2023 INCITE Proposal Submission

Proposal

Title: Understanding protein allostery and druggable locations from the unified perspective of energy flows

Principal Investigator: Ao Ma

Organization: University of Illinois at Chicago

Date/Time Generated: 6/15/2022 11:30:55 PM

Section 1: PI and Co-PI Information

Question #1

Principal Investigator: *The PI is responsible for the project and managing any resources awarded to the project. If your project has multiple investigators, list the PI in this section and add any Co-PIs in the following section.*

Principal Investigator

First Name

Ao

Last Name

Ma

Organization

University of Illinois Chicago

Email

aoma@uic.edu

Work Phone

312-996-7225

Address Line 1

851 South Morgan Street

Address Line 2

(No answer given.)

City

Chicago

State

Illinois

Zip Code

60607

Question #2

Co-PI (s)

Question #3

Institutional Contact: *For the PI's institution on the proposal, identify the agent who has the authority to review, negotiate, and sign the user agreement on behalf of that institution. The person who can commit an organization may be someone in the contracts or procurement department, legal, or if a university, the department head or Sponsored Research Office or Grants Department.*

Institutional Contact

Institutional Contact Name

Chelsea Matthews

Institutional Contact Phone

312-355-3874

Institutional Contact Email

cheli@uic.edu

Section 2: Project Information

Question #1

Select the category that best describes your project.

Research Category

Biological Sciences: Biophysics

Question #2

Please provide a project summary in two sentences that can be used to describe the impact of your project to the public (50 words maximum)

Project Summary

We will identify the allosteric pathways in HIV-1 protease and the druggable locations of SARS-COV-2 main protease. These findings will help drug discovery for these two systems.

Section 3: Early Career Track

Question #1

Early Career

Starting in the INCITE 2022 year, INCITE is committing 10% of allocatable time to an Early Career Track in INCITE. The goal of the early career track is to encourage the next generation of high-performance computing researchers. Researchers within 10 years from earning their PhD (after December 31st 2012) may choose to apply. Projects will go through the regular INCITE Computational Readiness and Peer Review process, but the INCITE Management Committee will consider meritorious projects in the Early Career Track separately.

***Who Can Apply**: Researchers less than 10 years out from their PhD that need LCF-level capabilities to advance their overall research plan and who have not been a previous INCITE PI.*

How to Apply:

In the regular application process, there will be a check-box to self-identify as early career.

- *The required CV should make eligibility clear.*
- *If awarded, how will this allocation fit into your overall research plan for the next 5 years?*

Projects will go through the regular INCITE review process. The INCITE Program is targeting at least

10% of allocatable time. When selecting the INCITE Career Track, PIs are not restricted to just competing in that track.

- *What is the Early Career Track?*
    - *The INCITE Program created the Early Career Track to encourage researchers establishing their research careers. INCITE will award at least 10% of allocatable time to meritorious projects.*
- *Will this increase my chances of receiving an award?*
    - *Potentially, this could increase chances of an award. Projects must still be deemed scientifically meritorious through the review process INCITE uses each year.*
- *What do I need to do to be considered on the Early Career Track?*
    - *In the application process, select 'Yes' at 'If you are within 10 years of your PhD, would you like to be considered in the Early Career Track?' You will need to write a paragraph about how the INCITE proposal fits into your 5-year research and career goals.*
- *What review criteria will be used for the Early Career Track?*
    - *The same criteria for computational readiness and scientific merit will be applied to projects in the Early Career Track as will be applied to projects in the traditional track. The different will be manifest in awards decisions by the INCITE management committee.*

Early Career Track

If you are within 10 years of your PhD, would you like to be considered in the Early Career Track? Choosing this does not reduce your chances of receiving an award.

No

If 'yes', what year was your PhD? If 'no' enter N/A

N/A

If 'yes', how will this allocation fit into your overall research plan for the next 5 years? If 'no' enter N/A.

N/A

Section 4: INCITE Allocation Request & Other Project Funding/Computing Resources

Question #1

OLCF Summit (IBM / AC922) Resource Request - 2023

Question #2

OLCF Frontier (Cray Shasta) Resource Request – 2023

Question #3

OLCF Frontier (Cray Shasta) Resource Request – 2024

Question #4

OLCF Frontier (Cray Shasta) Resource Request – 2025

Question #5

ALCF Theta (Cray XC40) Resource Request - 2023

Question #6

ALCF Polaris Resource Request - 2023

Question #7

ALCF Polaris Resource Request - 2024

Question #8

ALCF Polaris Resource Request - 2025

Question #9

ALCF Aurora (Intel X^e) Resource Request – 2023

Question #10

ALCF Aurora (Intel X^e) Resource Request – 2024

Question #11

ALCF Aurora (Intel X^e) Resource Request – 2025

Question #12

List any funding this project receives from other funding agencies.

Funding Sources

Question #13

List any other high-performance computing allocations being received in support of this project.

Other High Performance Computing Resource Allocations

Section 5: Project Narrative and Supplemental Materials

Question #1

Using the templates provided here, please follow the INCITE Proposal Preparation Instructions to prepare your proposal. Elements needed include (1) Project Executive Summary, (2) Project Narrative, (3) Personnel Justification and Management Plan, (4) Milestone Table, (5) Publications Resulting from prior INCITE Awards (if appropriate), and (6) Biographical Sketches for the PI and all co-PI's. Concatenate all materials into a single PDF file. Prior to submission, it is strongly recommended that proposers review their proposals to ensure they comply with the proposal preparation instructions.

Concatenate all materials below into a single PDF file.

1. **Project Executive Summary (One Page Max)**
2. **Project Narrative (15 Pages Max)**
3. **Personnel Justification and Management Plan (1 Page Max)**
4. **Milestone Table**
5. **Publications resulting from prior INCITE Awards (if appropriate)**

6. Biographical Sketches for the PI and all co-PI's.

INCITE_proposal_Ma.pdf
The attachment is on the following page.

Project Executive Summary

Title: Understanding protein allostery and druggable locations from the unified perspective of energy flows

PI: Ao Ma

Applying Institution/Organization: Department of Bioengineering, University of Illinois at Chicago

Number of Processor Hours Requested: 120 Million

Amount of Storage Requested: 60 TB

Executive Summary:

The overarching goal of the proposed research is to understand two important dynamic phenomena, 1) protein allostery, 2) druggable locations in proteins, from the unified view of collective functional protein dynamics and energy flows. Although these processes appear very diverse, they can be unified from the perspective of energy flows, which is the fundamental principle governing dynamic processes in complex molecular systems.

Allostery is a fundamental aspect of protein function intrinsic to all proteins. It has important implications in enzyme catalysis, signal transduction, drug design and protein engineering. It is a consequence of the collective and global nature of functional protein dynamics that is intrinsic to all proteins. Understanding allostery based on first-principle physics at mechanistic level requires rigorous understanding of the pathway and mechanism of functional protein dynamics. The 'energy flow' method we developed provides such a methodological framework. We propose to study: **1)**: **Identify the allosteric pathway in PDZ domain and elucidate its mechanism**. **2)**: **Identify the allosteric pathway responsible for the non-active-site drug-resistant mutations in HIV-1 protease and elucidate the mechanism of the allosteric interactions between these residues and the active site dynamics/conformation**.

Identifying druggable locations, i.e. patches in the binding pocket that a drug should complementarily occupy, is crucial for effective drug discovery. The traditional lock-and-key picture of ligand binding suggests that a drug that occupies all the fillable space in the binding pocket may lead to optimal binding, a premise often contradicted by practical experiences from drug development. Based on the biophysical principle emerged from our recent study on the ligand binding process of HIV-1 protease, optimal druggable locations are binding site patches that, upon drug occupation, do not disrupt the collective and cooperative global protein motions that are displayed in the ligand binding pathway**.** Within the framework of "energy flow", these collective and cooperative global motions correspond to concerted global energy flow channels of protein functional dynamics. Herein, a patch is druggable if its occupation facilitates global energy flows during ligand binding; it is not druggable if its occupation blocks the energy flows. We propose to use energy flow analysis to map out the energy flow channels during their ligand binding processes, aiming at three objectives: **1) Map out druggable locations of SARS-CoV-2 main protease**. **2) Identify functional roles of different moieties of drug candidates: α-ketoamides for main protease**. **3) Identify non-binding-site residues that are important for energy flows in drug binding of these two proteins**. They are potential sites for drug resistant mutations.

Outcomes: We expect to identify the pathway for signal transduction and allostery in PDZ domain, the allostery pathway responsible for the drug-resistance of non-active-site mutations in HIV-PR and insights into the general physical principle governing allostery. In addition, we expect to identify the physical principle that determines druggable locations in proteins and its relation to collective functional protein dynamics and allostery.

PROJECT NARRATIVE

1 Significance of Research

1.1 Significance of Protein Allostery. Allostery is an intrinsic property of proteins [1-12]. It is involved in enzyme catalysis, drug design and protein engineering. Understanding the mechanism of allostery is of fundamental importance. The gist of allostery is that a change at the allosteric site causes a change at a remote site. If this change is conformational, it is conformational allostery. If it is in the dynamics instead of the structure, it is the so called dynamic allostery [13-16]. A fundamental question in understanding allostery is the **allosteric pathway**—the chain of events that transmits the change at the allosteric site to the remote site. Pioneering works in this regard included network analysis on spatial contacts or correlations between residues, and correlational analysis of interactions between residues [17-33]. However, it remains a challenge to show that the observed correlations are indeed the cause of the observed allostery.

Based on our studies of the conformational dynamics of alanine dipeptide [34, 35] and HIV-1 protease (**HIV-PR**), we propose that allostery is a byproduct of the global and collective nature of functional protein dynamics. Collectiveness means motions of all the reaction coordinates [34-45]—the coordinates that completely determine the progress of a reaction process—are concerted, thus change in one reaction coordinate causes change in other reaction coordinates. Globalness means coordination exists between reaction coordinates that are far apart. Therefore, allosteric site and remote site are residues containing reaction coordinates that are far apart. The allosteric pathway for conformational allostery is the spatial arrangements of reaction coordinates of protein conformational dynamics. Key residues for allostery are residues that contain reaction coordinates. Consequently, rigorous understanding of the mechanism of protein dynamics provides us with rigorous understanding of the mechanism of allostery.

The energy flow method [34, 35] we developed can rigorously determine reaction coordinates—they are coordinates with high magnitude energy flows. This enables us to determine allosteric pathway rigorously. The inter-coordinate energy flows between different reaction coordinates shows a network structure, which explains why reaction coordinates move in concerted manner and consequently provides the rigorous mechanism for allostery. This kind of information goes beyond existing studies of allostery.

1.1.1 Significance of PDZ domain. PDZ domains are common protein modules involved in communications between partners on signal transduction pathways [46, 47]. It is a representative system for dynamic allostery and attracted intensive activities [48-61]. NMR experiments showed significant reduction in sidechain fluctuations [48, 50, 55] in the ligand-bound state; time dependent FTIR experiment monitoring a photo-switchable construct that mimics ligand-binding alluded to a series of small conformational changes [16, 57, 62, 63]. Computationally, allosteric networks were identified with co-evolution analysis, correlation analysis of interaction energies and network analysis of spatial contacts between residues [18, 49, 51, 54, 58]. However, it is unclear if signal transduction is achieved by the small conformational changes during ligand binding or the altered sidechain fluctuations. It is also unclear how the computationally identified allosteric networks can lead to changes in structure or dynamics of PDZ domains. We propose to conduct energy flow analysis on ligand-binding and energy relaxation of PDZ2 domain. From these analyses, we can determine reaction coordinates for both processes and their underlying mechanisms, providing rigorous mechanism for signal transduction in PDZ domain.

1.1.2 Significance of non-active-site drug-resistant mutations of HIV-PR. By 2016, HIV has infected 78 million people and killed 39 million. There is no drug that can cure HIV infected people. The most formidable challenge in HIV treatment is drug resistant mutations, which can render current therapies ineffective within months [64, 65]. HIV-PR is a major drug target—9 out of the 25 FDA-approved HIV drugs are PR inhibitors. HIV-PR has a high mutation rate: mutations at 45 out of its 99 residues have drug resistance [66-79]. Among them, 11 are active-site mutations, the rest are non-active-site mutations, which can relegate the detrimental effects of the active-site mutations on the binding of HIV-PR to natural substrates, but not inhibitors. Measurements of thermodynamic, structural and kinetic parameters of various combinations of active-site and non-active site mutations have shown that the effect of active-site mutations is highly diminished without paired non-active-site mutations [73, 74, 78, 80-86]. Understanding the mechanism of the drug resistance of non-active-site mutations poses a formidable challenge. The enormous number of combinations of mutations means one can get little from case-by-case studies unless a reliable principle that can unify all these mutations can be identified. The actions of non-active-site mutations show the signature feature of allostery: change at the non-active-site caused changes in the active site. Although HIV-PR is not known as allosteric, its ligand binding involves flap motion [87, 88], a large-scale conformational dynamics. Our study on the flap opening of HIV-PR showed that this process is highly collective and global. Many residues for important active-site and non-active-site mutations contain reaction coordinates of this process; change in a non-active-site residue can easily induce correlated changes in the active site. Therefore, the reaction pathway for the flap dynamics is likely the allosteric pathway for the drug resistance of non-active-site mutations, which can be unraveled by energy flow analysis.

1.2 Significance of determining druggable locations in SARS-CoV-2 main protease and RNA dependent RNA polymerase. The global pandemic of SARS-CoV-2 has created an urgent need for effective drugs against this virus, but conventional processes for drug development take too long. This creates an urgent demand for unprecedented deep information that can effectively guide drug design and speed up drug development.

Identifying druggable locations, i.e. patches in the binding pocket that a drug should complementarily occupy, is crucial for effective drug discovery and therefore immediately needed. The traditional lock-and-key picture of ligand binding suggests that a drug that occupies all the fillable space in the binding pocket may lead to optimal binding, a premise often contradicted by practical experiences from drug development. This conceptual picture, therefore, cannot provide the critically needed guiding information for mapping druggable locations around a target site.

Based on the biophysical principle emerged from our recent study on the ligand binding process of HIV-1 protease, optimal druggable locations are binding site patches that, upon drug occupation, do not disrupt the collective and cooperative global protein motions that are displayed in the functional ligand binding pathway**.** Within the methodological framework of "energy flow" that we developed, these collective and cooperative global motions correspond to concerted global energy flow channels of protein functional dynamics.

Herein, a patch is druggable if its occupation facilitates global energy flows during ligand binding; it is not druggable if its occupation blocks the energy flows. For example, our analysis of the ligand-binding process of HIV-1 protease showed that the drug resistant mutations documented in the literature occur on residues that carry high-magnitude energy flows in the global protein dynamics that open up the binding site. This suggests that these mutations impede drug binding by disrupting the global energy flows in which they are prominent players. Therefore, identifying the important energy flow channels during drug binding will enable us to, for the first time, determine accurate druggable locations. The energy flow analysis method we developed allows us to map out all the energy flows of a dynamic process. Moreover, from energy flow analyses on different drug candidates, we can determine how interactions due to different moieties affect energy flows. This enables us to assess how different chemical moieties affect drug binding. Finally, non-binding-site residues that are important for energy flows are potential sites for drug resistant mutations, analogous to secondary mutations in HIV-1 protease. Such information is unprecedented and is expected to drastically enhance the success rate of drug design and shorten the time of drug development.

Atomic structures of two major drug targets, SARS-CoV-2 main protease has recently become available. We propose to use energy flow analysis to map out the energy flow channels during their ligand binding processes, aiming at three objectives: **1)** Map out druggable locations of SARS-CoV-2 main protease. **2)** Identify functional roles of different moieties of drug candidates: α-ketoamides for main protease. **3)** Identify non-binding-site residues that are important for energy flows in drug binding of these two proteins. They are potential sites for drug resistant mutations. The crucial information for drug design obtained from our proposed studies will be made publicly available to fill the urgent demand.

The proposed research focuses on delivering precise and reliable information urgently needed to guide drug design and speed up drug development against the SARS-CoV-2 virus. The map of druggable locations provides urgently needed guidance for designing the size and overall shape of drug candidates. Information on how different chemical moieties affect drug binding provides urgently needed guidance for optimizing the combination of moieties in a drug candidate. The map of potential sites for drug resistant mutations outside the binding pocket provides guidance for dealing with drug resistant mutations that are deemed to show up down the road of the continued fight against this virus.

1.3 Expected outcomes and plans for federal grant submissions**.** Results from the proposed studies will be part of preliminary studies for our incoming grant applications to federal agencies. The understanding of protein allostery will be used as foundations for funding opportunities at NIGMS via an investigator initiated R01 mechanism. The understanding of druggable locations in SARS-COV-2 main protease will be used as foundations for funding opportunities at NIAID via R01 mechanism.

2 Innovation

We provide a novel view that unifies allostery, druggable locations and functional protein dynamics. Therefore, rigorous understanding of the mechanism of the protein dynamics responsible for an allosteric process automatically unravels the rigorous mechanism of allostery and provide rigorous physical principles that determine druggable locations. This view naturally explains how the allostery shown in the non-active-site drug-resistant mutations of HIV-PR appears under the evolution pressure imposed by PR inhibitors. Our method of energy flow is built on Lagrange-Hamiltonian mechanics [34, 35], the fundamental mechanical law that governs protein dynamics, and provides rigorous answers to the three fundamental questions on reaction mechanisms. 1) What are the reaction coordinates? [37-40, 89] 2) How reaction coordinates move in concerted manner? 3) Why they move in concerted manner? We have shown that energy flows from fast into slow coordinates during activation, until adequate energy accumulates in the slow coordinates so they can cross the activation barrier [35]. Therefore, reaction coordinates are coordinates with high energy flows because they are the slow coordinates in the system during activation. Because energy flows show a network structure, reaction coordinates need to move in concerted manner in order to maximize energy flow efficiency. These answers go beyond the scopes of other existing studies on reaction mechanism.

3 Research Objectives and Milestones

3.1 Preliminary Results

We first demonstrate with example that the energy flow method can identify rigorously defined reaction coordinates and explain how and why reaction coordinates move in concerted manner, before explaining how to use it to determine allosteric pathways and elucidate mechanisms of allostery.

3.1.1 Basic framework of the energy flow theory. We define both potential (**PEF**) and kinetic (**KEF**) energy flows. The PEF through a coordinate q_i is its work: $\Delta W_i = \int_{q_i(t_1)}^{q_i(t_2)} F_i dq_i(t) =$

$-\int_{q_i(t_1)}^{q_i(t_2)} \frac{\partial U}{\partial q_i} dq_i(t)$, here F_i is the force on q_i and $U(\vec{q})$ is the system potential energy. This definition allows a rigorous decomposition of the change in system potential energy: $\Delta U = \sum_i \Delta W_i$. The value of ΔW_i indicates the potential energy cost of the motion of q_i costs. Furthermore, we define the PEF from q_j to q_i as $\Delta W_{j\to i} = -\int_{q_i(t_1)}^{q_i(t_2)} dq_i(t) \int_{q_j(t_1)}^{q_j(t)} \frac{\partial^2 U}{\partial q_i \partial q_j} dq_j(\sigma)$, which is the change in the PEF through q_i caused by motion of q_j. This definition leads to a rigorous decomposition of the PEF through a coordinate: $\Delta W_i = \sum_j \Delta W_{j\to i}$. The KEF through q_i is: $\Delta_v K_i = \int_{q_i(t_1)}^{q_i(t_2)} \frac{\partial K}{\partial q_i} dq_i(t) + \int_{\dot{q}_i(t_1)}^{\dot{q}_i(t_2)} \frac{\partial K}{\partial \dot{q}_i} d\dot{q}_i(t) = \int_{q_i(t_1)}^{q_i(t_2)} \frac{\partial K}{\partial q_i} dq_i(t) + \int_{\dot{q}_i(t_1)}^{\dot{q}_i(t_2)} p_i(t) d\dot{q}_i(t)$. It is the change in the system kinetic energy caused by changes in $(q_i, \dot{q}_i)$, which fully characterizes the motion of q_i. The PEF and KEF defined here are along individual trajectories. To determine reaction mechanism, we need to examine their averaged behavior in a reactive trajectory ensemble. We define the ensemble averaged energy flow projected onto a projector $\xi(\Gamma)$:

$$\langle \delta A(\xi^*) \rangle = \frac{\int d\Gamma \rho(\Gamma) \delta A(\xi(\Gamma) \to \xi(\Gamma) + d\xi) \delta(\xi(\Gamma) - \xi^*)}{\int d\Gamma \rho(\Gamma) \delta(\xi(\Gamma) - \xi^*)}; \ \langle \Delta A(\xi_1 \to \xi_2) \rangle = \int_{\xi_1}^{\xi_2} \langle \delta A(\xi) \rangle \quad (1).$$

Here, Γ is the phase space, ΔA can be ΔW_i, $\Delta \mathrm{W}_{\mathrm{j}\to\mathrm{i}}$ or $\Delta_v K_i$, $\delta(\cdots)$ is Dirac δ-function, and $\langle \cdots \rangle$ denotes average over the reactive trajectory ensemble. The projector $\xi(\Gamma)$ needs to adequately characterize the progress of the reaction process. A convenient choice is $\xi = p_B$, the so-called committor [37, 38], defined as the probability that a dynamic trajectory initiated from a given configuration, with momenta sampled from equilibrium distribution, to reach the product state. By definition, p_B is the reaction probability in configuration space, the default probabilistic language for specifying the progress of a reaction process.

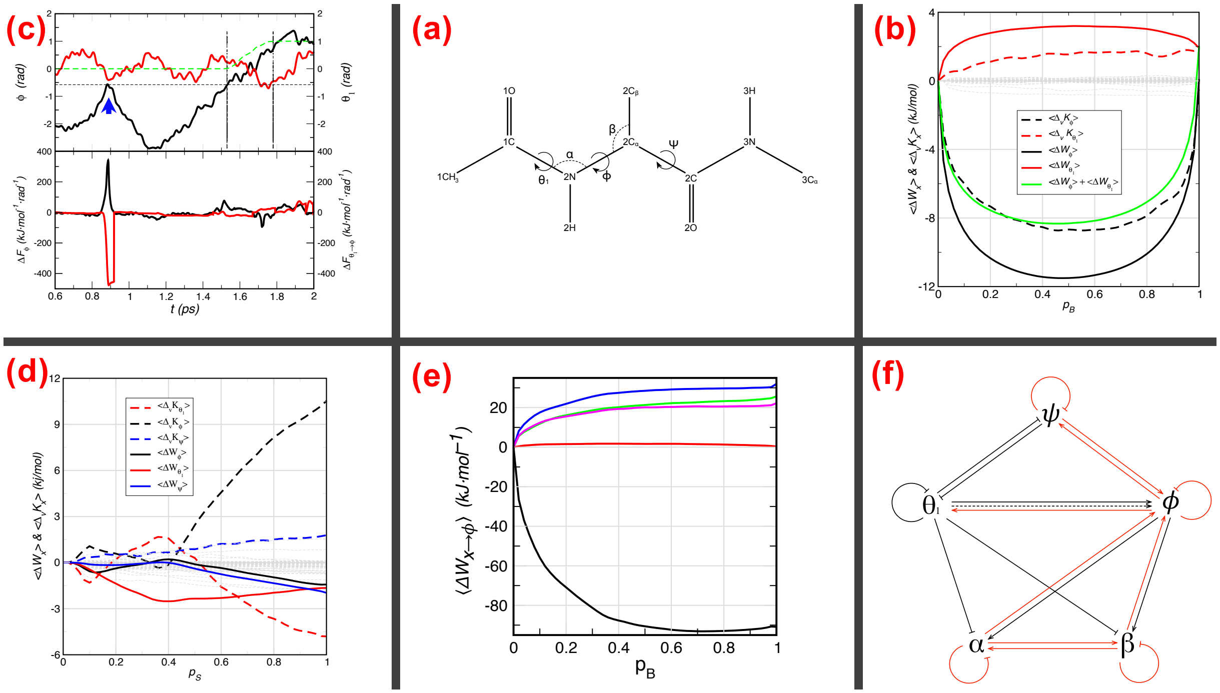


Fig. 1: Energy flows in isomerization of alanine dipeptide. (**a**) Schematic representation of an alanine dipeptide, with important coordinates marked. (**b**) PEFs (**solid**) and KEFs (**dashed**) through ϕ and θ_1 as a function of p_B. (**c**) Interactions between θ_1 and ϕ during a failed attempt of barrier crossing. The position of the failed attempt is marked by a blue arrow in the upper panel. (***upper***) Time evolution of ϕ (black), θ_1 (red) and p_B (green). (***lower***) Time evolution of F_ϕ (black) and $F_{\theta\to\phi}$ (red). (**d**) PEFs and KEFs through ϕ (black), θ_1 (red) and ψ (blue) as a function of p_S. (**e**) Energy flows from other coordinates to ϕ: $\langle \Delta \mathrm{W}_{\phi\to\phi} \rangle$ (black), $\langle \Delta W_{\theta_1\to\phi} \rangle$ (red), $\langle \Delta W_{\psi\to\phi} \rangle$ (green), $\langle \Delta W_{\alpha\to\phi} \rangle$ (blue), and $\langle \Delta W_{\beta\to\phi} \rangle$ (magenta). (**f**) A schematic diagram showing the network structure of PEFs among $\phi, \theta_1, \psi, \alpha, \beta$ during barrier crossing. PEFs are distinguished by their sign: positive (→); negative (⊣), and magnitude: high (red); small (black).

3.1.2 Energy flows can identify rigorously defined reaction coordinates. Reaction coordinates are the physical coordinates that can adequately determine the value of p_B for any configuration. The number of reaction coordinates is small compared to the total number of coordinates in the system. We use the $C_{7eq} \rightarrow C_{7ax}$ isomerization dynamics of an alanine dipeptide in vacuum as an example. This system is a prototype for protein dynamics because it is the smallest molecule with enough degrees of freedom to provide a heat bath adequate for activation. This is the defining feature that distinguishes a complex molecule such as a protein from a simple molecule such as $\mathrm{CO_2}$. Although this system appears small, its isomerization dynamics posed challenging questions when the objective is a rigorous understanding of its mechanism [35, 38, 40].

The major reaction coordinates for this process are two backbone dihedrals ϕ and θ_1 (Fig. 1a) [35, 38, 40]. Figure 1b showed that the PEFs through them are large, whereas the PEFs through the other coordinates are vanishingly small. This result demonstrates that coordinates with high magnitude energy flows are the reaction coordinates. The importance of different coordinates to a reaction process can be ranked by the magnitude of the energy flows through them.

The motions of ϕ and θ_1 are concerted during the isomerization. Figure 1b showed that $\langle \Delta W_\phi \rangle < 0$, suggesting that ϕ confronts the activation barrier. In contrast, $\langle \Delta W_{\theta_1} \rangle > 0$, suggesting that θ_1 gathers energy from other coordinates. On the other hand, Fig. 1b showed that $\langle \Delta_v K_\phi \rangle < 0, \langle \Delta_v K_{\theta_1} \rangle > 0$, but $|\langle \Delta_v K_\phi \rangle| < |\langle \Delta W_\phi \rangle|, |\langle \Delta_v K_{\theta_1} \rangle| > |\langle \Delta W_{\theta_1} \rangle|$ and $\langle \Delta_v K_\phi \rangle \simeq \langle \Delta W_\phi \rangle + \langle \Delta W_{\theta_1} \rangle$. Our analyses [35] showed that θ_1 transfers all the energy it receives in PEFs to ϕ by directly transferring kinetic energy to ϕ–θ_1 assists ϕ to cross the activation barrier. But this is one aspect of the coordination between the motions of θ_1 and ϕ.

Some trajectories in the reactive trajectory ensemble contain failed attempt of ϕ to cross the activation barrier. Figure 1c shows such an example. At time ~0.9ps, ϕ already reached the position the same as the beginning of the eventual successful barrier crossing. However, instead of proceeding to cross the barrier, it reversed its course. The reason is that θ_1 was at an incorrect position. Because of this mis-location of θ_1, θ_1 exerts a strong repelling force on ϕ to push away from crossing the barrier (Fig. 1c lower panel). Besides, this repelling force is responsive—even though the force that pushes ϕ to cross the barrier is much higher than what ϕ normally needs in a successful barrier crossing, the repelling force from θ_1 increases with it until exceeding it and ending up in pushing ϕ away. This observation suggests that there is a precise coordination between the motions of ϕ and θ_1—ϕ cannot start barrier crossing until θ_1 allows it to. Once θ_1 is ready, however, not only it will not prevent ϕ from barrier crossing, but it provides essential assistance to ϕ, as discussed above. This rigorous demonstration of the precise coordination between ϕ and θ_1 and the reason behind it cannot be obtained from the free energy surface of ϕ and θ_1, which is the standard bearer for demonstrating reaction mechanism in conventional approaches.

To understand the interaction between ϕ and θ_1 at this stage, we need systematic analysis. A reaction process consists of two steps, energy activation and barrier crossing [90]. The former is the process in which the reaction coordinates acquired sufficient energy to cross the activation barrier; it has to precede or be concurrent with barrier crossing. Committor is the proper quantity for parameterizing barrier crossing, but it is inadequate for describing energy activation. We define a new quantity, the reaction stability p_S, to parameterize the progress of the energy activation phase. The reaction stability is defined as the probability that a dynamic trajectory initiated from a phase space point on a reactive trajectory, with the momenta perturbed by a small amount ε (20% of the original values in this study), to remain a reactive trajectory. For instance, a phase space point with $p_S = 0.5$ means that if we initiate 100 trajectories from this point, each one with the momentum of each coordinate perturbed randomly by 20% of the original value, then about 50 trajectories will be reactive, whereas the other 50 will be non-reactive. By definition, p_S is the likelihood that a reactive trajectory will pass through the neighborhood surrounding a phase space point. In the energy flow analysis on the energy activation phase, we use p_S as the projector.

Figure 1d shows that, during energy activation phase, θ_1 needs to cross a small barrier around $p_S = 0.4$ before energy flow in ϕ can start. This result explains why θ_1 inhibits ϕ during its failed attempt for barrier crossing--θ_1 will inhibit ϕ motion before it crosses its own barrier at $p_S \simeq 0.4$ because energy flow through ϕ is not yet allowed. In addition, there is significant amount of kinetic energy accumulation in ϕ during the energy activation phase (Fig. 1d). The total amount roughly equals to the energy cost for its barrier crossing.

In addition to identifying the dominant reaction coordinates, the energy flow analysis also identified coordinates ψ, α, β as important players. Although they do not experience significant energy flows, they have significant energy exchange with ϕ and θ_1. Our analysis also demonstrated that the energy flows between different coordinates (i.e. $\Delta W_{j \to i}$) are closely related to the time scales of these coordinates—energy flows from the fast to the slow coordinates [35]. Through energy flows, the fast coordinates are slaved to the slower coordinates. In conclusion, although ψ, α, β are not reaction coordinates, they are essential for energy flows through the dominant reaction coordinates. We call them facilitators.

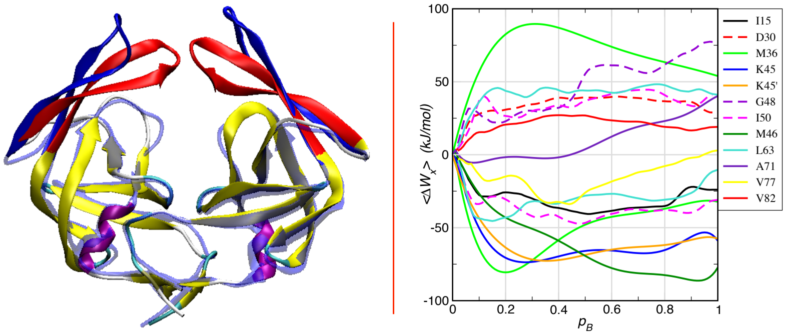


Fig. 2: (***Left***) Comparison of typical structures of HIV-PR in semi-open (flap in red) and open (flap in blue) states from TPS simulations of the flap-opening process. The two structures are overlapped by the part that change little during flap motion. (***Right***) Dependence of PEFs through some typical active-site (dashed) and non-active-site (solid) drug-resistant mutations on p_B. The PEFs are averaged over 350 reactive trajectories. We did not show all the PEFs due to limited space and emphasized more on non-active-site mutations.

The other coordinates in the system, in contrast, have neither significant energy flow through them nor significant energy exchange with the reaction coordinates. Instead, they provide energy to the reaction coordinates and the facilitators like ψ, α, β in small amounts. Because of their large number, the total energy they provided is significant. Therefore, they are the thermal bath in the standard model of reaction dynamics, analogous to buffer gas in gas-phase reactions and solvents in solution-phase reactions.

3.1.3 Residues of HIV-PR with drug-resistant mutations contain important reaction coordinates for the flap-opening process. Figure 2 shows PEFs for the flap opening of HIV-PR without ligand. An initial reactive trajectory was generated using MD simulation in implicit solvent following the procedure by Hornak et al [88]. Transition path sampling (**TPS**) was then used to proliferate reactive trajectories [37, 91]. Our results showed that this process is highly collective and global, as a large number of backbone dihedrals carry high PEFs. In contrast, PEFs through bonds, bond angles, or sidechain dihedrals are small. Interestingly, most residues with well-known drug-resistant mutations contain dihedrals with high energy flow, suggesting that their motions are concerted based on our understanding from the alanine dipeptide example. The PEFs also show more features than those from alanine dipeptide, suggesting more complicated mechanism in HIV-PR. The sidechain dihedrals are most likely facilitators like ψ, α, β in the alanine dipeptide example, but we need to finish the other steps of the energy flow analysis before we can pin down the detailed mechanisms.

3.2 Research Plans

3.2.1 Aim 1: Identifying allosteric pathway and mechanism for signal transduction in PDZ domain. It is unclear whether signal is transmitted by the PDZ domain through: 1) the small conformational change induced by ligand binding, or 2) the altered sidechain dynamic in the ligand-bound state. For the first possibility, the signal should be communicated during the ligand-binding process. For the second possibility, the signal should be from the ligand and energetic in nature. We propose to determine the mechanisms for signal transduction for both scenarios using energy flow analysis on two different types of dynamic processes.

Case 1: The allosteric signal is the small structural change induced by ligand binding. In this case, we need to analyze the energy flow and mechanism of the ligand binding process. We chose not to simulate the conformational dynamics in the photo-switchable construct because MD simulations in refs. [57, 63] indicated that it differs from the actual ligand-binding process, which is the real allosteric process used by the PDZ domain. This consists of three steps: 1) generate an unbiased reactive trajectory for the ligand binding process, 2) use TPS to generate reactive trajectory ensemble using the initial trajectory from step one, 3) perform energy flow analysis on reactive trajectory ensemble to determine the reaction coordinates and reaction mechanism in a way similar to what we demonstrated in alanine dipeptide above.

We will use the PDZ2 domain from human tyrosine-phosphatase 1E (PDB ID: 3LNX, 3LNY). The bias annealing method developed by the PI will be used to generate the initial natural reactive trajectory in step one [39, 92]. It was successfully applied to generating a natural reactive trajectory for the flipping of a nucleotide out of the active site of a DNA repair enzyme AGT. That process is a ligand-binding process, and AGT is a much larger protein than the PDZ domain. Therefore, we expect this method will be adequate for our purpose. Specifically, one or more order parameters (e.g. the distance between ligand and the active site, order parameters discussed in refs. [57, 63]) will be used to define two stable basins, one for the ligand-free and one for the ligand bound state. Initially, a biasing potential will be applied to these order parameters to realize a transition between the two states. In subsequent steps, the biasing potential will be gradually reduced, until transition between the two basins can happen without any biasing potential, which gives the initial natural reactive trajectory. This initial trajectory is the input to TPS, which have been used to proliferate reactive trajectories in many complex systems [37, 39, 91, 93-96]. The procedure for energy flow analysis is similar to the alanine dipeptide example, with details given in refs. [34, 35]. From the energy flow analysis, we expect to identify the correct reaction coordinates, and how and why they move together in a concerted manner, similar to what we have shown in the alanine dipeptide example.

Case 2: The allosteric signal is from the ligand and energetic in nature. In this case, the most relevant process is energy relaxation--the transmission of excess energy from the ligand to the other part of the PDZ domain. Here, the reduced thermal fluctuations in the ligand-bound state will likely lead to more efficient energy transmission, and consequently signal transduction, compared to the ligand-free state. We can determine the signal transduction pathway by analyzing the energy flows in the energy relaxation process in both ligand-bound and ligand-free states. Specifically, we will first perform equilibrium simulations in both ligand-bound and ligand-free states. From these simulations, we harvest an ensemble of representative equilibrium conformations for each state. For each equilibrium conformation, we first deposit extra kinetic energy into the ligand, then propagate the system using standard MD simulation. As simulation proceeds, the extra kinetic energy will propagate from the ligand through the PDZ domain. We will prepare an ensemble of energy relaxation trajectories. Then we will use energy flow analysis on them to identify the key coordinates and how energy propagates through them. The procedure for energy flow analysis is similar to that in the alanine dipeptide example. But the projector (ξ in Eq. (1)) we use will be: 1) the time lapse since the beginning of the relaxation process, or 2) the amount of excess kinetic energy left in the ligand. Both quantities provide a good parameterization of the progress of energy relaxation. From this analysis, we can identify the major coordinates responsible for transferring the excess kinetic energy of the ligand to the rest of the PDF domain, and how and why they move together in a concerted manner, similar to what we showed in the alanine dipeptide example.

As a fundamentally important phenomenon, energy relaxation in proteins have been examined both experimentally and computationally in the past [97-103]. However, previous studies used Cartesian coordinates in their analyses, which is not appropriate for unraveling mechanisms. The reason is that motions of Cartesian coordinates are dominated by strong constraint forces from bonded interactions, which are not related to energy transfer. In contrast, internal coordinates are the natural coordinates for protein dynamics, as they automatically took into account the constraints from chemical bonds. In fact, if we perform energy flow analysis on alanine dipeptide in Cartesian coordinates, the PEFs will be smeared over all coordinates and no mechanism can be extracted. In addition, the alanine dipeptide example demonstrated that PEFs and KEFs play distinct roles, but they are identical in Cartesian coordinates, another evidence that Cartesian coordinates are inappropriate for mechanistic analysis. Therefore, we can extract mechanisms of energy transmission that are more concrete and specific than what previous studies achieved.

3.2.2 Aim 2: Determine the mechanism for non-active-site drug resistant mutations in HIV-PR.

Non-active-site drug-resistant mutations show the signature feature of allostery. This must be a consequence of mutation-induced alterations in the ligand-binding process of HIV-PR. Therefore, understanding the mechanism of ligand-binding and the roles of different residues should unravel how non-active-site mutations can alter this process and lead to drug-resistance. There are two major components for ligand binding in HIV-PR: 1) flap-opening of the ligand-free HIV-PR and 2) flap closing when the ligand is already in the active site. Our initial analysis on the flap-opening process already showed that many sites of non-active-site mutations contain essential reaction coordinates for flap-opening. This means mutations of these residues can significantly alter ligand binding, though we do not yet know the exact effects of a particular mutation. Therefore, our first step is to finish the entire set of energy flow analysis similar to what we have shown in the alanine dipeptide example. This will enable us to completely determine the role of each coordinate and residue in the flap opening process. In addition, we can decompose ΔW_i into contributions from bonded, van der Waals and electrostatic interactions: $\Delta W_i = \Delta W_i^{bond} + \Delta W_i^{vdw} + \Delta W_i^{ele}$, because we can depose $F_i = F_i^{bond} + F_i^{vdw} + F_i^{ele}$. In this way, we can understand the effects of mutations. For example, if the PEF through a coordinate is dominated by electrostatic contribution, a mutation that changes the charges on the corresponding residue is likely to have a significant effect on ligand binding.

Based on results from this analysis, we will pick two combinations of non-active-site and active-site mutations that show strong drug resistance. We will then simulate the flap-opening in these two systems and perform energy flow analysis to determine the specific mechanisms of how mutations change flap opening.

In addition, we will perform energy flow analysis for the flap-closing in the presence of ligands. It has been shown that the flap-closing can be realized with straightforward MD simulations [104, 105]. From such simulations, we can obtain the initial reactive trajectories. Transition path sampling will then be used to proliferate reactive trajectories for the energy flow analysis. From these analyses, together with understandings from the flap opening process, we expect to be able to determine the exact mechanism how these specific non-active-site mutations achieve drug resistance. Furthermore, by comparing results of wild type and these two sets of mutations, we might be able to infer the general principle that governs non-active-site mutations. Otherwise, more mutation sets will be examined in the same way, so that we can conduct systematic induction to determine the mechanism for non-active site drug resistant mutations.

3.2.3 Aim 3: Map out the druggable locations for SARS-COV-2 main protease. Drug binding is controlled by the ligand-binding process and druggable locations are intrinsically related to the energy flow channels of the ligand binding process. Our basic assumption is that druggable locations are the sites that, when occupied, will facilitate energy flows for ligand binding dynamics. Therefore, the first step is to pin down the detailed mechanism of the ligand binding process. This consists of three steps, similar to discussed above: 1) use biased annealing method to generate an initial natural reactive trajectory for the ligand dissociation from the active site, 2) use TPS to proliferate the reactive trajectory to prepare a reactive trajectory ensemble, 3) use energy flow analysis on the reactive trajectory ensemble to identify the reaction coordinates and the relationships between different reaction coordinates. The starting structure for step one will be the crystal structure (Fig. 3) of the main protease with an inhibitor α-ketoamides in its active site. We will generate the natural reactive trajectory for the dissociation of one ligand from its active site. Since ligand binding and dissociation are reverse of each other, they share the same natural reactive trajectory.

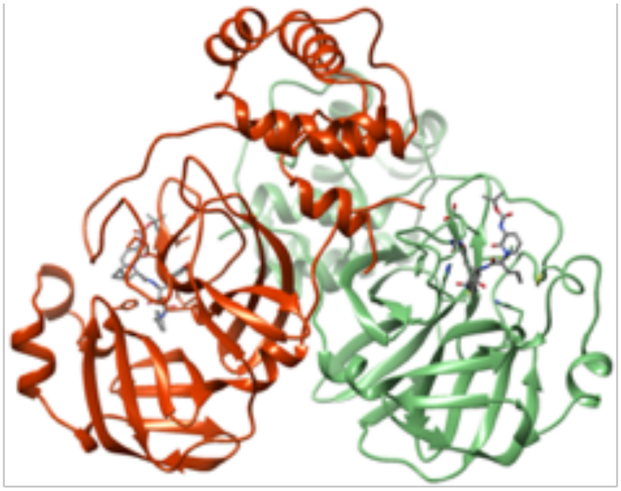


Fig. 3: Crystal structure of SARS-COV-2 main protease with the inhibitor α-ketoamides at the active sites.

From these results, we will first identify roles of active site residues in the ligand binding process. There are three possible situations: 1) A residue contains either reaction coordinates or the facilitator coordinates, and these coordinates contribute constructively to the overall energy flow—contributions from these coordinates are of the same sign as the overall energy flow. 2) A residue contains reaction coordinates, but these coordinates contribute negatively to the overall energy flow—contributions from these coordinates are of the opposite sign as the overall energy flow. 3) a residue does not contain reaction coordinates or important facilitator coordinates.

For the first case, the coordinates are important for proper energy flows during the ligand binding process. If their motions are constrained by interactions with moiety groups on a drug, then disruption in the energy flow is likely to happen as a result. The extent of this potential disruption of the energy flow is determined by the magnitude of energy flows through this coordinate. Therefore, these sites should not be occupied by a drug. For example, if a backbone dihedral of an active site residue is an important reaction coordinate in the ligand-binding process, then a good inhibitor for the main protease should not interact with atoms involved in this backbone dihedral. Similarly, if a sidechain dihedral is an important reaction coordinate, or an important facilitator coordinate that has significant energy exchange with reaction coordinates, then a good inhibitor for the main protease should not have strong interactions with atoms involved in this sidechain dihedral.

For the second case, the coordinates are making contributions opposite to the overall energy flow. If their motions are constrained by interactions with a drug, the effects need to be evaluated. For this purpose, we will harvest natural reactive trajectories for the ligand dissociation process with these coordinates constrained to their values in the ligand-bound state. Then energy flow analysis will be conducted to determine their effect on the overall energy flow. If their constraints lead to disruption of the overall energy flow, then atoms associated to these coordinates should not be occupied by a drug. On the other hand, if their constraints do not impact the overall energy flow, then atoms associated with these coordinate are druggable locations.

For the third case, a coordinate of a residue in the active site is not a reaction coordinate or important facilitator, but instead belongs to the heat bath, then its occupation by a drug will not disrupt energy flows during ligand binding. Atoms associated with these coordinates are druggable locations.

Determine the role of different moiety groups of drug candidate α-ketoamides. For this purpose, we need to decompose the energy flows that involve interactions with α-ketoamides into contributions from its different moiety groups. This can be easily achieved since the interaction between α-ketoamides and the main protease is a sum of interactions between its moiety groups and the main protease. In addition, the contribution from each moiety group can be further decomposed into contributions from electrostatic and van der Waals interactions, similar to what we discussed in Aim 2.

From these results, we can determine how much each chemical feature of each moiety group contribute to the energy flows for the ligand binding process. For example, if the contribution from a moiety group is in the same sign as the overall energy flow and this contribution is due to electrostatic interactions with the active site atoms, then modifications to this moiety that increases its electrostatic interactions with the active site atoms will enhance the binding affinity of the drug. In contrast, if the contribution from this moiety group is of the opposite sign as the overall energy flow, then modifications that enhances its electrostatic interactions with the active site atoms will lead to a less effective drug compared to the original compound.

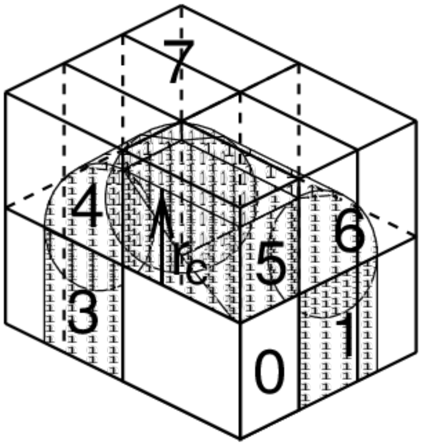
Identify potential non-active-site residues for potential drug resistant mutations. Similar to the case of HIV-PR, we expect a number of sidechain dihedrals of non-active-site residues will either be reaction coordinates or important facilitator coordinates for the ligand binding process. To determine potential effects of mutations on these residues, we will decompose the contributions from these sidechains to the overall energy flows into electrostatic and van der Waals components, similar to what we described in Aim 2. From these results, we can estimate how mutations to these residues may impact drug binding. For example, if the contribution from a non-active-site residue sidechain is of the same sign as the overall energy flow and it is mainly due to van der Waals interactions, the a mutation that reduces the size of the side chain will likely damage the overall energy flow of the drug binding process. Such a mutation will be potentially drug resistant. In contrast, if the contribution from this residue sidechain is of the opposite sign as the overall energy flow, then a mutation that decreases its van der Waals interactions with other residues will likely have no significant detrimental effect on drug binding.

4 Computational Readiness

Transition path sampling, committor and energy flow calculations

The basic procedure for TPS calculation is to shoot independent trajectories from different locations of an existing reactive trajectory that can connect the reactant and the product states. Among these shooting trajectories, only the reactive ones are harvested. Therefore, the basic units of TPS calculations are independent MD trajectories. The basic procedure for committor (p_B) calculations is to run independent MD trajectories from configurations on a TPS trajectory and count the number of trajectories that reached the product state. Therefore, the basic units of solvation dynamics calculations are also independent MD trajectories. The basic procedure for energy flow calculations is to compute the work and kinetic energy change during each MD step, using formula given in the Preliminary Results section. The basic units of energy flow calculations are independent system configurations from TPS calculations. The codes for solvation dynamics calculations are standard MD simulations. The codes for both TPS and energy flow calculations are implemented in the MD simulation software GROMACS. These implementations were developed and tested using time from our ALCF DD allocation and will be thoroughly utilized to achieve our Milestone.

GROMACS provides extremely high performance compared to other programs. A lot of algorithmic optimizations have been introduced in the code; they have for instance extracted the calculation of the virial from the innermost loops over pairwise interactions, and they use their own software routines to calculate the inverse square root. In GROMACS 4.6 and up, on almost all common computing platforms, the innermost loops are written in C using intrinsic functions that the compiler transforms to SIMD machine instructions, to utilize the available instruction-level parallelism. These kernels are available in either single and double precision, and in support all the different kinds of SIMD support found in x86-family (and other) processors. The following summarizes the multi-level single node parallelization capability of Gromacs:



Thread-MPI, OpenMP

Used in parallelization within a *node*, multithreading enables efficient use of multicore CPUs. Multithreading was first introduced in GROMACS 4.5 based on thread-MPI library which provides a threading-based MPI implementation. OpenMP-based multithreading is supported with GROMACS 4.6 and can be combined with (thread-)MPI parallelization.

Accelerated code, SSE, AVX, CUDA

To achieve high computational efficiency, GROMACS uses both CPU- and GPU-based acceleration. The most compute-intensive parts of the code are implemented as accelerated compute kernels for CPU using SSE or AVX and for GPUs using CUDA.

Heterogeneous parallelization (CPU + accelerator)

A hybrid or heterogeneous parallelization makes use of multiple different computational units, typically CPUs and GPUs. With the native GPU acceleration support, GROMACS 4.6 introduces hybrid parallelization.

Hybrid/multi-level parallelization

Consists of the use of multiple parallelization schemes on different hardware levels typically separating intra- and inter-node parallelization. GROMACS uses OpenMP multithreading for intra-node multicore-targeting parallelization and MPI for inter-node.

Gromacs employs efficient eighth-shell domain/force decomposition and MPI communication library, summarized as following:

Parallelization

The CPU time required for a simulation can be reduced by running the simulation in parallel over more than one core. Ideally, one would want to have linear scaling: running on *N* cores makes the simulation *N* times faster. In practice this can only be achieved for a small number of cores. The scaling will depend a lot on the algorithms used. Also, different algorithms can have different restrictions on the interaction ranges between atoms.

Domain decomposition

Since most interactions in molecular simulations are local, domain decomposition is a natural way to decompose the system. In domain decomposition, a spatial domain is assigned to each rank, which will then integrate the equations of motion for the particles that currently reside in its local domain. With domain decomposition, there are two choices that have to be made: the division of the unit cell into domains and the assignment of the forces to domains. Most molecular simulation packages use the half-shell method for assigning the forces. But there are two methods that always require less communication: the eighth shell and the midpoint method. GROMACS currently uses the eighth shell method, but for certain systems or hardware architectures it might be advantageous to use the midpoint method. Therefore, we might implement the midpoint method in the future. Most of the details of the domain decomposition can be found in the GROMACS 4 paper.

Coordinate and force communication

In the most general case of a triclinic unit cell, the space in divided with a 1-, 2-, or 3-D grid in parallelepipeds that we call domain decomposition cells. Each cell is assigned to a particle-particle rank. The system is partitioned over the ranks at the beginning of each MD step in which neighbor searching is performed. The minimum unit of partitioning can be an atom, or a charge group with the (deprecated) group cut-off scheme or an update group. An update group is a group of atoms that has dependencies during update, which occurs when using constraints and/or virtual sites. Thus different update groups can be updated independenly. Currently update groups can only be used with at most two sequential constraints, which is the case when only constraining bonds involving hydrogen atoms. The advantages of update groups are that no communication is required in the update and that this allows updating part of the system while computing forces for other parts. Atom groups are assigned to the cell where their center of geometry resides. Before the forces can be calculated, the coordinates from some neighboring cells need to be communicated, and after the forces are calculated, the forces need to be communicated in the other direction. The communication and force assignment is based on zones that can cover one or multiple cells. An example of a zone setup is shown in the figure.

On Theta, we utilize the binary of latest version Gromacs that provides all the above parallelization capability. ALCF has provided detailed online documentation to run Gromacs jobs optimally on Theta. Basically, both TPS and energy flow calculations are loosely coupled ensemble calculations and fully utilizes the massive parallel nodes of Theta. They involve thousands of uncoupled concurrent trajectories or configurations, each one is an independent instance of GROMACS job. Therefore, the computational performance exhibits perfect scaling for up to 2,048 nodes. We are already able to run 2048 nodes jobs on Theta by utilizing the multiple concurrent tasks mode. Below are the benchmark data of one trajectory generated within TPS and energy flow calculations for the HIV-PR system.

Number of nodes used	**Computational time per ps of TPS trajectory**
1	15.84 s
2	7.92 s
3	4.4 s
4	2.7 s

Table 1: Computational time per picosecond of TPS trajectory for HIV-PR

Number of nodes used	Computational time per system configuration
1	19.17 s
128	0.150 s
256	0.075 s
512	0.0375 s
2048	0.0095 s

Table 2: Computational time per system configuration for energy flow calculation for HIV-PR

Job Characterization

The TPS is the standard calculation to be performed. There are two types of standard calculations to be performed: 1) TPS/p_B calculation, 2) energy flow calculation. For any given system a baseline computation time can be calculated. For a system, 1000 trajectories will be used and computation is performed on Theta. The trajectory requirements are discussed above in single trajectory benchmark. In addition, we have three systems: 1) PDZ domain, 2) HIV-PR, 3) SARS-COV-2 main protease. The computational cost per TPS trajectory for PDZ domain and HIV-PR are similar. The computational cost for SARS-COV-2 main protease, due to its larger size, is about 4 times that of the HIV-PR, which is our baseline for estimating computational cost. Therefore, our run totals on Theta require 6 TPS calculations of approximately the same system size and number of trajectories. For the current systems each TPS trajectory can strongly scales to 8 nodes or 512 cores.

For each TPS trajectory, there are two types of associated calculations: 1) energy flow calculations, 2) p_B calculations. The energy flow calculation has perfect scaling for any number of nodes, so its computational cost can be conveniently incorporated into the cost for each TPS trajectory to make the overall counting of the computational more convenient. Each TPS trajectory will take 40 ps production runs. Among this 40 ps, 10 ps will be used in energy flow calculations, which amounts to energy flow calculation for 10,000 configurations per trajectory, as there is 1 configuration per femtosecond. In addition, for this 10 ps, there will be a p_B calculation every 50 fs, which amounts to 200 p_B calculations. Each p_B calculation requires shooting 40 MD trajectories, each one ~40 ps in length on average. This means each TPS trajectory will involve MD simulations of 1,000 trajectories, each one is ~40 ps in length. Taking into consideration all these factors and based on data given in Tables 1 and 2, the total computational cost associated with each TPS trajectory amount to the calculation of 200 x 40 = 8,000 MD trajectories with 40 ps in length. Therefore, each TPS job amounts to 8,000,000 trajectories, each trajectory needs 0.004882 x 8 node hours. The following formula lists the characterization time for the 6 TPS calculations:

6 TPS runs * 8,000,000 trajectories/TPS run * 8 nodes * 0.004882 hours = 1,874,688 node hours

The size and properties of the TPS runs could cause a small runtime fluctuation of +/- 5%.
In summary, totally we request 1,875,000 node hours including testing/debugging time.

Development Work

As GROMACS already has powerful GPU implementation on Both Nvidia (CUDA) and AMD GPU (OpenCL), we don't have plan to do low level GPU coding work.

Data Storage

The total space needed is 60 Terabytes. We estimate that this is needed to store trajectory data for two systems at a time. For the energy flow analysis, we need to store the positions and velocities of all the coordinates of a system at 1 or 2 fs time interval. This amounts to 20 GB for the HIV-PR and 40 GB for the SARS-COV-2 main protease system. In each case, we need about 500 reactive trajectories to obtain converged results in energy flow analysis. We plan to simultaneously simulate two systems at a time, thus we need about 60 TB storage space to store the trajectory data.

Restart I/O

The pipeline is particularly efficient with respect to restart I/O. Because of its loosely-coupled nature, there is no massive state that needs to be saved or coordinated at any stage. Each member of each workflow stage is independent, and restarts are based on the job completion state information. Small numbers of in-flight completion messages may be lost in certain failure scenarios, but in the workflows described here such jobs can be re-executed with no ill effects.

Analysis I/O

The present protein-ligand binding free energy calculation doesn't generate any special analysis files. Analysis is performed based on standard text output files from each independent serial or parallel application invocation. These text files will be efficiently batched and passed on for analysis using the WHAM postprocessing in Linux shell.

Data Transfer I/O

We don't have special needs for data transfer I/O. The size of input files for FEP/REMD calculations can be transferred to Mira via general scp. We do not anticipate any challenges with relocating data for long-term storage. The outputted files will be a combination of small text and binary files.

Personnel Justification

PI: Dr. Ao Ma is a theoretical biophysicist with significant expertise in developing theoretical and computational methods for understanding functional protein conformational dynamics. He is responsible for designing research plan and simulations, supervising the progress of simulations, analyzing the simulation data and revise manuscripts.

Three Research Assistants: The three graduate assistants (Shanshan Wu, Kenneth Tsui and Huiyu Li) will participate in and learn designing research, develop computer programs to carry out the simulations, perform simulations, analyze the simulation data and write the initial manuscripts.

Milestone Table

Year 1		
Milestone:	**Details** (as appropriate):	**Dates**:
Energy flow analysis on the allostery in PDZ domain	**Resource**: Theta **Node-hours**: 0.3125 M **Filesystem storage** (**TB**): 20 TB **Software Application**: GROMACS 2020.2 **Tasks**: loosely coupled ensemble computations on 2048 nodes of Theta **Dependencies**: No dependencies, start immediately	07/01/2020-07/01/2021
Energy flow analysis on the allostery in HIV-PR	**Resource**: Theta **Node-hours**: 0.3125 M **Filesystem storage** (**TB**): 20 TB **Software Application**: GROMACS 2020.2 **Tasks**: loosely coupled ensemble computations on 2048 nodes of Theta **Dependencies**: No dependencies, start immediately	07/01/2020-07/01/2021
Energy flow analysis on the druggable locations in SARS-COV-2 main protease	**Resource**: Theta **Node-hours**: 1.25 M **Filesystem storage** (**TB**): 20 TB **Software Application**: GROMACS 2020.2 **Tasks**: loosely coupled ensemble computations on 2048 nodes of Theta **Dependencies**: No dependencies, start immediately	07/01/2020-07/01/2021

Literature Cited

1. Tsai, C.J., A. del Sol, and R. Nussinov, *Allostery: absence of a change in shape does not imply that allostery is not at play.* J Mol Biol, 2008. **378**(1): p. 1-11.
2. Tsai, C.J., A. Del Sol, and R. Nussinov, *Protein allostery, signal transmission and dynamics: a classification scheme of allosteric mechanisms.* Mol Biosyst, 2009. **5**(3): p. 207-16.
3. Tsai, C.J. and R. Nussinov, *A unified view of "how allostery works".* PLoS Comput Biol, 2014. **10**(2): p. e1003394.
4. Wodak, S.J., et al., *Allostery in Its Many Disguises: From Theory to Applications.* Structure, 2019. **27**(4): p. 566-578.
5. Guo, J. and H.X. Zhou, *Protein Allostery and Conformational Dynamics.* Chem Rev, 2016. **116**(11): p. 6503-15.
6. Gunasekaran, K., B. Ma, and R. Nussinov, *Is allostery an intrinsic property of all dynamic proteins?* Proteins, 2004. **57**(3): p. 433-43.
7. Hilser, V.J., J.O. Wrabl, and H.N. Motlagh, *Structural and energetic basis of allostery.* Annu Rev Biophys, 2012. **41**: p. 585-609.
8. Motlagh, H.N., et al., *The ensemble nature of allostery.* Nature, 2014. **508**(7496): p. 331-9.
9. Tzeng, S.R. and C.G. Kalodimos, *Protein dynamics and allostery: an NMR view.* Curr Opin Struct Biol, 2011. **21**(1): p. 62-7.
10. Goodey, N.M. and S.J. Benkovic, *Allosteric regulation and catalysis emerge via a common route.* Nat Chem Biol, 2008. **4**(8): p. 474-82.
11. Kern, D. and E.R. Zuiderweg, *The role of dynamics in allosteric regulation.* Curr Opin Struct Biol, 2003. **13**(6): p. 748-57.
12. Taylor, S.S. and A.P. Kornev, *Protein kinases: evolution of dynamic regulatory proteins.* Trends Biochem Sci, 2011. **36**(2): p. 65-77.
13. Frederick, K.K., et al., *Conformational entropy in molecular recognition by proteins.* Nature, 2007. **448**(7151): p. 325-9.
14. Cooper, A. and D.T. Dryden, *Allostery without conformational change. A plausible model.* Eur Biophys J, 1984. **11**(2): p. 103-9.
15. McLeish, T.C., T.L. Rodgers, and M.R. Wilson, *Allostery without conformation change: modelling protein dynamics at multiple scales.* Phys Biol, 2013. **10**(5): p. 056004.
16. Stock, G. and P. Hamm, *A non-equilibrium approach to allosteric communication.* Philos Trans R Soc Lond B Biol Sci, 2018. **373**(1749).
17. Suel, G.M., et al., *Evolutionarily conserved networks of residues mediate allosteric communication in proteins.* Nat Struct Biol, 2003. **10**(1): p. 59-69.
18. Lockless, S.W. and R. Ranganathan, *Evolutionarily conserved pathways of energetic connectivity in protein families.* Science, 1999. **286**(5438): p. 295-9.
19. Selvaratnam, R., et al., *Mapping allostery through the covariance analysis of NMR chemical shifts.* Proc Natl Acad Sci U S A, 2011. **108**(15): p. 6133-8.
20. Boulton, S. and G. Melacini, *Advances in NMR Methods To Map Allosteric Sites: From Models to Translation.* Chem Rev, 2016. **116**(11): p. 6267-304.
21. Hunenberger, P.H., A.E. Mark, and W.F. van Gunsteren, *Fluctuation and cross-correlation analysis of protein motions observed in nanosecond molecular dynamics simulations.* J Mol Biol, 1995. **252**(4): p. 492-503.
22. Sethi, A., et al., *Dynamical networks in tRNA:protein complexes.* Proc Natl Acad Sci U S A, 2009. **106**(16): p. 6620-5.
23. Amaro, R.E., et al., *A network of conserved interactions regulates the allosteric signal in a glutamine amidotransferase.* Biochemistry, 2007. **46**(8): p. 2156-73.

24. Vanwart, A.T., et al., *Exploring residue component contributions to dynamical network models of allostery.* J Chem Theory Comput, 2012. **8**(8): p. 2949-2961.
25. Palermo, G., et al., *Key role of the REC lobe during CRISPR-Cas9 activation by 'sensing', 'regulating', and 'locking' the catalytic HNH domain.* Q Rev Biophys, 2018. **51**.
26. Palermo, G., et al., *Protospacer Adjacent Motif-Induced Allostery Activates CRISPR-Cas9.* J Am Chem Soc, 2017. **139**(45): p. 16028-16031.
27. Rivalta, I., et al., *Allosteric pathways in imidazole glycerol phosphate synthase.* Proc Natl Acad Sci U S A, 2012. **109**(22): p. E1428-36.
28. Rivalta, I., G.W. Brudvig, and V.S. Batista, *Oxomanganese complexes for natural and artificial photosynthesis.* Curr Opin Chem Biol, 2012. **16**(1-2): p. 11-8.
29. Hertig, S., N.R. Latorraca, and R.O. Dror, *Revealing Atomic-Level Mechanisms of Protein Allostery with Molecular Dynamics Simulations.* PLoS Comput Biol, 2016. **12**(6): p. e1004746.
30. Aoto, P.C., B.T. Martin, and P.E. Wright, *NMR Characterization of Information Flow and Allosteric Communities in the MAP Kinase p38gamma.* Sci Rep, 2016. **6**: p. 28655.
31. Lu, H.M. and J. Liang, *Perturbation-based Markovian transmission model for probing allosteric dynamics of large macromolecular assembling: a study of GroEL-GroES.* PLoS Comput Biol, 2009. **5**(10): p. e1000526.
32. Kong, Y. and M. Karplus, *The signaling pathway of rhodopsin.* Structure, 2007. **15**(5): p. 611-23.
33. Kong, Y., et al., *The allosteric mechanism of yeast chorismate mutase: a dynamic analysis.* J Mol Biol, 2006. **356**(1): p. 237-47.
34. Li, W. and A. Ma, *A benchmark for reaction coordinates in the transition path ensemble.* J Chem Phys, 2016. **144**(13): p. 134104.
35. Li, W. and A. Ma, *Reaction mechanism and reaction coordinates from the viewpoint of energy flow.* J Chem Phys, 2016. **144**(11): p. 114103.
36. Du, R., et al., *On the transition coordinate for protein folding.* Journal of Chemical Physics, 1998. **108**(1): p. 334-350.
37. Bolhuis, P.G., et al., *Transition path sampling: throwing ropes over rough mountain passes, in the dark.* Annu Rev Phys Chem, 2002. **53**: p. 291-318.
38. Bolhuis, P.G., C. Dellago, and D. Chandler, *Reaction coordinates of biomolecular isomerization.* Proc Natl Acad Sci U S A, 2000. **97**(11): p. 5877-82.
39. Hu, J., A. Ma, and A.R. Dinner, *A two-step nucleotide-flipping mechanism enables kinetic discrimination of DNA lesions by AGT.* Proc Natl Acad Sci U S A, 2008. **105**(12): p. 4615-20.
40. Ma, A. and A.R. Dinner, *Automatic method for identifying reaction coordinates in complex systems.* J Phys Chem B, 2005. **109**(14): p. 6769-79.
41. Li, W. and A. Ma, *Recent developments in methods for identifying reaction coordinates.* Mol Simul, 2014. **40**(10-11): p. 784-793.
42. Antoniou, D., et al., *Computational and theoretical methods to explore the relation between enzyme dynamics and catalysis.* Chem Rev, 2006. **106**(8): p. 3170-87.
43. Antoniou, D., et al., *Mass Modulation of Protein Dynamics Associated with Barrier Crossing in Purine Nucleoside Phosphorylase.* J Phys Chem Lett, 2012. **3**(23): p. 3538-3544.
44. Basner, J.E. and S.D. Schwartz, *How enzyme dynamics helps catalyze a reaction in atomic detail: a transition path sampling study.* J Am Chem Soc, 2005. **127**(40): p. 13822-31.
45. Schwartz, S.D. and V.L. Schramm, *Enzymatic transition states and dynamic motion in barrier crossing.* Nat Chem Biol, 2009. **5**(8): p. 551-8.
46. Lee, H.J., et al., *Identification of tripeptides recognized by the PDZ domain of Dishevelled.* Bioorg Med Chem, 2009. **17**(4): p. 1701-8.

47. Lee, H.J. and J.J. Zheng, *PDZ domains and their binding partners: structure, specificity, and modification.* Cell Commun Signal, 2010. **8**: p. 8.
48. Gianni, S., et al., *The kinetics of PDZ domain-ligand interactions and implications for the binding mechanism.* J Biol Chem, 2005. **280**(41): p. 34805-12.
49. Ota, N. and D.A. Agard, *Intramolecular signaling pathways revealed by modeling anisotropic thermal diffusion.* J Mol Biol, 2005. **351**(2): p. 345-54.
50. Gianni, S., et al., *Demonstration of long-range interactions in a PDZ domain by NMR, kinetics, and protein engineering.* Structure, 2006. **14**(12): p. 1801-9.
51. Sharp, K. and J.J. Skinner, *Pump-probe molecular dynamics as a tool for studying protein motion and long range coupling.* Proteins, 2006. **65**(2): p. 347-61.
52. Gianni, S., et al., *A PDZ domain recapitulates a unifying mechanism for protein folding.* Proc Natl Acad Sci U S A, 2007. **104**(1): p. 128-33.
53. Ivarsson, Y., et al., *An on-pathway intermediate in the folding of a PDZ domain.* J Biol Chem, 2007. **282**(12): p. 8568-72.
54. Kong, Y. and M. Karplus, *Signaling pathways of PDZ2 domain: a molecular dynamics interaction correlation analysis.* Proteins, 2009. **74**(1): p. 145-54.
55. Petit, C.M., et al., *Hidden dynamic allostery in a PDZ domain.* Proc Natl Acad Sci U S A, 2009. **106**(43): p. 18249-54.
56. Gianni, S., et al., *Sequence-specific long range networks in PSD-95/discs large/ZO-1 (PDZ) domains tune their binding selectivity.* J Biol Chem, 2011. **286**(31): p. 27167-75.
57. Buchenberg, S., F. Sittel, and G. Stock, *Time-resolved observation of protein allosteric communication.* Proc Natl Acad Sci U S A, 2017. **114**(33): p. E6804-E6811.
58. Kumawat, A. and S. Chakrabarty, *Hidden electrostatic basis of dynamic allostery in a PDZ domain.* Proc Natl Acad Sci U S A, 2017. **114**(29): p. E5825-E5834.
59. Stucki-Buchli, B., et al., *2D-IR Spectroscopy of an AHA Labeled Photoswitchable PDZ2 Domain.* J Phys Chem A, 2017. **121**(49): p. 9435-9445.
60. Gautier, C., et al., *Seeking allosteric networks in PDZ domains.* Protein Eng Des Sel, 2018. **31**(10): p. 367-373.
61. Gulzar, A., et al., *Energy Transport Pathways in Proteins: A Non-equilibrium Molecular Dynamics Simulation Study.* J Chem Theory Comput, 2019. **15**(10): p. 5750-5757.
62. Buchli, B., et al., *Kinetic response of a photoperturbed allosteric protein.* Proc Natl Acad Sci U S A, 2013. **110**(29): p. 11725-30.
63. Buchenberg, S., et al., *Long-range conformational transition of a photoswitchable allosteric protein: molecular dynamics simulation study.* J Phys Chem B, 2014. **118**(47): p. 13468-76.
64. Ali, A., et al., *Structure-based design, synthesis, and structure-activity relationship studies of HIV-1 protease inhibitors incorporating phenyloxazolidinones.* J Med Chem, 2010. **53**(21): p. 7699-708.
65. Ghosh, A.K., H.L. Osswald, and G. Prato, *Recent Progress in the Development of HIV-1 Protease Inhibitors for the Treatment of HIV/AIDS.* J Med Chem, 2016. **59**(11): p. 5172-208.
66. Bastys, T., et al., *Non-active site mutants of HIV-1 protease influence resistance and sensitisation towards protease inhibitors.* Retrovirology, 2020. **17**(1): p. 13.
67. Prabu-Jeyabalan, M., E. Nalivaika, and C.A. Schiffer, *How does a symmetric dimer recognize an asymmetric substrate? A substrate complex of HIV-1 protease.* J Mol Biol, 2000. **301**(5): p. 1207-20.
68. Prabu-Jeyabalan, M., E. Nalivaika, and C.A. Schiffer, *Substrate shape determines specificity of recognition for HIV-1 protease: analysis of crystal structures of six substrate complexes.* Structure, 2002. **10**(3): p. 369-81.

69. Kurt, N., et al., *Cooperative fluctuations of unliganded and substrate-bound HIV-1 protease: a structure-based analysis on a variety of conformations from crystallography and molecular dynamics simulations.* Proteins, 2003. **51**(3): p. 409-22.

70. Mitsuya, Y., et al., *N88D facilitates the co-occurrence of D30N and L90M and the development of multidrug resistance in HIV type 1 protease following nelfinavir treatment failure.* AIDS Res Hum Retroviruses, 2006. **22**(12): p. 1300-5.

71. Altman, M.D., et al., *HIV-1 protease inhibitors from inverse design in the substrate envelope exhibit subnanomolar binding to drug-resistant variants.* J Am Chem Soc, 2008. **130**(19): p. 6099-113.

72. Lefebvre, E. and C.A. Schiffer, *Resilience to resistance of HIV-1 protease inhibitors: profile of darunavir.* AIDS Rev, 2008. **10**(3): p. 131-42.

73. Mahalingam, B., et al., *Structural implications of drug-resistant mutants of HIV-1 protease: high-resolution crystal structures of the mutant protease/substrate analogue complexes.* Proteins, 2001. **43**(4): p. 455-64.

74. Mahalingam, B., et al., *Combining mutations in HIV-1 protease to understand mechanisms of resistance.* Proteins, 2002. **48**(1): p. 107-16.

75. Chen, X., I.T. Weber, and R.W. Harrison, *Molecular dynamics simulations of 14 HIV protease mutants in complexes with indinavir.* J Mol Model, 2004. **10**(5-6): p. 373-81.

76. Tie, Y., et al., *Molecular basis for substrate recognition and drug resistance from 1.1 to 1.6 angstroms resolution crystal structures of HIV-1 protease mutants with substrate analogs.* FEBS J, 2005. **272**(20): p. 5265-77.

77. Louis, J.M., et al., *HIV-1 protease: structure, dynamics, and inhibition.* Adv Pharmacol, 2007. **55**: p. 261-98.

78. Liu, F., et al., *Effect of flap mutations on structure of HIV-1 protease and inhibition by saquinavir and darunavir.* J Mol Biol, 2008. **381**(1): p. 102-15.

79. Sayer, J.M., et al., *Effect of the active site D25N mutation on the structure, stability, and ligand binding of the mature HIV-1 protease.* J Biol Chem, 2008. **283**(19): p. 13459-70.

80. Liu, F., et al., *Kinetic, stability, and structural changes in high-resolution crystal structures of HIV-1 protease with drug-resistant mutations L24I, I50V, and G73S.* J Mol Biol, 2005. **354**(4): p. 789-800.

81. Clemente, J.C., et al., *Comparing the accumulation of active- and nonactive-site mutations in the HIV-1 protease.* Biochemistry, 2004. **43**(38): p. 12141-51.

82. Svicher, V., et al., *Novel human immunodeficiency virus type 1 protease mutations potentially involved in resistance to protease inhibitors.* Antimicrob Agents Chemother, 2005. **49**(5): p. 2015-25.

83. Velazquez-Campoy, A., et al., *Thermodynamic dissection of the binding energetics of KNI-272, a potent HIV-1 protease inhibitor.* Protein Sci, 2000. **9**(9): p. 1801-9.

84. Todd, M.J., et al., *Thermodynamic basis of resistance to HIV-1 protease inhibition: calorimetric analysis of the V82F/I84V active site resistant mutant.* Biochemistry, 2000. **39**(39): p. 11876-83.

85. Luque, I., et al., *Molecular basis of resistance to HIV-1 protease inhibition: a plausible hypothesis.* Biochemistry, 1998. **37**(17): p. 5791-7.

86. Mahalingam, B., et al., *Crystal structures of HIV protease V82A and L90M mutants reveal changes in the indinavir-binding site.* Eur J Biochem, 2004. **271**(8): p. 1516-24.

87. Heugen, U., et al., *Solute-induced retardation of water dynamics probed directly by terahertz spectroscopy.* Proceedings of the National Academy of Sciences of the United States of America, 2006. **103**(33): p. 12301-12306.

88. Hornak, V., et al., *HIV-1 protease flaps spontaneously open and reclose in molecular dynamics simulations.* Proceedings of the National Academy of Sciences of the United States of America, 2006. **103**(4): p. 915-920.

89. Ma, A., A. Nag, and A.R. Dinner, *Dynamic coupling between coordinates in a model for biomolecular isomerization.* J Chem Phys, 2006. **124**(14): p. 144911.

90. Berne, B.J., M. Borkovec, and J.E. Straub, *Classical and Modern Methods in Reaction-Rate Theory.* Journal of Physical Chemistry, 1988. **92**(13): p. 3711-3725.

91. Dellago, C., P.G. Bolhuis, and P.L. Geissler, *Transition path sampling.* Advances in Chemical Physics, Vol 123, 2002. **123**: p. 1-78.

92. Hu, J., A. Ma, and A.R. Dinner, *Bias annealing: a method for obtaining transition paths de novo.* J Chem Phys, 2006. **125**(11): p. 114101.

93. Juraszek, J. and P.G. Bolhuis, *(Un)Folding mechanisms of the FBP28 WW domain in explicit solvent revealed by multiple rare event simulation methods.* Biophys J, 2010. **98**(4): p. 646-56.

94. Juraszek, J. and P.G. Bolhuis, *Sampling the multiple folding mechanisms of Trp-cage in explicit solvent.* Proc Natl Acad Sci U S A, 2006. **103**(43): p. 15859-64.

95. Dellago, C. and P.G. Bolhuis, *Transition path sampling simulations of biological systems.* Atomistic Approaches in Modern Biology: From Quantum Chemistry to Molecular Simulations, 2007. **268**: p. 291-317.

96. Quaytman, S.L. and S.D. Schwartz, *Reaction coordinate of an enzymatic reaction revealed by transition path sampling.* Proc Natl Acad Sci U S A, 2007. **104**(30): p. 12253-8.

97. Fujisaki, H. and J.E. Straub, *Vibrational energy relaxation in proteins.* Proc Natl Acad Sci U S A, 2005. **102**(19): p. 6726-31.

98. Lim, M., T.A. Jackson, and P.A. Anfinrud, *Ultrafast rotation and trapping of carbon monoxide dissociated from myoglobin.* Nat Struct Biol, 1997. **4**(3): p. 209-14.

99. Sagnella, D.E. and J.E. Straub, *A study of vibrational relaxation of B-state carbon monoxide in the heme pocket of photolyzed carboxymyoglobin.* Biophys J, 1999. **77**(1): p. 70-84.

100. Sagnella, D.E., et al., *Vibrational population relaxation of carbon monoxide in the heme pocket of photolyzed carbonmonoxy myoglobin: comparison of time-resolved mid-IR absorbance experiments and molecular dynamics simulations.* Proc Natl Acad Sci U S A, 1999. **96**(25): p. 14324-9.

101. Davarifar, A., D. Antoniou, and S.D. Schwartz, *The promoting vibration in human heart lactate dehydrogenase is a preferred vibrational channel.* J Phys Chem B, 2011. **115**(51): p. 15439-44.

102. Backus, E.H., et al., *Energy transport in peptide helices: a comparison between high- and low-energy excitations.* J Phys Chem B, 2008. **112**(30): p. 9091-9.

103. Botan, V., et al., *Energy transport in peptide helices.* Proc Natl Acad Sci U S A, 2007. **104**(31): p. 12749-54.

104. Appadurai, R. and S. Senapati, *Dynamical Network of HIV-1 Protease Mutants Reveals the Mechanism of Drug Resistance and Unhindered Activity.* Biochemistry, 2016. **55**(10): p. 1529-40.

105. Hornak, V., et al., *HIV-1 protease flaps spontaneously close to the correct structure in simulations following manual placement of an inhibitor into the open state.* J Am Chem Soc, 2006. **128**(9): p. 2812-3.

Personnel Justification

PI: Dr. Ao Ma is a theoretical biophysicist with significant expertise in developing theoretical and computational methods for understanding functional protein conformational dynamics. He is responsible for designing research plan and simulations, supervising the progress of simulations, analyzing the simulation data and revise manuscripts.

Three Research Assistants: The three graduate assistants (Shanshan Wu, Kenneth Tsui and Huiyu Li) will participate in and learn designing research, develop computer programs to carry out the simulations, perform simulations, analyze the simulation data and write the initial manuscripts.

Milestone Table

Year 1		
Milestone:	**Details** (as appropriate):	**Dates**:
Energy flow analysis on the allostery in PDZ domain	**Resource**: Theta **Node-hours**: 0.3125 M **Filesystem storage** (**TB**): 20 TB **Software Application**: GROMACS 2020.2 **Tasks**: loosely coupled ensemble computations on 2048 nodes of Theta **Dependencies**: No dependencies, start immediately	01/01/2023-12/31/2023
Energy flow analysis on the allostery in HIV-PR	**Resource**: Theta **Node-hours**: 0.3125 M **Filesystem storage** (**TB**): 20 TB **Software Application**: GROMACS 2020.2 **Tasks**: loosely coupled ensemble computations on 2048 nodes of Theta **Dependencies**: No dependencies, start immediately	01/01/2023-12/31/2023
Energy flow analysis on the druggable locations in SARS-COV-2 main protease	**Resource**: Theta **Node-hours**: 1.25 M **Filesystem storage** (**TB**): 20 TB **Software Application**: GROMACS 2020.2 **Tasks**: loosely coupled ensemble computations on 2048 nodes of Theta **Dependencies**: No dependencies, start immediately	01/01/2023-12/31/2023

BIOGRAPHICAL SKETCH

Provide the following information for the Senior/key personnel and other significant contributors.
Follow this format for each person. **DO NOT EXCEED FIVE PAGES.**

NAME: Ma, Ao

eRA COMMONS USER NAME (credential, e.g., agency login): aoma73

POSITION TITLE: Associate Professor

EDUCATION/TRAINING *(Begin with baccalaureate or other initial professional education, such as nursing, include postdoctoral training and residency training if applicable. Add/delete rows as necessary.)*

INSTITUTION AND LOCATION	DEGREE *(if applicable)*	Completion Date MM/YYYY	FIELD OF STUDY
Shandong University, Jinan, P. R. China	B.S.	07/1995	Chemistry
Brown University, Providence, Rhode Island	Ph.D.	05/2003	Theoretical Chemistry
University of Chicago, Chicago, Illinois	Postdoctoral	08/2006	Theoretical Chemistry

A. Personal Statement

I have a strong track-record of making theoretical contributions of fundamental importance. **1)** My PhD research was on the theory of dynamics and spectroscopy of liquids. I developed a novel algorithm that reduced the computational cost for simulating two-dimensional Raman spectroscopy of liquids by several orders of magnitude, which enabled molecular dynamics (**MD**) simulation of this spectroscopy. This was a major breakthrough because my simulation results preceded the correct experimental measurement and was instrumental in guiding theoretical and experimental progress on this subject when no experimental results were available. Suffering from technical challenges, incorrect experimental results were reported several times before Graham Fleming's group at Berkeley reported results that agreed with simulation, which was the critical evidence showing that their experiment was finally correct. **2)** The research during my postdoctoral fellowship was on biomolecular dynamics. During that period, I made a breakthrough and solved a challenging fundamental problem—identifying the critical solvent reaction coordinate for the isomerization reaction of alanine dipeptide in explicit water. This problem was formulated by David Chandler and co-workers at Berkeley in 2000s, but their intuition-based trial-and-error approach was not able to find the answer to this important problem. I solved this technically challenging problem by developing a systematic and general method for identifying reaction coordinates that used neural network. It was the first application of machine learning (**ML**) in studying protein dynamics, >10 years before the current enthusiasm in applying ML in chemistry and biophysics. **3)** Recently, my group developed the general framework of energy flow that explains the fundamental physics of the correct reaction coordinates in complex systems. Built on the Lagrange-Hamiltonian mechanics that is the fundamental physical law that governs protein dynamics, this unique approach enables fundamental understanding of the mechanism of conformational dynamics of biomolecules. This year we made another important progress. We developed a novel concept, the generalized work functional, which can be used to identify the rigorous reaction coordinate. This is the first time that the reaction coordinate in a complex system is identified by a physical operator without resorting to machine learning or human intuitions. Most importantly, the accuracy of the reaction coordinate identified by the generalized work functional far exceeds the accuracy of reaction coordinates identified by machine learning or human intuitions. This fact demonstrated the advantage of the physical-principle-based approach that we are currently pursuing.

B. Positions, Scientific Appointments, and Honors

Positions and Scientific Appointments

2013-present Associate Professor, The University of Illinois at Chicago, Department of Bioengineering, Chicago, Illinois

2006-2012 Assistant Professor, Albert Einstein College of Medicine, Dept. of Physiology & Biophysics, Bronx, New York

2003-2006 Postdoctoral Fellow, with Prof. Aaron R. Dinner, University of Chicago, Chicago, Illinois

2003-2003 Postdoctoral Fellow, with Prof. Richard M. Stratt, Brown University, Providence, Rhode Island

1998-2003 Research Assistant, with Prof. Richard M. Stratt, Brown University, Providence, Rhode Island.

Honors

2003: Sigma Xi – Graduate Research Award, Excellence in Research at Brown University
2002: Dissertation Fellowship at Brown University
1991-1995: First-Class scholarships at Shandong University (awarded to the top 5% students in the class)

C. Contributions to Science

1. Reaction coordinates, the few coordinates that can fully determine the progress of a reaction process, are central for understanding activated processes in complex systems, especially biomolecular dynamics. While easy to identify in small molecules, reaction coordinates are extremely challenging to determine in larger molecules due to their counter-intuitive nature. The conventional intuition-based trial-and-error method was met with stumbling challenges even for seemingly simple systems. I developed a machine learning based method that uses a combination of a genetic algorithm and a neural network (GNN) to select out the optimal combination of a pre-designated number of coordinates from a user-prepared candidate pool. I used this method to identify the reaction coordinates of a system that eluded previous investigations employing trial-and-error approaches. This method was the first application of machine learning to protein dynamics, >10 years before the current enthusiasm for applying machine learning in chemistry and biophysics.

a) J. Hu, A. Ma and A. R. Dinner, "A two-step nucleotide-flipping mechanism enables kinetic discrimination of DNA lesions by AGT", *Proc. Natl. Acad. Sci. USA*, **105**, 4615 (2008). PMID: 18353991; PMCID: PMC2290773. (**79 citations**)

b) J. Hu, A. Ma and A. R. Dinner, "Bias annealing: a method for obtaining transition paths de novo", *J. Chem. Phys*. 125, 114101 (2006). PMID: 16999460.

c) A. Ma, A. Nag and A. R. Dinner, "Dynamic coupling between reaction coordinates in a model for biomolecular isomerization", *J. Chem. Phys.* **124**, 144199 (2006). PMID: 16626249.

d) A. Ma and A. R. Dinner, "Automatic method for identifying reaction coordinates in complex systems", *J. Phys. Chem. B*, **109**, 6769 (2005). PMID: 16851762. (**358 citations**)

2. The machine learning based methods for identifying reaction coordinates are powerful in their own rights, but are limited in two essential aspects. First, their success critically relies on inputs from human intuition—the user-prepared candidate pool from which they choose the optimal solution has to include all the essential reaction coordinates. Yet, the counter-intuitive nature of reaction coordinates defies the strength of intuition and makes the candidate pool prone to incompleteness. Second, these methods do not provide any mechanistic insights in addition to the original human intuition input even when they succeeded in finding the correct reaction coordinates. Based on an energetic view of a reaction as stochastic energy flows biased towards preferred channels, which we deemed the reaction coordinates, we developed a rigorous procedure of decomposing the energy changes of a system into pairwise components. This method enables us to pin down reaction coordinates and understand how and why reaction coordinates move in concerted manner.

a) Wu, S., Li, H. and Ma, A. "A Rigorous Method for Identifying One-Dimensional Reaction Coordinate in Complex Molecules", J. *Chem. Theory Comp*, **18**, 2836 (2022).

b) Wu, S. and Ma, A. "Mechanism for the Rare Fluctuation that Powers Protein Conformational Change", *J. Chem. Phys*. **156**, 054119 (2022).

c) Li, H. and A. Ma. Kinetic energy flows in activated dynamics of biomolecules. *J. Chem. Phys*. **153**, 094109 (2020). PMID: 32891107.

d) W. Li and A. Ma, "Reaction mechanism and reaction coordinates from the viewpoint of energy flow", *J. Chem. Phys*. **144**, 114103 (2016). PMID: 27004858; PMCID: PMC4798989.

3. Microtubules are the backbone of the cytoskeleton and vital to numerous cellular processes. The central dogma of microtubules is that all their functions are driven by dynamic instability, but its mechanism has remained unresolved for over 30 years due to conceptual difficulties inherent in the dominant GTP-cap framework. We developed a physically rigorous structural mechano-chemical model: dynamic instability is driven by non-equilibrium transitions between the bent (B), straight (S) and curved (C) forms of tubulin monomers and longitudinal interfaces in the two-dimensional lattice of microtubule. All the different phenomena (growth, shortening, catastrophe, rescue and pausing) are controlled by the kinetic pathways for B↔S↔C transitions and corresponding energy landscapes. Different kinetics at minus-end are due to different B↔S↔C pathways imposed by the polarity of microtubule lattice. This model enables us to reproduce all the observed phenomena of dynamic instability of purified tubulins in kinetic simulations for the first time. This result is beyond the scope of other existing models of dynamic instability.

a) Stewman, S.F., K.K. Tsui, and A. Ma. "Dynamic Instability from Non-equilibrium Structural Transitions on the Energy Landscape of Microtubule". *Cell Syst*. 2020;11(6):608-624 e9. PMID: 33086051; PMCID: PMC7746586.

b) V. Mennella, D. Y. Tan, D. W. Buster, A. B. Asenjo, U. Rath, A. Ma, H. Sosa and D. J. Sharp, "A novel phospho-regulatory site on the Kinesin-13 motor domain", *J. Cell Biol*. **186**, 481 (2009). PMID: 19687256; PMCID: PMC2733746.

c) N. Fernandez, Q. Chang, D. W. Buster, D. J. Sharp and A. Ma, "A model for the regulatory network controlling the dynamics of kinetochore microtubule plus-ends and poleward flux in metaphase", *Proc. Nat. Acad. Sci*. **106**, 7846 (2009). PMID: 19416899; PMCID: PMC2683096.

4. Ultrafast laser spectroscopies are essential tools for studying dynamics of condensed phase systems, but signals from conventional one-dimensional spectroscopies are too congested to provide clear information. A promising approach to overcome this problem is to expand the signal onto two axes, in either time or frequency, inspired by the success of 2D NMR. The first of this new generation 2D spectroscopies is 2D Raman. Due to severe challenges in experimental setup, initial experimental results are contaminated and incorrect. On the other hand, theoretical studies using over-simplified models are difficult to compare with experiments. Molecular dynamics (MD) simulations can provide realistic results independent of experiments, but were prohibited by the high computational cost originated from high order correlation functions that are difficult to converge. I developed a novel computational algorithm to reduce the computational cost of MD simulations by orders of magnitude and performed the first MD simulation. My simulation results predated the first correct experiment and helped to validate its results. Using instantaneous normal mode analysis, I developed a theoretical model that can explain major features of the 2D Raman spectroscopies satisfactorily.

a) A. Ma and R. M. Stratt, "Selecting the information content of two-dimensional Raman spectra in liquids", *J. Chem. Phys*. **119**, 8500 (2003).

b) A. Ma and R. M. Stratt, "The molecular origins of the two-dimensional Raman spectrum of an atomic liquid. II. Instantaneous-normal-mode theory", *J. Chem. Phys*. **116**, 4972 (2002).

c) A. Ma and R. M. Stratt, “The molecular origins of the two-dimensional Raman spectrum of an atomic liquid. I. Molecular dynamics simulation”, *J. Chem. Phys*. **116**, 4962 (2002).

d) A. Ma and R. M. Stratt, “Fifth-order Raman spectrum of an atomic liquid: Simulation and instantaneous-normal-mode calculation”, *Phys. Rev. Lett*. **85**, 1004 (2000). PMID: 10991460. (**100 citations**)

5. Live cell fluorescence imaging techniques are major research tools for cell biology and provided the majority of information concerning dynamics of cellular processes. Microtubule dynamics in living cells are studied by fluorescently labeling tubulins and/or MAPs under different control and knockdown conditions, followed by imaging of the cellular processes of concern. It is important to extract quantitative information from the movies from the imaging processes and the conventional manual tracking is both labor intensive and subject to subtle human biases. My lab developed algorithms for automated tracking of both fluorescently labeled EB1 and tubulin imaging data. These algorithms were used in collaboration with experimental labs for analyzing fluorescent imaging movies to quantify microtubule dynamics and clarify functions of a few essential MAPs.

a) S. Solinet, K. Mahmud, S. Stewman, K. Ben EL Kadhi, B. Decellie, L. Talje, A. Ma, B. Kwok and S. Carreno, “The actin-binding ERM protein Moesin binds to and stabilizes microtubules at the cell cortex”, *J. Cell Biol*. **202**, 251 (2013). PMID: 23857773; PMCID: PMC3718980. (**88 citations**)

b) J. Currie, Shannon Stewman, G. Schimizzi, K. Slep, A. Ma and S. Rogers, “The microtubule lattice and plus-end association of Drosophila Mini spindles is spatially regulated to fine-tune microtubule dynamics”, *Mol. Biol. Cell*, **22**, 4343 (2011). PMID: 21965297; PMCID: PMC3216660.

c) D. Zhang, K. Grode*, Shannon Stewman*, D. Diaz*, E. Liebling, J. Curie, D. Buster, A. Asenjo, H. Sosa, J. Ross*, A. Ma*, S. Rogers*, D. Sharp, "Drosophila Katanin is a microtubule depolymerase that regulates cortical-microtubule plus-end interactions and cell migration”, *Nat. Cell Biol*. **13**, 361 (2011). * equal contribution. PMID: 21378981; PMCID: PMC3144748. (**127 citations**)

d) A. Wainman, D. W. Buster, T. Duncan, J. Metz, A. Ma, D. J. Sharp, and J. G. Wakefield, “A new Augmin subunit, Msd1, demonstrates the importance of mitotic spindle-templated microtubule nucleation in the absence of functioning centrosomes”, *Gene. Dev*. **23**, 1876 (2009). PMID: 19684111; PMCID: PMC2725934.

Complete List of Published Work in MyBibliography
https://scholar.google.com/citations?hl=en&user=1aoN6rUAAAAJ&view_op=list_works or
http://www.ncbi.nlm.nih.gov/sites/myncbi/ao.ma.1/bibliography/44239050/public/?sort=date&direction=descending

Section 6: Software Applications and Packages

Question #1

Please list any software packages used by the project, and indicate if they are on open source or export controlled.

Application Packages

Package Name

GROMACS

Indicate whether Open Source or Export Controlled.

Open Source

Section 7: Wrap-Up Questions

Question #1

National Security Decision Directive (NSDD) 189 defines Fundamental Research as "basic and applied research in science and engineering, the results of which ordinarily are published and shared broadly within the scientific community, as distinguished from proprietary research and from industrial development, design, production, and product utilization, the results of which ordinarily are restricted for proprietary or national security reasons." Publicly Available Information is defined as information obtainable free of charge (other than minor shipping or copying fees) and without restriction, which is available via the internet, journal publications, textbooks, articles, newspapers, magazines, etc.

The INCITE program distinguishes between the generation of proprietary information (deemed a proprietary project) and the use of proprietary information as input. In the latter, the project may be considered as Fundamental Research or nonproprietary under the terms of the nonproprietary user agreement. Proprietary information, including computer codes and data, brought into the LCF for use by the project - but not for generation of new intellectual property, etc., using the facility resources - may be protected under a nonproprietary user agreement.

Proprietary Information

Are the proposed project and its intended outcome considered Fundamental Research or Publicly Available Information?

Yes

Will the proposed project use proprietary information, intellectual property, or licensing?

No

Will the proposed project generate proprietary information, intellectual property, or licensing as the result of the work being proposed?

If the response is Yes, please contact the INCITE manager, INCITE@doeleadershipcomputing.org, prior to submittal to discuss the INCITE policy on proprietary work.

No

Question #2

The following questions are provided to determine whether research associated with an INCITE proposal may be export controlled. Responding to these questions can facilitate - but not substitute for - any export control review required for this proposal.

PIs are responsible for knowing whether their project uses or generates sensitive or restricted information. Department of Energy systems contain only data related to scientific research and do not contain personally identifiable information. Therefore, you should answer "Yes" if your project uses or generates data that fall under the Privacy Act of 1974 U.S.C. 552a. Use of high-performance computing resources to store, manipulate, or remotely access any national security information is prohibited. This includes, but is not limited to, classified information, unclassified controlled nuclear information (UCNI); naval nuclear propulsion information (NNPI); and the design or development of nuclear, biological, or chemical weapons or of any weapons of mass destruction. For more information contact the Office of Domestic and International Energy Policy, Department of Energy, Washington DC 20585, 202-586-9211.

Export Control

Does this project use or generate sensitive or restricted information?

No

Does the proposed project involve any of the following areas?

i. Military, space craft, satellites, missiles, and associated hardware, software or technical data

ii. Nuclear reactors and components, nuclear material enrichment equipment, components (Trigger List) and associated hardware, software or technical data

iii. Encryption above 128 bit software (source and object code)

iv. Weapons of mass destruction or their precursors (nuclear, chemical and biological)

No

Does the proposed project involve International Traffic in Arms Regulations (ITAR)?

No

Question #3

The following questions deal with health data.   PIs are responsible for knowing if their project uses any health data and if that data is protected. Note that certain health data may fall both within these questions as well as be considered sensitive as per question #2. Questions regarding these answers to these questions should be directed to the centers or program manager prior to submission.

Health Data

Will this project use health data?

No

Will this project use human health data?

No

Will this project use Protected Health Information (PHI)?

No

Question #4

The PI and designated Project Manager agree to the following:

Monitor Agreement

I certify that the information provided herein contains no proprietary or export control material and is correct to the best of my knowledge.

Yes

I agree to provide periodic updates of research accomplishments and to

acknowledge INCITE and the LCF in publications resulting from an INCITE award.

Yes

I agree to monitor the usage associated with an INCITE award to ensure that usage is only for the project being described herein and that all U. S. Export Controls are complied with.

Yes

I understand that the INCITE program reserves the right to periodically redistribute allocations from underutilized projects.

Yes

Section 8: Outreach and Suggested Reviewers

Question #1

By what sources (colleagues, web sites, email notices, other) have you heard about the INCITE program? This information will help refine our outreach efforts.

Outreach

Question #2

Suggested Reviewers

Suggest names of individuals who would be particularly suited to assess the proposed research.

Wei Yang, Florida State University, email: yang@sb.fsu.edu

Steven Schwartz, University of Arizona, email: sschwartz@email.arizona.edu

Shi-jie Chen, University of Missouri Columbia, email: ChenShi@missouri.edu

Section 9: Testbed Resources

Question #1

The ALCF and OLCF have test bed resources for new technologies, details below. If you would like access to these resources to support the work in this proposal, please provide the information below. (1 Page Limit)

The OLCF Quantum Computing User Program is designed to enable research by providing a broad spectrum of user access to the best available quantum computing systems, evaluate technology by monitoring the breadth and performance of early quantum computing applications, and Engage the quantum computing community and support the growth of the quantum information science ecosystems. More information can be found here: https://www.olcf.ornl.gov/olcf-resources/compute-systems/quantum-computing-user-program/quantum-computing-user-support-documentation.

The ALCF AI Testbed provides access to next-generation of AI-accelerator machines to enable evaluation of both hardware and workflows. Current hardware available includes Cerebras C-2, Graphcore MK1, Groq, Habana Gaudi, and SambaNova Dataflow. New hardware is regularly acquired as it becomes available. Up to date information can be found here: https://www.alcf.anl.gov/alcf-ai-testbed.

Describe the experiments you would be interested in performing, resources required, and their relationship to the current proposal. Please note, these are smaller experimental resources and a large amount of resources are not available. Instead, these resources are to explore the possibilities for these technologies might innovate future work. This request does not contribute to the 15-page proposal limit.

testbed_resource.pdf
The attachment is on the following page.

I do not request Testbed Resources because the computations of the proposed project cannot take advantage of systems for quantum computing or AI accelerated machines.