5-2024

DEVELOPMENT OF AN ENHANCED SAMPLING WORKFLOW TO ACCELERATE MOLECULAR DOCKING WITH SPARSE BIOPHYSICAL INFORMATION

Zachary Stichter

DEVELOPMENT OF AN ENHANCED SAMPLING WORKFLOW TO ACCELERATE MOLECULAR DOCKING WITH SPARSE BIOPHYSICAL INFORMATION

A Thesis submitted in partial fulfillment
of the requirements for the degree
Master of Science

Department of Chemistry
Western Kentucky University
Bowling Green, Kentucky

By
Zachary B. Stichter

May, 2024

DEVELOPMENT OF AN ENHANCED SAMPLING WORKFLOW TO ACCELERATE MOLECULAR DOCKING WITH SPARSE BIOPHYSICAL INFORMATION

Zachary B Stichter

Date Recommended 4/9/2024

DocuSigned by:
Matthew Mc
68AB3E5491ED4DD...
Chair

DocuSigned by:
Jeremy Maddox
201D0DAAACC443F...
Committee Member

DocuSigned by:
Kevin Williams
1C82EF106C3846C...
Committee Member

Committee Member

DocuSigned by:
Jennifer Hammonds
EBE3858E068F42D...
Interim Director of the Graduate School

ABSTRACT

DEVELOPMENT OF AN ENHANCED SAMPLING WORKFLOW TO ACCELERATE MOLECULAR DOCKING WITH SPARSE BIOPHYSICAL INFORMATION

Rapid docking of flexible biological macromolecules remains a significant open challenge in protein structure determination. While rigid docking is relatively simple with toolkits such as TagDock, a key obstacle to rapid flexible docking is the complexity and roughness of the free energy surface associated with protein conformational motion (often termed the many-minima problem), meaning conventional molecular dynamics methods do not effectively sample protein conformations near the interaction complex in accessible timescales. Methods such as metadynamics and replica exchange molecular dynamics exist to ameliorate this obstacle, yet these methods use nonphysical biases or random swaps to enhance sampling. In contrast, high temperature molecular dynamics simulations using simulated annealing offer rapid sampling of a continuous trajectory, biased only by an imposed external temperature. Herein, work is performed to extend the rigid docking toolkit TagDock by implementing a simulated annealing workflow to sample protein conformational motion, extract relevant simulation frames, and perform TagDock analysis, yielding decoy structures as much as 39% closer to the target complex.

Keywords: computational chemistry; protein structure; enhanced sampling; molecular dynamics; flexible docking; method development

This thesis is dedicated to the Lord Jesus Christ, the creator and sustainer of all things.

ACKNOWLEDGEMENTS

I would like to acknowledge the support and assistance of several individuals and entities:

- Dr. Matthew Nee for his guidance and backing during the course of this research and for his willingness to oversee a project that was not in his wheelhouse.
- My committee, Dr. Jeremy Maddox and Dr. Kevin Williams, who provided guidance and support at various times throughout my journey.
- The graduate cohort who helped maintain my mental state and motivation.
- My parents and brothers who have sacrificed many things to encourage and support me over the last two years.
- WKU's High Performance Computing Cluster which allowed me to perform many more calculations than would have been otherwise accessible.
- WKU's Graduate School which provided funding for equipment and travel which allowed me to explore these subjects more deeply.
- Dr. Sarah Edwards for encouraging me to be ambitious and to attempt difficult things. It is my hope that this work fulfils some of the vision she had for this project before she passed away.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

1. Introduction

Current research in biochemistry seeks to understand human physiology by identifying and studying chemical pathways in the body. These pathways carry out biological processes such as signal transduction, gene expression, and enzyme inhibition. It is well established that problems arise when dysfunction occurs within such fundamental processes.[1–4] For example, errors in gene expression and cell regulation have been linked to many different cancers.[5–7]

Of course, most of these chemical pathways are mediated, regulated, or activated by proteins, implying that it is necessary to understand protein function to fully grasp biology. Anfinsen's classic ribonuclease reduction/oxidation experiment[8] is an early example of protein research, and it indicates a correlation between proteins' structures and their functions. In fact, Anfinsen identifies the comparative importance of distinct disulfide bonds, and he suggests the existence of "active centers" (active sites) within the ribonuclease. Newer work has cemented and expanded understanding of the protein structure-function relationship to the extent that it is considered a fundamental principle of molecular biology.[9–11]

1.1 Rational Drug Design and Protein Structure

It logically follows from the previous paragraph that understanding protein structure is key to understanding the biological pathways affecting human health. This knowledge is of particular interest to researchers working in drug discovery because detailed protein structures enable knowledge-based drug design.[12] One example of this strategy is found in the design of protease inhibitors for SARS-CoV-2 in 2021.[13] Researchers used protein structures of the main SARS-CoV-2 protease to focus and optimize a screening search, resulting in a 54-fold potency improvement over a parent compound originally identified by high throughput screening. It is notable that conventional screening methods had been unable to provide significant optimization for the drug lead. Instead, rational, structure-based design helped to identify the necessary

modifications to the inhibitor.

1.1.1 Instrumental and Algorithmic Sources for Protein Structures

The authors of that modeling study utilized protein structures from the Research Collaboratory for Structural Bioinformatics' (RCSB) Protein Data Bank (PDB), available online at rcsb.org.[14] The PDB is a key repository for solved and proposed protein structures. As of December 2023, the PDB contained more than 210,000 experimental models and 1,070,000 computed structure molecules. Of the computed models, the vast majority were proposed by the AlphaFold2 (AF2)[15] or RoseTTA fold (RF)[16] algorithms. These algorithms utilize machine learning (ML) methods to propose protein structures, as opposed to classical physics (e.g. molecular dynamics) or quantum mechanics (e.g. electronic structure) approaches to propose protein models. Accordingly, attempts to validate ML algorithms have become a major recent theme in structural biochemistry.[17–21]

These validation studies generally support the accuracy of ML-proposed structures for protein monomers. However, AF2 structures tend to be significantly less accurate for oligomeric protein complexes than for monomers,[17,19] indicating the continued need for methods to accurately predict protein quaternary structure. Two preeminent instrumental methods exist for determining that structure: x-ray crystallography[22,23] and NMR spectroscopy.[24–26] While modern advances in crystallization,[27–30] crystallography,[31–33] and structure-solving computational software[34–36] have improved the state of the art for instrumental structure determination, crystallographic methods are still time-consuming and expensive. Further, many protein complexes are transient and do not crystallize easily. This hinders progress and underscores the need for novel approaches.[23,37,38] Likewise, improvements in NMR methodology and instrumentation provide access to some large proteins and protein complexes, yet complexes that are disordered, flexible, or very large typically

remain inaccessible without the addition of complementary techniques.[26]

1.1.2 Integrative Sources for Protein Structures

Hybrid methods that integrate biochemical or instrumental information with high-performance computing (HPC) offer promising alternatives to fully-experimental methods. For example, RosettaDock[39] uses a Monte Carlo-based search algorithm to identify likely protein-protein interfaces and then optimize the backbones and side chains of the interacting monomers. The authors use a free energy scoring function to discriminate between reasonable and unfavorable docking poses. A recent version of the algorithm[40] performs admirably (77% success rate) for rigid-docking proteins, though it struggles with proteins that dock flexibly (31% success rate). In comparison, the HADDOCK toolkit[41] uses biophysical information like titrated NMR perturbation data to dock proteins. A newer revision of HADDOCK[42] performs very well (100% success rate) for easy cases, but, like RosettaDock, it struggles with flexible cases (60% success rate). It is important to note the benchmark sizes used here: RosettaDock utilized a benchmark with 13 easy cases and 32 flexible cases while HADDOCK used 14 easy and 5 flexible cases. The reduced size of the HADDOCK flexible benchmark means the 60% success rate must be accepted with caution. It is possible that the true success rate of the toolkit could be much lower if the toolkit were evaluated against a larger benchmark.

One factor contributing to these toolkits' success with rigid complexes is an energy-based scoring function. In each case, the authors use a model of intermolecular interaction energies to rank decoys. While RosettaDock attempts to model the free energy of binding using realistic all-atom force calculations, the authors do include several approximations, citing concerns about computational complexity. Likewise, HADDOCK calculates the average interaction energy between the components by summing energy contributions from electronic effects, van der Waals

interactions, and an attractive term that biases the simulation towards decoys that accurately model experimental data.

A key drawback of this type of energy-based scoring function is high computational complexity. Smith *et al.* reduced this complexity with TagDock,[37] which makes use of a distance-based method to rank decoys solely by their agreement with experimental data. This simplified approach enumerates the search space very quickly. In a benchmarking trial using the CDB3 homodimer, TagDock produced an intermediate-resolution decoy over 125 times faster than HADDOCK 2.1 and more than 220 times faster than RosettaDock 3.5. While TagDock's intermediate-resolution structures require further optimization, it is notable that the TagDock decoys were about 2Å closer to the CDB3 crystal structure than those produced by HADDOCK 2.1, demonstrating the toolkit's potential for accuracy.

1.2 Flexible Docking and Enhanced Sampling

TagDock is highly effective for complexes that interact via rigid docking, but, like the others, it struggles with flexible targets. When understood in context, these challenges highlight the continued need for improvements in the state of the art – protein-protein docking is considered "easy" for targets that have excellent templates and maintain rigidity during interactions, but toolkit performance is greatly reduced for flexible targets or those with no strong templates. This continues to be a trend across the discipline.[43] Figure 1-1 demonstrates the trend in more detail: it summarizes a recent round of the CAPRI community docking challenge,[44] and it reports submitted model quality by research group.[43] While 44 of 52 groups reported at least one high quality structure for complexes in the "easy" category, only 22 of 52 groups reported a high quality structure among the "difficult" targets. Further, the average number of acceptable-or-better structures is notably lower for "difficult" targets (20% of submitted models were of acceptable

quality or better) than for "easy" targets (75% of submitted models were of acceptable quality or better). This discrepancy between "easy" and "difficult" targets indicates a knowledge gap in the study of flexible protein interactions and demonstrates the need for tools to improve accuracy in difficult cases.

1.2.1 Modeling Protein Flexibility During Docking

Because of the recent explosion in proposed computational models and improvements in

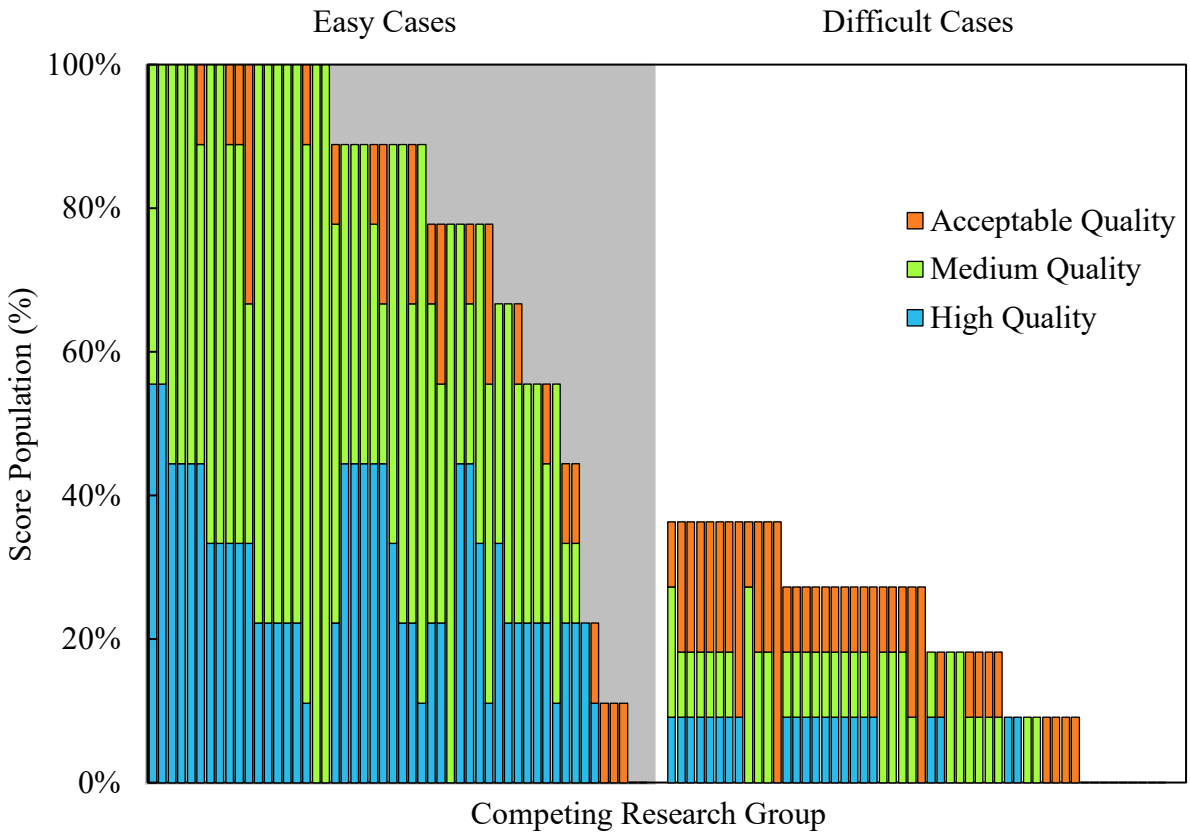


Figure 1-1: Competitive performance of docking approaches in CAPRI[44] round 46. Each column on the horizontal axis represents attempts by competing research groups to dock easy (left) or difficult (right) protein complexes. Groups attempted to dock complexes in each category and their scores are reported on the vertical axis, ranked by total structure quality. High quality structures are stacked on bottom in blue, medium quality in the middle in green, and acceptable quality on top in orange. Structures of unacceptable quality are omitted from the chart. Easy cases encompass rigid complexes with high quality template while difficult cases either lack monomer templates or undergo significant backbone rearrangement upon interaction. The reduced number of acceptable or better structures and decreased average score in the difficult category indicates a gap in current modeling practices. Adapted from ref. 43.

instrumental techniques, it seems clear that finding or creating excellent templates for molecular docking will soon become trivial. However, no comparable solution exists to rapidly model protein flexibility during docking. One promising choice is to extend TagDock to work effectively in cases such as those. The resulting intermediate models for flexible proteins can then be used as starting structures for slower, more accurate atomistic algorithms. One established approach to model this flexibility is molecular dynamics software.[45,46] Molecular dynamics (MD) uses classical force calculations to model a macromolecule or chemical system. In comparison to quantum mechanical electronic structure calculations, MD trades accuracy for computational efficiency; MD can model much larger systems than electronic structure methods, and it can model them over longer periods of time.

1.2.2 Enhanced Sampling

While these tradeoffs enable some rapid modeling of macromolecules, an in-depth understanding of protein flexibility during docking interactions is severely limited by the rarity and relative complexity of conformational transitions inherent in molecular docking. Because of the high energy barriers to coordinated motion, conventional MD simulations repeatedly sample the same "conformational space," rarely simulating novel structures.[47,48] Likewise, this highly coordinated nature exacerbates the kinetic effect of those free energy barriers, meaning that domain movements and other long-range conformational shifts typically occur in microseconds to milliseconds, as highlighted in Figure 1-2. This presents a significant challenge for MD simulations because of the sheer number of calculations necessary to model proteins.[49,50] Recent improvements in MD-specific hardware are helping to alleviate this challenge,[51] yet it is clear that algorithmic improvements to optimize sampling are necessary to enable modeling of protein docking using MD.

While a detailed review of these enhanced sampling methods is outside of the scope of this thesis, references 43 and 52 are excellent sources that review the topic extensively. It is important to note that reference 52 implies the reviewed methods are primarily for protein folding. However, sampling methods effective for protein folding may also be effective for modeling protein flexibility, as folding and flexibility are both special cases of protein dynamics.

Two modern themes among enhanced sampling protocols are Monte Carlo algorithmic modifications[53] and non-physical external biases.[54] Simulations utilizing these methods trade aspects of physical realism for algorithmic enhancement. However, this tradeoff can be risky:

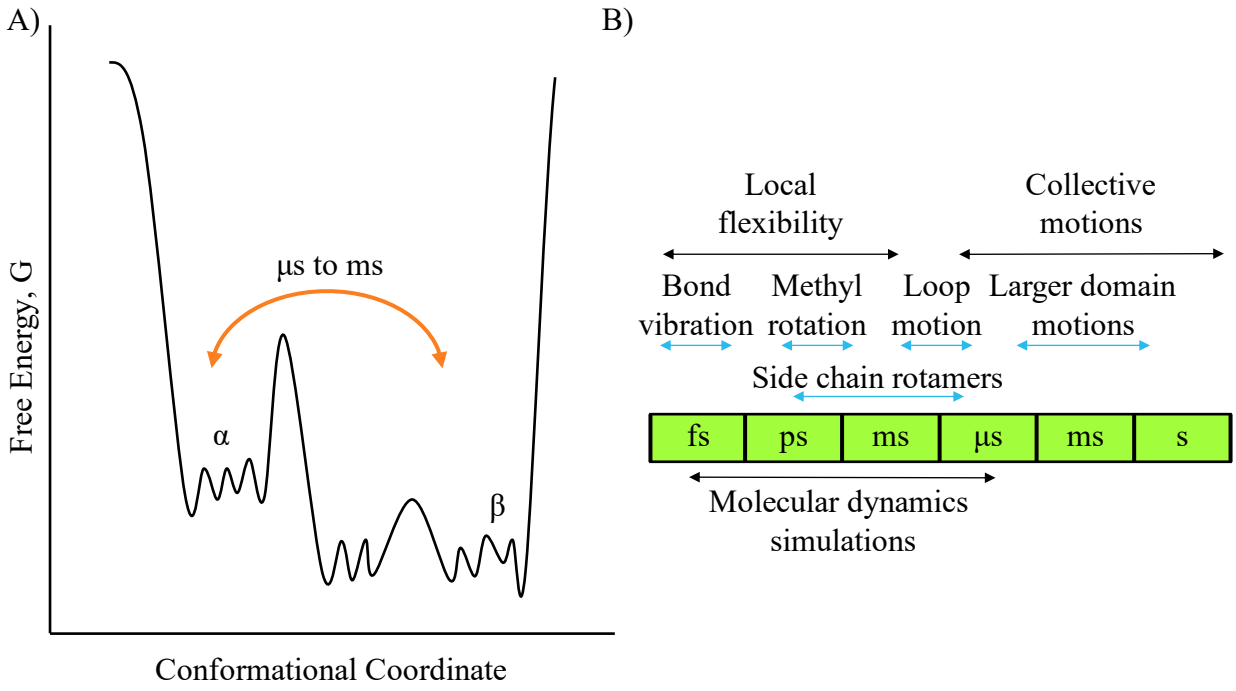


Figure 1-2: Illustration of the energy barriers to protein conformational shifts and a comparison of protein modeling time resolutions. A) A hypothetical one-dimensional protein free energy surface plotted against its spatial conformation. A coordinated transition from a local minimum in region α to the global minimum in region β is analogous to a large-scale domain motion, like the motion of flexible protein docking. *In vivo*, such a transition typically requires 10^{-6}-10^{-3} seconds. B) Comparison of time resolutions of protein motions (top) against the time scales simulated by molecular dynamics (MD). MD simulations excel at modeling local protein flexibility including methyl rotations and loop motions, but many larger domain motions are beyond the capabilities of conventional molecular dynamics, underscoring the need for enhanced simulation techniques. Adapted from reference 50.

simulations that deviate too broadly from natural physical processes risk generating protein conformations that do not have physiological relevance.

For example, replica exchange molecular dynamics (REMD),[55] sometimes called parallel tempering,[56] simulates a set of replicas at different states (traditionally, different temperatures - though some variations exchange other functions like the Hamiltonian, e.g.[57,58]). The algorithm swaps replicas between the different states randomly according to a Metropolis criterion, allowing each to access a wide swath of the conformational landscape while preserving a canonical distribution at each temperature. The approach is commonly used across protein folding and biomolecular dynamics studies,[59] and it is typically effective for proteins that spontaneously fold. One significant challenge for REMD schemes, however, is their tendency to sample proportionally to the probability of finding a protein in a given state at a given temperature. Because proteins denature as temperature increases, the probability of finding a protein in an unfolded conformation, p(unfolded), is proportional to T as described in eqn. (1).[52]

$$p(unfolded) \propto e^{\frac{1}{k_b T}} \quad (1)$$

Simulated annealing faces a similar challenge, meaning simulation temperature must be carefully controlled to avoid selecting denatured protein states. On the other hand, an advantage of simulated annealing is that the trajectory is simulated under continuous conditions instead of a series of discrete simulations, meaning the effects of changes in simulation parameters can be closely monitored and reproduced as needed.

In comparison, metadynamics[60] performs on-the-fly modifications to the free energy surface of the molecule to achieve motion along target collective variables (CVs). This modification takes the form of a Gaussian penalty, artificially increasing the free energy of a given conformation[61,62] and disfavoring commonly visited states. One advantage of the method is its

efficacy – by design, commonly visited states are disfavored, meaning simulations rapidly explore novel conformational space. However, rapid exploration does not necessarily correlate to effective sampling, as the bias potential imposed by the method can favor high energy states.[63] Further, selection of suboptimal CVs can result in ineffective sampling. As a result, it is beneficial to develop a different method that can achieve similar rapid exploration while avoiding many of these pitfalls.

1.3 Simulation Monitoring

It is useful to maintain some form of collective variable (CV) to monitor the state of the simulation. The simplest approach is to utilize the root-mean-square distance (RMSD) between some initial structure and the simulation trajectory.[64] While this enables rapid evaluation of the current state of the system (e.g. by monitoring the change from the beginning of the simulation as a method to evaluate the extent of conformational motion), the applicability of the approach is severely limited by its nature, relying on an isotropic function of distance, rather than a true evaluation of collective motion along any spatial or abstract axis. In comparison, it is often valuable to have in-depth motion information along such an axis, to facilitate data compression or anisotropic directory evaluation. One popular method to develop effective CVs is by using principal components analysis (PCA).[65–67] PCA uses statistical correlations between atoms to develop a set of CVs where motion along each principal component (PC) is a correlated set of atomic motions in three dimensions. Because each PC is an eigenvalue of the variance-covariance matrix of a trajectory, any protein motion may be constructed using a linear combination of PCs from a sufficiently large initial simulation. This property is especially valuable when evaluating conformational motion, as it is possible to select frames that contain motion along some collection of PCs, even when a target structure is unknown.

1.4 Project Goals: Development of an Enhanced Sampling Method to Accelerate Flexible Molecular Docking

Here, work has been performed to enhance conformational sampling in molecular docking simulations. The rigid docking toolkit TagDock was used as a starting point, and simulated annealing molecular dynamics was used to encourage conformational motion. Simulation parameters were optimized to improve efficacy, and principal components analysis was implemented to monitor conformational motion. Finally, the PCs were used to develop a novel metric for predicting conformational motion towards the bound complex of a structure, and the metric was utilized in an enhanced sampling simulation as a proof of concept to predict favorable simulation frames in protein 1F6M,[68,69] showing promise for a future method that might rapidly and automatically predict flexible protein complexes when given only sparse experimental data.

2. Methods and Workflow Development

To solve the problems with enhanced sampling outlined in Chapter 1, a novel sampling workflow was developed and benchmarked to determine protein quaternary structure when some sparse biophysical data is known *a priori*. The details of this computational method development project follow.

2.1 TagDock and Definition of Problem

As discussed previously, the problem of docking flexible proteins *in silico* continues to be an important open question in computational biochemistry. Hardware improvements help to accelerate simulations, but because biomolecules have many degrees of freedom, conventional algorithms are prohibitively slow. The TagDock toolkit[37] ameliorates this challenge by using a small set of *a priori* biophysical information (6-12 distances or distance distributions) to limit its scope. Instead of treating the protein as a set of N interacting atoms with 3N degrees of freedom, TagDock reduces the problem to a rigid-body orientation problem, which it solves by randomly enumerating possible orientations. Models are ranked by their respective agreement with the *a priori* distances, and top models are refined and reported to the user.

Because TagDock treats proteins as fully rigid bodies, the toolkit cannot effectively replicate flexible protein docking without some modification. Extending the tool, then, encompasses four key challenges:

- Sourcing component structures and *a priori* distances
- Sampling conformational space
- Identifying unique conformations
- Performing protein docking and optimizing native-like structures

Further, the many changes in hardware, software, and protein structure availability since

the tool was released in 2013 meant it was necessary to develop a modern structure benchmark. These were addressed in the order outlined above. Initial investigation used the toolkit's original functionality before slowly expanding the proposed method to encompass solutions to each challenge in turn.

2.2 Sourcing Component Structures and *a priori* Distances

TagDock requires all-atom three-dimensional coordinate files for each input component in PDB 3.30 data format.[70] Crystal structures or computed structure models of many proteins are available in this format from the RCSB PDB (PDB) at rcsb.org,[14,71] though in some situations it may be advantageous for a group to collect this data independently. Since the major goal of this study was to extend the existing toolkit, the original reference was mirrored by sourcing crystal structures from the ZDock benchmark.[72] While Smith et al.[37] used benchmark 4.0,[73] structures for this study were sourced from the updated benchmark 5.5.[74] Benchmark 5.5 is significantly expanded from the previous version, and 123 of 261 complexes are maintained between the two databases. These 123 retained sets of structures were the complete source reference set used in this study unless explicitly noted, and they are individually listed in Appendix A. Utilizing the same resource as the original enabled direct comparison between the results of this study and the results of the TagDock paper, allowing quick discrimination between effective and ineffective methods.

As with coordinate files, TagDock requires a set of distances or distance distributions to perform rigid docking. In practice, these could be derived from existing literature, though the toolkit advertises itself primarily to groups that collect this data experimentally. Once again, since the goal of the study was to extend TagDock, distances used in the study were identical to the ones the authors used in the original paper. These restraint sets were obtained from the supporting information of the original publication,[37] where they were initially determined by heuristically

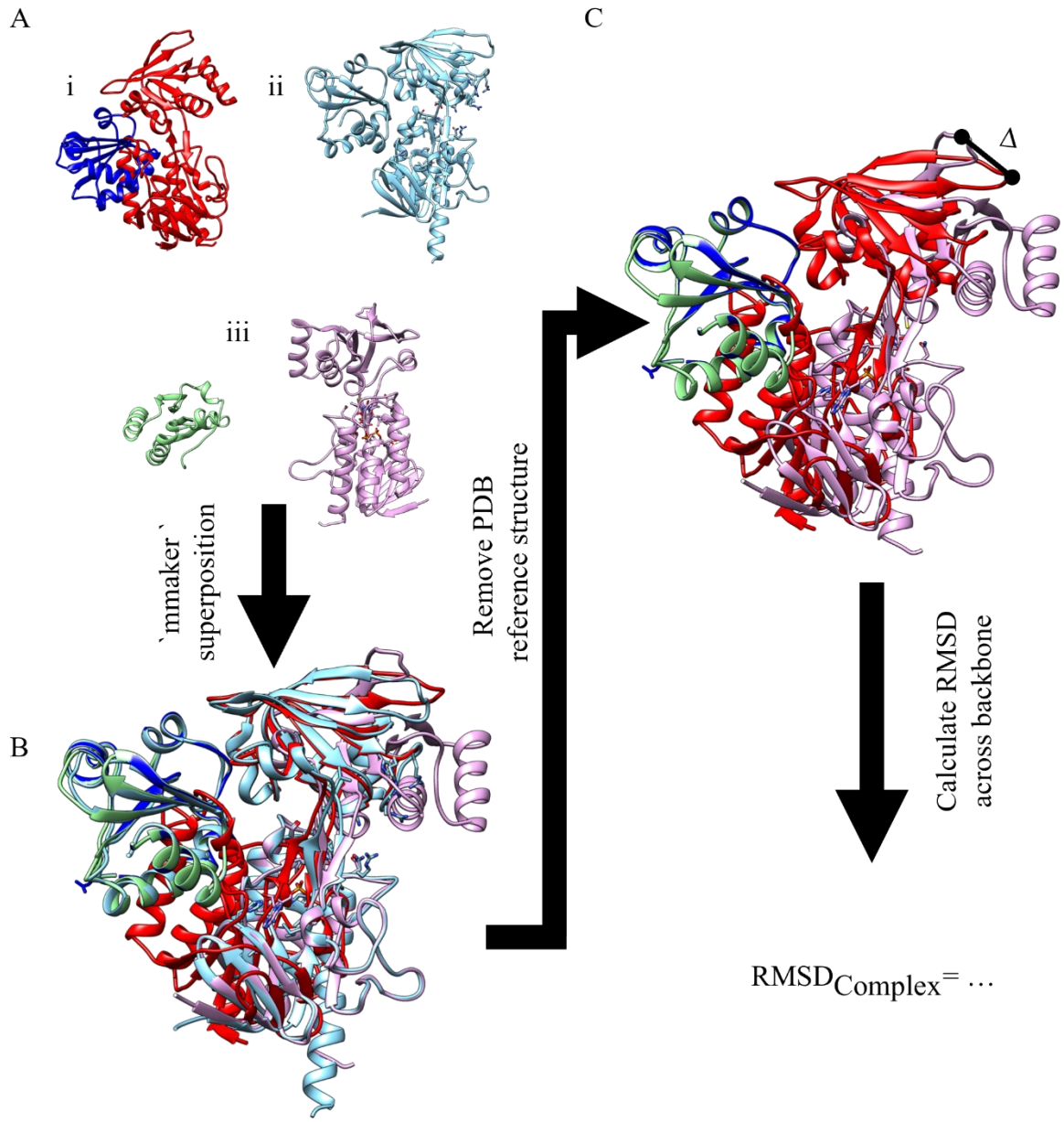
analyzing PDB entries for each benchmark complex *in silico*.

2.3 Replicating the *TagDock* Publication

Next, a replication study was performed to confirm and assess TagDock's fitness for docking protein complexes and to obtain some benchmark values to reference in further study. TagDock version 0.73 was instantiated on a high-performance computing cluster running Linux CentOS 6.10 and the job manager OpenPBS. Docking simulations were performed using the ZDock "unbound" structures, each enumerating 500,000 initial decoys and refining the best 100. Structures were refined according to the steps suggested in the TagDock user manual, available with the release version of the toolkit. This refinement first performed 100,000 low-resolution perturbations of up to 3.0 Å or 15° with a short circuit condition if no score improvement greater than 0.5 was realized within 20,000 steps, followed by 50,000 perturbations of 1.5 Å or 5°, short circuiting after 10,000 steps at a threshold of 0.1. Finally, 10,000 perturbations of 1.0 Å and 1° were performed, short circuiting after 2000 steps with no improvement.

The best-scoring decoy from each simulation was then compared to the bound structure from the PDB. For each complex, the best TagDock decoy and its corresponding undocked partners were loaded into UCSF Chimera 1.17,[75] and the molecule's accepted PDB structure was fetched electronically using the `open pdbID:` command. Solvent and ligand molecules were deleted to isolate the macromolecules, and each component was superimposed upon the PDB reference structure using the `mmaker` command. This method was used in an attempt to control for the effects of conformational change on the reported structure quality. Next, the root-mean-square deviation (RMSD) between each strand of the decoy and its superimposed, undocked component was calculated. The two resulting component RMSDs were summed, yielding a combined RMSD score for the complex. Scheme 2-1 illustrates this process. Once RMSDs were

obtained for each complex in the benchmark, the scores were normalized to 100 residues according to the method proposed by Carugo and Pondor[76] and used in Smith.[37] Notably, the work reported here deviated from Smith's method, performing further analysis with only the top scoring decoy for each complex in triplicate. In comparison, the original study averaged the RMSD across all



Scheme 2-1: Complex RMSD calculation for the TagDock replication experiment. A.i) TagDock best scoring decoy. A.ii) RCSB PDB crystal target. A.iii) Unbound starting components. A.i-iii were superimposed using the mmaker tool in Chimera 1.16 to produce B. The PDB reference was removed, and the RMSD between components of the decoy and the superimposed, unbound structures were calculated (C).

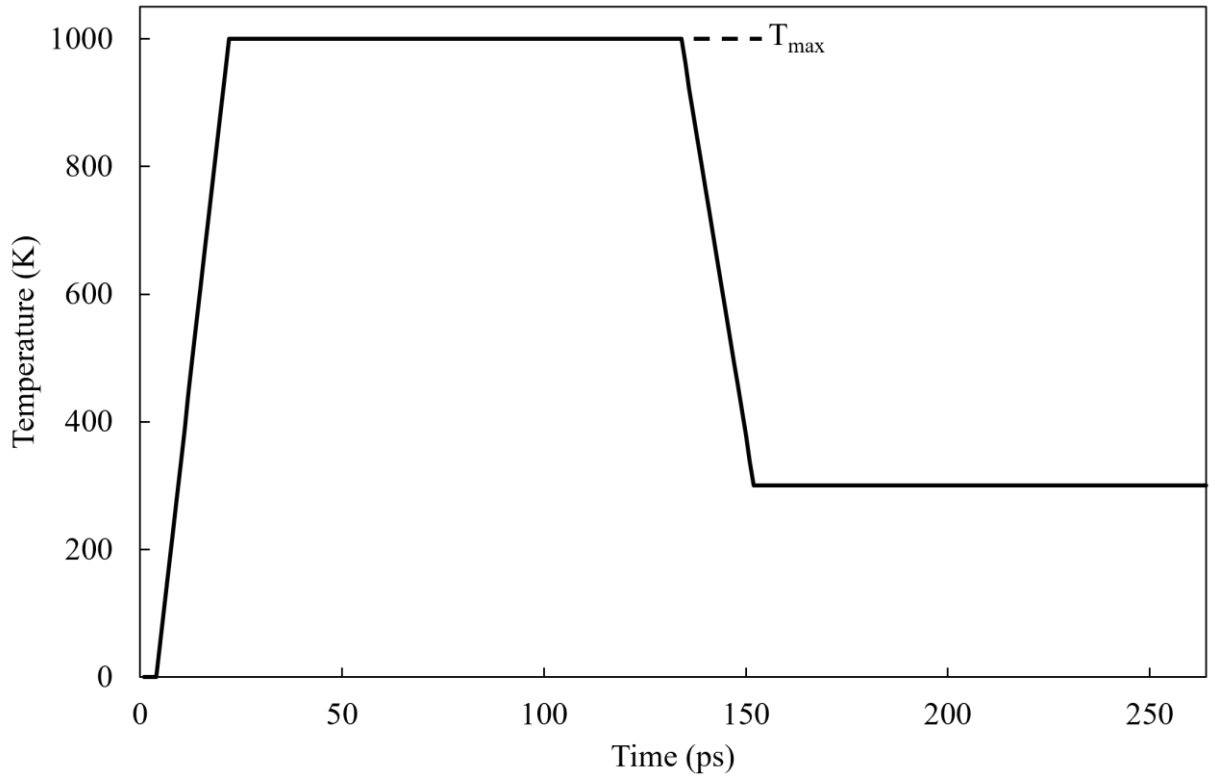
refined decoys for each complex. It was expected that while this approach might produce less consistent results than those published in Smith, any method that could effectively discriminate between two decoys of comparable TagDock score would either supersede such a tool or require some explicit knowledge of the structure these methods attempt to interrogate. Further, Smith established that TagDock decoys closely converge to an average structure in most cases, adding validity to the modification. Results of the replication study are discussed in detail in section 3-1.

2.4 Sampling Conformational Space

Once a working benchmark had been developed, it was possible to begin testing algorithms that searched conformational space. In previous work, members of our lab implemented a standard simulated annealing algorithm to accelerate protein conformational motion.[77] Initial efforts in this work sought to replicate the results of the previous study. Molecules of interest were analyzed using AMBER 14[46] installed on a high performance computing cluster running Linux CentOS 6.10 and OpenPBS. Molecules were parametrized using the ff14SB force field[78] and the AMBER tLEAP tool. Initial structures were sourced from the ZDock benchmarking database[74] as discussed in section 2.2 and minimized using 1,000 steps of steepest descent minimization, followed by 1,000 steps of conjugate gradient minimization. The system was then simulated using a Langevin thermostat,[79] approximating an NVT ensemble and enabling temperature control. The target molecule was considered *in vacuo* because the Langevin thermostat simulates atomic collisions between the molecule and a solvent. To replicate the previous work, the ZDock entry `1F6M_l_u.pdb` was heated from 0K to 1000 K over 18 ps, simulated at 1000 K for 124 ps, cooled to 300 K over 18 ps, and simulated at 300 K for 100 ps. This simulation is demonstrated in detail in Scheme 2-2, and annotated example submission scripts are included in Appendix B. All simulated annealing runs used these example scripts, with modifications as noted throughout.

RMSD analysis was performed in CPPTraj,[80] except as specifically indicated.

After initial replication trials qualitatively and quantitatively demonstrated deficiency in conformational sampling, a series of optimization experiments were performed to identify useful variables for further investigation. Simulated annealing was performed on the "r" component of 1F6M, according to the recipe in Appendix B, except as noted below. This molecule was chosen as a model system because of its highly flexible docking, poor TagDock score, and relative ease of use. The crystal structure was parametrized using the ff14SB force field and simulated in AMBER 14. A temperature optimization was performed by running a series of simulations varying the maximum temperature of the annealing protocol, T_{max} (see scheme 2-2). The temperatures considered were 300 K, 500 K, 750 K, 1000 K, 1500 K, 2000 K, 2500 K, and 3000 K. Resultant



Scheme 2-2: Simulated annealing temperature scheme as a function of time. A molecule of interest is first minimized, before being heated to an annealing temperature. The elevated temperatures allow the molecule access to more of its possible conformational space during a production run at T_{max} before being cooled down and allowed to relax at a physiological temperature.

structures were analyzed by calculating the RMSD between the trajectory and the bound complex and with TagDock, optimizing for best-scoring structures. A thermostat optimization was performed by varying the thermostat between the Langevin[79] and Berendsen[81] thermostats (ntt=3 and ntt=1, respectively). Finally, a simulation time step optimization was performed by varying the time step between 2 fs, 1 fs, and 0.5 fs (dt=0.002, dt=0.001, and dt=0.0005).

While these optimization studies were informative, qualitative analysis of the data produced in the experiments suggested a need for a more thorough investigation of protein dynamics at different temperatures. A rapid heating experiment was performed to evaluate protein denaturing dynamics as a function of temperature. "R" components of 1CGI, 1F6M, 1GPW, 1IBR, 1J2J, 1XD3, 1ZM4, 2AYO, and 2HRK were parametrized using the ff14SB force field and simulated in AMBER 14. The structures were minimized using 1000 steps of steepest descent followed by 1000 steps of conjugate gradient minimization. Structures were then heated to 3000 K over 300 ps. The Kabsch and Sander Define Secondary Structure of Proteins algorithm[82] as implemented in UCSF Chimera 1.16[75] was used to assign secondary structure, and the fraction of secondary structure residues retained was calculated as a function of temperature. Finally, the total energy of the simulation was plotted as a function of the temperature for each heating trial. This information was used to inform future decisions regarding annealing temperature. A synthesis of the temperature optimization data and the results of the secondary structure retention study suggested an optimal annealing temperature of 750 K. This is discussed in detail in section 3-2.

2.5 Evaluation of Structural Identification Techniques

The third challenge to proposing a novel workflow was identifying structures that were representative of native, native-like, or physiologically possible protein conformations (<2 Å RMSD from crystal structure). Early investigation sought to resolve this challenge by finding some

correlation between these structures and RMSD values that might be derived from *a priori* information. Simulations were initially analyzed by evaluating the RMSD between each frame of the simulation and some reference, as summarized in Table 2-1. Later, it was determined that an anisotropic metric was necessary to monitor conformational changes with some sort of directionality, and principal components analysis (PCA) was implemented to provide this metric.

PCA is an established statistical method[65,66] that correlates atomic motions in a protein trajectory and uses those correlations to reduce the complexity of a trajectory dataset. Suppose a protein structure, q, is a 3N vector of cartesian coordinates, where N is the number of atoms in the protein. A trajectory, A, is then a $q \times \frac{T}{dt}$ matrix of protein structures, where T is the simulation time and dt is the simulation time step. PCA first calculates the average structure, $\langle q \rangle = \frac{dt}{T} \sum_{dt=0}^{T} q$, and the deviation from the average for each atom at each simulation step, $\Delta A_{dt} = A_{dt} - \langle q \rangle$. Then, $\langle \Delta A \, \Delta A^T \rangle$ is the 3N×3N variance-covariance matrix. The eigenvectors of the variance-covariance are always positive, real vectors representing correlated sets of atomic motions. When sorted by decreasing eigenvalue, the first i eigenvectors describe collective motions of the protein, and the

Table 2-1: Summary of RMSD calculations and the corresponding reference.

Name	Reference	Atoms Included	Annealing Step	Available for Unknown?
RMSD1-1	First frame of annealing step	All atoms	High Temperature Production	Yes
RMSD1-2	First frame of annealing step	All atoms	Low Temperature Production	Yes
RMSDU-1	Crystal structure of unbound component	All backbone C_α	High Temperature Production	Yes
RMSDU-2	Crystal structure of unbound component	All backbone C_α	Low Temperature Production	Yes
RMSDB-1	Crystal structure of bound component	All backbone C_α	High Temperature Production	No
RMSDB-2	Crystal structure of bound component	All backbone C_α	Low Temperature Production	No

first j eigenvectors describe virtually all of the motion of a trajectory (j is typically <20% of 3N). This allows data set dimensionality reduction of ≥80% across trajectories, and it allows clear investigation of protein collective motions. Proteins 1IBR, 1CGI, 1F6M, 1J2J, 2AYO, and 2HRK were annealed as described in Appendix B, and results were analyzed with PCA to determine the method's fitness for this application.

2.6 Performing Protein Docking and Optimizing Native-Like Structures

After methods were implemented to accelerate and monitor protein conformational motion, it was necessary to optimize the workflow to select protein motions that bias the trajectory towards a physiological structure. For reasons discussed in Section 1.2.2, it was deemed valuable to limit nonphysical or arbitrary biases in any proposed method. A logical approach, then, was to model a hybrid theory of protein docking that synthesized the conformer selection and induced fit models of protein interactions. Using the tools implemented above, the first step was to perform rigid docking with TagDock. When atomic or chain overlaps existed, the chains of the decoy were translated to remove these interactions, and the refined decoy was then submitted for simulated annealing at the optimized temperature of 750 K. During the annealing process the stable "l" component was constrained using a harmonic restraint of 5.0 $\frac{\text{kcal}}{\text{mol}\cdot\text{Å}^2}$, allowing that component some flexibility while maintaining its initial conformation. The molecular ensemble was monitored using PCA and backbone RMSD, and promising structures were minimized and redocked with TagDock. Finally, all-atom RMSDs were calculated for the docked decoys to compare with initial benchmarks.

3. Results and Discussion

3.1 Replicating the *TagDock* Publication

Early work in this study created a modern TagDock benchmark by replicating the work reported by Smith et. al.[37] TagDock was used to perform molecular docking on those structures existing in common between ZDock benchmark 4.0[73] and 5.5.[74] The 100-residue normalized root-mean-square distance ($RMSD_{100}$) between each best-scoring resultant structure and its corresponding protein databank (PDB) crystal structure reference was then calculated according to scheme 2-1. The results of this benchmark were binned at 1 Å and are reported in Figure 3-1 adjacent to the results of the original study. While specific trends are not identical between the two sets of results, a few broad conclusions may be drawn. The results reported in this study show

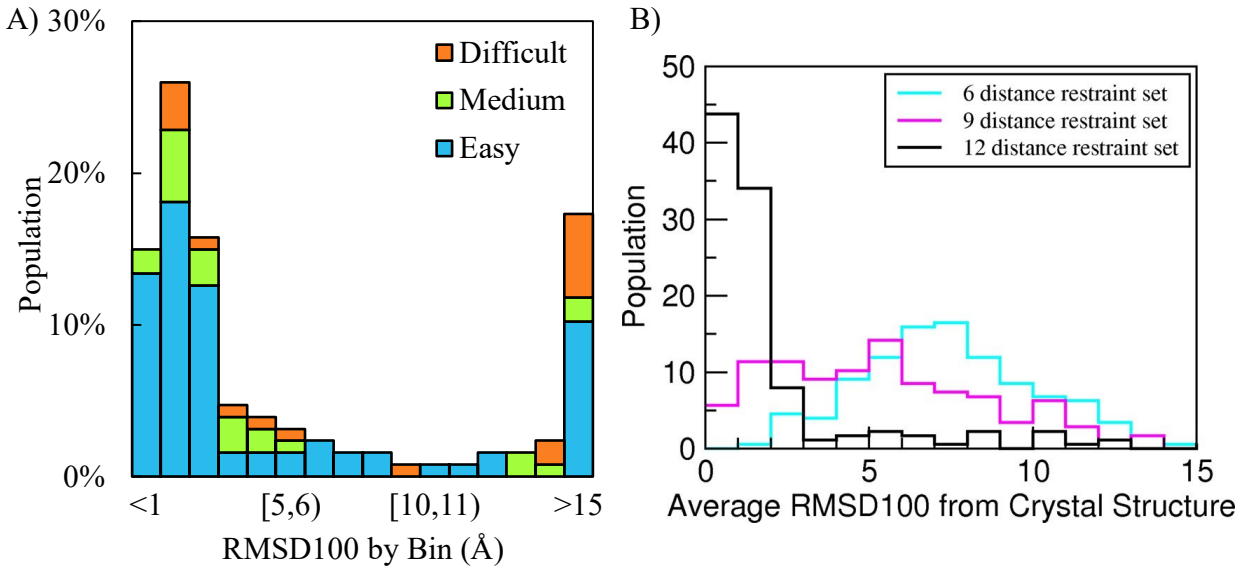


Figure 3-1: Comparison between the replicated TagDock benchmark (A) and (B) the original data as published in ref. 37. In each case, normalized root-mean-square differences (RMSD100) were calculated between the TagDock decoys and the reference complex crystal structure. These scores were binned to 1 Å and plotted as a fraction of the total set. While clear differences exist in the trends between (A) and (B), both benchmarks support that TagDock is capable of generating high quality decoys. This is evidenced by high population of scores in RMSD100 < 3 Å. Adapted with permission from Smith, J. A.; Edwards, S. J.; Moth, C. W.; Lybrand, T. P. TagDock: An Efficient Rigid Body Docking Algorithm for Oligomeric Protein Complex Model Construction and Experiment Planning. Biochemistry 2013, 52 (33), 5577–5584. Copyright 2013 American Chemical Society.

significantly worse $RMSD_{100}$ than those reported by the original authors. While the initial study reported 86% of structures with an $RMSD_{100}$ less than 3 Å for the 12-restraint set, this study reports only 56% of structures with such close matches in a similar benchmark.

Several possible sources are suspected for this inconsistency. For example, this work reports 17% of structures with an $RMSD_{100}$ greater than 15 Å, while the original reports none, suggesting some change in input or methodology. Such a change could come from different starting structures, problematic restraint files, or insufficient sampling size. Given sampling methodology and restraint files were exactly replicated from the original study while starting structures were only closely replicated, it seems reasonable that some inconsistency exists between the starting structures used in this work (ZDock benchmark 5.5) and those used in the original (from ZDock 4.0). In theory, this should be simple to examine. However, benchmark 4.0 is no longer freely available online. Further, TagDock requires custom preprocessing, meaning even if original benchmark 4.0 files could be obtained, the necessary preprocessing may have caused some meaningful change in order or structure that would only be indicated in the authors' original input files, meaning a direct comparison is no longer possible.

Another plausible reason this benchmark was less effective lies in a slight methodology change between this work and the original. Smith *et al.* reported $RMSD_{100}$ values that represented an average between all structures within two standard score deviations from the best scoring decoy. In this work, however, scores were only reported for the best scoring structure. While this change likely affected the reported scores, Smith established a tight convergence amongst the top scoring cluster, with decoys differing by less than 1 Å in over 90% of cases. Then, any difference resulting from this methodological change is likely to be small, except in a few special cases.

Alternatively, a change in methodology regarding $RMSD_{100}$ calculation could account for

the score difference. The authors did not report the method of RMSD calculation in the initial study, and it is reasonable to assume they performed a standard pairwise calculation of distances relative to the known crystal structure. In contrast, this work controlled for conformational change by orienting the best matching core region of an unbound reference structure to the PDB crystal structure and using the resulting pseudostructure as a reference, as described in section 2-3. In theory, this methodological choice should have eliminated the effects of conformational change while still demonstrating the toolkit's effectiveness in predicting docked partners when little conformational change is present. In practice, however, it is plausible this method instead biased simulations towards orientations that matched well in core regions but failed to effectively reduce overall distance to the reference structure. This is a known weakness of RMSD calculations[83] and it is supported when comparing the relative score by difficulty class. Though "difficult" structures should have benefitted most from the adjusted analysis protocol, 9 of 18 difficult cases reported scores over 10 Å, while only 22% of medium cases and 20% of easy cases scored the same.

Finally, since TagDock is a stochastic method, it is possible the parameters used did not fully enumerate the conformational space in a sufficient manner to converge, but given the substantial number of decoys considered, this seems unlikely. How, then, can this benchmark be validated? It was concluded that a direct comparison to the average $RMSD_{100}$ value as reported in Smith would be sufficient to exclude some cases, providing some modicum of validity to the remainder. In 37 cases (30%), the $RMSD_{100}$ value reported in this study was greater than 100% different than those values reported in Smith. These cases were classified as invalid, and the remaining 90 cases were accepted.

The trimmed benchmark was used to evaluate a core assumption of the early work. That is, the docking penalty score reported by TagDock was assumed to correlate to the actual RMSD

of the structures being evaluated, with respect to the unbound complex (RMSDU). Because Smith *et al.* established that decoys with very small docking penalty scores correlated to a native-like bound structure, it was determined that minimizing penalty score could be useful as a metric to approximate RMSD to a bound crystal structure (RMSDB). While this approach enables some estimation when a crystal structure is unavailable, running the TagDock analysis for every structure in a trajectory is nontrivial: in early trials, the toolkit required about 0.75 processor hours per cycle. Then, an analysis across a 1,000-frame simulation would require almost 100 computational node hours. This suggests a need for some rapidly accessible metric that can predict a trajectory's penalty score without requiring significant computational resources. The trajectory RMSDU was a likely option, and it was assumed this metric might vary in some predictable fashion with respect to the docking penalty score. TagDock penalty scores are a function of distance, being calculated using distance restraint violations, yet some initial simulations implied the correlation between penalty score and RMSDU was weak. To investigate this assumption, the docking scores of the trimmed benchmark were plotted on the horizontal axis and the corresponding RMSDU was reported on the vertical axis. This plot is reported in Figure 3-2. The trend in this data is not immediately apparent, though a Pearson correlation test (≈ 0.53) indicates a moderate correlation between docking score and RMSDU. However, these results are complicated by the nature of the penalty score. Though RMSD is an absolute, isotropic metric in cartesian space, TagDock calculates penalty scores in a one-dimensional distance space. This means solutions TagDock cannot distinguish may vary drastically in RMSDU or RMSDB, yielding significant variation in the predictive value of the docking metric (consider, e.g., points (61, 1.7) and (11, 13.3)). This highlights the need for stronger metrics to qualify structures when quaternary structure is not known.

Figure 3-3 presents an experiment to further characterize this core assumption. Each frame of an early trial was docked using TagDock, and the results of the trial were plotted to investigate the correlation between the two metrics. In panel A, the black line indicates a smoothed,

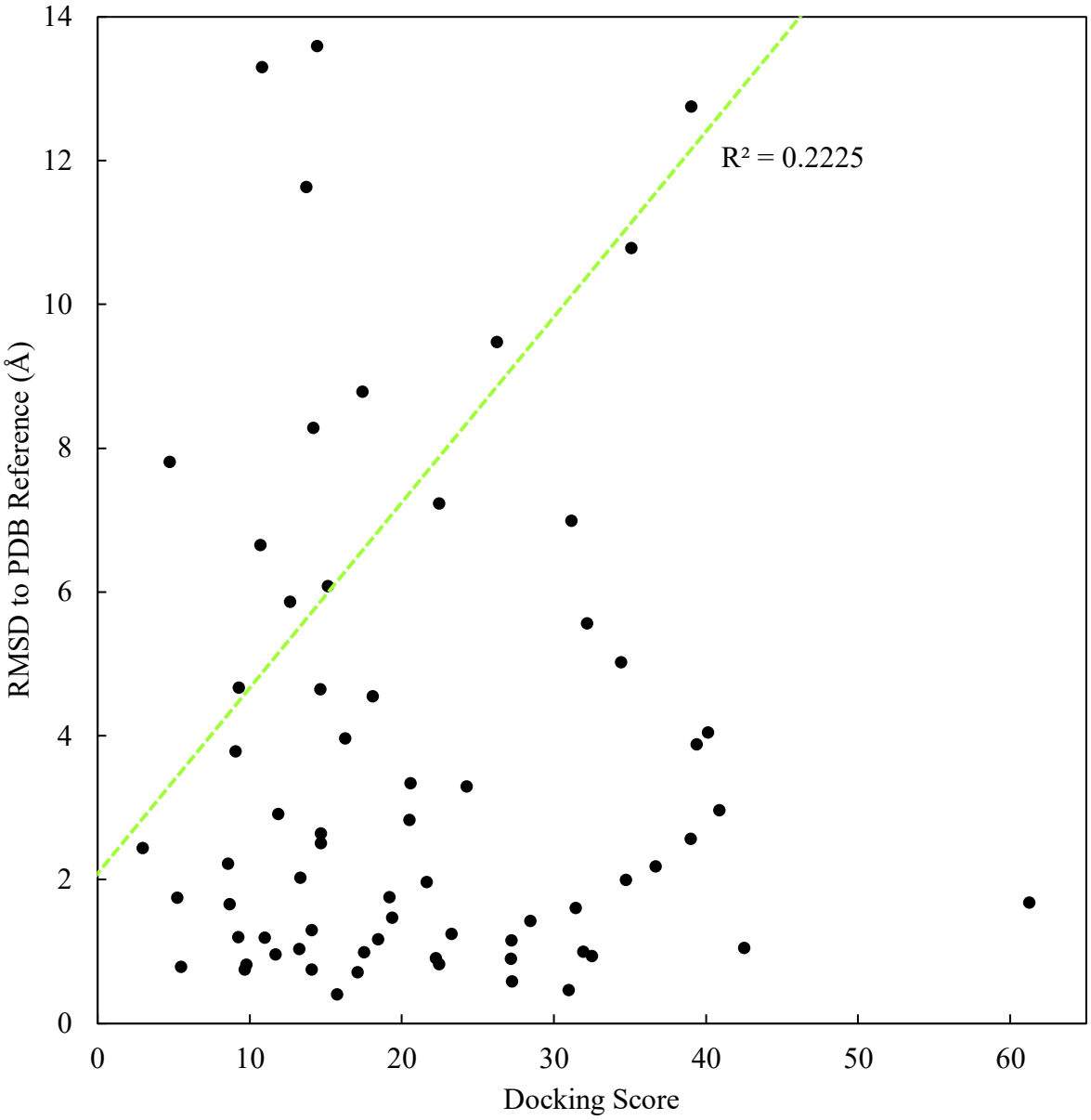


Figure 3-2: Demonstration of weak correlation between docking penalty score and RMSD to unbound reference across the ZDock benchmark. While there appears to be some correlation between TagDock score and RMSD, as evidenced by a Pearson correlation score of 0.53, the data shows no clear visual trend and is broadly dispersed across the field of the plot.

normalized docking score for each frame of a 1000-frame molecular dynamics (MD) simulation. Likewise, the green line indicates the smoothed, normalized RMSDU. These graphs appear to have

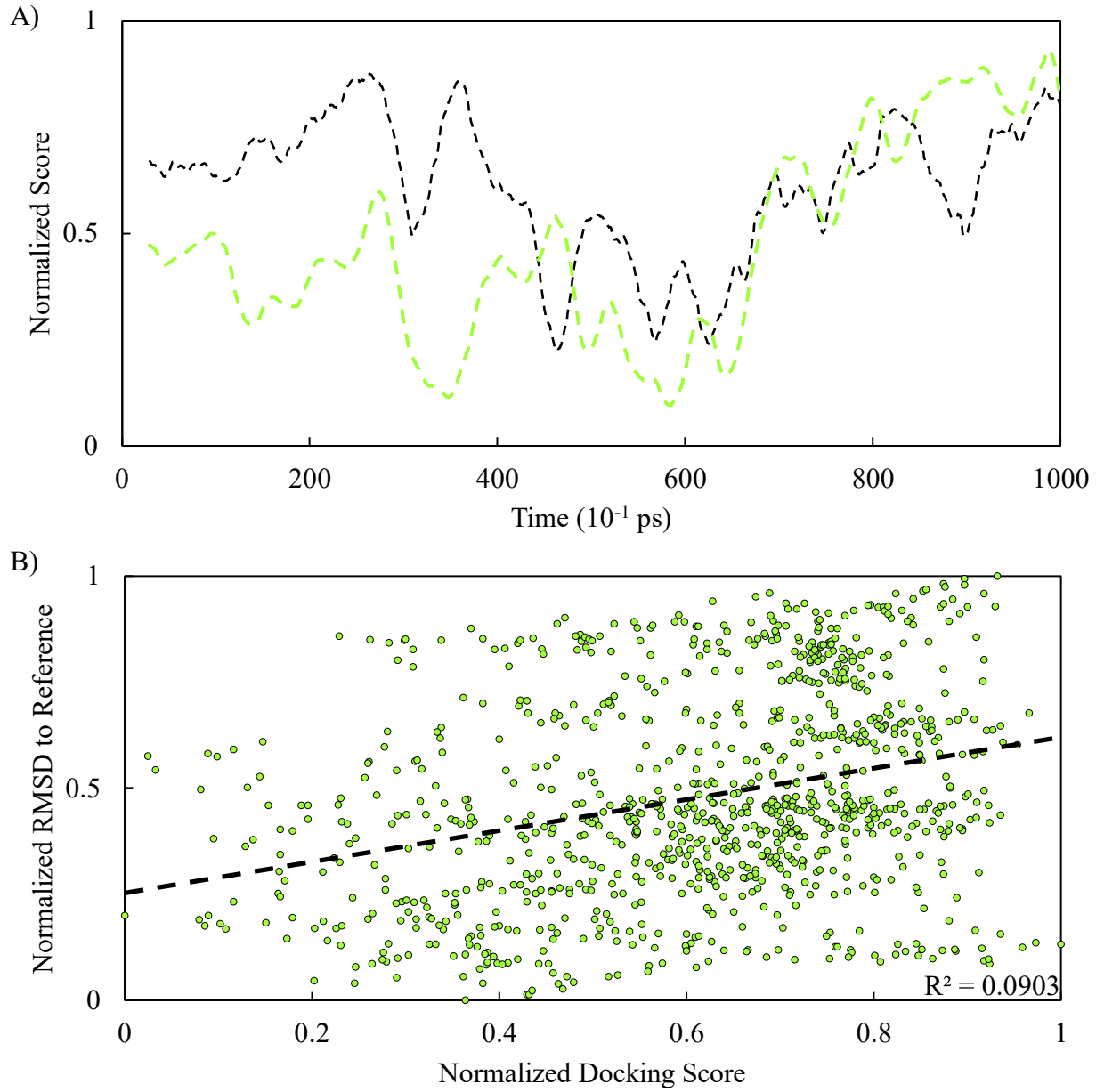


Figure 3-3: Further demonstration of weak correlation between docking penalty score and RMSD to unbound reference across a single molecular dynamics run. A) The RMSD between the trajectory and the unbound crystal structure (black) and the frame's corresponding TagDock penalty score (green) plotted as a function of time. While the metrics both trend upwards after t=60 ps, distinct differences exist in the behavior of the two curves, supporting the hypothesis of a weak correlation. B) The normalized RMSD plotted as a function of the normalized docking penalty score. While the data has a slight upward trend as docking score increases, the high variance suggests RMSD is a poor predictor of docking score.

some features in common, and when plotted directly against each other (panel B), it becomes clear that there exists a weak correlation between the two variables. A Pearson correlation value of 0.3 supports this conclusion. Notably, a marked change in the correlation occurs around frame 600, and the docking score becomes much more closely aligned to the RMSDU ($Pearson_{1\text{-}600} \approx 0.21$; $Pearson_{601\text{-}1000} \approx 0.57$). When considered in the context of the data (high penalty scores and RMSD values), the trend indicates that a high RMSDU may be used as a discriminator to separate denatured structures from candidates for further analysis.

3.2 Sampling Conformational Space

While significant value was found in characterizing the relationship between RMSDU and the complex's TagDock penalty score, the main body of the work reported here focused on the central challenge of the project: efficiently sampling conformational space. Previous work in the group had developed a recipe for simulated annealing[77] as the starting point for this sampling problem, but the results were only a marginal improvement against the results reported in Smith *et. al.*[37] It was logical, therefore, to replicate this study to confirm the results of the previous work. One production step of a simulated annealing run was extracted and the results are reported in Figure 3-4 with a direct comparison to the work performed by Gray.[77] Panel A is the replication study, and panel B indicates the results reported by Gray. The trends closely match between the two works, indicating a successful replication. In each case, the solid bottom line represents the RMSD between the trajectory and the first frame of the production cycle (RMSD1). The increasing trend in RMSD1 indicates a change in structure as the annealing step proceeds, signifying some conformational motion. Likewise, the dotted middle line represents the RMSDU and the increasing trend in this metric indicates conformational motion away from the unbound crystal structure. Because this work attempts to address the case of proteins undergoing significant backbone

conformational shift during interaction, it is expected that effective sampling in this metric should approach some $RMSD_{diff}$, defined as the backbone root-mean-square difference between the unbound and the bound structures. Such a trend, however, is not apparent in these trials, indicating a lack of convergence and ineffective sampling. The upper line, representing RSMDB, further supports this conclusion. The increasing trend in this metric opposes that of optimal sampling, which should trend to zero. In context, the successful replication of the previous work, then, served only to verify a chain of evidentiary continuity; this was used as a reference point for future work.

Assuming the classic simulated annealing algorithm was effective to enhance sampling of molecular dynamics simulations to optimize conformational motion upon protein docking, it was reasonable to perform an optimization study to interrogate changes to the annealing output as system parameters were varied. The first and easiest parameter to vary was the system heating temperature, T_{max}. The receptor component of ZDock entry 1F6M (1F6M-r) was simulated as indicated in Appendix A, varying T_{max} between 500 K, 750 K, 1000 K, 1500 K, 2000 K, 2500 K, and 3000 K. An additional set of trajectories at 300 K was utilized as a control experiment. Due to the rarity of the target events, these simulations were performed in nonuplicate. None of the trajectories converged to the bound, native-like structure, so representative trajectories from the experiment are summarized in Figure 3-5. Two parallel metrics are used to characterize the data: panel A indicates the RMSDU, indicating conformational motion away from the starting point. The trend in this data is clear: as simulation temperature increased beyond 500 K, the protein rapidly adopted novel structures, differing from the unbound component. The temperature-correlated increase in average annealed RMSDU ($t \geq 160$ ps) and maximum RMSDU support this conclusion. Likewise, panel B, indicating the RMSDB, provides further support for the conclusion that the simulation underwent increased motion. The trajectory at 750 K (T_{RMSD}) is especially

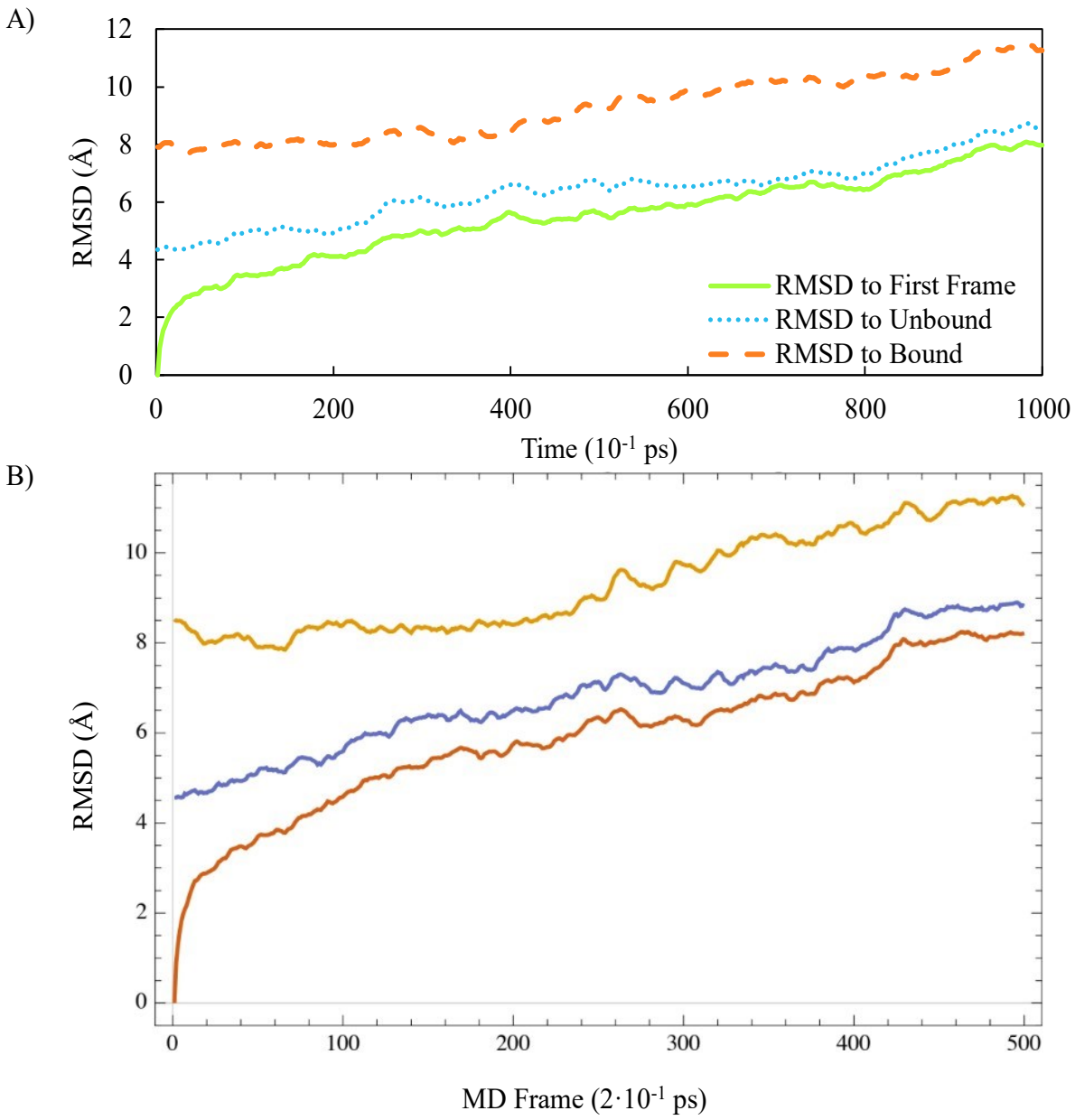


Figure 3-4: Demonstration of RMSD values as a function of time and comparison between this work and previous work in the lab. A) The replication study produced in this work. B) The original study as reported in ref. 77. In each case, the bottom line is the RMSD between the molecule and the first frame of the simulation; increased values in this metric indicate some conformational or structural motion away from the starting crystal structure. The middle line is the RMSD between the trajectory and the crystal structure of the unbound complex. A change in this metric is expected to correlate with effective sampling, though it is expected to converge to some value. The upward trend in the metric suggests protein denaturing. The upper line is the RMSD between the trajectory and the crystal structure of the bound complex. Applications for the methods proposed herein will not have access to this metric, but a decreasing trend indicates sampling towards the structure of the bound complex.

notable here. While the simulation indicates increased conformational space searching away from the unbound starting structure, the protein shows little indication of denaturing. This combination of behavior suggests an exceptionally fluid trajectory that maintains some biological relevance. In fact, T_{RMSD} outperformed the 300 K room temperature control experiment with respect to sampling towards the bound structure of the complex. This is encouraging, as it indicates sampling temperatures up to 750 K may be useful to accelerate conformational space searching. In comparison, literature values for simulated annealing schemes tend to vary broadly and do not often range above 500 K.[56,84,85] Since atomic root mean speed correlates to the square root of temperature, the 2.5-fold increase in temperature between 300 K and 750 K should correspond to a 35% increase in atomic speed in addition to the freedom of motion gained by increasing the kinetic and potential energy of the protein. If properly harnessed, these atomic speed and kinetics increases should result in a corresponding decrease in computational effort. However, it is nontrivial to make effective use of these properties of the high-temperature state. This is demonstrated by the nature of the plot of the RMSDB vs time, Figure 3-5 (B). While the trajectories at 300 K, 500 K, and 750 K accessed space that was isotropically similar to the starting structure (indicated by the constant trend in RMSDB), the structures did not converge to the bound complex, suggesting incomplete or ineffective sampling. This behavior stands opposed to the behavior expected from an effective enhanced sampling method: a protocol producing rapid, repeated convergence in RMSDB would be an excellent enhanced sampling method, and the lack of convergence in this simulation suggests that while increasing T_{max} beyond temperatures used in typical schemes might improve the breadth of accessible sampling area and accelerate conformational motion, this simple optimization is insufficient to produce convergence to some global minimum.

Another compelling bit of data can be gleaned from these same thermal simulations. When comparing the average results of the cooled RMSDU reported in Figure 3-5 (A) ($t \geq 160$ ps), the curves seem to vary with some regularity. While this behavior (denaturing as a function of T_{max}) is expected, it is less clear why structures simulated at T_{max} = 300 K and T_{max} = 500 K do not demonstrate the same behavior in the metric. Instead of denaturing as a function of temperature like the $T_{max} \geq 750$ K runs, these two trajectories trend to RMSDU ≈ 3.5. Describing the nature of this change in behavior might reasonably propose an optimal T_{max}, so an attempt was made to linearize the average RMSDU during the cooled production run (RMSDU-2; $t \geq 160$ ps). The quantity ln(RMSDU-2) was found to vary linearly with inverse temperature for $T_{max} \geq 750$ K, and this linearization is demonstrated in Figure 3-6. Figure 3-6 clearly demonstrates that above some flexible temperature, T_{flex}, trajectories vary linearly, yet below T_{flex}, the nature of the variation is far different. Interpolation of the line of fit between structures simulated above T_{flex} (750 K to 3000 K, as described above) and structures simulated below T_{flex} (300 K and 500 K), suggests T_{flex} for 1F6M is 602.5 K. While the exact cause of the change in behavior above T_{flex} is not clear from this experiment, it may be that the behavior is a function of the folded state of the protein: as the target structure degrades to a linear, unfolded chain of residues, the RMSDU will asymptotically approach the RMSD between the bound structure and the amino acid chain, matching the logarithmic behavior predicted by the linearization; no comparable behavior is immediately apparent for structures simulated below T_{flex}. For this reason, it is assumed that structures simulated above T_{flex} for long periods of time will consistently denature, while structures simulated below T_{flex} will trend to native-like.

The results of the temperature optimization made it clear that it was necessary to understand the denaturing behavior of proteins in a different manner, so an experiment was

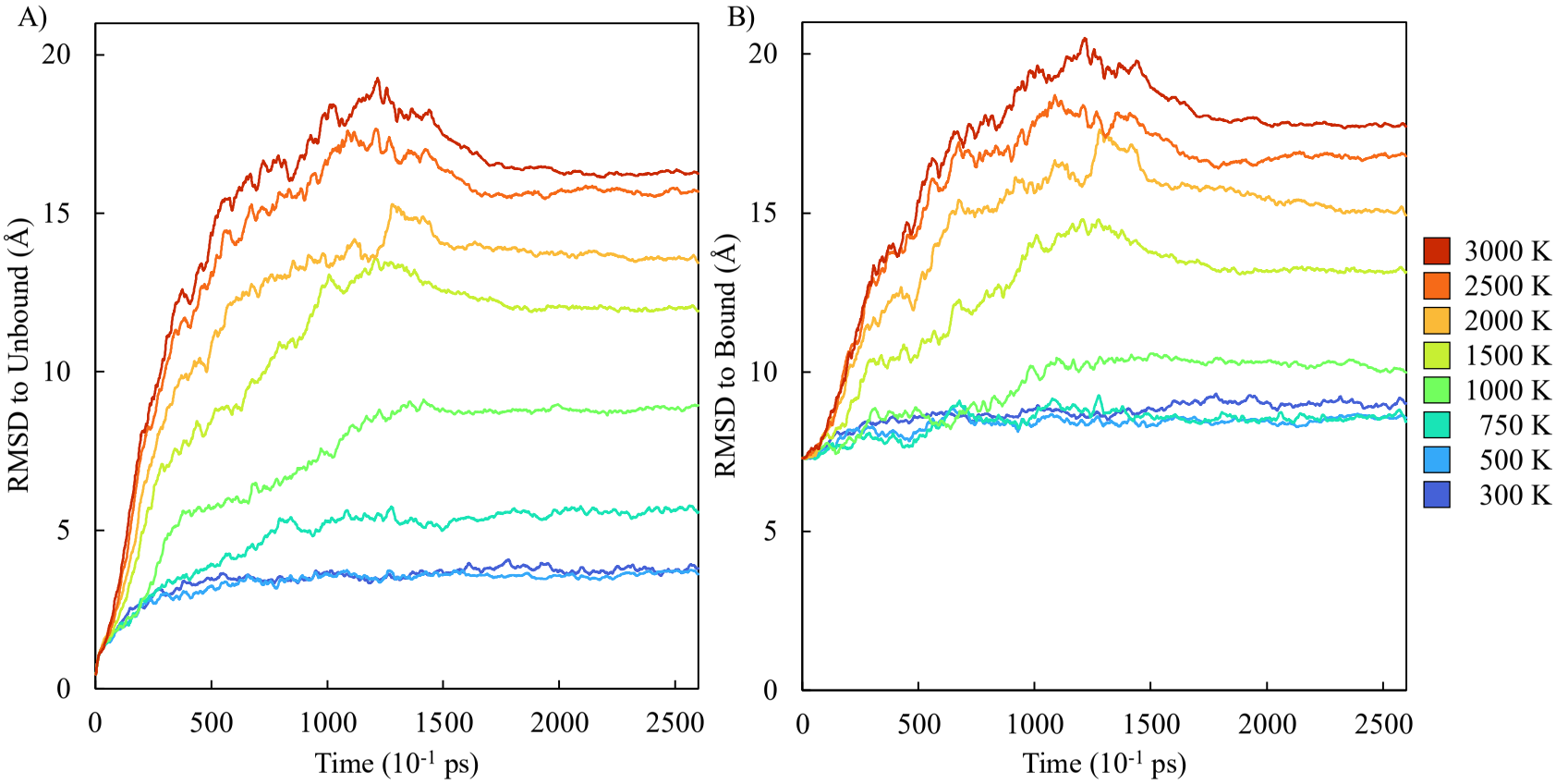


Figure 3-5: Conformational Space Searching of the receptor component of 1F6M as a function of temperature. Each line represents the RMSD between A) the crystal structure of the component in its unbound state or B) the crystal structure of the component in its bound state. The correlated increase between RMSD and T indicates protein denaturing. To avoid denaturing, it is necessary to select temperatures where there is a negative trend in panel B as a function of time or where there is a large increase in panel A with no corresponding increase in panel B. Conformational motions matching these criteria are likely to select favorable structures.

performed to quantify the denaturing process as a function of temperature. Nine different proteins from each difficulty class of the ZDock database were rapidly heated, and the fraction of secondary structure retained between the simulation and its starting crystal structure was plotted as a function of temperature, as reported in Figure 3-7. While theoretical approaches exist to model this process,[86,87] it was suspected that direct structural analysis could provide sufficient evidence to

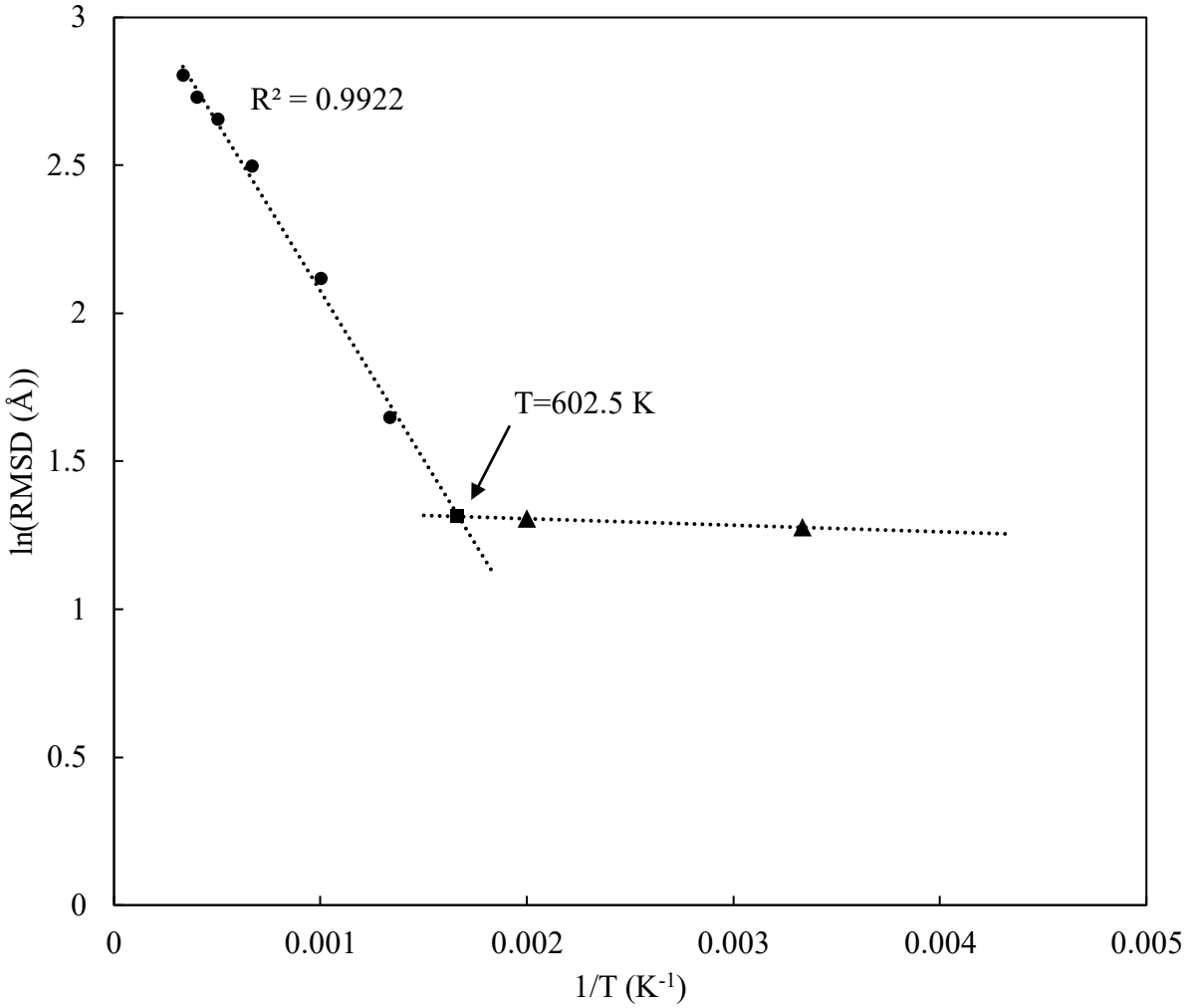


Figure 3-6: Variation of the average annealed RMSD between the trajectory and the unbound crystal structure vs inverse temperature. For structures annealed to a temperature higher than 500 K, the inverse temperature varies with the natural logarithm of the root-mean-square distance between the trajectory of the relaxed structure and the unbound crystal structure. Unexpectedly, simulations at 500 K and 300 K do not vary in the same manner, suggesting enhanced stability. Points indicated in triangles are $T \leq 500$ K, points indicated with circles are $T \geq 750$ K. The point indicated at T = 602.5 K is interpolated between the lines of fit, and it is suspected to correlate to some minimal temperature where flexible motion might first be accessed.

approximate the trajectory's status as folded or denatured at a given temperature. This might provide support for either T_{RMSD} or T_{flex} as an optimal simulation temperature. Given that a totally denatured protein has no secondary structure and that a native, globular protein has significant secondary structure, the retention of secondary structure as a function of temperature was used to indicate the folded status of the protein. It was inferred that because secondary structure elements provide stability, a trajectory maintaining its secondary structure was more likely to adopt a native-like bound structure than one lacking secondary structure. This approach also assumed that few residues gained or lost secondary structure when interconverting between conformations. The degradation visible in Figure 3-7 was modeled by numerically minimizing the sum of squared residuals to a logistic curve, eqns. (2-4):

$$S(T) = a\left(1 - \frac{1}{1 + e^{-b(T-T_0)}}\right) \tag{2}$$

$$R(T) = F_{ss} - S(T) \tag{3}$$

$$RSS = \sum\left(R(t)\right)^2 \tag{4}$$

where a and b are vertical and horizontal scaling factors, respectively, T is the Langevin temperature, T_0 is a horizontal translation factor indicating the inflection point of the curve, and F_{ss} is the fraction of secondary structure predicted by the trajectory at a given temperature. The curve can be utilized to approximate three stages of thermal denaturation. The first is a slow decrease where the structure might be considered native-like. This region ends at approximately 615 K, as early reductions in secondary structure decrease the heat capacity of each molecule,[88] accelerating the loss of secondary structure. Finally, beyond about 1200 K, the protein can be considered totally denatured, only adopting a few secondary structure residues by random association. This experiment, then, is consistent with selection of either T_{RMSD} or T_{flex} as optimal annealing temperatures for performing conformational space search for 1F6M. While it is useful

to maximize temperature to access novel regions and to accelerate atomic motion, it is also necessary to produce structures that might reasonably be part of the physiological conformational ensemble. Because of the many local energetic minima available to totally denatured proteins (see section 1.2.2), trajectories completely lacking secondary structure are less likely to adopt physiological conformations than trajectories that retain a larger degree of their original secondary structure. This is because the retained secondary structure helps to create energetic funnels[89,90] that favor physiological states. Then, it is implied that one must necessarily strike a balance between the thermal simulation bias prescribed herein and the practical consideration of the biological relevance of the resulting structure.

This implication may be intuited in another manner: suppose a protein may be crudely modeled in three dimensions by a two-variable free energy function. The protein may first be considered in the unbound state, a local energetic minimum. Alternate states, including the bound global energetic minimum, may be accessible by overcoming some energy barrier. While conventional molecular dynamics simulations may ultimately achieve the target motion, the high energetic barriers mean any observed transitions will be slow, as discussed in section 1.2.2. Then, a large increase in thermal energy enables conformational motion towards nearby local minima, but it does not direct or distinguish between those minima, as the elevated temperature enables motion to many potential states, meaning motion towards denatured states is far more probable than motion towards native-like states. However, if it is assumed that a physiological conformational transition path follows an optimal, low-energy path from the unbound to the bound state in the presence of a docking partner, it becomes clear that thermal tuning of the free energy surface might selectively facilitate target conformational transitions *a priori,* reducing the likelihood of conformational transitions towards denatured states.

This hypothesis prompts another review of the degradation data, this time through the lens of protein energetics. If some clear relationship exists between temperature and protein potential

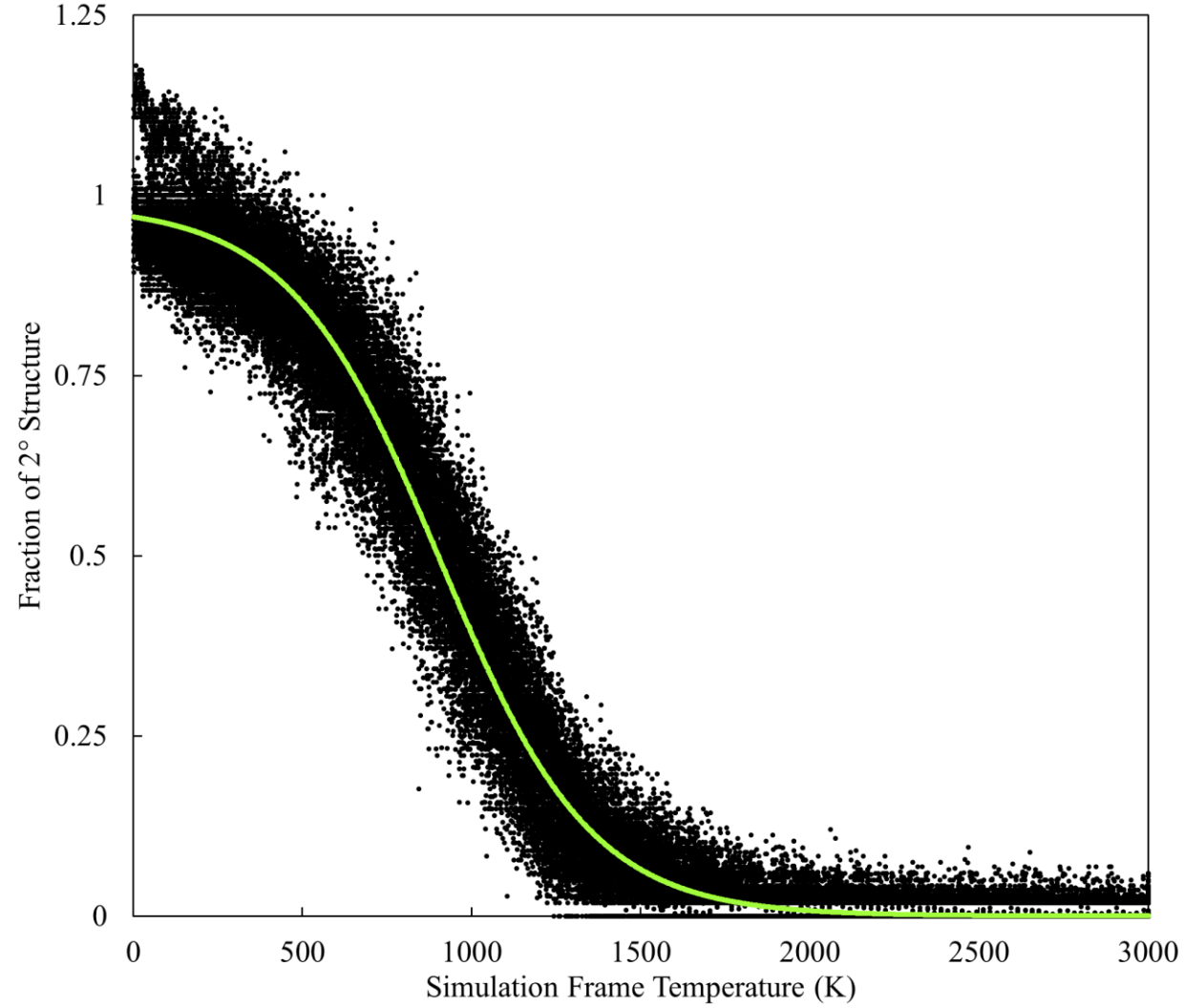


Figure 3-7: Thermal degradation of secondary structure across a subset of the ZDock database. The vertical axis indicates the fraction of residues in secondary structure (as defined by the DSSP algorithm) for each step in the trajectory with respect to the number of secondary structure residues in the crystal structure of the unbound component. The horizontal axis indicates the simulation frame temperature, as calculated by the Langevin thermostat. Each dot represents a simulation frame from either 1CGI-R, 1F6M-R, 1GPW-R, 1IBR-R, 1J2J-R, 1XD3-R, 1ZM4-R, 2AYO-R, or 2HRK-R. The curve of fit is a sigmoid of the form indicated in eqn. (2), where T is the system temperature, T_0 is some inflection temperature, and a and b are vertical and horizontal scaling factors. Values above 1.0 indicate a protein adopting greater secondary structure when minimized than was present in the PDB coordinates file. The decreasing trend indicates thermal degradation of secondary structure that varies with temperature. This was used as a qualitative measure of protein denaturation, supporting the assignment of 750 K as a reasonable annealing temperature for further study.

energy at the interface between the first and second regions indicated by the secondary structure retention experiment (around 600 K), then such a relationship might suggest the appropriate temperature for thermal tuning, implying an optimal T_{max}. However, when protein free energy is plotted against simulation temperature, it becomes clear that optimizing T_{max} through energetics is more complex than the secondary structure retention method implied. This is likely because the kinetics of the retention simulation act to limit the rate at which the proteins denature. Figure 3-8, panel A plots the actual AMBER molecular potential energy, E_{ptot}, as a function of simulation temperature. For simplicity, panel B presents a linear fit of the same data. Because the simulated proteins vary widely in heat capacity and initial potential energy, the slopes (a function of heat capacity) and intercepts (a function of minimum potential energy) vary, meaning no single annealing temperature is optimal for all proteins. Instead, this data suggests it is necessary to tune the annealing temperature for each intended application. Since a protein with low potential energy exists in an energetic funnel, the optimal simulation temperature projected by energetics, T_E, should be some temperature where the total potential energy of the molecule, E_{ptot}, is zero or positive. To minimize the number of available states and increase the likelihood of sampling towards the bound state, T_E should be minimized, suggesting T_E should be the temperature where $E_{ptot} = 0$. Such a temperature provides thermodynamic impetus for the conformational shift in a manner emulating the established metadynamics simulation technique,[60] except that the simulation is biased by thermal tuning from the Langevin thermostat rather than by a Gaussian penalty while maximizing the probability of sampling towards the target. Figure 3-8, panel C reports relevant temperatures correlating to $E_{ptot} = 0$ for each trial of the secondary structure retention experiment, determined by calculating the x-intercept of the line of fit. Of particular interest is the trial for 1F6M, as it has a direct comparison to the thermal optimization data reported in Figures 3-5 and

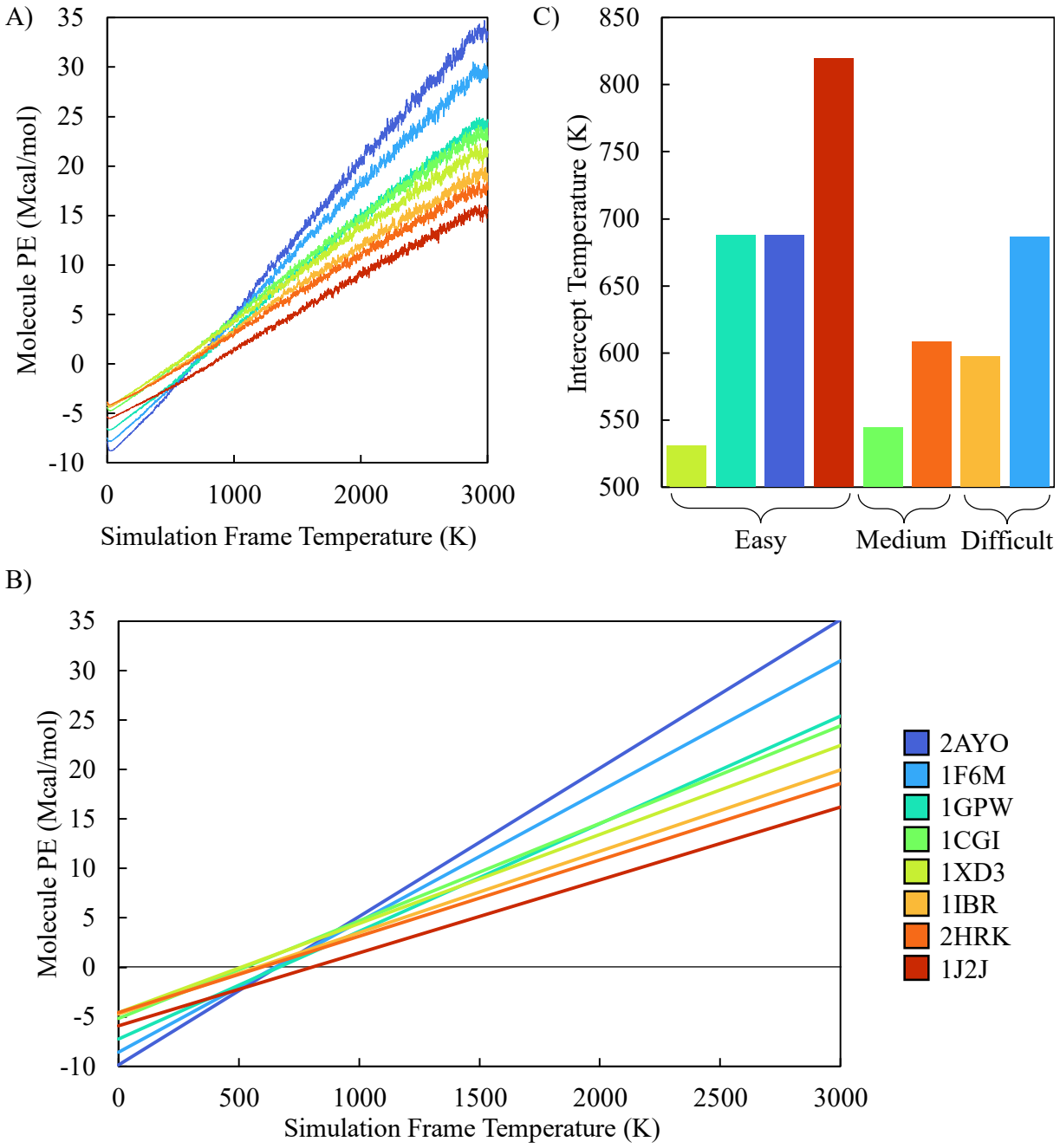


Figure 3-8: Plots of AMBER molecular potential energy (PE) as a function of temperature. A) Molecular dynamics potential energy as a function of frame temperature, demonstrating the effect of thermal modification on the energy states available to a protein. Increased potential energy allows access to less favorable conformational states. B) Trend lines of panel A, indicating the temperature at which the potential energy goes from negative to positive. It was hypothesized that proteins simulated above PE = 0 would trend towards denatured. C) The intercept temperature indicated for each protein. Note the vertical axis starts at T = 500 K to maximize vertical resolution.

3-6. Notably, the reported temperature of 686 K is higher than T_{flex} (602.5 K), suggesting that protein trajectories begin to access fluid conformational space before reaching $E_{ptot} = 0$, though further efforts are necessary to characterize and understand the nature of this relationship.

In summary, two different experiments were performed to interrogate the optimal temperature for performing simulated annealing on 1F6M, yielding disparate results as summarized in Table 3-1. Because of the disparity of suggested temperatures between the approaches, the annealing temperature for future trials was determined based on a conservative interpretation of the direct thermal optimization. Since T_{RMSD} enhanced conformational motion relative to the starting structure while maintaining similar structural integrity to that of the lower temperature simulations (see Figure 3-5), T_{max} was set at 750 K for future trials.

Though temperature was the most thoroughly optimized parameter, trials were also performed varying the temperature control between Langevin[79] and Berendsen[81] thermostats. While an optimized isokinetic Nosé-Hoover chain thermostat[91] may present another well-suited candidate for thermostat optimization, it was not implemented in the AMBER package until the 2016 release, two releases after the software used in this research. The Andersen thermostat[92] was excluded from the optimization because it modifies individual particle motions randomly, meaning a macromolecule simulated using the Anderson scheme would not experience even temperature regulation. This makes the scheme ill-suited for simulations of proteins. While thermostat options were somewhat limited, more control was available over the simulation time step, which was

Table 3-1: Comparison of disparate temperatures considered for T_{max} and the methods used to find those temperatures.

Name	Temperature (K)	Method
T_{RMSD}	750	Direct inspection of RMSD curves as a function of temperature
T_{flex}	603	Interpolation between lines of fit for the two linear regions of Figure 3 -6
T_E	686	Temperature corresponding to simulation $E_{ptot} = 0$

varied between 2 fs, 1 fs, and 0.5 fs. The results of these two optimization studies are reported in Figure 3-9. Panel A reports the results of the thermostat optimization, and panel B reports the results of the time step optimization. The trends from these simulations are clear: the time step optimization has little effect on the quality of the simulation, provided dt ≥ 0.001. Since this had no substantial effect on the quality of the simulation, dt = 0.001 was chosen as the time step for enhanced temporal resolution and because this value is recommended in the AMBER documentation.

The thermostat optimization, however, had slightly different results. In the trajectory for the Berendsen thermostat, the trajectory destabilizes and denatures the structure early in the simulation before rubber banding back around 11.5 ps. This unusual behavior suggests that the Berendsen thermostat is an inferior thermostat for this purpose, and the Langevin thermostat was

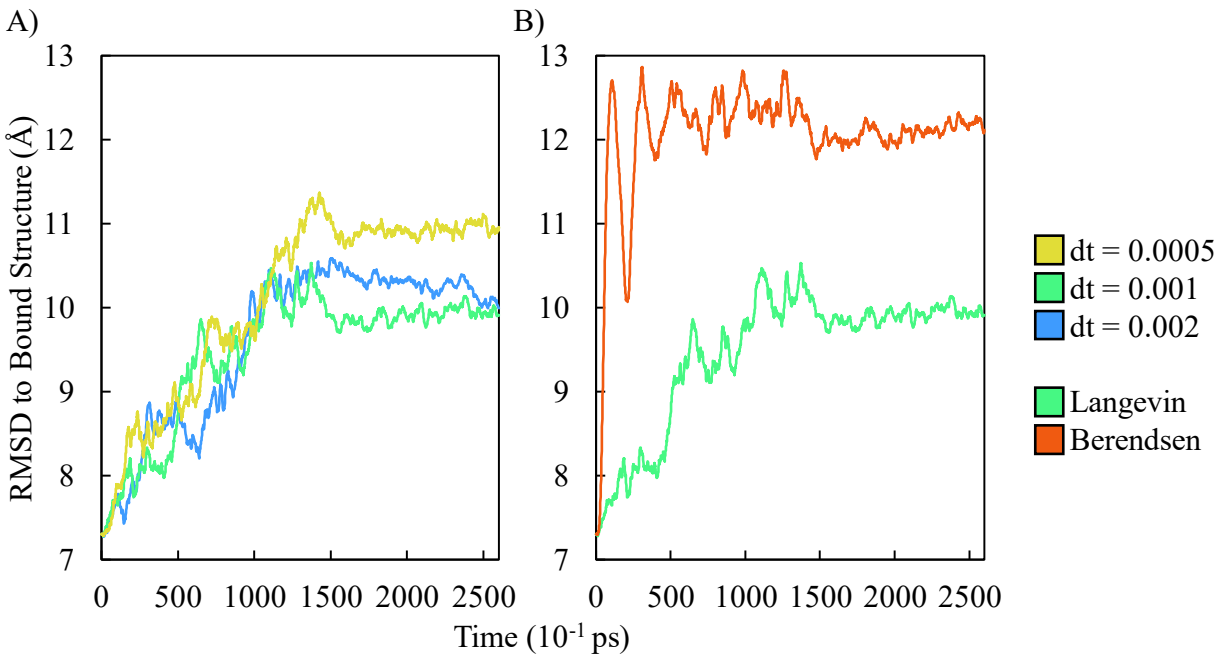


Figure 3-9: Graphs of RMSD vs time for time step (A) and thermostat (B). The lack of a broad disparity in (A) indicates that any of the proposed time steps might reasonably be used for simulations. The time step dt = 0.001 was used because it was recommended in the AMBER manual and to balance resolution with efficiency. The rapid spike-recoil pattern at t=150-350 in the Berendsen thermostat in panel B suggests the method is less effective for simulated annealing, so the Langevin thermostat was chosen for future runs.

maintained as a superior choice for this application. This is expected because the Berendsen thermostat performs only a velocity rescaling step as opposed to the atomistic acceleration and damping of the Langevin thermostat, meaning velocity changes unevenly through the macromolecule.

3.3 Implementation and Evaluation of Principal Components Analysis

Once some optimization studies had been performed, it was necessary to implement an anisotropic metric for evaluating trajectory motion. While RMSD between the trajectory and the bound structure of the complex (RMSDB) is an excellent metric for evaluating conformational motion between the states, this metric requires *a priori,* atomistic knowledge of the bound structure of the protein multimer. The intended use case for this method, however, is for situations where only limited knowledge of the multimer is available (e.g., only several point-to-point distances between residues). Then, this use case is mutually exclusive with usage of RMSDB. Likewise, as discussed in section 3.1, it was found that RMSD to the unbound structure (RMSDU) does not strongly correlate to any "correct" structure. This means in the absence of some additional metric, it will not be possible to evaluate conformational motion of a trajectory in any predictive manner.

One generally accepted method to overcome this challenge is to use principal components analysis (PCA) to reduce the dimensionality of atomistic trajectories by evaluating the covariance between correlated atomic positional fluctuations in a trajectory.[65–67] In a sense, PCA can be considered a descriptive, statistical relative of normal mode analysis (NMA), using observed simulated positions instead of forces to calculate eigenvectors representing correlated motions. Kitao[65] contains a short, informative section further describing the relationship between PCA and NMA.

A key value of PCA is its descriptive nature. Since the method uses observed trajectories,

any sufficiently large trajectory should indicate all relevant classes of physiological conformational motion. However, many of these motions will be random or insignificant; those insignificant motions will have a miniscule eigenvalue and may be neglected. For most simulations, much of the trajectory may be described with approximately 100 principal components (PCs). An example best indicates the power of the method: in a protein with 300 residues, backbone PCA can provide a dimensionality reduction of over 90% with less than 5% loss of precision. Further, the method is especially valuable in this case because each PC, extracted solely from the observed trajectory, is anisotropic. Because conformational motion towards a target structure will move along a linear combination of these PCs, possible transition paths can be identified by tracing conformational motion towards PC extrema, like in the frontier expansion sampling method proposed by Zhang and Gong.[93] Other relevant implications of this method will be discussed in detail in the next section.

To validate PCA as a viable method for tracking protein conformational motion at different temperatures, an experiment was performed varying the temperature between 300 K and T_{max}. The trajectories of the simulations were processed according to the principal components analysis method in section 2.5, and the first two principal components were normalized and plotted against each other. This creates a two-dimensional scatter plot representing relative motions of the trajectories along the two lowest frequency correlated sets of atomic fluctuations. An example of one such experiment to implement PCA is reported in Figure 3-10. The horizontal axis reports normalized motion along the first principal coordinate, and the vertical axis reports normalized motion along the second principal coordinate. Due to the nature and complexity of the data set, the precise locations of points or lines on this plot may be considered arbitrary; instead, line density at a given point represents a much more informative metric for evaluating conformational motion.

The black, solid lines, representing the 300 K simulation, are closely packed and often revisit locations on the plot. This is indicative of conformational equilibrium around some average structure, matching the expected motion of a simulated protein at physiological temperatures. The red, dashed lines, however, have a much lower line density and cover a much larger region of the graph. These features of the 750 K data set clearly indicate enhanced conformational motion. Further, conformational motion towards the target structure (green diamond, (0.33, 1)) is readily apparent in one trajectory at high temperature, and it is found near a conformational extreme, implying the viability of frontier expansion sampling for further analysis cycles in this case. Other PCA implementation and validation experiments were performed for 1CGI, 1IBR, 1J2J, 1XD3, 2AYO, and 2HRK, and the results of those experiments demonstrated substantially similar trends, namely increased sampling breadth at elevated simulation temperature.

3.4 Complete Method Implementation and Proof of Concept

Finally, an experiment was performed that combined the three key concepts implemented in the previous portions of this work. First, TagDock was used to propose a docking decoy for 1F6M. The decoy was checked for chain or atom conflicts, and upon finding an overlap, the chains were translated by the minimum distance necessary to separate the components and stabilize simulation energies. The translated decoy was annealed at 750 K as described in Appendix B. Principal components analysis was performed on the trajectory, and the first 100 PCs were analyzed for correlation to the RMSDB. These first 100 components were chosen as a set to describe 96.00% of the motions of the protein while reducing the size of the data set by 89.45%. The eigenvalue accumulation data for this set is reported in Figure 3-11, where the horizontal axis indicates each ν_i and the vertical axis reports the accumulated contribution of the eigenvalues as described in eqn. (5):

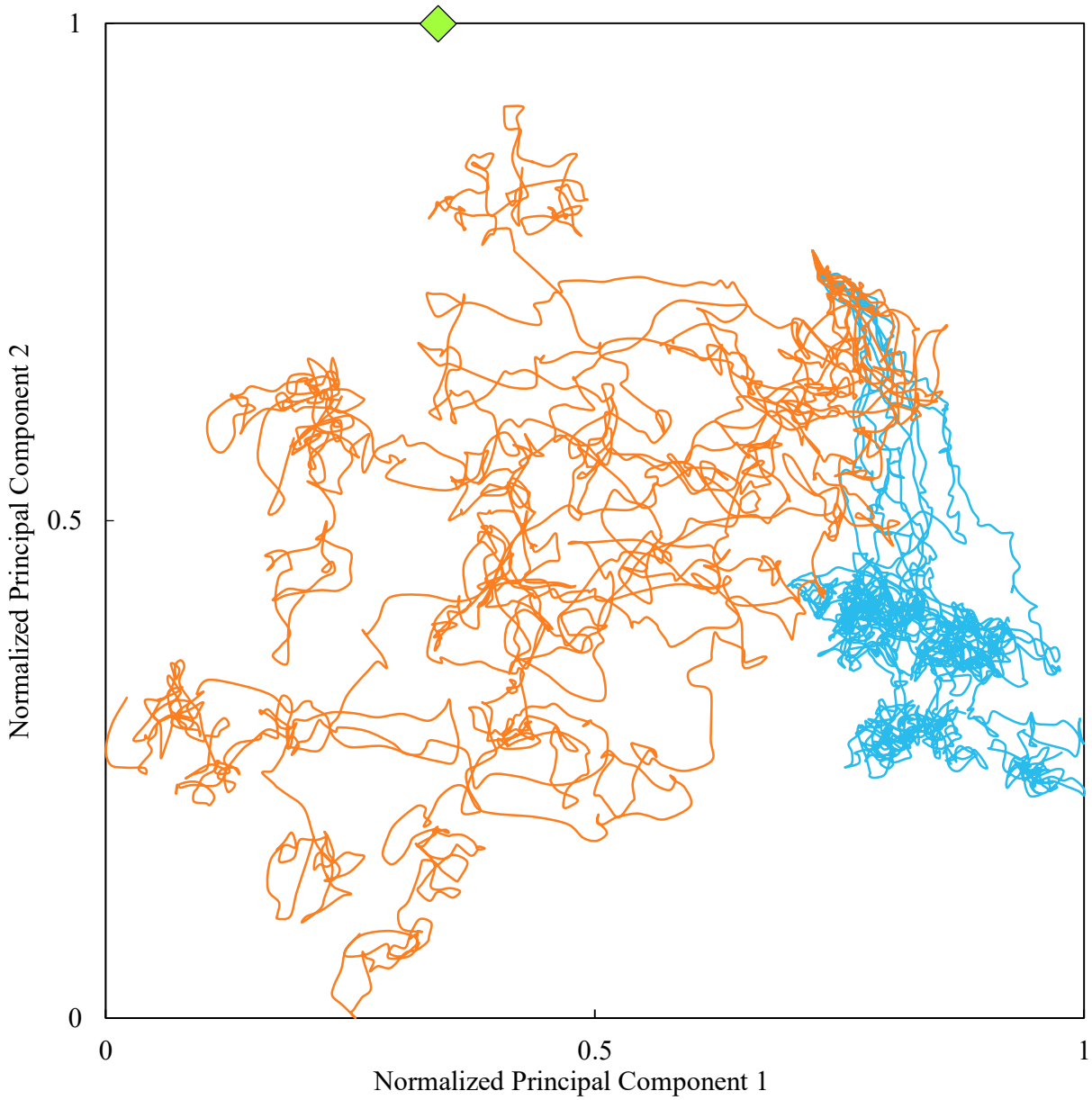


Figure 3-10: Conformational motion of 1F6M along the first and second lowest frequency motions (principal components 1 and 2, respectively) at different temperatures. Axes are normalized for clarity. The line density indicates sampling of a region. The blue lines represent trials at T=300 K, while the orange lines indicate trials at T=750 K. The green diamond is the projection of the target structure along the selected principal components. The lower temperature trials demonstrate tight line density, indicating repeated sampling of a few regions. Meanwhile, the orange lines demonstrate a much lower density, indicating increased sampling along the slowest motions and suggesting increased sampling along all motions. This result is expected and suggests principal components analysis is a reasonable method to interpret simulated annealing trajectories.

$$Accumulated\ Contribution = \frac{\sum_{i=1}^{n} \lambda_i}{\sum \lambda} \tag{5}$$

To find a correlation between the principal components and the bound structure, a heuristic was developed. First, the initial 40 ps of the annealing simulation were neglected to account for thermalization and minor rearrangements not reflecting the high-temperature conformational motion of the protein. Next, it was assumed that desirable conformational transitions would be simultaneously located at the extremes of several PCs. Therefore, the PCs were rescaled to the range [0,1] at each simulation frame according to eqn. (6) and then transformed according to eqn. (7):

$$M_{norm}(t) = \frac{\mathrm{PC_{i,t}} - \min(\mathrm{PC_i})}{\max(\mathrm{PC_i}) - \min(\mathrm{PC_i})} \tag{6}$$

$$M_i(t) = 2\left|M_{norm}(t) - \frac{1}{2}\right| \tag{7}$$

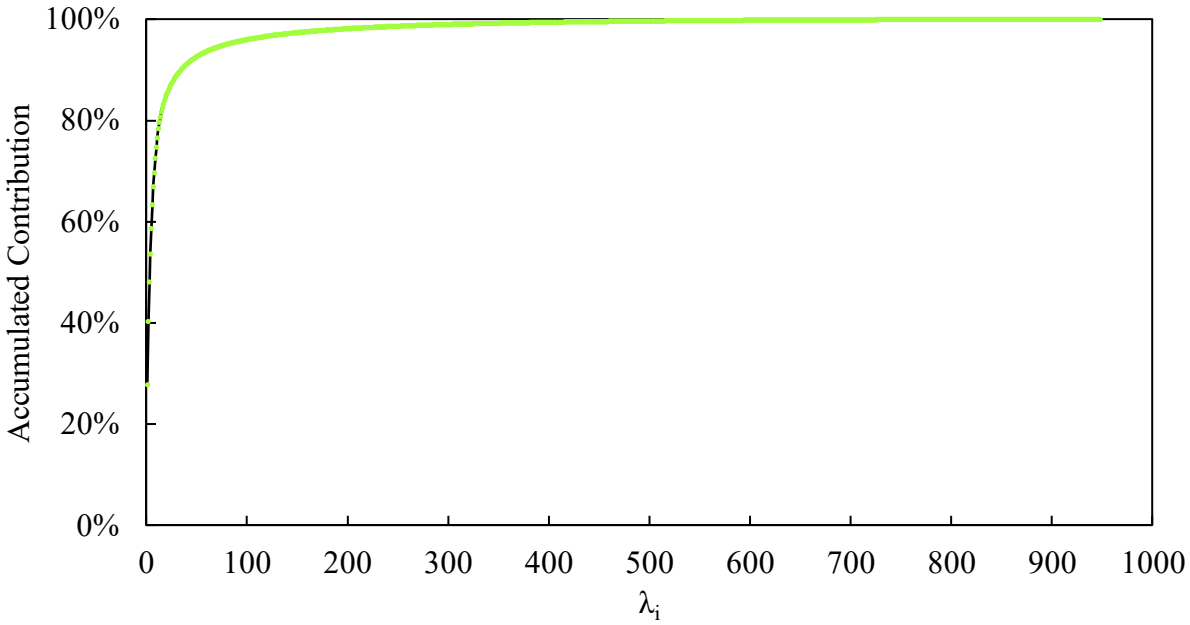


Figure 3-11: Accumulated contribution from principal components (PCs) for 1F6M. The eigenvalues (λ) associated with PCs in the range [1, *i*] are summed and presented as a fraction of the sum of all eigenvalues. This metric allows rapid determination of the percent of motions captured by the first *i* PCs. In this case, the first 100 PCs represent 96% of the motion of the protein backbone. Utilizing only the first 100 PCs means the trajectory experiences a corresponding dimensionality reduction of 89%, from 948 initial PCs.

Where $PC_{i,t}$ is the value corresponding to the simulation frame projected onto PC i at time t, and $M_{norm}(t)$ is the rescaled PC. Because of the complex and descriptive nature of PCs, positive and negative directions were considered arbitrary. The transformation in eqn. (7), then, penalizes states in the center of the PC while considering states at either extreme to be equivalent. Summing across the first 100 M_i at a given t produced an extrema score, which allowed distinction between states unlikely to be desirable because they were commonly visited and those states that were simultaneously at the edge of many PCs. However, this heuristic failed to account for totally denatured states and states that were identical to the crystal structure, so the metric was multiplied both by the normalized RMSD to the unbound crystal structure ($RMSDU_n$), produced by rescaling RMSDU according to eqn. (6), and by the quantity (1 - $RMSDU_n$). This modified heuristic, S_e, penalized states that failed to undergo conformational shifts and states that denatured, respectively. Finally, the product was rescaled and inverted for convenience. This inversion was performed to align the metric with the RMSD to the bound structure that trends to zero as the conformation approaches the global minimum bound state. This yielded an extremity metric (EM) score that ranked frames by estimated likelihood to undergo a conformational shift towards the desired bound structure, with a minimal score indicating a favorable shift.

$$S_e\,(t) = \left(\sum_{i=1}^{100} M_i\right)(RMSDU_n)(1 - RMSDU_n) \tag{8}$$

$$EM(t) = 1 - \left(\frac{\mathrm{S_e(t)} - \min(\mathrm{S_e})}{\max(\mathrm{S_e}) - \min(\mathrm{S_e})}\right) \tag{9}$$

The results of this metric are very encouraging. Table 3-2 compares the minima predicted by the EM to the actual minimum RMSDB, while Figure 3-12 presents the EM at each t>40 ps directly compared to the RMSDB rescaled according to eqn. (6). In two out of three cases, the EM predicted simulation RMSDB minima within 3 ps, while in the third case, the metric failed to

Table 3-2: Comparison of extremity metric (EM) predictions against minimum simulation RMSD to bound complex. The close agreement in trials 1 and 3 suggests the validity of the metric for predicting conformational motion towards a target structure, even when the target structure is unknown.

Trial	Simulation RMSDB Minimum Predicted by Extremity Metric (MD Frame)	Actual Simulation RMSDB Minimum (MD Frame)	Difference (ps)
1	1003	1015	1.2
2	410	1160	75.0
3	768	796	2.8

clearly predict any significant minimum. While it is unclear the exact cause of the discrepancy in the second trial, the corresponding trajectory exhibited a higher amount of denaturing than the other trials, so it is possible that this failure is because the heuristic predicted no valuable structures, rather than because there exists a non-correlation between the heuristic and the absolute metric.

The implications of such a heuristic are significant: a method that rapidly assesses and filters conformational motion toward some target or absolute minimum is widely valuable in cases where the crystal structure of the complex is well characterized.[94–96] A method capable of discriminating these motions without knowledge of the bound state, then, is exceptionally valuable. As discussed in section 3.3, finding a predictor of RMSDB is difficult when the target structure is not known *a priori*. Yet, EM seems to be a computationally efficient heuristic that relies only on simulation parameters to predict frames that have recently or will soon undergo significant RMSDB decreases. As a final check, predicted minima were extracted and redocked with TagDock. The results of the redocking experiment are reported in Table 3-3. Notably, in the two cases where the heuristic closely predicted the simulation RMSDB minimum, the docking score of the complex decreased by 41% for trial 1 and 35% for trial 3, reinforcing the implication of validity for the EM heuristic. Further, this decreased docking score correlated to an absolute RMSD reduction of 44% for trial 1 and 23% for trial 3, indicating the method can successfully improve docking predictions compared to the toolkit in the absence of the workflow.

Table 3-3: Comparison of results obtained by extracting frames predicted with the extremity metric (EM) heuristic. The close match predicted by the EM in trials 1 and 3 correlated to a notable decrease in TagDock penalty score and in RMSD to the bound structure (RMSDB), further supporting the hypothesis that the EM may be useful in predicting conformational motion towards the bound structure.

Trial	Simulation Frame Used for Redock	Starting TagDock Penalty Score	Final TagDock Penalty Score	Change	Starting RMSDB (Å)	Final RMSDB (Å)	Change
1	1003	41.72	24.60 ± 4.54	-41.04%	16.29	9.80 ± 0.028	-39.85%
2	410	41.72	45.04 ± 0.89	7.96%	16.29	14.20 ± 0.003	-12.79%
3	768	41.72	27.20 ± 1.65	-34.80%	16.29	13.61 ± 0.044	-16.45%

Figure 3-12, panel D also reports a compelling result from the simulated annealing experiment. The global minimum with respect to the bound structure in the third trial (located at t=79.6 ps) is about 5% closer to the bound structure than that of the starting conformation. This improvement suggests the potential of the method and its magnitude suggests capacity for a cyclic approach seeking convergence to some structure, though further experiments are necessary to characterize such a relationship.

3.5 Summary

In summary, a series of experiments were performed to characterize and optimize a simulated annealing workflow for use in an enhanced sampling context. First, the relationships between available metrics were interrogated. It was determined that the experimentally derived RMSD to the unbound complex has a weak correlation to the TagDock penalty score, which in turn has a moderate correlation to the RMSD between the structure and the bound conformation. Armed with this information, it was possible to interrogate the intricacies of conformational space sampling with simulated annealing. A series of optimization experiments were performed to evaluate the optimal temperature, thermostat, and simulation time resolution. It was determined that while simulation temperature and thermostat have a significant impact on sampling area, sampling resulting from the time resolution (time step) is negligible. Next, principal component

analysis (PCA) was performed on the trajectories and the conformational area explored along the first two principal components (PCs) was evaluated to confirm the value of the statistical method in this use case. Because these simulations demonstrated an increase in sampling area when analyzed with PCA, it was determined that the method was acceptable for use in this context. Finally, a series of simulations were performed to propose and benchmark a novel workflow for simulated annealing. After a protein decoy was predicted using rigid docking, the decoy was translated to remove chain and atomic overlaps, annealed, and evaluated using PCA. A heuristic

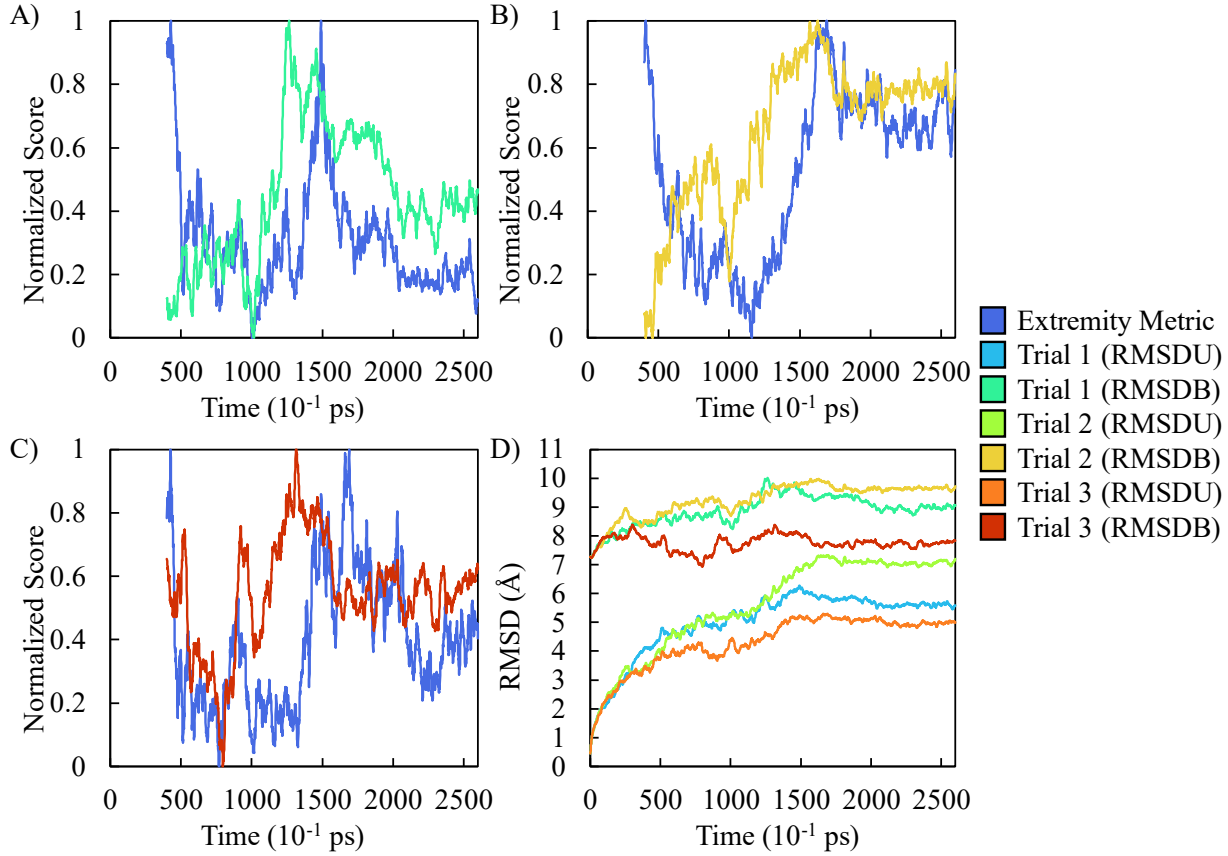


Figure 3-12: Heuristic predictions and normalized RMSDs to bound structure for trials 1-3 and total RMSD for all trials, compared. A-C) Normalized RMSD vs extremity metric (EM) for trials 1-3, respectively. EM minima in trials 1 and 3 closely predict the minimal RMSD to bound for the trajectory without knowledge of the bound structure *a priori*. Trial 2, however, does not predict any specific minimum. D) Total RMSD for all trials, compared. The increasing trend in Trial 2 (to Unbound) suggests the failure of the EM to predict an RMSD minimum may be due to increased denaturing associated with the trial. Notably, Trial 3 (to Bound) briefly dips below its initial value, and the EM recognizes this, suggesting the EM may be useful for predicting conformational motion towards an unknown structure in a cyclic application.

using the first 100 PCs and the root mean square difference between the trajectory and the unbound structure was developed to predict conformational transitions, and the metric successfully predicted motion towards the bound conformation within 3 ps in two out of three cases. Finally, the structure predicted by the heuristic was extracted and redocked with TagDock, yielding a clear decrease in RMSD to the bound structure.

4. Conclusion

As novel technologies and machine learning approaches have expanded the use cases for atomistic protein structures, it has become clear that methods to obtain these structures for flexible protein complexes continue to struggle when compared to cutting-edge approaches for monomers and rigid complexes. It is necessary, therefore, to develop unique methods to efficiently interrogate these structures while minimizing necessary *a priori* information and computational complexity. Rigid docking algorithms such as TagDock address these questions by rapidly performing molecular docking from a few pieces of biophysical information, provided the input components are structurally similar to the "bound" conformation of the protein. Yet, finding a "bound" conformation is often more difficult than performing molecular docking. Simulated annealing had been proposed as a simple method to understand protein flexibility, but work in the field has trended towards methods requiring external biases or rapid thermal fluctuation to achieve more effective results. The goal of this research was to propose a method that can span the gap, performing rapid molecular docking and efficiently optimizing conformational motion towards the bound state.

Ultimately, working toward this goal meant performing several broad evaluation and optimization studies before focusing on statistical techniques to predict favorable conformational motion. It was determined that extra care was necessary when evaluating initial docking decoys, as even though TagDock penalty score strongly correlates to how close a complex is to the "right" answer, existing metrics did not strongly correlate to the TagDock penalty score. Because TagDock penalty score is still comparatively time-intensive, it is not feasible to simply calculate the score for every frame of a simulation, highlighting the need for another metric. Likewise, optimization studies revealed the complexity of temperature-induced protein conformational

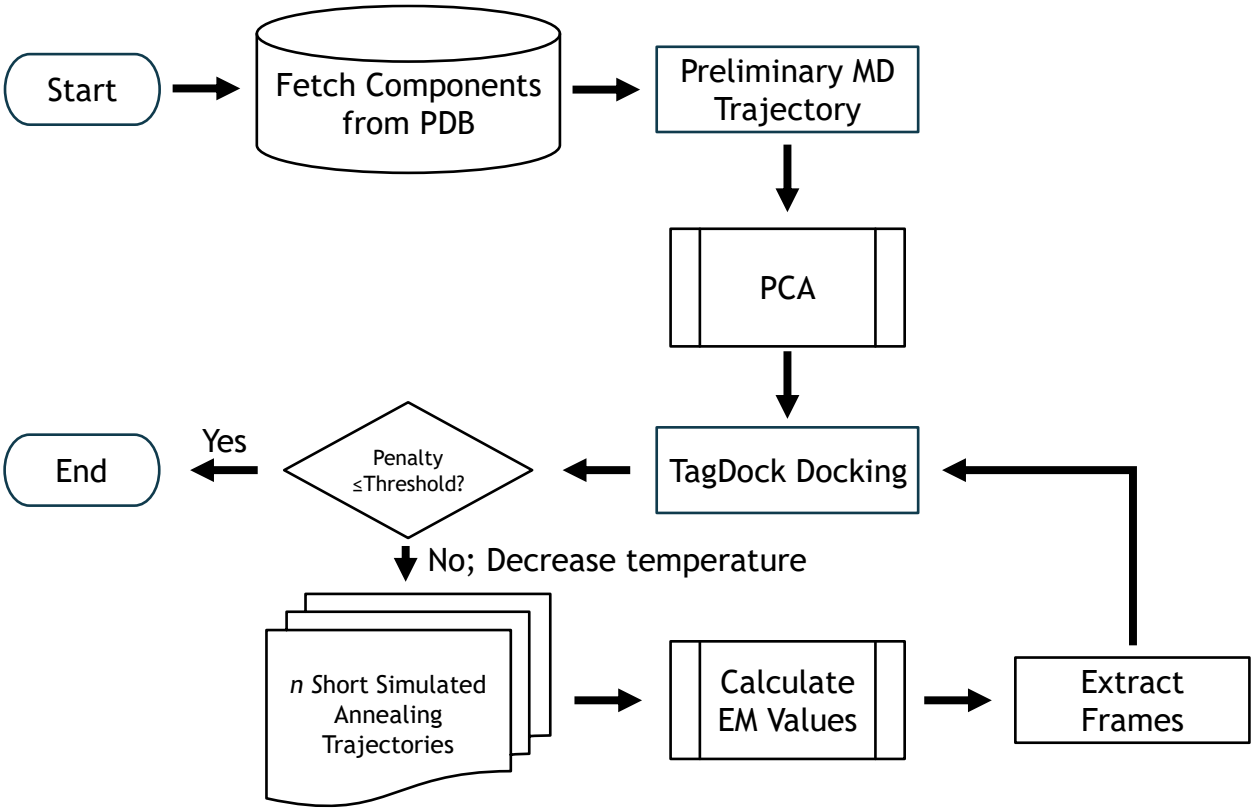
motion, and it was determined that no single optimal annealing temperature exists due to the differences in secondary structure and loop regions between different monomers. The temperature of 750 K, Langevin thermostat, and time step of 0.001 ps were chosen for 1F6M because of superior performance (750 K, Langevin) or ease of use and source recommendation (0.001 ps). Accordingly, it is suspected that the thermostat and time step may be held constant across any protein studied. The annealing temperature, however, will likely require a calibration for each new protein complex.

The implementation of principal components analysis to track conformational motion also provided interesting new ways to monitor the trajectory. By comparing motion along many principal components and by biasing the metric towards motion diverging from the initial structure but not denaturing, it became apparent that it was possible to predict conformational motion towards the bound component, resulting in a commensurate decrease in TagDock penalty score and RMSD with reference to the bound structure. In one case, this decrease was nearly 40%, suggesting the metric may be highly effective.

Future work is necessary, however, to characterize the reliability and applicability of the novel extremity metric (EM). While the EM was at least marginally effective in each of the three trials, it is conceivable that such a correlation may be a rare occurrence or result from some unforeseen correlation or property of the trial protein, 1F6M. For example, the hinge-twist binding motion inherent to 1F6M may lead to the observed behavior. Therefore, establishing the efficacy of the EM will require a more extensive study across several proteins of different classes. Further, though the EM predicts most favorable frames from the trajectory, the annealing method does not select the “bound” state explicitly, instead relying on conformational motion down the potential energy surface as a selection criterion. This behavior is unsuited for single-trial processing; instead,

it will be necessary for future work to implement a cyclic approach for more effective results. It is unknown whether the EM will maintain efficacy as the simulation approaches convergence. An example workflow is proposed in scheme 4-1.

As the simulation approaches convergence, it will be necessary to lower the annealing temperature to reduce the likelihood of conformational motion away from the target. However, caution will be necessary to minimize false convergence from the many free energy minima (see section 1.2.2). An optimization study will eventually be necessary, then, to evaluate the rate of this temperature decrease, as well as to predict the order of magnitude of computational time needed to reach convergence. This computational time will be a key metric in evaluating the overall strength of the workflow in comparison to other methods such as RosettaDOCK. The niche of the



Scheme 4-1: Proposed workflow implied by the results of this thesis. Combining the extremity metric with TagDock docking and simulated annealing in a cyclic workflow may provide an effective method to flexibly dock proteins, as indicated by the score improvements that were realized using a single cycle of the method.

proposed workflow is to rapidly develop low-to-medium-resolution decoys for use with these high-resolution methods, so it is important that the production of novel decoys requires much less computational effort than the effort needed to produce similar decoys with higher-resolution methods.

Another compelling line of potential research lies in characterizing the critical temperature necessary to access flexible conformational space. The thermal optimization of 1F6M produced three different optimal temperatures using differing metrics, and understanding the relationship between these metrics has the potential to yield rich rewards for computational efficiency using high-temperature MD simulations. For example, replica exchange methods sampled across a larger thermal range might enable acceleration of conformational motion without increased denaturing. While the data from the thermal optimization suggest some critical temperature where the protein can efficiently access conformational space without trending towards denaturing, the exact nature and extent of this behavior is not readily apparent. Therefore, studies to characterize this behavior and examine its reproducibility might be valuable in gaining fine optimization for molecular dynamics simulations of protein structures.

References:

(1) Zhang, L.; Yousefzadeh, M.; Suh, Y.; Niedernhofer, L.; Robbins, P. Signal Transduction, Ageing and Disease. In *Subcellular Biochemistry*; 2019; Vol. 91, pp 227–247.

(2) Emilsson, V.; Thorleifsson, G.; Zhang, B.; Leonardson, A. S.; Zink, F.; Zhu, J.; Carlson, S.; Helgason, A.; Walters, G. B.; Gunnarsdottir, S.; Mouy, M.; Steinthorsdottir, V.; Eiriksdottir, G. H.; Bjornsdottir, G.; Reynisdottir, I.; Gudbjartsson, D.; Helgadottir, A.; Jonasdottir, A.; Jonasdottir, A.; Styrkarsdottir, U.; Gretarsdottir, S.; Magnusson, K. P.; Stefansson, H.; Fossdal, R.; Kristjansson, K.; Gislason, H. G.; Stefansson, T.; Leifsson, B. G.; Thorsteinsdottir, U.; Lamb, J. R.; Gulcher, J. R.; Reitman, M. L.; Kong, A.; Schadt, E. E.; Stefansson, K. Genetics of Gene Expression and Its Effect on Disease. *Nature* **2008**, *452* (7186), 423–428.

(3) Awatef, O.; Ouertani, R.; Mosbah, A.; Masmoudi, A.; Cherif, A. Effectiveness of Enzyme Inhibitors in Biomedicine and Pharmacotherapy. **2019**.

(4) Saltiel, A. R. Insulin Signaling in Health and Disease. *J. Clin. Invest.* **2021**, *131* (1).

(5) Duncavage, E. J.; Schroeder, M. C.; O'Laughlin, M.; Wilson, R.; MacMillan, S.; Bohannon, A.; Kruchowski, S.; Garza, J.; Du, F.; Hughes, A. E. O.; Robinson, J.; Hughes, E.; Heath, S. E.; Baty, J. D.; Neidich, J.; Christopher, M. J.; Jacoby, M. A.; Uy, G. L.; Fulton, R. S.; Miller, C. A.; Payton, J. E.; Link, D. C.; Walter, M. J.; Westervelt, P.; DiPersio, J. F.; Ley, T. J.; Spencer, D. H. Genome Sequencing as an Alternative to Cytogenetic Analysis in Myeloid Cancers. *N. Engl. J. Med.* **2021**, *384* (10), 924–935.

(6) Audrito, V.; Messana, V. G.; Moiso, E.; Vitale, N.; Arruga, F.; Brandimarte, L.; Gaudino, F.; Pellegrino, E.; Vaisitti, T.; Riganti, C.; Piva, R.; Deaglio, S. NAMPT Over-Expression Recapitulates the BRAF Inhibitor Resistant Phenotype Plasticity in Melanoma. *Cancers* **2020**, *12* (12), 3855.

(7) Hanahan, D.; Weinberg, R. A. Hallmarks of Cancer: The Next Generation. *Cell* **2011**, *144* (5), 646–674.

(8) Sela, M.; White, F. H.; Anfinsen, C. B. Reductive Cleavage of Disulfide Bridges in Ribonuclease. *Science* **1957**, *125* (3250), 691–692.

(9) Wild, D.; Saqi, M. Structural Proteomics: Inferring Function from Protein Structure. *Curr. Proteomics* **2004**, *1* (1), 59–65.

(10) Hvidsten, T. R.; Lægreid, A.; Kryshtafovych, A.; Andersson, G.; Fidelis, K.; Komorowski, J. A Comprehensive Analysis of the Structure-Function Relationship in Proteins Based on Local Structure Similarity. *PLOS ONE* **2009**, *4* (7), e6266.

(11) Ferrie, J. J.; Karr, J. P.; Tjian, R.; Darzacq, X. "Structure"-Function Relationships in Eukaryotic Transcription Factors: The Role of Intrinsically Disordered Regions in Gene Regulation. *Mol. Cell* **2022**, *82* (21), 3970–3984.

(12) Kuntz, I. D. Structure-Based Strategies for Drug Design and Discovery. *Science* **1992**, *257* (5073), 1078–1082.

(13) Kitamura, N.; Sacco, M. D.; Ma, C.; Hu, Y.; Townsend, J. A.; Meng, X.; Zhang, F.; Zhang, X.; Ba, M.; Szeto, T.; Kukuljac, A.; Marty, M. T.; Schultz, D.; Cherry, S.; Xiang, Y.; Chen, Y.; Wang, J. Expedited Approach toward the Rational Design of Noncovalent SARS-CoV-2 Main Protease Inhibitors. *J. Med. Chem.* **2022**, *65* (4), 2848–2865.

(14) 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 Res.* **2000**, *28* (1), 235–242.

(15) Jumper, J.; Evans, R.; Pritzel, A.; Green, T.; Figurnov, M.; Ronneberger, O.; Tunyasuvunakool, K.; Bates, R.; Žídek, A.; Potapenko, A.; Bridgland, A.; Meyer, C.; Kohl, S. A. A.; Ballard, A. J.; Cowie, A.; Romera-Paredes, B.; Nikolov, S.; Jain, R.; Adler, J.; Back, T.; Petersen, S.; Reiman, D.; Clancy, E.; Zielinski, M.; Steinegger, M.; Pacholska, M.; Berghammer, T.; Bodenstein, S.; Silver, D.; Vinyals, O.; Senior, A. W.; Kavukcuoglu, K.; Kohli, P.; Hassabis, D. Highly Accurate Protein Structure Prediction with AlphaFold. *Nature* **2021**, *596* (7873), 583–589.
(16) Baek, M.; DiMaio, F.; Anishchenko, I.; Dauparas, J.; Ovchinnikov, S.; Lee, G. R.; Wang, J.; Cong, Q.; Kinch, L. N.; Schaeffer, R. D.; Millán, C.; Park, H.; Adams, C.; Glassman, C. R.; DeGiovanni, A.; Pereira, J. H.; Rodrigues, A. V.; van Dijk, A. A.; Ebrecht, A. C.; Opperman, D. J.; Sagmeister, T.; Buhlheller, C.; Pavkov-Keller, T.; Rathinaswamy, M. K.; Dalwadi, U.; Yip, C. K.; Burke, J. E.; Garcia, K. C.; Grishin, N. V.; Adams, P. D.; Read, R. J.; Baker, D. Accurate Prediction of Protein Structures and Interactions Using a Three-Track Neural Network. *Science* **2021**, *373* (6557), 871–876.
(17) van Breugel, M.; Rosa e Silva, I.; Andreeva, A. Structural Validation and Assessment of AlphaFold2 Predictions for Centrosomal and Centriolar Proteins and Their Complexes. *Commun. Biol.* **2022**, *5* (1), 1–10.
(18) Stevens, A. O.; He, Y. Benchmarking the Accuracy of AlphaFold 2 in Loop Structure Prediction. *Biomolecules* **2022**, *12* (7), 985.
(19) Bæk, K. T.; Kepp, K. P. Assessment of AlphaFold2 for Human Proteins via Residue Solvent Exposure. *J. Chem. Inf. Model.* **2022**, *62* (14), 3391–3400.
(20) Liang, T.; Jiang, C.; Yuan, J.; Othman, Y.; Xie, X.-Q.; Feng, Z. Differential Performance of RoseTTAFold in Antibody Modeling. *Brief. Bioinform.* **2022**, *23* (5), bbac152.
(21) Peñas-Utrilla, D.; Marcos, E. Identifying Well-Folded de Novo Proteins in the New Era of Accurate Structure Prediction. *Front. Mol. Biosci.* **2022**, *9*.
(22) Smith, C. W. J. *RNA-Protein Interactions : A Practical Approach: A Practical Approach*; Oxford University Press, UK, 1998.
(23) Gisriel, C.; Coe, J.; Letrun, R.; Yefanov, O. M.; Luna-Chavez, C.; Stander, N. E.; Lisova, S.; Mariani, V.; Kuhn, M.; Aplin, S.; Grant, T. D.; Dörner, K.; Sato, T.; Echelmeier, A.; Cruz Villarreal, J.; Hunter, M. S.; Wiedorn, M. O.; Knoska, J.; Mazalova, V.; Roy-Chowdhury, S.; Yang, J.-H.; Jones, A.; Bean, R.; Bielecki, J.; Kim, Y.; Mills, G.; Weinhausen, B.; Meza, J. D.; Al-Qudami, N.; Bajt, S.; Brehm, G.; Botha, S.; Boukhelef, D.; Brockhauser, S.; Bruce, B. D.; Coleman, M. A.; Danilevski, C.; Discianno, E.; Dobson, Z.; Fangohr, H.; Martin-Garcia, J. M.; Gevorkov, Y.; Hauf, S.; Hosseinizadeh, A.; Januschek, F.; Ketawala, G. K.; Kupitz, C.; Maia, L.; Manetti, M.; Messerschmidt, M.; Michelat, T.; Mondal, J.; Ourmazd, A.; Previtali, G.; Sarrou, I.; Schön, S.; Schwander, P.; Shelby, M. L.; Silenzi, A.; Sztuk-Dambietz, J.; Szuba, J.; Turcato, M.; White, T. A.; Wrona, K.; Xu, C.; Abdellatif, M. H.; Zook, J. D.; Spence, J. C. H.; Chapman, H. N.; Barty, A.; Kirian, R. A.; Frank, M.; Ros, A.; Schmidt, M.; Fromme, R.; Mancuso, A. P.; Fromme, P.; Zatsepin, N. A. Membrane Protein Megahertz Crystallography at the European XFEL. *Nat. Commun.* **2019**, *10* (1), 5021.
(24) Kaptein, R.; Boelens, R.; Scheek, R. M.; Van Gunsteren, W. F. Protein Structures from NMR. *Biochemistry* **1988**, *27* (15), 5389–5395.
(25) Wüthrich, K. Protein Structure Determination in Solution by NMR Spectroscopy. *J. Biol. Chem.* **1990**, *265* (36), 22059–22062.

(26) Hu, Y.; Cheng, K.; He, L.; Zhang, X.; Jiang, B.; Jiang, L.; Li, C.; Wang, G.; Yang, Y.; Liu, M. NMR-Based Methods for Protein Analysis. *Anal. Chem.* **2021**, *93* (4), 1866–1879.
(27) Maveyraud, L.; Mourey, L. Protein X-Ray Crystallography and Drug Discovery. *Molecules* **2020**, *25* (5), 1030.
(28) O'Reilly, M.; Cleasby, A.; Davies, T. G.; Hall, R. J.; Ludlow, R. F.; Murray, C. W.; Tisi, D.; Jhoti, H. Crystallographic Screening Using Ultra-Low-Molecular-Weight Ligands to Guide Drug Design. *Drug Discov. Today* **2019**, *24* (5), 1081–1086.
(29) Collins, P. M.; Ng, J. T.; Talon, R.; Nekrosiute, K.; Krojer, T.; Douangamath, A.; Brandao-Neto, J.; Wright, N.; Pearce, N. M.; von Delft, F. Gentle, Fast and Effective Crystal Soaking by Acoustic Dispensing. *Acta Crystallogr. Sect. Struct. Biol.* **2017**, *73* (3), 246–255.
(30) Gelin, M.; Delfosse, V.; Allemand, F.; Hoh, F.; Sallaz-Damaz, Y.; Pirocchi, M.; Bourguet, W.; Ferrer, J.-L.; Labesse, G.; Guichou, J.-F. Combining `dry' Co-Crystallization and *in Situ* Diffraction to Facilitate Ligand Screening by X-Ray Crystallography. *Acta Crystallogr. D Biol. Crystallogr.* **2015**, *71* (8), 1777–1787.
(31) Brändén, G.; Neutze, R. Advances and Challenges in Time-Resolved Macromolecular Crystallography. *Science* **2021**, *373* (6558), eaba0954.
(32) Zhu, L.; Chen, X.; Abola, E. E.; Jing, L.; Liu, W. Serial Crystallography for Structure-Based Drug Discovery. *Trends Pharmacol. Sci.* **2020**, *41* (11), 830–839.
(33) Ziegler, S. J.; Mallinson, S. J. B.; St. John, P. C.; Bomble, Y. J. Advances in Integrative Structural Biology: Towards Understanding Protein Complexes in Their Cellular Context. *Comput. Struct. Biotechnol. J.* **2021**, *19*, 214–225.
(34) Franke, D.; Petoukhov, M. V.; Konarev, P. V.; Panjkovich, A.; Tuukkanen, A.; Mertens, H. D. T.; Kikhney, A. G.; Hajizadeh, N. R.; Franklin, J. M.; Jeffries, C. M.; Svergun, D. I. ATSAS 2.8: A Comprehensive Data Analysis Suite for Small-Angle Scattering from Macromolecular Solutions. *J. Appl. Crystallogr.* **2017**, *50* (4), 1212–1225.
(35) Giri, N.; Roy, R. S.; Cheng, J. Deep Learning for Reconstructing Protein Structures from Cryo-EM Density Maps: Recent Advances and Future Directions. *Curr. Opin. Struct. Biol.* **2023**, *79*, 102536.
(36) Seffernick, J. T.; Lindert, S. Hybrid Methods for Combined Experimental and Computational Determination of Protein Structure. *J. Chem. Phys.* **2020**, *153* (24), 240901.
(37) Smith, J. A.; Edwards, S. J.; Moth, C. W.; Lybrand, T. P. TagDock: An Efficient Rigid Body Docking Algorithm for Oligomeric Protein Complex Model Construction and Experiment Planning. *Biochemistry* **2013**, *52* (33), 5577–5584.
(38) Karaca, E.; Bonvin, A. M. On the Usefulness of Ion-Mobility Mass Spectrometry and SAXS Data in Scoring Docking Decoys. *Acta Crystallogr. D Biol. Crystallogr.* **2013**, *69* (5), 683–694.
(39) Gray, J. J.; Moughon, S.; Wang, C.; Schueler-Furman, O.; Kuhlman, B.; Rohl, C. A.; Baker, D. Protein–Protein Docking with Simultaneous Optimization of Rigid-Body Displacement and Side-Chain Conformations. *J. Mol. Biol.* **2003**, *331* (1), 281–299.
(40) Marze, N. A.; Roy Burman, S. S.; Sheffler, W.; Gray, J. J. Efficient Flexible Backbone Protein–Protein Docking for Challenging Targets. *Bioinformatics* **2018**, *34* (20), 3461–3469.
(41) Dominguez, C.; Boelens, R.; Bonvin, A. M. J. J. HADDOCK: A Protein−Protein Docking Approach Based on Biochemical or Biophysical Information. *J. Am. Chem. Soc.* **2003**, *125* (7), 1731–1737.

(42) Roel-Touris, J.; Don, C. G.; V. Honorato, R.; Rodrigues, J. P. G. L. M.; Bonvin, A. M. J. J. Less Is More: Coarse-Grained Integrative Modeling of Large Biomolecular Assemblies with HADDOCK. *J. Chem. Theory Comput.* **2019**, *15* (11), 6358–6367.
(43) Harmalkar, A.; Gray, J. J. Advances to Tackle Backbone Flexibility in Protein Docking. *Curr. Opin. Struct. Biol.* **2021**, *67*, 178–186.
(44) Janin, J.; Henrick, K.; Moult, J.; Eyck, L. T.; Sternberg, M. J. E.; Vajda, S.; Vakser, I.; Wodak, S. J. CAPRI: A Critical Assessment of PRedicted Interactions. *Proteins Struct. Funct. Bioinforma.* **2003**, *52* (1), 2–9.
(45) Salomon-Ferrer, R.; Case, D. A.; Walker, R. C. An Overview of the Amber Biomolecular Simulation Package. *WIREs Comput. Mol. Sci.* **2013**, *3* (2), 198–210.
(46) Case, D. A.; Babin, V.; Berryman, J. T.; Betz, R. M.; Cai, Q.; Cerutti, D. S.; Cheatham, III, T. E.; Darden, T. A.; Duke, R. E.; Gohlke, H.; Goetz, A. W.; Gusarov, S.; Homeyer, N.; Janowski, P.; Kaus, J.; Kolossváry, I.; Kovalenko, A.; Lee, T. S.; LeGrand, S.; Luchko, T.; Luo, R.; Madej, B.; Merz, K. M.; Paesani, F.; Roe, D. R.; Roitberg, A.; Sagui, C.; Salomon-Ferrer, R.; Seabra, G.; Simmerling, C. L.; Smith, W.; Swails, J.; Walker, R. C.; Wang, J.; Wolf, R. M.; Wu, X.; Kollman, P. A. AMBER 14, 2014.
(47) Yang, L.; Liu, C.-W.; Shao, Q.; Zhang, J.; Gao, Y. Q. From Thermodynamics to Kinetics: Enhanced Sampling of Rare Events. *Acc. Chem. Res.* **2015**, *48* (4), 947–955.
(48) Harada, R. Simple, yet Efficient Conformational Sampling Methods for Reproducing/Predicting Biologically Rare Events of Proteins. *Bull. Chem. Soc. Jpn.* **2018**, *91* (9), 1436–1450.
(49) Dror, R. O.; Dirks, R. M.; Grossman, J. P.; Xu, H.; Shaw, D. E. Biomolecular Simulation: A Computational Microscope for Molecular Biology. *Annu. Rev. Biophys.* **2012**, *41* (1), 429–452.
(50) Wei, G.; Xi, W.; Nussinov, R.; Ma, B. Protein Ensembles: How Does Nature Harness Thermodynamic Fluctuations for Life? The Diverse Functional Roles of Conformational Ensembles in the Cell. *Chem. Rev.* **2016**, *116* (11), 6516–6551.
(51) Shaw, D. E.; Adams, P. J.; Azaria, A.; Bank, J. A.; Batson, B.; Bell, A.; Bergdorf, M.; Bhatt, J.; Butts, J. A.; Correia, T.; Dirks, R. M.; Dror, R. O.; Eastwood, M. P.; Edwards, B.; Even, A.; Feldmann, P.; Fenn, M.; Fenton, C. H.; Forte, A.; Gagliardo, J.; Gill, G.; Gorlatova, M.; Greskamp, B.; Grossman, J. P.; Gullingsrud, J.; Harper, A.; Hasenplaugh, W.; Heily, M.; Heshmat, B. C.; Hunt, J.; Ierardi, D. J.; Iserovich, L.; Jackson, B. L.; Johnson, N. P.; Kirk, M. M.; Klepeis, J. L.; Kuskin, J. S.; Mackenzie, K. M.; Mader, R. J.; McGowen, R.; McLaughlin, A.; Moraes, M. A.; Nasr, M. H.; Nociolo, L. J.; O'Donnell, L.; Parker, A.; Peticolas, J. L.; Pocina, G.; Predescu, C.; Quan, T.; Salmon, J. K.; Schwink, C.; Shim, K. S.; Siddique, N.; Spengler, J.; Szalay, T.; Tabladillo, R.; Tartler, R.; Taube, A. G.; Theobald, M.; Towles, B.; Vick, W.; Wang, S. C.; Wazlowski, M.; Weingarten, M. J.; Williams, J. M.; Yuh, K. A. Anton 3: Twenty Microseconds of Molecular Dynamics Simulation before Lunch. In *Proceedings of the International Conference for High Performance Computing, Networking, Storage and Analysis*; ACM: St. Louis Missouri, 2021; pp 1–11.
(52) Zuckerman, D. M. Equilibrium Sampling in Biomolecular Simulations. *Annu. Rev. Biophys.* **2011**, *40* (1), 41–62.
(53) Hénin, J.; Lelièvre, T.; Shirts, M. R.; Valsson, O.; Delemotte, L. Enhanced Sampling Methods for Molecular Dynamics Simulations. *Living J. Comput. Mol. Sci.* **2022**, *4* (1).

(54) Chen, H.; Chipot, C. Enhancing Sampling with Free-Energy Calculations. *Curr. Opin. Struct. Biol.* **2022**, *77*, 102497.
(55) Sugita, Y.; Okamoto, Y. Replica-Exchange Molecular Dynamics Method for Protein Folding. *Chem. Phys. Lett.* **1999**, *314* (1–2), 141–151.
(56) Hansmann, U. H. E. Parallel Tempering Algorithm for Conformational Studies of Biological Molecules. *Chem. Phys. Lett.* **1997**, *281* (1–3), 140–150.
(57) Sugita, Y.; Kitao, A.; Okamoto, Y. Multidimensional Replica-Exchange Method for Free-Energy Calculations. *J. Chem. Phys.* **2000**, *113* (15), 6042–6051.
(58) Ostermeir, K.; Zacharias, M. Accelerated Flexible Protein-Ligand Docking Using Hamiltonian Replica Exchange with a Repulsive Biasing Potential. *PLOS ONE* **2017**, *12* (2), e0172072.
(59) Wang, W. Recent Advances in Atomic Molecular Dynamics Simulation of Intrinsically Disordered Proteins. *Phys. Chem. Chem. Phys.* **2021**, *23* (2), 777–784.
(60) Bussi, G.; Laio, A. Using Metadynamics to Explore Complex Free-Energy Landscapes. *Nat. Rev. Phys.* **2020**, *2* (4), 200–212.
(61) Barducci, A.; Bussi, G.; Parrinello, M. Well-Tempered Metadynamics: A Smoothly Converging and Tunable Free-Energy Method. *Phys. Rev. Lett.* **2008**, *100* (2), 020603.
(62) Laio, A.; Parrinello, M. Escaping Free-Energy Minima. *Proc. Natl. Acad. Sci.* **2002**, *99* (20), 12562–12566.
(63) Invernizzi, M.; Parrinello, M. Rethinking Metadynamics: From Bias Potentials to Probability Distributions. *J. Phys. Chem. Lett.* **2020**, *11* (7), 2731–2736.
(64) Maiorov, V. N.; Crippen, G. M. Significance of Root-Mean-Square Deviation in Comparing Three-Dimensional Structures of Globular Proteins. *J. Mol. Biol.* **1994**, *235* (2), 625–634.
(65) Kitao, A. Principal Component Analysis and Related Methods for Investigating the Dynamics of Biological Macromolecules. *J* **2022**, *5* (2), 298–317.
(66) Jolliffe, I. T.; Cadima, J. Principal Component Analysis: A Review and Recent Developments. *Philos. Trans. R. Soc. Math. Phys. Eng. Sci.* **2016**, *374* (2065), 20150202.
(67) Yasuda, T.; Shigeta, Y.; Harada, R. Efficient Conformational Sampling of Collective Motions of Proteins with Principal Component Analysis-Based Parallel Cascade Selection Molecular Dynamics. *J. Chem. Inf. Model.* **2020**, *60* (8), 4021–4029.
(68) Lennon, B. W.; Williams Jr., C. H.; Ludwig, M. L. CRYSTAL STRUCTURE OF A COMPLEX BETWEEN THIOREDOXIN REDUCTASE, THIOREDOXIN, AND THE NADP+ ANALOG, AADP+. **2000**.
(69) Lennon, B. W.; Williams, C. H.; Ludwig, M. L. Twists in Catalysis: Alternating Conformations of Escherichia Coli Thioredoxin Reductase. *Science* **2000**, *289* (5482), 1190–1194.
(70) Protein Data Bank Contents Guide: Atomic Coordinate Entry Format Description Version 3.30, 2012. https://files.wwpdb.org/pub/pdb/doc/format_descriptions/Format_v33_Letter.pdf (accessed 2024-02-15).
(71) Berman, H.; Henrick, K.; Nakamura, H. Announcing the Worldwide Protein Data Bank. *Nat. Struct. Mol. Biol.* **2003**, *10* (12), 980–980.
(72) Chen, R.; Li, L.; Weng, Z. ZDOCK: An Initial-Stage Protein-Docking Algorithm. *Proteins Struct. Funct. Bioinforma.* **2003**, *52* (1), 80–87.

(73) Hwang, H.; Vreven, T.; Janin, J.; Weng, Z. Protein–Protein Docking Benchmark Version 4.0. *Proteins Struct. Funct. Bioinforma.* **2010**, *78* (15), 3111–3114.
(74) Vreven, T.; Moal, I. H.; Vangone, A.; Pierce, B. G.; Kastritis, P. L.; Torchala, M.; Chaleil, R.; Jiménez-García, B.; Bates, P. A.; Fernandez-Recio, J.; Bonvin, A. M. J. J.; Weng, Z. Updates to the Integrated Protein–Protein Interaction Benchmarks: Docking Benchmark Version 5 and Affinity Benchmark Version 2. *J. Mol. Biol.* **2015**, *427* (19), 3031–3041.
(75) Pettersen, E. F.; Goddard, T. D.; Huang, C. C.; Couch, G. S.; Greenblatt, D. M.; Meng, E. C.; Ferrin, T. E. UCSF Chimera--a Visualization System for Exploratory Research and Analysis. *J. Comput. Chem.* **2004**, *25* (13), 1605–1612.
(76) Carugo, O.; Pongor, S. A Normalized Root-Mean-Spuare Distance for Comparing Protein Three-Dimensional Structures. *Protein Sci.* **2001**, *10* (7), 1470–1473.
(77) Gray, P. Implementation of Simulated Annealing to Derive Alternative Conformations and Improve Docking Compatibility in Protein Dimer Complexes. Honors Thesis, Western Kentucky University, Bowling Green, Kentucky, 2020.
(78) 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* (8), 3696–3713.
(79) Davidchack, R. L.; Handel, R.; Tretyakov, M. V. Langevin Thermostat for Rigid Body Dynamics. *J. Chem. Phys.* **2009**, *130* (23), 234101.
(80) Roe, D. R.; Cheatham, T. E. I. PTRAJ and CPPTRAJ: Software for Processing and Analysis of Molecular Dynamics Trajectory Data. *J. Chem. Theory Comput.* **2013**, *9* (7), 3084–3095.
(81) Berendsen, H. J. C.; Postma, J. P. M.; Van Gunsteren, W. F.; DiNola, A.; Haak, J. R. Molecular Dynamics with Coupling to an External Bath. *J. Chem. Phys.* **1984**, *81* (8), 3684–3690.
(82) Kabsch, W.; Sander, C. Dictionary of Protein Secondary Structure: Pattern Recognition of Hydrogen-bonded and Geometrical Features. *Biopolymers* **1983**, *22* (12), 2577–2637.
(83) Carugo, O. Statistical Validation of the Root-Mean-Square-Distance, a Measure of Protein Structural Proximity. *Protein Eng. Des. Sel.* **2007**, *20* (1), 33–37.
(84) Eastman, P.; Grønbech-Jensen, N.; Doniach, S. Simulation of Protein Folding by Reaction Path Annealing. *J. Chem. Phys.* **2001**, *114* (8), 3823–3841.
(85) Harada, R.; Nakamura, T.; Shigeta, Y. A Fast Convergent Simulated Annealing Algorithm for Protein-Folding: Simulated Annealing Outlier FLOODing (SA-OFLOOD) Method. *Bull. Chem. Soc. Jpn.* **2016**, *89* (11), 1361–1367.
(86) Oliveberg, M.; Tan, Y. J.; Fersht, A. R. Negative Activation Enthalpies in the Kinetics of Protein Folding. *Proc. Natl. Acad. Sci.* **1995**, *92* (19), 8926–8929.
(87) Zheng, W.; Andrec, M.; Gallicchio, E.; Levy, R. M. Simulating Replica Exchange Simulations of Protein Folding with a Kinetic Network Model. *Proc. Natl. Acad. Sci.* **2007**, *104* (39), 15340–15345.
(88) Makhatadze, G. I.; Privalov, P. L. Energetics of Protein Structure. In *Advances in Protein Chemistry*; Anfinsen, C. B., Richards, F. M., Edsall, J. T., Eisenberg, D. S., Eds.; Academic Press, 1995; Vol. 47, pp 307–425.
(89) Yang, A.-S.; Honig, B. Free Energy Determinants of Secondary Structure Formation: I. α-Helices. *J. Mol. Biol.* **1995**, *252* (3), 351–365.
(90) Yang, A.-S.; Honig, B. Free Energy Determinants of Secondary Structure Formation: II. Antiparallel β-Sheets. *J. Mol. Biol.* **1995**, *252* (3), 366–376.

(91) Omelyan, I.; Kovalenko, A. Multiple Time Step Molecular Dynamics in the Optimized Isokinetic Ensemble Steered with the Molecular Theory of Solvation: Accelerating with Advanced Extrapolation of Effective Solvation Forces. *J. Chem. Phys.* **2013**, *139* (24), 244106.
(92) Andrea, T. A.; Swope, W. C.; Andersen, H. C. The Role of Long Ranged Forces in Determining the Structure and Properties of Liquid Water. *J. Chem. Phys.* **1983**, *79* (9), 4576–4584.
(93) Zhang, J.; Gong, H. Frontier Expansion Sampling: A Method to Accelerate Conformational Search by Identifying Novel Seed Structures for Restart. *J. Chem. Theory Comput.* **2020**, *16* (8), 4813–4821.
(94) Ferrara, P.; Apostolakis, J.; Caflisch, A. Targeted Molecular Dynamics Simulations of Protein Unfolding. *J. Phys. Chem. B* **2000**, *104* (18), 4511–4518.
(95) Izrailev, S.; Stepaniants, S.; Isralewitz, B.; Kosztin, D.; Lu, H.; Molnar, F.; Wriggers, W.; Schulten, K. Steered Molecular Dynamics. In *Computational Molecular Dynamics: Challenges, Methods, Ideas*; Deuflhard, P., Hermans, J., Leimkuhler, B., Mark, A. E., Reich, S., Skeel, R. D., Eds.; Griebel, M., Keyes, D. E., Nieminen, R. M., Roose, D., Schlick, T., Series Eds.; Lecture Notes in Computational Science and Engineering; Springer Berlin Heidelberg: Berlin, Heidelberg, 1999; Vol. 4, pp 39–65.
(96) Harada, R.; Kitao, A. Parallel Cascade Selection Molecular Dynamics (PaCS-MD) to Generate Conformational Transition Pathway. *J. Chem. Phys.* **2013**, *139* (3), 035103.

Appendix A

Each four-character code listed below refers to an RCSB PDB[14] structure. These structures were used as reference structures throughout this work, and the corresponding unbound or bound components in the ZDock Benchmark 5.5[74] were used as starting crystal structures or target crystal structures, respectively.

1A2K	1FQJ	1JWH	1R0R	1ZM4	2J7P
1ACB	1GCQ	1JZD	1R6Q	2A5T	2JEL
1AHW	1GHQ	1K5D	1R8S	2A9K	2MTA
1AY7	1GL1	1KAC	1S1Q	2ABZ	2O3B
1B6C	1GLA	1KLU	1SBB	2AJF	2O8V
1BKD	1GP2	1KTZ	1SYX	2AYO	2OOB
1BUH	1GXD	1M10	1T6B	2B42	2OOR
1CGI	1H9D	1MAH	1TMQ	2B4J	2OT3
1D6R	1HE1	1ML0	1UDI	2C0L	2OZA
1DFJ	1HE8	1MLC	1US7	2CFH	2PCC
1DQJ	1I2M	1MQ8	1WDW	2FD6	2SIC
1E6E	1IB1	1NCA	1WEJ	2FJU	2SNI
1E6J	1IBR	1NW9	1WQ1	2G77	2UUY
1EAW	1IJK	1OC0	1XD3	2H7V	3BP8
1EER	1IQD	1OFU	1XQS	2HLE	3D5S
1EFN	1IRA	1OPH	1Y64	2HQS	3SGQ
1EWY	1J2J	1OYV	1Z0K	2HRK	4CPA
1F34	1JIW	1PPE	1Z5Y	2I25	7CEI
1F6M	1JK9	1PVH	1ZHI	2I9B	9QFW
1FLE	1JPS	1PXV	1ZLI	2J0T	BOYV
1FQ1	1JTG	1QA9			

Appendix B

Annotated submission scripts for simulated annealing trials.

a. Minimization:
Minimize # *script name*
&cntrl # *start of script*
imin=1, # *flag for minimization*
ntx=1, # *read input coordinates; do not read velocities*
irest=0, # *run as a new simulation; ignore input file velocities*
maxcyc=2000, # *set maximum minimization cycles to 2000*
ncyc=1000, # *switch from steepest descent to conjugate gradient minimization after 1000* cycles
ntpr=100, # *print human-readable progress every 100 steps*
ntwx=0, # *do not print intermediate trajectory coordinates*
ntb=0, # *do not use periodic boundaries and disable Particle Mesh Ewald calculations*
cut=999, # *do not sever bonds within 999 Å*
/ # *end of script*

b. Heating:
Heating
&cntrl
imin=0, # *do not perform minimization*
ntx=1,
irest=0,
nstlim=30000, # *perform 30,000 simulation steps*
dt=0.001, # *use 0.001 ps of time per simulation step*
ntf=2, # *omit bond-proton interactions*
ntc=2, # *constrain protons for use with SHAKE algorithm*
tempi=0.0, # *set initial temperature to 0 K*
temp0=750.0, # *set final temperature to 750 K*
ntpr=100,
ntwx=100, # *write trajectory coordinates every 100 steps*
cut=999,
ntb=0,
ntp=0, # *do not monitor or constrain pressures*
ntt=3, # *use Langevin dynamics to control temperature*
gamma_ln=2.0, # *Langevin collision frequency γ. Denotes frequency of collisions in* ps^{-1}
nmropt=1, # *use NMR restraints and weight changes*
ig=-1, # *seed Langevin dynamics from system time*
/
&wt type='TEMP0', istep1=0, istep2=18000, value1=0.0, value2=750.0 / # *increase system temperature from 0 K to 750 K over 18 ps of simulation*
&wt type='TEMP0', istep1=18001, istep2=30000, value1=750.0, value2=750.0 / #

simulate at 750 K for 12 ps
&wt type='END' / # *end step control*

c. Production I:
Production
&cntrl
imin = 0,
irest = 1, # *perform a "restart" simulation where velocities are read from input file*
ntx = 7, # *read coordinates and velocities from input file (identical to ntx=5)*
ntb = 0,
pres0 = 1.0, # *set system pressure to 1.0 bar (can be deprecated; identical to system default value)*
ntp = 0,
taup = 2.0, # *set pressure relaxation time to 2.0 ps (can be deprecated; not applicable when ntp=0)*
cut = 10.0, # *cut bonds between atoms more than 10 Å apart*
ntr = 0, # *do not use harmonic restraints to hold atoms in 3D space*
ntc = 2,
ntf = 2,
tempi = 750.0, # *set initial temperature to 750 K*
temp0 = 750.0,
ntt = 3,
gamma_ln = 1.0,
nstlim = 100000, # *simulate for 100000 steps*
dt = 0.001,
ntpr = 100,
ntwx = 100,
ntwr = 1000 # *write a restart file every 1000 steps*
/

d. Cooling:
Cooling
&cntrl
imin=0,
ntx=1,
irest=0,
nstlim=30000,
dt=0.001,
ntf=2,
ntc=2,
tempi=750.0,
temp0=300.0, # *set final temperature to 300 K*
ntpr=100,

ntwx=100,
cut=999,
ntb=0,
ntp=0,
ntt=3,
gamma_ln=2.0,
nmropt=1,
ig=-1,
/
&wt type='TEMP0', istep1=0, istep2=18000, value1=750, value2=300.0 / # *change system temperature from 750 K to 300 K over 18 ps*
&wt type='TEMP0', istep1=18001, istep2=30000, value1=300, value2=300.0 / # *simulate at 300 K for 12 ps*
&wt type='END' /

e. Production II:

Production
&cntrl
imin = 0,
irest = 1,
ntx = 7,
ntb = 0,
pres0 = 1.0,
ntp = 0,
taup = 2.0,
cut = 10.0,
ntr = 0,
ntc = 2,
ntf = 2,
tempi = 300.0, # *set initial temperature to 300 K*
temp0 = 300.0,
ntt = 3,
gamma_ln = 1.0,
nstlim = 100000,
dt = 0.001,
ntpr = 100,
ntwx = 100,
ntwr = 1000
/

Copyright Permission

Name: Stichter, Zachary

Email (to receive future readership statistics): zstichter@gmail.com

Type of document: ['Thesis']

Title: Development of an Enhanced Sampling Workflow to Accelerate Molecular Docking with Sparse Biophysical Information

Keywords (3-5 keywords not included in the title that uniquely describe content): computational chemistry; protein structure; enhanced sampling; molecular dynamics; flexible docking

Committee Chair: Matthew Nee

Additional Committee Members: Kevin Williams Jeremy Maddox

Select 3-5 TopSCHOLAR® disciplines for indexing your research topic in TopSCHOLAR®: Physical Chemistry; Bioinformatics; Other Chemistry

Copyright Permission for TopSCHOLAR® (digitalcommons.wku.edu) and ProQuest research repositories:

I hereby warrant that I am the sole copyright owner of the original work.
I also represent that I have obtained permission from third party copyright owners of any material incorporated in part or in whole in the above described material, and I have, as such identified and acknowledged such third-part owned materials clearly. I hereby grant Western Kentucky University the permission to copy, display, perform, distribute for preservation or archiving in any form necessary, this work in TopSCHOLAR® and ProQuest digital repository for worldwide unrestricted access in perpetuity.
I hereby affirm that this submission is in compliance with Western Kentucky University policies and the U.S. copyright laws and that the material does not contain any libelous matter, nor does it violate third-party privacy. I also understand that the University retains the right to remove or deny the right to deposit materials in TopSCHOLAR® and/or ProQuest digital repository.

['I grant permission to post my document in TopSCHOLAR and ProQuest for unrestricted access.']

The person whose information is entered above grants their consent to the collection and use of their information consistent with the Privacy Policy. They acknowledge that the use of this service is subject to the Terms and Conditions.

['I consent to the above statement.']