors to have an inhibitory effect on CoV2 PLpro through a similar mechanism as in CoV1 PLpro.

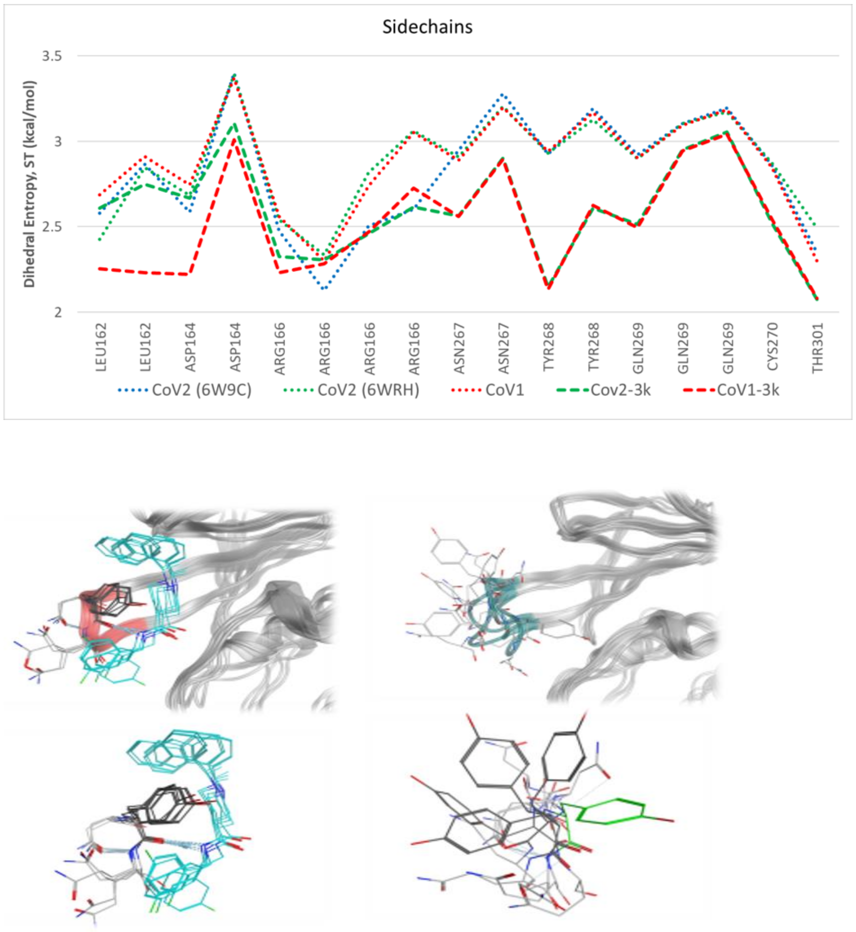


Figure 6.5. Plot of entropy for 3k binding site residues and pictures of their conformations. a) Sidechain dihedral angle entropy for 3k binding site residues in ligand-free and 3k-bound CoV1 and CoV2 PLpro shows the stabilization of these residues after ligand binding. b) An overlay of several MD frames shows the range of conformations adopted by BL2 loop (dark green) residues in the ligand-free state. Tyr268 in the ligand-bound conformation is shown in light green. c) The conformational sampling of these residues is dramatically reduced upon binding of 3k (teal). Entropy calculated at 298.

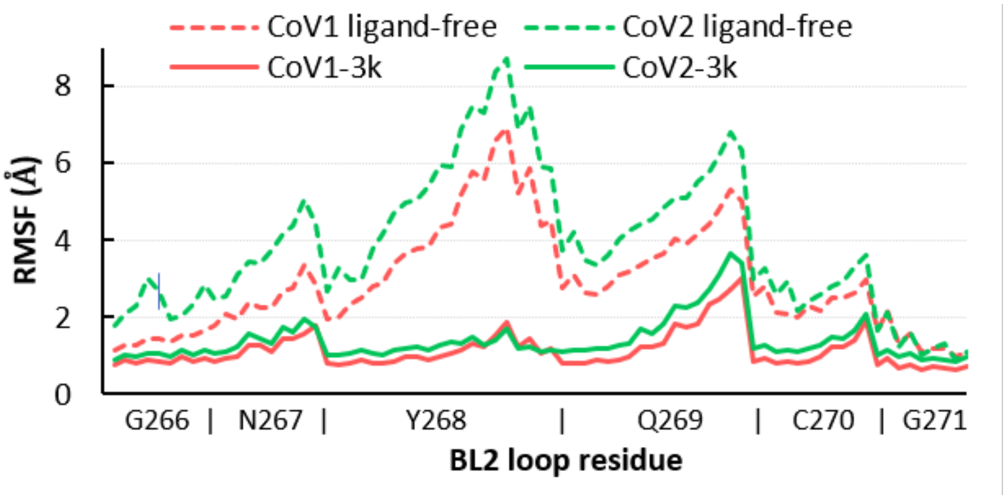


Figure 6.6. RMSF of all atoms in the BL2 loop in ligand-free and 3k-bound CoV1 and CoV2 PLpro. In both systems, ligand binding induces a closed, ordered BL2 loop conformation resulting in dramatically reduced mobility of this region. The x-axis indicates the range of atoms in each BL2 loop residue.

6.3.3 Ligands Binding Modes in CoV2 PLpro and Strategies for Drug Design

Because the putative naphthalene inhibitor binding site of CoV2 PLpro is comprised by the same residues as in CoV1 PLpro, we examined one of the most effective second generation naphthalene inhibitors of CoV1 PLpro (Baez-Santos et al., 2014), 3k, to reveal structural information regarding binding to CoV2 PLpro for future structure-based drug development. After analyzing free and ligand-bound CoV1 PLpro simulations, we docked 3k to one CoV2 PLpro conformation to obtain four different binding poses (Figure 6.7) and ran three simulations for each pose (MD5a-d). Poses A and B were nearly the same, except B was docked with unconstrained side chain rotations allowed, so 3k starts slightly rotated with respect to A. MD5c and MD5d start with a 180° rotation of the naphthalene or piperidine moiety, respectively, compared to MD5a. Ultimately, MD5a, b and d all establish the same major interactions with PLpro. The majority of our discussion focuses on MD5a , where the initial conformation (Figure 7a) is the most similar to the CoV1 PLpro-3k crystal structure. Our results indicated that 3k binds strongly and suggest that the ligand can inhibit the enzymatic function of CoV2 PLpro. Experimental results have confirmed 3k binding by showing the NMR spectrum for ligand-free and -bound CoV2 PLpro (Figure S10).

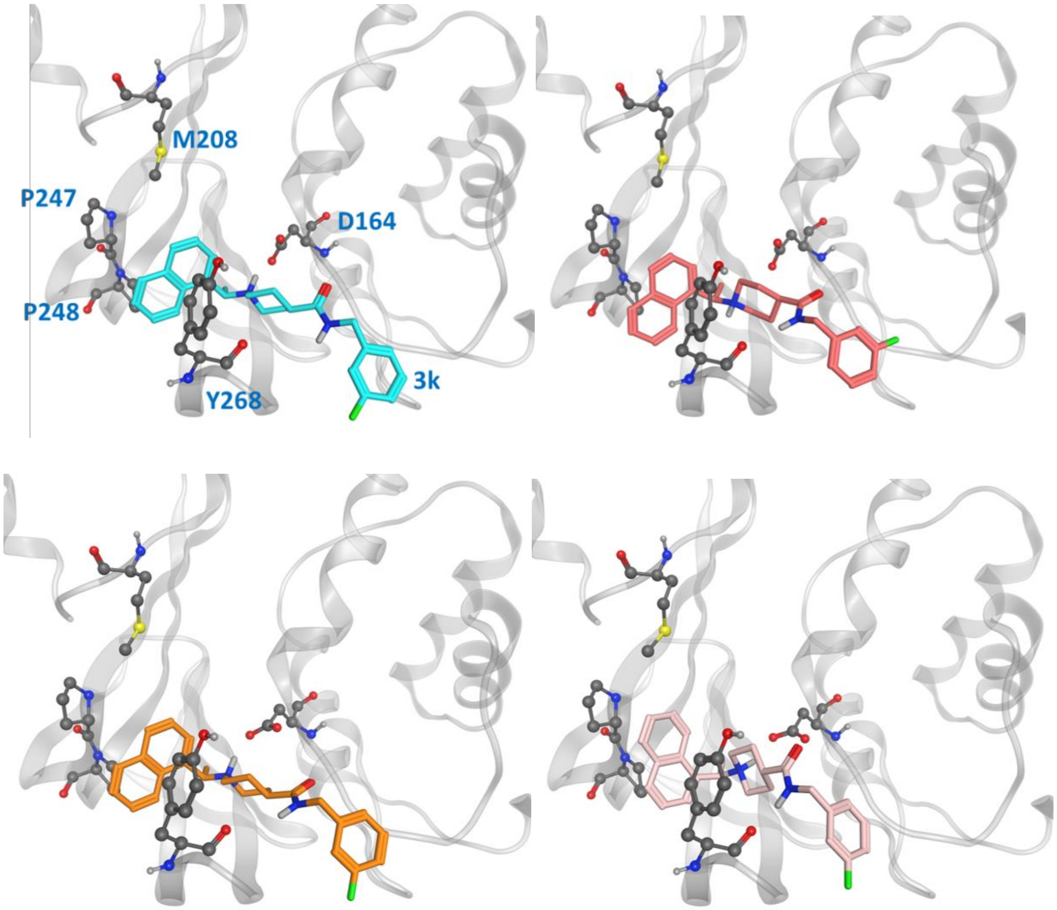


Figure 6.7. Ligand poses A-D from which the CoV2 PLpro-3k complex simulations began. Key binding site residues are shown in grey to show the difference in relative orientation of the ligand in each pose.

In the hydrophobic portion of the binding site, the naphthalene moiety sits stably between residues Pro247 and Pro248 to one side and Tyr268 on the other (Figure 6.8) However, additional space exists in this pocket to engage more residues. Specifically, it may be possible to increase the hydrophobic interactions here with a methyl (or larger) substitution on the naphthalene to further engage in vdW interactions with Pro248 or

Tyr264 (Figure 6.8, blue dots). Pro248 and Met208 can also be further engaged in hydrophobic interactions with substitutions at the appropriate positions on naphthalene (Figure 6.8, yellow dots), or potentially even a substitution of the entire naphthalene moiety for a larger aromatic structure such as anthracene or phenanthrene (Figure 6.8). MD5c (Figure 6.7c) , in which the naphthalene in the initial 3k conformation is flipped 180° relative to MD5a, provides support for this idea, as the flipped moiety is seen making closer contact with residues Met208 and Pro248, resulting in greater attraction to these residues (Figure S11) and slightly lower overall binding energy (Table 2) than in MD5a. Lastly, a hydrogen bond donor or acceptor substitution at the correct naphthalene position (Figure 6.8, blue dots) may be able to engage with the Gly266 backbone.

3k engages in two strong hydrogen bonds with the protein: one to Asp164 and the other to the backbone of Tyr268. Notably, even in MD5d, which began with the piperidine nitrogen and its hydrogen pointed in the opposite direction of Asp164, the entire moiety rotates after 25-140 ns (varying between the three runs) to establish the hydrogen bond with this residue. Previous studies found that bulky ligand substitutions that occupy this portion of the pocket decreased inhibition1. A possible explanation is that the specific ligand orientation needed to maintain both of these strong hydrogen bonds was not attainable due to the additional bulk. Moreover, we observed a consistent intra-protein hydrogen bond between Tyr264 and Asn267 in ligand-bound CoV2 that could be disrupted by larger ligand substitutions here, which may destabilize the closed BL2 loop. Indeed, analysis of this interaction shows very high correlation between formation of the hydrogen bond and a closed loop conformation (Figure S12). A small hydrophobic pocket formed by Tyr264, Tyr273 and Thr301 accommodates the methyl group at the benzylic-naphthyl position.

The fluorophenyl ring of 3k appears to interact mostly with the hydrocarbon portion of Gln269, but also engages in vdW interactions with Tyr268 (Figure 6.8). However, because of the openness of PLpro in this region, the position of the ring fluctuates widely, and it rotates freely with the fluorine observed at several positions consistent with 360-degree rotation. Increasing ligand engagement with PLpro residues is achievable in this region, although previous attempts at doing so in CoV1 PLpro had mixed results. Substitutions on the benzyl ring in first generation naphthalene inhibitors found that anything bulkier than methyl at the ortho position decreased inhibition (Baez-Santos et al., 2014); however, the linkage between the amide and benzene ring was one carbon shorter than in the second generation, possibly causing the added bulk to disrupt one of the two important hydrogen bonds with Tyr268 or Asp164. With a longer linkage to the benzene in second-generation naphthalene inhibitors, various benzene substitutions were tested, but showed no clear trend in effectiveness. Ultimately, the fluorine substitution at the meta position, as seen in 3k, showed the best result. Extending the linkage between the amide and benzene ring by one additional carbon was found to weaken inhibition, providing evidence that benzene ring primarily contributes to binding through vdW interactions with Tyr268 and Gln269, and so needs to be close to those residues. This is consistent with our observations and residue-wise force calculations as well (Figure S11).

One method to increase binding affinity in this region may be through increasing the hydrophobic surface area of the benzyl end of the ligand. This can be achieved either through substitution of methyl or larger hydrocarbon groups onto the benzene ring, or by replacing the benzene with a bulkier group, such as naphthalene. Although, as previously stated, it has been shown that both increasing ligand bulkiness near the benzene end and extending the linkage to benzene can sometimes lead to decreased inhibition, changing these two factors simultaneously has not been tested. A longer linkage may accommodate increased bulk, while the added hydrophobic mass can still reach residues Tyr268 and Gln269 for attractive interactions. Moreover, since no clear trend in ligand effectiveness from substitutions on the benzyl ring has been found, we suggest exploration of the available space in this portion of the binding site.

To validate some of our proposed modifications to the current naphthalene-based scaffold, we docked these modified ligands to the same CoV2 PLpro conformation used to dock 3k. First, to investigate the potential for making additional hydrophobic contacts in the cleft near SUb1, we substituted anthracene or phenanthrene to the naphthalene position. Results for both substitutions show more favorable docking scores than for 3k, with anthracene showing slightly better performance than phenanthrene (Figure 6.9). The favorable contacts arise from interaction with Asp166, a residue that rotates freely in the CoV2-3k simulations, indicating that it may be available to form a stable interaction with ligands capable of reaching it. Additionally, both the anthracene and phenanthrene-substituted ligands maintained all the other essential interactions we identified for 3k. In an attempt to increase polar interactions, we added a hydroxyl to the naphthalene moiety to form a hydrogen bond with Gly266 or other hydrogen bond acceptors in the area. We found that the hydrogen bond with Gly266 does indeed form as expected, with a distance of just 1.74 Å. However, in this conformation, the important hydrogen bonding group on the ligand that usually interacts with Asp164 is slightly out of position (Figure 6.9). Despite this, the binding is still more favorable than that of 3k. Also, notably, our MD simulations of CoV2-3k that started without the 3k-Asp164 hydrogen bond quickly formed that hydrogen bond after the simulation began, providing evidence that the same will likely happen in the case of the new ligand. Taken altogether, the MD and docking results show that the 3k scaffold should be capable of exploiting bulkier hydrophobic groups or polar groups at the naphthalene end to establish both favorable hydrophobic and polar contacts while maintaining the essential residue interactions that made 3k successful in CoV1. These are but a few of many possible enhancements to the naphthalene-based scaffold, and they require further validation through MD simulation or experimental assays. However, this serves as a strong proof-of-concept for future CoV2 PLpro design directions.

Pair-wise force distribution analysis (Figure S11) and interaction energies (Table S1) indicate that the binding mode of 3k in CoV2 is dependent on both strong vdW interactions and hydrogen bonds (Figure S13), highly similar to that of 3k in CoV1 (Table 2). The interaction with Tyr268 is a dominant one in all ligand-bound simulations. The residue engages in a hydrogen bond donated by the amide nitrogen of 3k and a Tshaped pi-stacking interaction with the naphthalene moiety (Figure 6.8), which is seen with all naphthalene-based inhibitors (Baez-Santos et al., 2015). The hydrogen bond between 3k and Asp164 (Figure 6.1c) is another strong protein-ligand interaction shared in CoV1 and CoV2. All major ligand interactions with binding site residues are shown in Figure S11 and listed in Table S1.

In addition to 3k, 6-mercaptopurine (6MP) from the thiopurine class of inhibitors has been reported to reversibly inhibit CoV1 (Chou et al., 2008) and MERS-CoV (Cheng et al., 2015) PLpro activity. To assess its potential for inhibiting CoV2 PLpro, we docked 6MP in the active site of the enzyme and the putative ligand binding site (Figure 6.1a) based on the proposed binding poses from existing studies. We ran three independent simulations of the two complexes for 200 ns each (MD7a and MD7b). The compound dissociated from the putative ligand binding site within 80 ns or less and no stable intermolecular interactions were established. The compound stayed within the active site during two of the three MD7b simulations; however, it remained highly mobile and unstable in the pocket (Figure S14). Because the ligand was unstable in both binding sites, we did not compute interaction energies between 6MP and Cov2 PLpro. Our analysis suggests that ligand 6MP is a weak binder and may be a poor inhibitor of Cov2 PLpro.

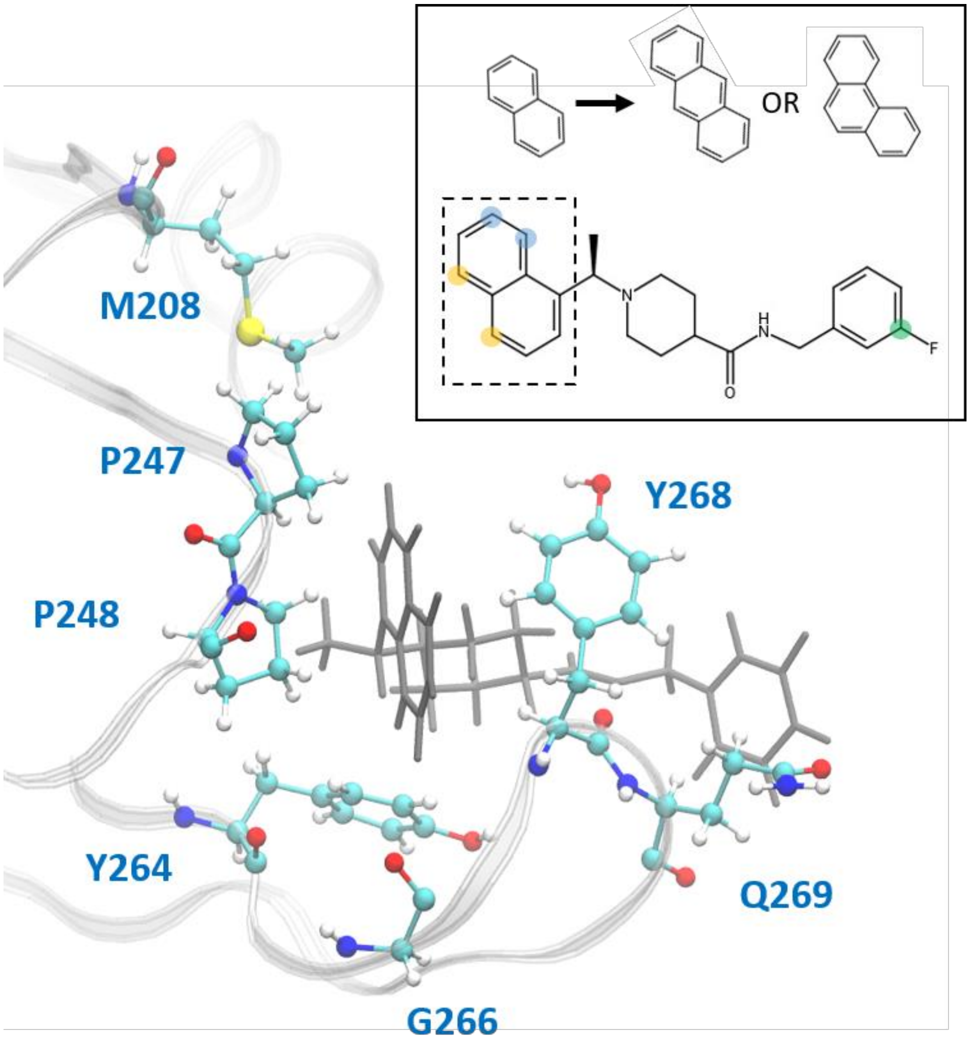
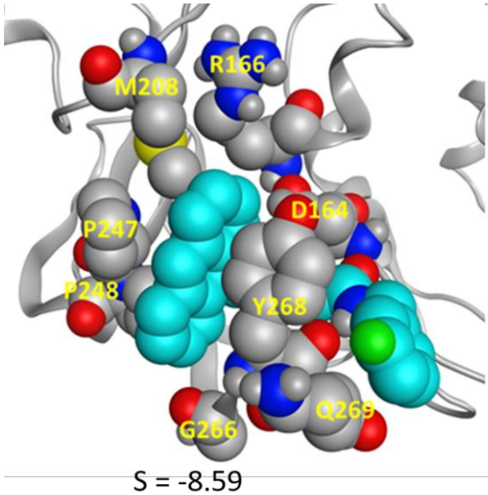
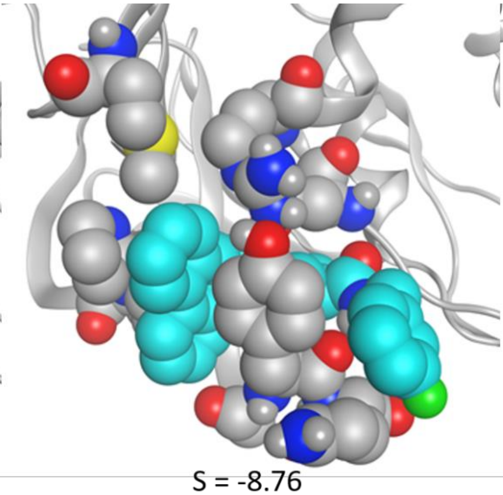


Figure 6.8. Ligand 3k (grey) in the CoV2 PLpro binding site. Residues with which new interactions are achievable or current ones can be strengthened are labeled and shown in ball-and-stick representation. Top-right: 2D molecular structure of ligand 3k indicating proposed substitution positions for increased binding affinity. Substitutions at the yellow positions may be capable of additional hydrophobic contacts with Pro247, Pro248, or Met208. Substitutions at blue positions may be capable of additional hydrophobic

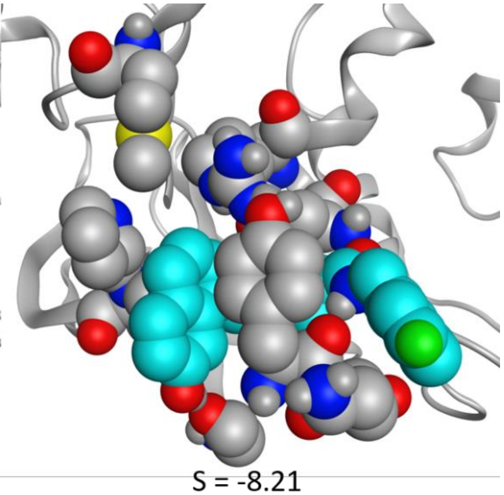
contacts with Pro248 or Tyr264, or hydrogen bonds with the backbone carboxyl of Gly266. Finally, substitutions at the green position in combination with an extended benzene linkage may be capable of increased attractive interactions with Gln269 or other nearby residues. The naphthalene moiety is indicated by the dashed box, with the proposed anthracene or phenanthrene substitutions indicated above.



S = -8.59
[-0.39]



S = -8.76
[-0.56]



S = -8.21
[-0.01]

Figure 6.9. The modified ligands (teal) docked to CoV2 PLpro. a) 3k with anthracene substituted for naphthalene, b) 3k with phenanthrene substituted for naphthalene, and c) 3k with a hydroxyl substituted on the napthyl moiety; the hydrogen bond with Gly266 is clearly visible. S is the docking score given by *MOE*, and the difference between the pictured docked conformation and the best 3k score is show in brackets.

MM/PBSA Binding Energies					
	CoV1-3k	**CoV2-3k A**	**CoV2-3k B**	**CoV2-3k C**	**CoV2-3k D**
$\Delta\mathbf{E}_{\mathbf{elec+PB}}$	11.9	11.8	11.6	11.0	14.7
$\Delta\mathbf{E}_{\mathbf{vdW+np}}$	-28.9	-26.2	-28.1	-28.7	-27.0
$\Delta\mathbf{E}_{\mathbf{MM/PBSA}}$	-17.0 ± 3.9	-14.5 ± 4.3	-16.5 ± 4.5	-17.7 ± 3.6	-12.3 ± 5.4

Table 6.2. MM/PBSA energy breakdowns for the binding energy from simulations of the four different starting poses (A-D) of CoV2 PLpro-3k and CoV1 PLpro-3k. $\Delta E_{elec+PB}$ is the electrostatic plus polar solvation energy contributions, and ΔE_{vdW+np} is the van der Waals plus non-polar solvation energy contribution. $\Delta E_{MM/PBSA}$ is the binding energy predicted by MM/PBSA. Energies are in kcal/mol; values are ± SD.

6.4 Conclusions

By analyzing the dynamics of ligand-free and ligand-bound CoV2 PLpro, we have gained insight to the important dynamics and intramolecular interactions relevant to its function and the development of small molecule inhibitory drugs. The BL2 loop, zinc binding region, and UBL domain are the most mobile protein regions, and CoV2 PLpro overall dynamics are extremely similar to those of CoV1 PLpro. SUb1 contains hydrophobic residues that contact the ligand in the 3k binding site, while SUb2 is adjacent to the highly mobile UBL domain and is affected by contacts to its Ub-interacting residues brought about by UBL domain rotation.

We docked two ligands, 3k and 6MP, known to inhibit CoV1 and MERS-CoV PLpro, respectively, into CoV2 PLpro to assess their ability as CoV2 inhibitors and identify opportunities for further optimization of the ligand scaffolds. We found that not only can 3k bind strongly to CoV2 PLpro, but that there is room for further optimization of binding affinity by exploitation of space in the small hydrophobic cleft near Pro247, Pro248 and Tyr264, or by making additional residue contacts in the open pocket region at the opposite end of the binding site. By docking newly designed ligands based on 3k, we validated our optimization suggestions, which showed better docking scores than 3k and exhibited binding modes in agreement with our proposed concepts that still maintained the intermolecular interactions that characterize successful naphthalene-based inhibitors. 6MP was unable to bind stably in the 3k site and dissociated quickly in all three simulations. It associated for longer to the active site; however, even when it remained bound, the compound was unstable.

Our results show that naphthalene-based inhibitors or similar compounds should have an inhibitory effect on CoV2 PLpro, and we have provided detailed suggestions for how this ligand scaffold can be furthered improved by engaging residues in underutilized space of the binding site. This study also generates an ensemble of CoV2 PLpro conformations that illustrate potential inhibitor-protein interactions for structure-based inhibitor design and elucidates protein dynamics relevant to Ub or Ub-like protein binding.

6.5 Methods

6.5.1 MD Simulation Protocol

MD simulations were prepared and run using the Amber18 molecular dynamics package with GPU acceleration (D.A. Case, 2018; Götz et al., 2012). Force fields ff14SB (Maier et al., 2015) and the general Amber force field (GAFF2) (Wang et al., 2004) were used on proteins and ligands, respectively. Ligands 3k are 6MP were parameterized using Amber's antechamber program with the AM1-BCC charge assignment method (Jakalian et al., 2002). All systems were solvated with a rectangular box of explicit TIP3P water extending 12 Å beyond the solute edges, and each contained no more than 1-3 Na+ or Cl- counterions, which were added only to neutralize overall system charge. Systems were minimized in four steps. First, using Generalized Born implicit solvent (Chen et al., 2008), we minimized the hydrogen atoms, then protein sidechains, and finally the entire protein for 500, 1000, and 5000 steps, respectively. Next, the entire solvated structure was minimized for 5000 steps. Solvated systems were equilibrated in the isothermic-isobaric (NPT) ensemble from 50 to 275 K in 25 K increments for 100 ps each, and finally at 298 K for 500 ps. Production simulations were performed in the NPT ensemble at 298K using the Langevin thermostat with a 2 fs timestep. A 12 Å cutoff distance was used for direct non-bonded energy calculations and long-range electrostatics were calculated by the particle mesh Ewald method (Sagui et al., 2003). The SHAKE algorithm (Ryckaert et al., 1977) was employed to constrain all bonds involving hydrogen. Raw trajectories were saved every 2 ps and then processed using Amber's cpptraj (Roe and Cheatham, 2013) for analysis.

6.5.2 Selection of Initial Structures for MD Simulation

Initial coordinates for CoV2 PLpro simulations were obtained from two crystal structures of the ligand-free protein, PDB IDs 6W9C and 6WRH . The ligand-bound complexes for CoV2 were obtained by docking ligands into a protein conformation selected from MD1; details are provided in the following subsection. CoV1 PLpro simulations began from a crystal structure of a 3k-bound complex, PDB 4OW0 (Baez-Santos et al., 2014). Ligand 3k was manually removed from the binding site for our ligand-free CoV1 PLpro simulation. MERS-CoV PLpro was simulated only in the ligand-free state, starting from crystal structure 4RNA (Lee et al., 2015). For simplicity, we have indexed these simulations as shown in Table 1.

6.5.3 Ligand Docking to CoV2 PLpro

Force distribution analysis tool (FDA) (Stacklies et al., 2011) was used to identify the residues interacting with 3k in CoV1 PLpro (Figure S15), and since these residues are identical in CoV2 PLpro, we used them as a ligand docking site. To choose a CoV2 PLpro conformation that was highly similar to the minimized CoV1 3k-bound crystal structure, we found the CoV2 frame from MD1-1 with minimum RMSD between key binding site residues to use for docking. (Figure S15). The ligands were docked to this single PLpro conformation using Molecular Operating Environment (MOE) (2018). Four poses (Figure 6.7) of 3k in this site were selected for MD simulations. To obtain poses A and D, we used the induced fit docking option with constrained/tethered side chain rotations allowed; poses B and C were obtained using the same induced fit option with free sidechain rotation allowed. Poses A and B closely resemble the Cov1 PLpro–3k crystal structure, PDB 4OW0. Pose C is a rotamer of A with a 180° rotation of the naphthalene moiety, while in pose D the piperidine moiety is rotated 180° with respect to A. In addition to the 3k binding site, 6MP was also docked to the active site following the same method. The designed ligands reflecting our suggested modifications to the 3k scaffold were docked in the 3k site by the same protocol as above to the same CoV2 PLpro conformation as 3k.

6.5.4 Simulation Analysis

Trajectory Visualization and Dihedral Analysis. The simulations were visualized using Visual Molecular Dynamics (VMD) (Humphrey et al., 1996) and MOE. Dihedral angle populations and entropy were calculated using T-Analyst (Ai et al., 2010). MD simulations trajectories have been made available on our group website: http://chemcha-gpu0.ucr.edu/software/ and are deposited at the COVID-19 Molecular Structure and Therapeutics Hub: https://covid.molssi.org/. The trajectories there have been stripped of water and counterions and were saved every 10 ps. Trajectories with water are available upon request.

Cartesian Principal Component Analysis. To observe major protein motions, we performed

principal component analysis (1933; Hotelling, 1992) of α-carbon atoms in the 1 μs trajectory of ligand-free CoV2 PLpro. PCA reduces the high-dimensional data set of all α-carbon motions throughout the MD trajectory to its principal components (PCs), the directions which contain the largest motions. We used the average α-carbon positions as references. The first and second largest PCs were analyzed to reveal the dominant motions.

MM/PBSA. We used the MM/PBSA method (Wang et al., 2018) to evaluate the intermolecular interactions between ligands and PLpro. From a total of 20,000 MD frames making up the 200 ns ligand-bound trajectories, system conformations were analyzed every 2 ns. This method computes the energy (E) of a system from the protein, ligand, and protein-ligand complex, and computes the interaction energy as Δ<E> = <Ecomplex> - <Eprotein> - <Eligand>. <E> denotes the computed average energy from a given MD trajectory. The default values of a solute dielectric of 1.0 and solvent dielectric of 80.0 were used. The total binding energy term was computed as EMM/PBSA = EMM + GPB + Gnp, where EMM includes standard molecular mechanics force field terms, GPB is the solvation energy computed by solving the Poisson Boltzmann (PB) equation, and Gnp is the nonpolar energy estimated from the solvent accessible surface area (A) as γA + b + Gdisp. Here γ is the surface tension, b is a correction term, and Gdisp is the free energy of forming attractive solute-solvent van der Waals interactions. In this work, γ = 0.03780 kcal mol-1 Å-2 and b = -0.5692 kcal mol-1.

References

1. Osipiuk, J., Tesar, C., Jedrzejczak, R., Endres, M., Welk, L., Babnigg, G., Kim, Y., Michalska, K., Joachimiak, A.. The crystal structure of Papain-Like Protease of SARS CoV-2 , C111S mutant FAU. Center for Structural Genomics of Infectious Diseases (CSGID). (2020) http://dx.doi.org/10.2210/pdb6wrh/pdb.

2. Osipiuk, J., Jedrzejczak, R., Tesar, C., Endres, M., Stols, L., Babnigg, G., Kim, Y., Michalska, K., Joachimiak, A.. The crystal structure of papain-like protease of SARS CoV-2 FAU. Center for Structural Genomics of Infectious Diseases (CSGID). (2020) http://dx.doi.org/10.2210/pdb6w9c/pdb.

3. *Molecular Operating Environment (MOE)*. (2018) (1010 Sherbrooke St. West, Suite #910, Montreal, QC, Canada, H3A 2R7, Chemical Computing Group ULC).

4. Centers for Disease Control and Prevention (2020). Provisional Death Counts for Coronavirus Disease (COVID-19). https://www.cdc.gov/nchs/nvss/vsrr/covid19/index.htm [Accessed May 28, 2020].

5. Ai, R., Qaiser Fatmi, M., and Chang, C.-e.A. (2010). T-Analyst: a program for efficient analysis of protein conformational changes by torsion angles. J. Compu. Aided Mol. Des. *24*, 819.

6. Baez-Santos, Y.M., Barraza, S.J., Wilson, M.W., Agius, M.P., Mielech, A.M., Davis, N.M., Baker, S.C., Larsen, S.D., and Mesecar, A.D. (2014). X-ray Structural and Biological Evaluation of a Series of Potent and Highly Selective Inhibitors of Human Coronavirus Papain-like Proteases. J. Med. Chem. *57*, 2393-2412.

7. Baez-Santos, Y.M., St John, S.E., and Mesecar, A.D. (2015). The SARS-coronavirus papain-like protease: Structure, function and inhibition by designed antiviral compounds. Antivir. Res. *115*, 21-38.

8. Bekes, M., van Noort, G.J.V., Ekkebus, R., Ovaa, H., Huang, T.T., and Lima, C.D. (2016). Recognition of Lys48-Linked Di-ubiquitin and Deubiquitinating Activities of the SARS Coronavirus Papain-like Protease. Mol. Cell *62*, 572-585.

9. Chaudhuri, R., Tang, S., Zhao, G., Lu, H., Case, D.A., and Johnson, M.E. (2011). Comparison of SARS and NL63 papain-like protease binding sites and binding site dynamics: inhibitor design implications. J. Mol. Biol. *414*, 272-288.

10. Chen, J., Brooks, C.L., and Khandogin, J. (2008). Recent advances in implicit solvent-based methods for biomolecular simulations. Curr. Opin. Struc. Biol. *18*, 140-148.

11. Cheng, K.W., Cheng, S.C., Chen, W.Y., Lin, M.H., Chuang, S.J., Cheng, I.H., Sun, C.Y., and Chou, C.Y. (2015). Thiopurine analogs and mycophenolic acid synergistically inhibit the papain-like protease of Middle East respiratory syndrome coronavirus. Antivir. Res. *115*, 9-16.

12. Chou, C.Y., Chien, C.H., Han, Y.S., Prebanda, M.T., Hsieh, H.P., Turk, B., Chang, G.G., and Chen, X. (2008). Thiopurine analogues inhibit papain-like protease of severe acute respiratory syndrome coronavirus. Biochem. Pharmacol. *75*, 1601-1609.

13. Clasman, J.R., Báez-Santos, Y.M., Mettelman, R.C., O'Brien, A., Baker, S.C., and

14. Mesecar, A.D. (2017). X-ray Structure and Enzymatic Activity Profile of a Core Papain-like Protease of MERS Coronavirus with utility for structure-based drug design. Sci. Rep. *7*, 1-13.

15. D.A. Case, I.Y.B.-S., S.R. Brozell, D.S. Cerutti, T.E. Cheatham, III, V.W.D. Cruzeiro, T.A. Darden, R.E. Duke, D. Ghoreishi, M.K. Gilson, H. Gohlke, A.W. Goetz, D. Greene, R Harris, N. Homeyer, S. Izadi, A. Kovalenko, T. Kurtzman, T.S. Lee, S. LeGrand, P. Li, C. Lin, J. Liu, T. Luchko, R. Luo, D.J. Mermelstein, K.M. Merz, Y. Miao, G. Monard, C. Nguyen, H. Nguyen, I. Omelyan, A. Onufriev, F. Pan, R. Qi, D.R. Roe, A. Roitberg, C. Sagui, S. Schott-Verdugo, J. Shen, C.L. Simmerling, J. Smith, R. Salomon-Ferrer, J. Swails, R.C. Walker, J. Wang, H. Wei, R.M. Wolf, X. Wu, L. Xiao, D.M. York and P.A. Kollman (2018). Amber 2018. (University of California, San Francisco.).

16. Devaraj, S.G., Wang, N., Chen, Z., Chen, Z., Tseng, M., Barretto, N., Lin, R., Peters, C.J., Tseng, C.-T.K., Baker, S.C., *et al.* (2007). Regulation of IRF-3-dependent Innate Immunity by the Papain-like Protease Domain of the Severe Acute Respiratory Syndrome Coronavirus. J. Biol. Chem. *282*, 32208-32221

17. Götz, A.W., Williamson, M.J., Xu, D., Poole, D., Grand, S.L., and Walker, R.C. (2012). Routine Microsecond Molecular Dynamics Simulations with AMBER on GPUs. 1. Generalized Born. J. Chem. Theory Comput. *8*, 1542-1555

18. Harcourt, B.H., Jukneliene, D., Kanjanahaluethai, A., Bechill, J., Severson, K.M., Smith,

19. C.M., Rota, P.A., and Baker, S.C. (2004). Identification of Severe Acute Respiratory Syndrome Coronavirus Replicase Products and Characterization of Papain-Like Protease Activity. J. Virol. *8*, 13600-13612.

20. Hotelling, H. (1933). Analysis of a complex of statistical variables into principal components. *Journal of Educational Psychology, 24*(6), 417-441.

21. Hotelling, H. (1992). Relations Between Two Sets of Variates. In Breakthroughs in Statistics: Methodology and Distribution, S. Kotz, and N.L. Johnson, eds. (New York, NY: Springer New York), pp. 162-190.

22. Humphrey, W., Dalke, A., and Schulten, K. (1996). VMD: Visual molecular dynamics. J. Mol. Graph. *14*, 33-38.

23. Jakalian, A., Jack, D.B., and Bayly, C.I. (2002). Fast, efficient generation of high-quality atomic charges. AM1-BCC model: II. Parameterization and validation. J. Comput. Chem. *23*, 1623-1641.

24. Kemp, M. (2016). Chapter Three - Recent Advances in the Discovery of Deubiquitinating Enzyme Inhibitors. In Progress in Medicinal Chemistry, G. Lawton, and

25. D.R. Witty, eds. (Elsevier), pp. 149-192.

26. Lee, H., Lei, H., Santarsiero, B.D., Gatuz, J.L., Cao, S.Y., Rice, A.J., Patel, K., Szypulinski, M.Z., Ojeda, I., Ghosh, A.K*., et al.* (2015). Inhibitor Recognition Specificity of MERS-CoV Papain-like Protease May Differ from That of SARS-CoV. ACS Chem. Biol. *10*, 1456-1465.

27. Maier, J.A., Martinez, C., Kasavajhala, K., Wickstrom, L., Hauser, K.E., and Simmerling, C. (2015). ff14SB: Improving the Accuracy of Protein Side Chain and Backbone Parameters from ff99SB. J. Chem. Theory Comput. *11*, 3696-3713

28. Ratia, K., Kilianski, A., Baez-Santos, Y.M., Baker, S.C., and Mesecar, A. (2014). Structural Basis for the Ubiquitin-Linkage Specificity and deISGylating Activity of SARS-CoV Papain-Like Protease. PLoS Pathog. *10*, e1004113.

29. Roe, D.R., and Cheatham, T.E. (2013). PTRAJ and CPPTRAJ: Software for Processing and Analysis of Molecular Dynamics Trajectory Data. J. Chem. Theory Comput. *9*, 3084-3095.

30. Ryckaert, J.-P., Ciccotti, G., and Berendsen, H.J.C. (1977). Numerical integration of the cartesian equations of motion of a system with constraints: molecular dynamics of n-alkanes. J. Comput. Phys. *23*, 327-341.

31. Sagui, C., Pedersen, L.G., and Darden, T.A. (2003). Towards an accurate representation of electrostatics in classical force fields: Efficient implementation of multipolar interactions in biomolecular simulations. J. Chem. Phys. *120*, 73-87.

32. Stacklies, W., Seifert, C., and Graeter, F. (2011). Implementation of force distribution analysis for molecular dynamics simulations. BMC Bioinformatics *12*, 101.

33. Sulea, T., Lindner, H.A., Purisima, E.O., and Ménard, R. (2005). Deubiquitination, a New Function of the Severe Acute Respiratory Syndrome Coronavirus Papain-Like Protease? In J Virol. *79*, 4550-4551.

34. Tan, C., Tan, Y-H., and Luo, R. (2007). Implicit nonpolar solvent models. J. Phys. Chem. B *111* 12263-12274.

35. Wang, C., Greene, D.A., Xiao, L., Qi, R., and Luo, R. (2018). Recent Developments and Applications of the MMPBSA Method. Frontiers in Molecular Biosciences *4*, 87.

36. Wang, J., Wolf, R.M., Caldwell, J.W., Kollman, P.A., and Case, D.A. (2004). Development and testing of a general amber force field. J. Comput. Chem. *25*, 1157-1174.

37. Wertz, I.E., and Murray, J.M. (2019). Structurally-defined deubiquitinase inhibitors provide opportunities to investigate disease mechanisms. Drug Discov. Today Technol. *31*, 109-123.
38. Ye, Y., Scheel, H., Hofmann, K., and Komander, D. (2009). Dissection of USP catalytic domains reveals five common insertion points. Mol. Biosyst. *5*, 1797-1808.

Chapter 7. Future Work

7.1 Development of GeomBD3 Brownian Dynamics Simulation Software

My plan is to expand and improve the GeomBD3 Brownian dynamics simulation program. My goal is to make the program as versatile as possible. This will entail programming additional functionality so that it is of use to all kinds of researchers. For example, future functions will include a variety of ligand starting conditions and different simulation cell shapes and boundary conditions. These options will allow users to replicate natural or engineered systems and to make better comparison to theoretical and experimental results. Another addition will be to make the simulations restart-able. This allows simulations of very large systems more practical, as they may take many days to produce converged results, in which case saving simulation data to restart later is useful – or even required – when working on remote supercomputing servers. Additionally, we will add ligand and receptor desolvation energy calculations to the software. The desolvation energy generally opposes binding of two partner molecules as they have to strip away water from each other, breaking favorable interactions. This addition should make comparisons between simulation and experiment see better agreement. Finally, another challenge facing the program is the simulation of very large systems (> 25000 nm^3). Such systems require potential energy grids that can be many tens of GB in size, which becomes impractical on most computers. To circumvent this problem, I will create a more efficient grid shape, that better follows the contours of the system, potentially producing a many-fold decrease in grid size.

7.2 Simulation Force Fields and Sampling Techniques

One major challenge in molecular mechanics-based computational chemistry is to design a force field that is highly accurate and generally applicable. This is because much of the fundamental physics is too costly to calculate, so it is replaced with effective terms that attempt to capture net effects. Thus, force field inaccuracy will always be the limiting factor of simulation reliability. I would like to devise better ways to capture the important physics, such as polarizable atoms, halogen bonds, and hydrophobic effects, which are currently lacking. A second challenge is to adequately sample all conformational states of a system in a reasonable amount of time. For example, in order to calculate accurately the free energy change of a ligand-protein binding process, or drug association/dissociation rates, full sampling of relatively long events is required. That is why I would like to continue developing and applying enhanced sampling techniques in the future. Milestoning is one method that simulates long-scale events by dividing them into many shorter trajectories. Also, machine learning shows some promise as a data-driven method of conformational generation that is worthy of further exploration.