2023 INCITE Proposal Submission
Proposal

Title: Lipid shuttling molecular machines enabling functions of human cell membranes

Principal Investigator: George Khelashvili

Organization: Weill Cornell Medical College

Date/Time Generated: 6/14/2022 10:13:38 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

Harel

Last Name

Weinstein

Organization

Weill Cornell Medicine

Email

haw2002@med.cornell.edu

Work Phone

212 746 6358

Address Line 1

Box 75, Room E-509, 1300 York Avenue

Address Line 2

(No answer given.)

City

New York

State

NY

Zip Code

10065

Question #2

Co-PI (s)

First Name

George

Last Name

Khelashvili

Organization

Weill Cornell Medicine

Email

gek2009@med.cornell.edu

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

Harel Weinstein

Institutional Contact Phone

212 746 6358

Institutional Contact Email

haw2002@med.cornell.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 use computation to learn how specific proteins in the membranes of our cells work to maintain the properties needed for life and health. Understanding how they function will enable repair in disease, and the engineering of proteins with new functional applications.

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

Node Hours

733000

Storage (TB)

278

Off-Line Storage (TB)

0

Question #2

OLCF Frontier (Cray Shasta) Resource Request – 2023

Question #3

OLCF Frontier (Cray Shasta) Resource Request – 2024

Node Hours

697000

Storage (TB)

248

Off-Line Storage (TB)

0

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 Xe) Resource Request – 2024

Question #11

ALCF Aurora (Intel Xe) 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.**

Assembled_application_2023.pdf
The attachment is on the following page.

PROJECT EXECUTIVE SUMMARY

Title: Lipid shuttling molecular machines enabling functions of human cell membranes; **PI and Co-PI(s):** Profs. Harel Weinstein, George Khelashvili**; Applying Institution/Organization:** Weill Cornell Medicine**; Node-Hours on Summit Requested:** Year 1 – 733,000, Year 2 – 697,000**; Amount of Storage Requested:** Year 1 – 278 TB, Year 2 – 248 TB.

The growing ability to measure the activities and structures of the many different proteins inserted in the cell's diverse membranes revealed that these membranes are dynamic components that perform a variety of necessary physiological functions. Understanding how a cell maintains the mechanistically relevant properties and compositions of its membranes in the changing environment produced by its physiological function is considered a current **grand challenge** in biomedicine and biophysics. A key element of this challenge is the recognition that the membrane lipids are both the substrate that is being shuffled by the molecular machines that maintain the membrane composition, and the components of the medium in which these molecular machines function. The discovery and quantifications of these molecular mechanisms thus a central component of the ***grand challenge. It is addressed here by molecular dynamics (MD) simulations that are coordinated and combined iteratively with experimental research.*** The focus is on the TMEM16 protein family of phospholipid scramblases (PLS) that catalyze the fast diffusion of lipids between membrane leaflets, and on the highly selective lipid transporter MFSD2A which mediates Na^+-dependent uptake of the ω-3 docosahexaenoic fatty acid (DHA) into the brain. These molecular machines represent the two aspects of the grand challenge of the membrane being both the medium and the substrate, which lends this work a very high biological and technological significance. Thus, our goals are to ***discover, quantify and develop blueprints for practical uses of, molecular and functional properties of key types of lipid-shuttling molecular machines: lipid scramblase and transporter proteins***. Compelling reasons for achieving these goals include (1)-the great biological importance of what these molecular machines achieve as evidenced by their misfunction being involved in recognized genetic disorders of tissues and entire organs; (2)-their experimentally determined role in normal cell physiology based on membrane regulation; and (3)-their ability to serve as mechanistic templates for the biomimetic engineering of synthetic regulators of lipid membrane systems, lipid transport machines, and the creation of specific environments for biological function. To attain the **objectives and Milestones of this study,** the major emphasis is on the collection and analysis of massive amounts of data from computational simulations and analysis/interpretation with machine-learning based approaches to trajectory analysis. The computation will focus on the determinant (1)-conditions and special modes of activation of the mammalian TMEM16 PLS by Ca^{2+}; (2)-molecular mechanisms underlying Na^+-dependent MFSD2A-mediated lipid transport; and (3)-the formulation of specific testable mechanistic hypotheses for the structure-based functional mechanisms and how these relate to dysfunction of these molecular machines produced by mutations and membrane conditions in disease. On this basis, the **Study Objectives** are designed and detailed to provide mechanistic information and quantitative data culminating in the Milestones, to (i)-answer the key open questions about physiological mechanisms, (ii)-enable the mitigation of disease caused by dysfunction of the TMEM16 and MFSD2A proteins, and (iii)-guide protein engineering efforts to design regulatable biomimetic machines performing the same types of function for a variety of endpoints.

In addressing these ambitious goals we will leverage several advantages including (1) Our documented expertise (see Bibliography) in developing and applying novel theoretical concepts and computational tools, including machine learning-based approaches, for quantifying the dynamics of and the allosteric mechanisms of molecular machines, and (2) Our specific experience in the investigation of both TMEM16 PLS and MFSD2A mechanisms in closely integrated protocols of iterative functional, structural, and computational experimentation we gained in the ongoing collaborations with Profs. Alessio Accardi and Filippo Mancia working on collaborative NIH-funded projects on these specific subjects. The computational work designed is dependent on the requested resources available to us only on the Summit machine for this first-in-kind study of very large dynamic protein-membrane systems. Our experience from our current INCITE allocation (BIP109) and previous ALCC and INCITE awards (BIP124) demonstrates both the absolute need for Summit-level resources and beyond (e.g., Frontier) to accomplish the **Milestones**. It also underscores the strong expertise of our team in efficiently handling and running the massive MD simulations on the available platform, and the ability to adapt them rapidly and seamlessly (e.g., from Summit to Frontiers) with new and constantly improved workflows.

PROJECT NARRATIVE

1 MOTIVATION AND SIGNIFICANCE OF THE RESEARCH

1.1 The grand challenge addressed by computational simulations described in the application

After decades in which the multitude of functions performed by the cell's diverse membranes remained underappreciated, the ability to measure the activities and structures of the many different proteins inserted in the membranes brought to light the important functions that the cell's membranes perform. The answer to the big question of how these necessary physiological functions are accomplished by membranes that were long considered oily barriers of the cell, depends on understanding (1)-the dynamic properties imparted by the different lipid compositions and their physicochemical properties; (2)-the way in which these properties are regulated, and (3)-the role played by the molecular machines embedded in these membranes. This is because both experiments and theory have demonstrated that the mechanistically relevant properties and compositions of cell membranes are maintained in spite of the constantly changing environment of a living cell. The answers to key mechanistic questions about how the compositional asymmetry of the *outer* vs *inner* bilayers of the membranes is achieved, and what happens to the cell if these properties change is one of the **grand challenges** in current biomedicine and biophysics. The realization that specific molecular machines embedded in the membrane are responsible for the physiological processes establishing and maintaining the composition and properties, prompted focused efforts to determine their 3D molecular structure. This was difficult in part because the membrane lipids are both the substrate that is being shuffled by the molecular machines that maintain the membrane composition, and at the same time they constitute the medium that permits and regulates the functions of these molecular machines. Our work aims to discover and quantify these membrane-supported and membrane-dependent molecular mechanisms as major steps in addressing the ***grand challenge.***

Our project tackles this difficult challenge ***as described in this application, with computational molecular dynamics (MD) simulations that are coordinated and combined iteratively with experimental research.*** Specifically, we focus on the TMEM16 protein family of phospholipid scramblases (PLS) that catalyze the fast and passive diffusion of lipids between membrane leaflets, and on the highly selective lipid transporter MFSD2A which mediates Na^+-dependent uptake of the ω-3 fatty acid docosahexaenoic acid (DHA) into the brain. The research strategy is designed to reveal how the TMEM16 PLS functions to regulate membrane composition by shuffling components, and how the Major Facilitator Superfamily (MFS) member protein MFSD2A, mediates selective uptake of specific lipids. This information is essential for the ability to evaluate modes of cell membrane participation in physiology, and to intervene in the mechanisms of its dysfunction with target-specific therapies. In specific sections below we describe in detail the approaches we employ to (1)- discover conditions and special modes of activation of the mammalian TMEM16 phospholipid scramblases (PLS) by Ca^{2+}; (2)-quantify the molecular mechanisms underlying Na^+-dependent MFSD2A-mediated lipid transport; and (3)-formulate specific testable mechanistic hypotheses for the structure-based functional mechanisms and how these relate to dysfunction of these molecular machines produced by mutations and membrane conditions in disease. The detailed mechanistic information and quantitative data probed and validated in iterations between collaboratively designed experimental and computational investigations, lead to the important Milestones detailed in the Milestones Table. Together, the results will (i)-answer the key open questions about physiological mechanisms determined by the activities of the proteins in the membrane, (ii)-help in the mitigation of disease caused by dysfunction of the TMEM16 and MFSD2A proteins, and (iii)-guide protein engineering efforts to design regulatable biomimetic machines performing the same types of function for a variety of endpoints.

Our project pursues these aims and the Milestones as described herein through ***a) the discovery and quantification of molecular and functional properties of the family of TMEM16 scramblases that enable Ca^{2+}-dependent rapid "swapping" of components of the inner and outer layers of the cell membrane***, ***and b) determination of molecular mechanisms underlying the Na^+-dependent lipid transport and substrate specificity in MFSD2A transporter.*** This is enabled by the structure determination of members of these two protein families, which is enabling a growing understanding of structure/function relationships, to which we have contributed significantly [1-4].

1.2 Specifics of the study targets and approach

We focus on vital molecular machines that catalyze and regulate the exchange of lipids between leaflets of cell membranes: phospholipid scramblases (PLS) in the family TMEM16 proteins and the lipid transporter in the MSF family of proteins, MFSD2A. The computational investigations are the basis for iterative experimental probing and refinement in established NIH-supported collaborations. We leverage published, experimentally derived data generated for these systems in view of their great biological importance and their role in human physiology and their established involvement in human disease. At the same time, we explore these unique biophysical systems as templates of molecular machines for which cell membranes serve as both reservoir of substrates and a mechanistically responsive environment [3].

1.2.1 *The mammalian TMEM16 lipid scramblases*: TMEM16 proteins are Ca^{2+}-activated and capable of rapidly swapping lipid components of the inner and outer layers of the cell membrane to enable key physiological mechanisms[5]. The regulated lipid swapping has been shown to affect the state of the cell and specific actions of embedded proteins while keeping the cell in balance structurally and biochemically. In the plasma membrane (PM) the TMEM16 PLS can dissipate the lipid asymmetry across the cell membrane bilayer that is set by energy (ATP)-driven lipid transporters and, for example, bring to the cell's surface phosphatidylserine (PS) lipids which are usually confined to the interior-facing leaflet. This change is interpreted by the cell's environment as a signal of the state of the cell, with significant consequences [6]. Thus, the mechanistic understanding of the structure-function relations of PLS is necessary not only to enable the mitigation of disease caused by dysfunction of the TMEM16, but also for the ability to use this knowledge as mechanistic templates for the biomimetic engineering of synthetic regulators of lipid membrane systems.

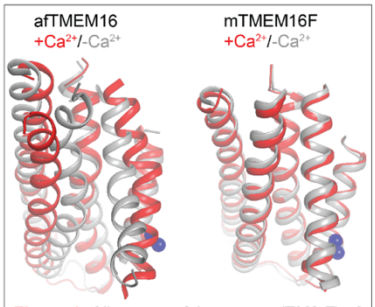

Figure 1: Alignment of the groove (TM3-7) of Ca2+-bound (red) and Ca2+-free (grey) of afTMEM16 (PDBID: 6DZ7, 6E0H), and of mTMEM16F (PDBID: 6QP6, 6QPB)

A current challenge in gaining the needed molecular understanding is that the main mechanistic considerations are based on studies on fungal homologs, nhTMEM16 and afTMEM16 [1-3,7-10,11-13] of the *mammalian* TMEM16 PLS. **For the latter, structure-based functional insights are only now starting to emerge[14-16], bearing surprises.** Significant differences have come to light that despite the generally high structural similarity between the fungal and mammalian homologs, and it is still unclear to what extent the molecular mechanisms underlying their activities or regulation are similar. Moreover, in the human TMEM16K (hTMEM16K)[16], as in the fungal nhTMEM16 and afTMEM16[7,17], Ca^{2+} binding induces global rearrangements that result in the opening of a membrane-exposed pathway through which lipid headgroups can move between membrane leaflets (**Fig. 1**, *left*). But structural studies in another mammalian system – the mouse TMEM16F (mTMEM16F) PLS – have suggested that Ca^{2+} binding does not induce opening of the putative lipid pathway in this homolog[14,15] (**Fig. 1**, *right*). This led to the proposal that lipid permeation might occur outside a closed scrambling pathway in mTMEM16F PLS. As the human TMEM16E (hTMEM16E) bears ~50% sequence identity to mTMEM16F [14,15], this was considered a possibility for this PLS as well. While our findings thus far do not supportsuch a drastic departure, it remains an important challenge in biology to understand how, and to what extent *the mammalian, and specifically human molecular machines with highly similar structures to those of fungal homologs have evolved to differ from them* [18]. *We must address it here as part of the goal (see section **1.1**) to reveal how disease-related mutations (see **1.2.2**) impact the* function of mammalian TMEM16 PLS. Therefore, specific computational experiments are designed to elucidate mutations-induced

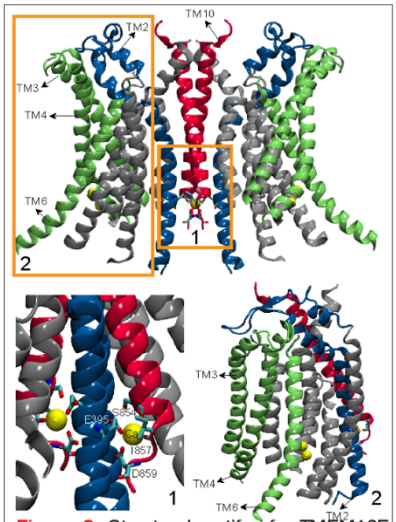

Figure 2: Structural motifs of mTMEM16F (PDBID 6QP6). Rectangle 1 highlights the third (distal) Ca2+ binding site at the dimer interface. Shown in licorice and labeled are the 4 residues coordinating the Ca2+ ion in its binding site. Rectangle 2 highlights the groove regions in mTMEM16F. and shows TMs 3, 4, and 6 (in green) lining the groove region, TM10 (in red) and TM2 (in blue).

changes in the molecular mechanisms, at a level that can clarify functional differences introduced during the sequence evolution of the fungal to mammalian PLS.

One of the most important insights we can offer in the coming year relates to the discovery in mammalian mTMEM16F PLS, of an additional (third) Ca^{2+} ion binding site not observed in the fungal structures (**Fig. 2**). As described below (see **2.1.2**), we have uncovered an allosteric coupling pathway between this additional Ca^{2+} binding site located at the dimerization interface, and the hydrophilic groove. In the context of the membrane, the allosteric mechanism dynamics seem to to enable the transformation of the closed groove into the open 'Lipid-Conductive' conformation observed in the fungal homologous. We will probe this central mechanistic hypothesis regarding the allosteric mechanism (sections **2.2 & 2.3)** in the mTMEM16F and hTMEM16E PLS, and enable experimental validation in functional studies by ***predicting from our analysis a set of mutants that can modulate the allosteric channel between the 3rd Ca^{2+} and the groove*** and probing their structural, dynamic, and mechanistic consequences in computational simulations. The ***parallel experiments testing*** of the functional phenotypes of these constructs *in vitro should also* probe/validate the mechanistic predictions regarding the conditions for activation by the complete set of Ca^{2+} binding modes to PLS **to be delivered in Milestone 1.**

1.2.2 *Interpreting the functional effects of disease mutants in hTMEM16E:* **Achieving Milestone 2** involves the mechanistic interpretation of known and measurable functional effects of disease-linked mutations in mammalian TMEM16 PLS (see **1.2.1)**. Such mutations are particularly abundant in the hTMEM16E PLS, and several of them have been well-characterized functionally[19,20] – e.g., mutations causing Gnathodiaphyseal dysplasia (GDD) and muscular dystrophy (MD) found to exhibit, respectively, gain- and loss-of-function phenotypes[19-21]. The mechanisms remain unknown. **Section 2** describes our extensive expertise with TMEM16 PLS structure-function-dynamics[1,3,7] that will serve to extract the required mechanistic understanding (**Milestone 2**).

1.2.3 *The MFSD2A membrane protein is a transporter,* highly enriched in endothelial cells of the blood-brain (BBB) and blood-retinal (BRB) barriers where it mediates Na^+-dependent uptake, into the brain and eyes, of ω-3 fatty acid docosahexaenoic acid (DHA) in the form of lysophosphatidylcholine (LPC-DHA) [22]. This lipid transport process is vital to human physiology as DHA accounts for ~20% of the total membrane fatty acids in the central nervous system and is essential for brain and eye function and development. These organs cannot synthesize DHA, and therefore rely on the uptake of this nutrient from systemic and dietary sources across the BBB and BRB via MFSD2A transporter. *Our work to determine the molecular mechanisms underlying the functional dynamics and substrate specificity of MFSD2A seeks* ***a first time understanding*** *of such specific lipid transport,* opening heretofore unavailable opportunities for drug delivery across the BBB and BRB **(see section 1.3 – Broader implications).**

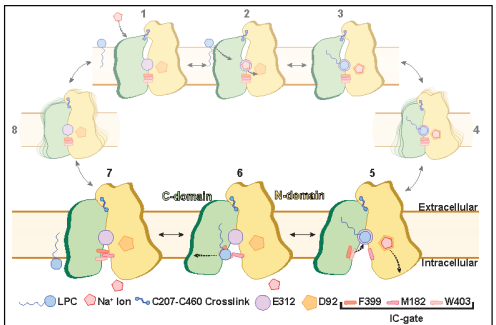

Figure 3: Schematic of the conformational states visited by MFSD2A throughout its transport cycle. The larger panels (5-7), depicting inward-facing (IF) states, are derived from our data, whereas the remaining panels are hypothesized on the basis of our observations and previous knowledge of related transporters. Outward facing (OF) conformations are shown in panels 1-3; an occluded state is depicted in panels 4, 8.

Our preliminary results [4] from exploring the fundamental structure-function relationships in MFSD2A by combining large-scale atomistic molecular dynamics (MD) simulations with structure determination using single-particle cryo-electron microscopy (cryoEM) and functional analyses have already illuminated how MFSD2A releases the substrate from the inward-facing (IF) state into the cytoplasm in a Na^+-dependent manner (see states 5-7 in **Fig. 3**). These successes lay a path to probe the major unanswered mechanistic questions (see section **2.1.3**): (1) how the MFSD2A transporter cycles through experimentally identified functional states, including the recently structurally characterized outward-facing (OF) state [23] (panels 1-3 in **Fig. 3**), and (2) how binding of Na^+ ions and substrates facilitate these transitions. It is also unknown how MFSD2A preferentially selects specific lipids for transport while excluding others [22]. The physico-chemical properties of lysolipids (head-group identity, charge, lipid tail length) are critical selectivity determinants [22], but structural and dynamic features of the transporter responsible for their recognition remain unknown. As illustrated by the **Milestones** we expect that answering these questions

will facilitate bioengineering efforts for modifying the MFSD2A's specificity to leverage it for therapeutic purposes. Thus, bringing to light currently unknown structure-function relationships in MFSD2A will advance basic and translational aims of developing compounds that can hijack the MFSD2A transport mechanism to penetrate the brain. In addressing the ambitious goals of the proposed *investigations of TMEM16 PLS and MFSD2A transporter mechanisms we will leverage several important advantages including:* ***1)*** Our documented experience and expertise in developing and applying novel theoretical concepts and tools, including state-of-art machine learning based approaches for quantifying the dynamics of membrane proteins and their allosteric mechanisms (see **bolded** entries in **Bibliography**); and **2)** Our specific experience in the investigation of these membrane-involved molecular machines with the integrated protocol of iterative functional, structural, and computational approaches, gained in the context of the ongoing collaborations with the leading experimental researchers in the respective fields (Dr. Alessio Accardi, TMEM16 structure and function [1,2,5,7-10,16], including our NIH-funded collaboration [1-3]; Dr. Filippo Mancia, MFSD2A structure and function [4]). Accordingly, the large-scale computational effort we propose here is combined with synergistic experiments. Moreover, the **Milestones** indicate how the mechanistic findings will be translated into guides for molecular engineering of analogs of the molecular machines with specifically engineered properties and functions. As detailed below **(sections 2.2&2.3),** the computational work designed to attain the **Milestones** corresponding to these research goals is entirely dependent on the requested resources that are available only on the Summit machine and the critical biomedical, biophysical, and biomimetic goals of this grand challenge study require us to leverage the established and growing pipeline of advanced methods for enhanced sampling in molecular dynamics (MD) simulations that we have implemented and continue to refine (see **bolded** entries in **Bibliography)**, and the workflow and trajectory analysis approaches implemented under our NSF award number:1740990 (e.g., see [24-28]).

1.3 The broader implications of the project

The broad biological and biomedical significance of revealing the molecular mechanisms of mammalian TMEM16 PLS is underscored by their established role in human physiology, demonstrated by their continuously newly discovered physiological roles as well as by their recognized involvement in genetically inherited disorders of muscle[29,30], bone[21,31,32], blood[33,34] and brain[35-38], and by their disease-linked mutations[19,20,39-41]. Experience in the field shows that not much can be done about any of these even now, when structures have become available, as long as the mechanisms are based on inferences from rigid structures in detergent or nanodiscs. The mechanisms are dynamic, involve membrane-regulated activation and allostery, and the breakthrough structures must be investigated in terms of their dynamic properties in cell membrane-like environments. Our project is designed to address this challenge and reveal the dynamic mechanisms. Adding to the broad implication of the expected results is the expected answer to the specific question of the source and mechanisms of the differences between fungal and mammalian proteins (discussed in **section 1.2**). This will address the **biological grand challenge** of understanding the modes of *structure-based adaptation to different physiological conditions.* In the specific case of the TMEM16 scramblases the salient change is brought by the mammalian homologs having acquired an additional occupied Ca^{2+} binding site. Given the nature and functions of the TMEM16 proteins we describe throughout, the implications of the project are far reaching, covering eagerly sought mechanistic information about fundamental biological questions in the evolution of molecular machines in the membrane, to attaining new tools for disease mitigation, and impacting directly the field of biomimetic systems engineering.

Similarly, the other membrane-dependent molecular machine we study, MFSD2A, is implicated in several neurological disorders, such as autosomal recessive primary microcephaly, intracranial hemorrhage and tumor, sepsis-associated encephalopathy, and Alzheimer's disease (AD) [42]. Broad implications for uncovering the molecular mechanisms underlying function of MFSD2A lipid transporter stem from the importance of its position at the Blood-Brain-Barrier. **The expected impact of this progress** is underscored by the current consideration of MFSD2A as a promising new molecular target for drug delivery across the Blood-Brain-Barrier [43]. Owing to its strategic localization and thus far unique properties of MFSD2A, our results would enable our team and others to take ***advantage of*** MFSD2A-mimetic ***machinery to develop a novel platform for drug delivery into the brain across the blood-brain barrier.*** The goal of our

computational studies is to overcome the major challenge represented by the currently very incomplete mechanistic understanding of the molecular basis for MFSD2A substrate specificity and transport. The steps to mitigate this knowledge gap using advanced computational tools is illustrated by the recent promising results of our collaboration with the lab of Filippo Mancia (Columbia University) summarized in our recent article in Nature [4]. MFSD2A features a 12-transmembrane domain architecture arranged into two pseudosymmetric six-helix bundles (the N and C domains) characteristic of the MFS family. However, because MFSD2A transports lysolipids rather than water-soluble molecules (unlike most MFS proteins), the molecular mechanism of MFSD2A-mediated transport *cannot be the same as for other MFS*. The elucidation of the transport characteristics will therefore impact several transporter fields and will be "a first in several different respects. The practical implications, discussed in section **1.2.3** above, are profound but represent only what can be planned at the current (poor) level of understanding.

From a computational perspective, the broad impact of the project relates to the goals of **NSF-supported collaborative projects** IIS 1741057/1841758, 1740990, and 1741040 **entitled BIGDATA: IA: Collaborative Research: In Situ Data Analytics for Next Generation Molecular Dynamics Workflows (H. Weinstein PI of NSF Award Number:1740990)** because the massive amounts of data resulting from extensive atomistic MD simulations and analyses in the project described here are used to test and refine the performance of new workflows and algorithms. The synergy benefits the goals of both the project proposed here and the NSF project for which the entire team develops efficient approaches to data challenges posed by MD simulations at the Exascale. This is demonstrated in our publications [25-28] reporting workflow and algorithmic developments that include (1)-novel data analytics algorithms ideal for in situ data analysis of relevant structural molecular properties (e.g., see [28]), and (2)- definitions of MD-based machine learning (ML) techniques to automatically identify rare events in the trajectory and molecular domains where the properties reside at runtime[25,44]. As discussed in **sections 2.4, 2.5 and 3** below**, our team continues to integrate** new algorithms and techniques into MD workflows. Thus, the planned effort drives the implementation of the most appropriate high-performance tools, fostering new developments of workflows (see section **2.5**) and continuous refinement of computational approaches that can work best on, and with, the specific computational resources requested in this INCITE application.

Of further importance for the broader implication of this project is its documented record of close integration of the computational effort with collaborative acquisition, analysis, and interpretation of data obtained experimentally in sustained thematic collaborations with labs studying molecular structure and dynamics of the same systems. The multiscale, interdisciplinary nature of this research covers and illustrates for the community collaborative efforts in a variety of contexts – from single molecules to biochemical and cell systems. This powerful combination supported by our continuous NIH funding ***advances progress for computation by offering direct means for valuable feedback from testing, validation, and refinement. It also fosters new ways for quantitative experimentation based on the new and challenging hypotheses and models generated from computational modeling and simulation (e.g., see recent developments in [45]).***

We elaborate further in sections below **why** attaining the research objectives leading to the Milestones **will be possible only** (i)-by accessing the major computational resources made available by INCITE on the Summit system, and (ii)-by using these resources to run continuously developed and refined computational protocols that enable the rigorous quantitative analyses algorithms we describe and document[46-50]. *On this basis we note that an even* **broader and more general impact** *outcome of this project* **than discussed above can be expected** *in view of its proffering (1)-approaches applicable to structure/function studies of a wide range of molecular machine-like systems; and (2)-novel insights for molecular engineering.* This expectation is based on the examples of approaches that have yielded such impact from descriptions in our publications of similar solutions for other molecular machines in the cell membrane that are governed by the type of allosteric mechanisms we identify in the TMEM16 PLS and MFSD2A. Included are the G protein-coupled receptors (GPCRs)[26,51-53], and solute transporters across membranes[25,26,54-57].

2 RESEARCH OBJECTIVES AND MILESTONES

2.1 Preliminary Results

In this Section we present the molecular systems, mechanistic questions, and some key methodology in the context of ongoing and published work (more extensive Methodology is in section **2.4**). The findings

and discoveries documented in our publications were enabled in part by our previous INCITE allocation (BIP109) on Summit, and the results frame and illustrate the approaches we use to achieve the scientific goals and Milestones of the proposed project.

2.1.1 *A gating mechanism at the extracellular entry to the lipid pathway in the fungal nhTMEM16 PLS enables lipid scrambling:* The computational approach to revealing the key mechanistic differences between the fungal and mammalian systems relates directly to the discovery of the additional (third) Ca^{2+} ion binding site not observed in the fungal structures (**Fig. 2**). In our preliminary MD simulations of mTMEM16F described in the subsequent section, **2.1.2**, we discuss its role in the rearrangement of the closed groove into the open 'Lipid-Conductive' conformation observed in the fungal homologous. The investigation of the specifics of the scrambling mechanism through such a groove, will follow the methodological approach to simulation and analysis we used for the fungal scramblase, which is therefore reviewed briefly below. For the fungal nhTMEM16 PLS, our extensive atomistic MD simulations revealed a molecular gating mechanism of the constricted extracellular (EC) entry to the lipid pathway of fungal nhTMEM16 (from PDBID: 4WIS[58]) that connects further into the groove to rearrange the mid-groove constriction. This transforms the "Membrane-exposed" conformation of the groove observed in the frozen structures, into a "Lipid-Conductive" state[1]. We used the ensemble MD simulations and the adaptive sampling protocol described in section **2.4.** to simulate the full-length nhTMEM16 dimer in membranes matching the experimental conditions (POPC or 3:1 POPE:POPG)[1]. The mechanistic information about the scrambling process observed in multiple trajectories was extracted using the tICA dimensionality-reduction approach (section **2.4**) in a space defined by collective variables describing the time-evolution of: i)-the position of the scrambled lipid along the groove; and ii)-pairwise interactions between the residues constituing the EC gates during the translocation process. Following our established protocols (section **2.4**)[1,2], the space described by the first two tIC vectors (describing 80% of the total variance) was then discretized into 50 microstates[1]. These microstates cover the configurational space of the entire system as *lipid translocation occurs from the intracellular (IC) to EC leaflets, and their corresponding structures reveal the structural context of the key mechanistic stages in the translocation process (**Fig. 4**).* **Fig. 4** shows the structural states of nhTMEM16 that correspond to key intermediate conformations along the transition pathway and describes major mechanistic steps in lipid translocation between the two leaflets ("Sequence of Mechanistic Events"). Thus, discrete stages in the lipid flipping process include simultaneous destabilization of the interactions that tether TM3/TM4 to TM6 (between E313-R432, E318-R432 and T333-Y439), which results in a widening of the EC vestibule to allow formation of the continuous pathway for the release of the flipped lipid into the external leaflet [1-3,11,12] (Microstates 5-6 in **Fig. 4**). ***Together, these published mechanistic findings of groove opening and lipid translocation triggered by modulation of the EC gate in the Ca^{2+}-loaded fungal nhTMEM16 serve to illustrate how we plan to study: i)-functional mechanisms in mammalian TMEM16 PLS; and ii)-the transition from functionally defined states to the complete mechanism. The results will be used in the combined computational and experimental studies to evaluate and understand the mechanisms of disease-related mutations*** [1,19,20,59].

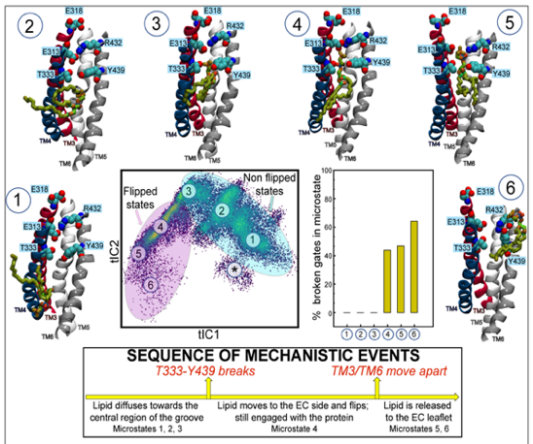

Figure 4: Molecular mechanism of lipid scrambling by nhTMEM16. tICA analysis results in the central panel show a mapping of all the MD trajectories (each dot) to the two-dimensional landscape in the space of the first two tICA eigenvectors. Microstates 1–6 locations indicate the translocation of the lipid through the nhTMEM16 groove, and representative structural representation of the microstates are the snapshots surrounding the central panel. In these structures, TMs 3-7 lining the nhTMEM16 groove are shown in different colors, advancing lipid is rendered in licorice, and mechanistically relevant groove residues are shown in space fill and labeled. The rare state with a lipid tail inserted in the groove, interfering with translocation (see section D2) is indicated by the '*'. The plot next to the tICA landscape compares the percentage of trajectory frames in which the three EC gates, T333-Y439, E313-R432, and E318-R432, are simultaneously broken in the microstates. The lower panel summarizes the corresponding sequence of mechanistic events leading to the lipid flip.

2.1.2 *The allosteric mechanism leading to an open-groove lipid conductive state of the TMEM16F scramblase*: We have performed Markov State Model (MSM) analysis on a set of ensemble atomistic MD trajectories (~400 μs aggregate sampling) collected in our preliminary computational studies of the mammalian homolog mTMEM16F PLS following the protocol described in section **2.4**). While the initial

results suggested similar mechanistic steps in lipid translocation as seen in nhTMEM16, we observed local rearrangements in the sidechains coordinating the 3rd Ca^{2+} ion bound at the distal site at the IC end of the protein dimer interface, which appeared to trigger the conformational changes of the lipid translocation groove (**Fig. 5**). The fungal TMEM16 PLS lack this distal Ca^{2+} binding site and analogous local structural changes were not observed in these PLS. In the *mTMEM16F*, however, these local dynamics led to concerted motions radiating in the direction of the EC vestibule of the groove and appeared to cause the eventual breaking open of the EC gates. ***These early mechanistic insights suggest a specific mechanism of groove opening in the mammalian TMEM16 PLS that is dependent on the unique structural elements present in these types of scramblases.*** *To achieve* ***Milestone 1****, we will probe as described* ***in section 2.3*** *the mechanistic hypotheses arising from these preliminary observations with extensive simulations of mammalian scramblases (mTMEM16F and hTMEM16E) and mutant constructs. These will include mutations at the Ca^{2+} coordinating residues, as well as known gain-of-function and activity-impairing mutations.*

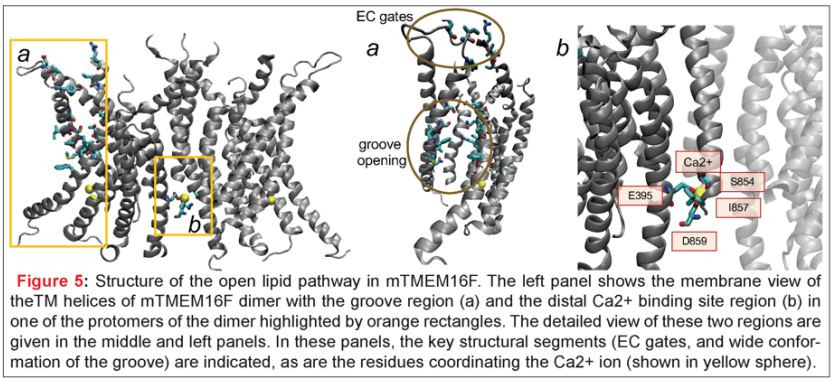

Figure 5: Structure of the open lipid pathway in mTMEM16F. The left panel shows the membrane view of theTM helices of mTMEM16F dimer with the groove region (a) and the distal Ca2+ binding site region (b) in one of the protomers of the dimer highlighted by orange rectangles. The detailed view of these two regions are given in the middle and left panels. In these panels, the key structural segments (EC gates, and wide conformation of the groove) are indicated, as are the residues coordinating the Ca2+ ion (shown in yellow sphere).

2.1.3 *Probing MFSD2A mediated lipid transport mechanisms*: Utilizing powerful integrative approach of combining structure determination with single-particle cryo-electron microscopy, functional analyses, and our large-scale MD simulations, we were able to uncover for the first time the mechanism by which the MFSD2A transporter interacts with lipid substrates, and how the Na^+-dependent conformational changes release these substrates into the membrane through a lateral gate (**Fig. 6**). There initial investigations are summarized in [4]. Thus, the initial simulations performed in the absence of substrate (i.e., MFSD2A immersed into a POPC membrane), showed the Na^+ ion entering the IC cavity and interacting with either E312 or D92. In the apo (**Fig. 6a**) and the E312-Na^+ bound states (**Fig. 6b**), residues M182, F399, and W403 form a gate between TMs 5, 8, and 10, which restricts access to the IC cavity from the inner leaflet of the membrane. When Na^+ binds to D92, this IC gate opens (**Fig. 6c**), allowing POPC headgroups to transiently penetrate the IC cavity. Preliminary results concerning lipid selectivity showed that when the POPC molecule adjacent to the open IC-gate was replaced with LPC-18:1, LPC-18:3, or LPC-DHA, these substrates inserted deeply into the hydrated IC cavity and sampled it with their polar headgroup while the hydrophobic tails remained predominantly in the membrane (**Fig. 6d**). Our simulations also revealed an intriguing dynamic interplay between the positioning of the lysolipid headgroup, the Na^+ binding modes, and the conformations of the residues in the charged central region (E312, R85, and K436, D88, and D92) [4]. These observations will be pursued in the planned studies to establish whether Na^+ binding to D88 and D92 constitutes a key mechanistic step that opens the IC-gate and repositions R85, which, in tandem with E312, coordinates the lysolipid headgroup as it travels through the IC cavity.

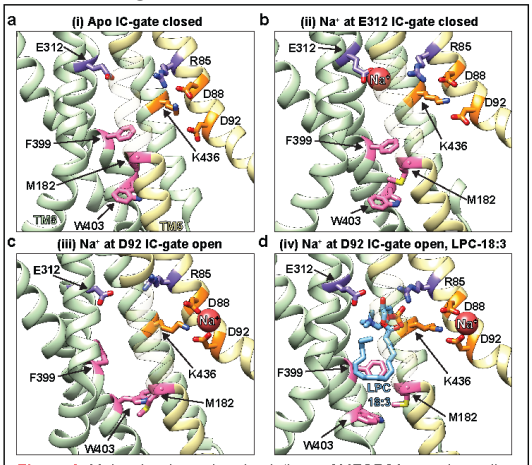

Figure 6: Molecular dynamics simulations of MFSD2A reveal coupling between Na+ binding and lysolipid movement through a dynamic intracellular gate. a–d, Structural representation of conformational states in an apo conformation with the intracellular gate closed (a); Na+ at E312 and the intracellular gate closed (b); Na+ at D92 and the intracellular gate open (c); and Na+ at D92, bound LPC-18:3 and the intracellular gate open (d). Select residues are shown in stick representation.

2.1.4 *Identification of a substrate-bound occluded state of MFSD2A*: Another essential step of the transport mechanism was suggested by our preliminary computational studies of the outward facing structure of MFSD2A (PDBID: 7N98) solved recently [23]. Thus, we observed spontaneous penetration of

LPC-18:1 from the bulk EC membrane leaflet into the central cavity of the transpoter in a process mediated by conformational changes in TM5 and TM8 helices. This was dependent on Na^+ binding to E312, and the simulations showed that, concomitant with the lysolypid penetration, the transporter isomerized into a state in which the substrate became occluded from both the EC and IC side. This sustrate-bound occluded state that is characteristic of alternate access transporters was stable for tens of microseconds of MD simulation. ***These early mechanistic insights suggest a specific mechanistic role for Na^+ ions regulating transitions between functionally relevant conformational states in MFSD2A.*** *We will explore the specific hypothesis that Na^+ binding to E312 promotes an outward open to occluded state transition, whereas Na^+ binding to D92 facilitates an occluded to inward opening transition.* As described in **sections 2.3 and 2.5** we will use the combination of biased and ensemble simulations to probe conformational transitions between various functional states, and will compare energetics and kinetics of these transitions between transport-supporting and transport-inhibiting conditions (i.e., wild type vs mutant protein constructs, replacing Na^+ with Li^+, which cannot support transport [22]). ***Our studies also point to groups of residues (e.g., E312, R85) that may play important roles in substrate specificity of MFSD2A***. We will determine structural and dynamic determinants of substrate specificity by comparing quantitatively interactions between various lysolipids (both substrates and non-transported) and the wild type and mutant protein constructs (sections **2.2-2.3**).

2.2 The Study Objectives determine the need for large-scale computational strategies and resources

The preliminary studies showed clearly that the computational approaches we employ are successful and efficient tools for discovering the mechanistic elements of the functional dynamics and properties of the TMEM16 PLS and MFSD2A systems and at the same time emphasize the need for very large-scale computational efforts to achieve rigorous and valid results. This is underscored by the results of our evaluations of quality, convergence, and consistency that we perform in parallel for each methodological component (see **sections 2.4.1-2.4.4**) that show the need for very extensive simulations of these large systems. The scale of the investigations needed to achieve the **Study Objectives listed below** is detailed in **Table 1 (section 2.3)** in the context of the *mechanistic information and quantitative structure dynamics data required to attain the* **Milestones** listed in the **Milestones Table** that is part of this application:

Study Objective 1: *Determine the mechanistic role of Ca^{2+} ions in the activation dynamics of mTMEM16F and hTMEM16E PLS by probing the hypothesis that Ca^{2+} ion binding at the dimerization interface of these scramblases (and not the fungal ones) allosterically triggers the opening of the lipid translocation pathway*.

Study Objective 2: *Identify structural elements in mTMEM16F and hTMEM16E PLS that regulate opening/closing of the lipid pathway and scrambling.*

Study Objective 3: *Discover molecular determinants for the effect of disease mutations in hTMEM16E on scrambling functions*.

Study Objective 4: *Determine molecular mechanisms underlying the functional dynamics in MFSD2A and establish the role of Na^+ ions in the transport mechanism in MFSD2A.*

Study Objective 5: *Identify structural and dynamic determinants of MFSD2A substrate specificity.*

The planning and design of these Objectives, and of the computations we propose in **section 2.3** (summarized in **Table 1** and the **Milestones Table)**, are informed by the experience and expertise in studying membranes and membrane proteins in general, and the TMEM16 PLS and MFSD2A systems specifically, as documented in our publications (please see citations denoted in **Bold** in the **Bibliography**). Importantly, the proposed computational studies are designed and performed in coordination with leading experimental labs in the field (*see Personnel Justification and Management Plan*), in the context of NIH-funded research as detailed in our publications (see **bolded** entries in **Bibliography**).

The results we and others have contributed to the literature have shown that the lipid translocation events in scramblases and transporter proteins require conformational changes that are: *i*) relatively **rare** molecular events overall as reflected in the simulations; and *ii*) depend on mechanistic involvement of **a set of factors** that need to be evaluated, and utilized in the molecular engineering efforts (e.g., regulation by ions and by lipids). This is why **the planned computations** designed to identify dynamic mechanisms ***require the extensive conformational sampling of a full set of appropriately chosen and designed conditions and controls*****,** and the ***probing with a variety of convergence criteria as detailed in*** **sections 2.3 to 2.5.** Our experience[52,55,57] agrees with the accepted view in the field (e.g. see Ref. [60] and citations therein)

that the most efficient way to sample the desired biomolecular processes and extract quantitative insights is to leverage the statistical mechanics basis of MD simulations to accumulate extremely long trajectories from *ensemble* MD simulations in which the time-propagation of the system is recorded in large numbers of independent replicates, much like in the cognate single-molecule experiments we describe in our publications – in which the status of many hundreds (in smFRET), and tens of thousands (in Cryo-EM), of individual molecules are monitored. Furthermore, as discussed [60], the efficiency of these ensemble approaches increases substantially when the simulations are carried out in *adaptive* protocols, where high-level algorithms are utilized to determine iteratively the next round of sampling based on the coverage of the configuration space of the system achieved with simulations in the previous steps, by the identification of productive trajectories[24] (see section **2.5** and **3-Computational Readiness**). The size of the molecular systems, taking into consideration atomistic representations of the membrane and water environments, make it impossible to pursue this work on any other accessible scientific computing platform. Indeed, the scale, efficiency, and the required software and hardware infrastructure [60] needed for our studies are available on **Summit**. The details of **research objectives and the Milestones** to be attained as described in **section 2.3**, further underscore specifically the conceptual framework and the advanced methodology underlying our pursuit of the **grand challenge** outlined in **section 1**, which compels the use of such significant computational resources for our project.

2.3 Specific Research Stages, Constructs, and Milestones

The **requested 2-year allocation on Summit** (Year 1 **– 733,000**; Year 2 **– 697,000 node hours**) will be used to carry out the sets of leading-edge atomistic MD simulations of the functional mechanisms of lipid scrambling by mammalian TMEM16 scramblases, i.e., mTMEM16F and hTMEM16E, and of the Na^+-dependent lipid transport by mammalian MFSD2A. To gain the information necessary for the conceptual advances and practical goals motivated and detailed in sections above, we will probe computationally the carefully chosen constructs of these molecular machines (*including wild types and a variety of mutants that include both those previously characterized, and the ones newly designed based on ongoing findings and predictions*). These constructs represent molecular systems and conditions that determine and modulate function (e.g., TMEM16 with +/- Ca^{2+} in 2 or 3 binding sites, and MFSD2A with +/- Na^+ or Li^+ bound). All of them are embedded in membrane environments mimicking experimental setups, as documented in our publications [57,61-70]. The investigated constructs detailed below in **Table 1** and in the **Milestones Table,** reflect stages in the functional processes of these systems, and other mechanism-related conditions discussed in previous sections, and in our publications [57,61-70].

YEAR 1										
System	Milestones	# of constructs	# of initial conditions	Simulation time (μs)		Node hours (x10³)			Storage (TB)	
				per construct	per system	per 1 μs	per construct	per system	per 1 μs	per system
mTMEM16F	1, 2, 5	8	50	250	2,000	0.16	40.00	320.00	0.05	90
hTMEM16E	1, 2, 5	6	50	250	1,500	0.16	40.00	240.00	0.05	68
MFSD2A	3, 4, 6	8	60	250	2,000	0.07	16.67	133.33	0.03	60
Workflow testing, management and enhancement				300.00		40			60	
TOTAL FOR YEAR 1				5,800		733			278	
YEAR 2										
System	Milestones	# of constructs	# of initial conditions	Simulation time (μs)		Node hours (x10³)			Storage (TB)	
				per construct	per system	per 1 μs	per construct	per system	per 1 μs	per system
hTMEM16E	2, 5	6	50	250	1,500	0.16	40.00	240.00	0.05	75
MFSD2A	4, 6	25	60	250	6,250	0.07	16.67	416.67	0.03	188
Workflow testing, management and enhancement				300.00		40			60	
TOTAL FOR YEAR 2				8,050		697			248	
OVERALL				13,850		1,430			525	

TABLE 1: Molecular systems to be studied with atomistic MD simulations and yearly breakdown of estimated simulation times, requested node hours (in thousands), and data storage. For each year and system listed, shown are: number of constructs, number of initial conditions per construct, Milestones corresponding to each set of simulations (see **Milestone Table**), estimated simulation times per construct and per system, corresponding node hours per 1 μs trajectory, per construct, and per system (based on benchmarks in **Fig. 8**, described in **section 3.1**), and storage requirements per 1μs of trajectory data and per system. Also included for each year are simulation times, node hours, and storage required for testing and management of **workflow** algorithms and their enhancement with additional automated components for various analysis steps. The overall estimated simulation times, node hour count, and requested storage is given in the bottom row (in **red**).

In Year 1, the major focus is on the mechanisms of mTMEM16F/hTMEM16E PLS and MFSD2A systems. For **Milestone 1**, we will consider both Ca^{2+}-bound and apo wild type constructs of hTMEM16E,

and mTMEM16F (a large set of MD simulations for Ca^{2+}-bound mTMEM16F has already been collected on a Summit-like system, as described in section **2.1.1**, and is being analyzed). From the comparative analyses of these systems, we will formulate specific mechanistic hypotheses regarding the role of Ca^{2+} ions in the activation dynamics of these phospholipid scramblases (PLS). Hypotheses will be probed computationally by simulating Ca^{2+} site mutants, including E395A, D859A (in mTMEM16F numbering), and interpreted in the functional context based on the parallel experiments with these constructs in the Accardi lab using *in vitro* scramblase assay (e.g., see our collaborative publications [1-3]). Still in **Year 1**, we will carry out simulations to identify structural elements in mTMEM16F and hTMEM16E PLS that regulate opening/closing of the lipid pathway and will also initiate the evaluation of known disease mutants in hTMEM16E. Specifically, we will simulate both gain-of-function (D409G, F518A, Y563A) and loss-of-function (Q559K, M522P) mTMEM16F mutants[11,71], as well as two hTMEM16E disease mutants: gain-of-function T513I, and loss-of-function R547Q[19-21]. Comparative analyses of these systems will allow us to accomplish **Milestone 2** by ***(i)***-determining a minimal set of structural components required for scrambling; and ***(ii)***-generating a hypothesis regarding the link of dysfunction to structure-based functional mechanisms in mammalian TMEM16 PLS. The evaluation of disease mutants will continue in **Year 2 for** the human PLS, hTMEM16E, (i.e., gain-of function S500F, C360Y, G518E, C356G, R215G mutants, and loss-of-function S555I mutant [19-21]) in order to attain **Milestone 5** (*deliver a structure-based framework for the design of modulation and repair strategies of mammalian TMEM 16 PLS*).

For the MFSD2A system, we will simulate **in Year 1** the wild type, and the mutants in the hydrophilic cavity (E312D, D92A, and R85A), in the environment of either Na^+Cl^- or Li^+Cl^-. As starting conformations for these systems we will use the outward-facing and inward-facing cryo-EM structures, as well as the *intermediate structures* between the *occluded state* we have discovered as described in our preliminary studies (**section 2.1.4)**, and the *inward facing conformation* (see workflow in section **2.5**). These computations will probe the molecular mechanisms of transport in MFSD2A and establish the mechanistic role for Na^+ ion in lipid transport. This will enable us to identify MFSD2A mutants predicted to enhance or reduce lipid transport phenotypes for evaluation in collaborative structure-function experiments in order to accomplish **Milestone 3**. Still **in Year 1**, we will initiate the studies aiming to uncover the structural basis for the substrate specificity of MFSD2A. To this end we will simulate the wild type and mutant MFSD2A interacting with various substrate lysolipids (LPC with tail lengths of 14 carbons or longer, LPE, and LPS) and non-transported compounds (including LPC with tail lengths between 6 and 12 carbons, LPA, and S1P). These efforts will continue **in Year 2** to fulfill **Milestone 4** of predicting mutations that enable previously non-transported lipids to become substrates and *vice versa*. Complete identification of a minimal set of structural components that are required for Na^+-dependent lipid transport in MFSD2A and for substrate specificity will enable us to address **Milestone 6** to initiate design of small molecule compounds that will act as substrates or inhibitors at MFSD2A.

2.4 Major methodology for achieving the objectives and Milestones from the analysis of massive trajectory data: Machine learning (ML) algorithms; Automated adaptive MD sampling and convergence; Kinetic models from MSMs; and Quantitative allosteric mechanisms.

The abbreviated (by space constraints) description of methods for the analysis of the extensive trajectory needed to describe the complex mechanisms, focuses on advanced approaches that can handle successfully the massive data on ample computational resources requested in this application.

2.4.1 *ML-based analyses of MD data*: We have developed and implemented state-of-the-art ML-based tools for: (i) identification of structurally and dynamically common/divergent features between various sets of MD data (e.g., wild type and mutant system, protein bound to different ligands); and (ii) detection of rare conformational events during MD trajectory and identification of temporal correlations between various structural loci [25,26]. We are developing these protocols to extract function-related, construct-specific information encoded in the MD trajectories and perform classification tasks with high accuracy to identify molecular determinants of mechanism in the context of protein structure and function [25,26]. The use such ML approaches to detect dynamic events in TMEM16 PLS and MFSD2A will enable us to classify **(i)**

mutations, **(ii)** structural motifs, and **(iii)** ligand/substrate binding effects according to the common dynamic mechanisms they induce. This information is an essential component for the protein engineering goals.

2.4.2 *Construction of Markov State Models (MSMs)*: The energy surfaces describing the complex processes of lipid translocation by the molecular machines we study contain **local minima and barriers that are not readily crossed in conventional MD simulations**. This makes it difficult to achieve the needed level of configurational space sampling (e.g., see[72] and citations therein). To overcome this sampling problem we use a computational strategy (section **2.5** and **Fig. 7**) that takes advantage of dimensionality reduction of the trajectory data with tICA (time independent component analysis) (see **2.4.2.1**, below), and the construction of MSMs[57,73-83]. We [52,55,57] and many others have shown the advantages of using the powerful capabilities of this strategy to describe quantitatively molecular mechanisms and kinetics of functional mechanisms of various membrane proteins[75,76], protein folding[74,84], and protein-protein interactions[80]. MSM implementation and validation is well documented in several reviews[77-79]. Briefly, in constructing the MSM from an ensemble of MD trajectories (see **2.4.2.2**, below), the conformational space **from all trajectories** is reduced by a transformation to the lower dimensionality tICA space (see section **2.4.2.1** and our publications[1,2,52,55,57]). The reduced space is discretized into multiple (~100-1000) zones (microstates) using automated clustering algorithms, and a transition count matrix (TCM) is constructed and symmetrized to satisfy detailed balance and local equilibrium[85]. Transition probabilities among microstates are calculated by normalizing the TCM to obtain the Transition Probability Matrix. Our implementations of the method with validity and quality controls are well documented [52,55,57].

2.4.2.1 *Dimensionality reduction using the tICA approach*: For a dynamic system with multiple kinetic modes, the dimensionality reduction used in mechanistic analysis and in the construction of an MSM removes the redundant information stored in the atomic coordinates and provides a framework for accurate clustering of conformations. This can be achieved by the tICA transformation which projects the conformations of the system on its slowest reaction coordinates [86-89]. This tICA space is spanned by vectors defined in terms of specific parameters which we choose as illustrated in detail in our recent publications [55,57], to be collective variables (CVs) that describe the functional dynamics of the molecular system (e.g., reflecting specific motions, distance parameters describing gate openings and conformational rearrangements, ligand binding/dissociation etc.). Notably, in our established protocols the CV selection process is strengthened by application of the ML approaches described in section **2.4.1** which reveal rare dynamic events in the trajectories that are functionally relevant and thus highly efficient as CVs.

2.4.2.2 *Parameters for MSM construction and Transition Path Theory (TPT) analysis*: As we demonstrated [55,57], an optimal combination of parameters can be achieved with the well-documented scoring method "generalized matrix Rayleigh quotient" (GMRQ) [77,85,90]. To obtain the most probable pathways for a kinetic process (e.g., lipid scrambling, selective lipid transport, binding/unbinding of functional ions) from the MSMs, we use the TPT approach. Microstates in the tICA are first grouped into macrostates based on their kinetic similarity, and TPT employs graph theory, specifically the Dijkstra algorithm, to extract and quantify the most probable pathways connecting the macrostates [91]. The Robust Perron Cluster Analysis (PCCA+) [92] algorithm is used to facilitate visualization of most probable pathways.

2.4.3 *Quantifying the allosteric coupling between functional sites of membrane proteins*: To address the special mechanistic challenges presented by allostery in complex molecular machines such as the TMEM16 PLS and MFSD2A transporter we employ a statistical mechanics framework we have developed [93] and employed [56] to rigorously define and quantify the coupling between allosteric sites of a protein. As we demonstrated [93], its central concept – the Thermodynamic Coupling Function (TCF) – provides a quantitative description of how particular states, or transitions between them, are favored or opposed by allosteric coupling. The quantitative formalism estimates the contribution of particular molecular interactions to the TCF. For example, in the mechanistic analysis of the dopamine transporter (DAT)[56]**,** *we demonstrated that TCFs could be constructed in the context of MSMs, using the first tIC eigenvectors as CVs, and the microstate free energies inferred by the MSM.* This allowed us to represent, for the first time, a detailed map of allosteric couplings between functional sites on DAT[56]. Here we will use TCF to quantify the allosteric coupling between: **(i)** the distal Ca^{2+} binding site and the EC vestibule residues in TMEM16 PLS, with the goal of determining the functional role of Ca^{2+} ions in activation (**Milestone 1**); and **(ii)** the

substrate and ion binding sites and the functional IC gates in MFSD2A to determine allosteric mechanisms enabling functional dynamics in this transporter (**Milestone 3**).

2.4.4 ***Convergence probing –*** *Identification of under-sampled regions of phase space and adaptive sampling using MSMs:* The algorithms described above were demonstrated to yield robust quantitative results provided the phase space of a system is well-sampled. Still, even very long MD trajectories may fail to provide adequate sampling as they can become trapped in locally stable states, leaving unexplored important quantitative details of the kinetic model. We are using **adaptive sampling strategies** to address and overcome such difficulties with additional MD simulations capable of enhancing the statistics, by leveraging the MSMs. One such strategy involves the ensemble-biased sampling/adaptive protocol Fluctuation Amplification of Specific Traits (FAST) that was shown to increase exploration of the configurational space by several orders of magnitude [94,95]. Briefly, we determine statistical uncertainties in the elements of an existing transition matrix, and evaluate which transitions contribute the most to the statistical error of properties of interests (e.g., lipid transport, vestibule opening, destabilization of ions, etc.). **The process runs on Summit** (see section **3.2)** in parallel with the simulations; it **automatically** identifies the states that affect the accuracy of the model and initiates new sets of trajectories from these poorly connected states while stopping trajectories in well-sampled regions. In our protocols, this procedure is further strengthened by application of the ML approaches (section **2.4.1)** to extract additional leads for selecting conformations as inputs for subsequent MD simulation in our adaptive sampling strategy.

2.5 Computational approach and workflow

Our project workflow employs the specialized software, OpenMM[96], which runs MD simulations efficiently on Summit's NVIDIA Tesla V100s, not only in a special protocol of massively parallel production (see specifics in **3.1**), but also adaptively in the framework of MSMs. **Fig. 7** shows the protocol for atomistic MD simulations, **tICA dimensionality reduction and analysis, and MSM construction "*on the fly*",** for the molecular systems of interest. Initially, ensemble MD simulations are run with hundreds of statistically independent replicates (see section **3.1**). CVs of interest are then extracted from the individual trajectories to serve as parameters in the global dimensionality reduction analysis with tICA in the MSM framework [52,55,57]. Convergence of the ensemble is evaluated by ***identifying under-sampled regions of the phase space (section 2.4.4), and as needed, new sets of simulations are initiated by re-seeding velocities on a set of trajectory frames selected*** the allosteric coupling between various collective variables of interest. ***As detailed further in Section 3.2, all simulations, including the iterative convergence tests and analyses, will be executed on Summit/Andes clusters in an automated manner, from the poorly sampled regions***. CVs for all MD trajectories are then combined for the same MSM analysis and convergence probing. **Iterations are repeated until robust sampling is achieved**. Then the resulting MSMs are used to build quantitative kinetic models (see **2.4**, and our publications[52,55,57]), and to construct TCF analyses (section **2.4.2**) that quantify the allosteric coupling between various collective variables of interest.

Estimates of the run time requirements for the **TMEM16 PLS systems** (**Table 1** in **2.3**) are based on our documented experience with ensemble MD simulations of the structurally similar fungal TMEM16 PLS, as well as the mammalian mTMEM16F (see **2.1.1**). These show that the adaptive sampling for a given molecular construct requires sampling on hundreds of microsecond timescales. As discussed in section **2.2**, extensive documented experience in the field concerning the sampling of configurational space of membrane-protein systems – which includes our own work (*see **bold** entries in Bibliography*) – shows that whenever possible, a more rigorous sampling of conformational dynamics induced by local structural perturbations (e.g., mutations, or interactions with lipids, or ions), is achieved by running massively parallel multiple replica MD simulations rather than collecting a single very long MD trajectory that might exhibit single occurrences of anharmonic modes of motion. Analysis of these simulations provides the benefit of reproducing specific sets of observations in different trajectories, which helps in the formulation of robust, experimentally testable mechanistic predictions. By deploying ***swarms of independent trajectories*** at each

stage of iteration in our protocol (**Fig. 7**; see also **Section 3**) we achieve statistically meaningful sampling of functionally important conformational dynamics of the systems in the most efficient manner.

The planned simulations on MFSD2A will take advantage of the conformational sampling we have already achieved for this system in our preliminary computations. Thus, we have created intermediate structural conformations connecting the occluded and the inward facing states by using steered MD simulations (as described in section **2.1.4**, the occluded state was stable during tens of microsecond unbiased MD sampling). Together with the pathway connecting the outward facing and occluded states (which our preliminary simulations revealed), we have created large number of intermediate structures along the transport pathway which we will use for launching the adaptive ensemble MD simulations described above (**Fig. 7**). We will introduce the mutations into these intermediate conformations and with subsequent large-scale simulations we will quantify how energetics of the functional states and the transition kinetics between them are modulated by the mutations. Similarly, we will establish the effect of Na^+ ion binding at different sites in the protein on the stability of the different functional states. Lastly, we will use the ensemble MD simulations to model dynamics and spontaneous binding of the lipidic ligands (see section **2.3**) to the various conformational states of MFSD2A. Thus, the trajectory timescales for these systems accumulated on the Summit machine will involve relaxation of the models due to the perturbations introduced (i.e., mutations, binding/unbinding of ions, ligands). We note that selected ion-bound and ligand-bound protein complexes (the wild type or mutants) will be subjected to further energy-based evaluation with biased MD approaches conducted on our local computational resources, such as Free Energy Perturbation (FEP), to quantify strength of the binding under specifically defined conditions. The process of selecting relevant molecular complexes for the FEP simulations will be guided by our ML based analyses algorithms which will identify protein features related to specific phenotypes (e.g., transported vs non-transported ligands, ions that support/inhibit transport and their mode of binding).

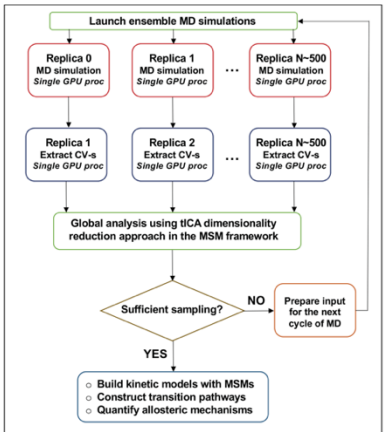

Figure 7: Sample workflow for an adaptive ensemble MD simulations leading to the construction of Markov State Models (MSMs). For each molecular construct, we consider hundreds of statistically independent replicates.Convergence of the simulations is assessed within tICA framework by identifying undersampled regions of the phase space.

3 COMPUTATIONAL READINESS

3.1 OpenMM is a highly efficient, GPU-based application for MD simulations on Summit

The requested allocation will be used for atomistic large-scale ensemble MD simulations using OpenMM 7.7.0 [96], a production software optimized to run on NVIDIA GPUs. As shown below, we found that OpenMM outperforms significantly other popular MD engines compiled on Summit when running ***massive swarms of independent simulations and collecting trajectory data from single GPUs***. Thus, the optimal strategy on Summit is to ***deploy large numbers (many 100s) of independent runs***, each taking full advantage of the computational capabilities of one GPU running OpenMM. This ***maximizes the output of the computational experiment in terms of aggregated μs/day***, which translates directly into the number of independent samples that contribute towards statistically meaningful estimation of physical observables from an MD simulation. (**Note that** statistical mechanics ensures that all data obtained from such independent runs of exactly the same systems and conditions are absolutely additive and enhance the statistical significance (sampling) of the simulations). **Using this strategy with our ongoing BIP109 allocation we are achieving persistent linear scaling across 1,000 nodes running the MD simulations.** As our strategy enables scaling over an arbitrarily large number of GPUs for the large number of independent simulations of the same system (**Fig. 8a**), the "Time to Solution" is minimized for a given construct (see inset in **Fig. 8a**). With the ***production-run-proven MD ensemble orchestration code developed with our ongoing allocation*** **(BIP109)** our work protocols utilize efficiently the ability to submit simulation swarms requiring **1,000 or more nodes of Summit for the systems described in Table 1 in section 2.3. This is essential in order to achieve the Milestones,** and running independent trajectories has several additional advantages: **i)**-optimal use of the Summit computing architecture without performance degradation due to internode communications; **ii)**-statistical independence of trajectories, which provides

improved sampling; and **iii)**-*on-the-fly* convergence monitoring, in which some trajectories can be stopped while others can be extended, and new ones can be started in poorly explored regions of configurational space (see **2.5**). **Fig. 8b** shows OpenMM 7.7.0 benchmarks obtained on a single GPU on Summit, illustrating strong performance even for the largest size systems considered. We found that the performance benefited only very modestly from increasing the number of GPUs per simulation from 1 to 2 or more (scaling across intranode GPUs was very poor). Furthermore, the performance of 6 independent OpenMM simulations on a **single Summit node** (with each simulation using 1 GPU) was identical to the performance of running just 1 OpenMM simulation on 1 GPU on a single node. Our workflow to run ***large swarms of trajectories initiated in a single batch job, with the option of including enhanced sampling algorithms*** is based on these benchmark conclusions: Our recent benchmark **(Fig. 8b)** for *single GPU per replica* achieved a performance of ~29ns/day for the TMEM16 PLS systems (~560,000 atoms), and ~60ns/day for the MFSD2A systems (~230,000 atoms). Consequently, our protocol uses **many 100s to 1,000 nodes** simultaneously to run different replicas of protein-membrane systems, each of them on a single GPU (i.e., 6 simultaneous jobs on one single Summit node). ***This approach takes full advantage of the massively parallel architecture of Summit's Volta GPUs without any of the common inefficiencies of multi-node parallel runs.*** Massive data can be accumulated from MD trajectories of otherwise unattainable time scales (save on the newest DE Shaw Anton), as needed for our studies (**see 2.2-2.4**). Thus, 1,000 statistically independent simulations of the mTMEM16F and MFSD2A systems would yield 29 and 60 µs, respectively, in just one day – an essential performance for appropriate sampling of these complex systems, and ***not achievable on any other platform***. Note that thematically different systems (e.g., wild type, mutants, w/o bound ions) can be grouped to run simultaneously in a single *batch* ***job to accumulate for each construct trajectories of tens-to-hundreds of microseconds. Results obtained with our ongoing BIP109 allocation demonstrate the need for such performance, our strong expertise in efficiently handling, running, and analyzing such large-scale MD trajectories, and the resulting high and impactful productivity (see List of publications from prior allocations).*** Molecular systems large and small, benefit greatly from such extensive sampling that generates rigorous, quantitative mechanistic models (**2.4**, and Refs. [57,52,55]) .

For performance comparisons of OpenMM 7.7.0 with other popular MD packages we benchmarked on Summit the same systems with NAMD 2.14 using identical run parameters (with the exception of the timestep: OpenMM uses 4fs, NAMD uses 2 fs). We found (**Fig. 8b**) NAMD to be significantly less efficient on a single GPU (regardless of the difference in the time-step). Importantly, we also found that NAMD did not scale well across the Summit GPUs. These benchmarks show that: **i)** running large ensembles of replica simulations is sampling-wise and computationally more efficient than running a single replica over many compute nodes, whether using OpenMM or NAMD; **ii)**-OpenMM outperforms NAMD significantly on a single GPU node. These arguments support our strategy, and illustrate that our project scales linearly to an arbitrary number of Summit nodes with OpenMM, for a typical "embarrassingly parallel" performance. Consequently, integration of our computational protocol in ***the advanced workflow we use*** **(**see **3.2)** ***will most efficiently take advantage of the Summit's GPU architecture*** for large-scale atomistic unbiased or enhanced sampling MD, as we have already demonstrated with our previous ALCC/INCITE allocations. We reiterate the broader implication of this planned massive data collected from the simulations in identifiable batches, for our NSF Award (see section **1.3**) in which **(1)-novel data analytics algorithms** are developed and tested for *in situ* data analysis of molecular properties from MD simulations, and **(2)-machine learning techniques are elaborated** to identify automatically the molecular domains where the properties reside at runtime, for integration into MD workflows at the extreme scale (see **1.3**).

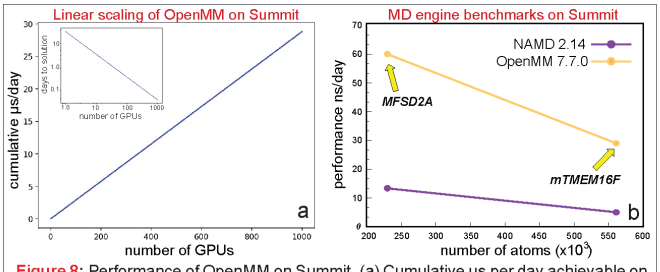

Figure 8: Performance of OpenMM on Summit. (a) Cumulative µs per day achievable on Summit vs number of GPUs with Openmm 7.7.0. The inset shows time to solution (in days) vs number of GPUs plotted on log-log scale. The time to solution was defined as cumulative number of days required to reach 1 µs for a system size of ~561,000 atoms: dimeric mTMEM16F in explicit lipid membrane and water solution.(b) Nanoseconds per day achieved on a single GPU core of Summit using OpenMM 7.7.0 (orange) and NAMD 2.14 (purple) for two different systems sizes: ~561,000 atoms (mTMEM16F) and ~230,000 million atoms (MFSD2A). The OpenMM simulations used 4fs integration time-step, whereas the NAMD runs used 2fs timestep.

3.2 Execution of the project workflow – OpenMM adaptive MD simulations on Summit

All the components of the protocol in **Fig. 7** have been rigorously tested in our current work on Summit. We have incorporated our adaptive ensemble workflow into RADICAL-Cybertools (RCT) software stack. This modular approach is based on: 1) the Ensemble Toolkit (EnTK)[97,98] – which provides the ability to create and execute ensemble-based applications with complex coordination and communication; 2) RADICALPilot (RP)[99], which provides resource management and task execution capabilities. Notably, RCT supports several INCITE/ALCC awards, many of which require adaptive sampling[100-103] and served in developing multiple ensemble-based adaptive sampling workflows on HPC platforms[100-106], similar to those used here for tICA and in other advanced sampling algorithms[60,105,106]. Briefly, once tasks have been executed, they may invoke adaptive capabilities in EnTK to add more simulations and analysis tasks to the current run based on the intermediate results generated. As shown [97,98], EnTK demonstrates negligible and constant overhead up to O(10,000) tasks, independent of task type and computing platform. In these workflows, the MD simulations carried out on Summit will use OpenMM as described in section **3.1**, and construction of MSMs and all ancillary analysis (see section **2.4.4**), will be done using components from MSMbuilder software which we have already installed on Summit.

As ***the scheme in Fig. 7 is applied to multiple molecular systems simultaneously***, a single workflow can incorporate multiple systems from **Table 1** by chaining them during MD simulation steps into a single batch job of 100s-to-1000s OpenMM replicas (each on a single GPU), and analyzing them individually *in real time* so that convergence tests and required analyses are performed as described (**2.4.4** and **2.5**) for separate molecular constructs. ***For ensembles, the runs will require 1,000 or more nodes simultaneously to achieve efficient progress to solution.*** Indeed, the RP abstraction achieves scalable execution of applications consisting of large ensembles of tasks by (i)-using a placeholder job to acquire resources via the local resource management system (LRMS), and (ii)-decoupling the initial resource acquisition from task-to-resource assignment. Once the pilot is scheduled via the LRMS, it can in turn schedule computational tasks on the available resources. This functionality allows for all the computational tasks to be executed directly on the resources, without being individually queued to the LRMS. Thus, this approach supports the requirements of task-level parallelism and high-throughput. RP utilizes IBM's Platform Load Sharing Facility (LSF) and its JSRUN tool to schedule both its Executor component and computational tasks on Summit's work nodes. RP can concurrently run multiple Executor components, enabling each Executor to execute tasks on a subset of the available work nodes. This allows RP to scale to increasing proportions of Summit, while guaranteeing fault-tolerance across Executor components.

3.3 Construction of molecular simulations and simulation force-fields

Our publications[1,52,55,57,107,108] contain details and illustrations of the preparatory stages including modeling the functionally relevant environments (lipid, solvent), and equilibration protocols. OpenMM production runs are performed under NPT ensemble, using semi-isotropic pressure coupling, and with all the default run parameters validated by OpenMM developers[96]: 4fs time step with hydrogen mass repartitioning; PME for electrostatics; switched Lennard-Jones interactions with an extended cut-off of 12Å, and switching distance set to 10Å. We use CHARMM36 force field for lipids, proteins, and ions.

3.4 Data storage requirement and software workflow solution

The generated raw data consists of MD trajectory files in compressed XTC format, which contain frame-by-frame coordinates of atoms in the simulated systems, outputted with a fine stride necessary to obtain sufficient sampling of the data for reliable quantitative analyses. It is our experience that original data must be retained on Summit for at least 1yr to enable comparative analyses with newer trajectories.

3.5 Computational readiness for the Frontier supercomputer

We are well-positioned for Summit to Frontier transition. Our workflow framework RADICAL-Cybertools supports multiple job schedulers, including LSF and Slurm. This workflow software stack has been successfully run on x86_64-based hardware by many groups; this also holds true for all MD simulation and analysis software (and related libraries) described in this application. Furthermore, the OpenMM MD engine supports both CUDA and OpenCL-based GPU acceleration—so we are prepared for the transition from NVIDIA to AMD GPUs. ***Altogether, all of our proposed work is 'lift and shift' ready, such that we do not anticipate any significant downtime when migrating from Summit to Frontier.***

BIBLIOGRAPHY

(Citations authored by Harel Weinstein lab members are **denoted in bold**)

1 **Lee, B. C. *et al.* Gating mechanism of the extracellular entry to the lipid pathway in a TMEM16 scramblase. *Nature Communications* 9, 3251 (2018).**

2 **Khelashvili, G. *et al.* Dynamic modulation of the lipid translocation groove generates a conductive ion channel in Ca(2+)-bound nhTMEM16. *Nat Commun* 10, 4972 (2019).**

3 **Khelashvili, G. *et al.* Membrane lipids are both the substrates and a mechanistically responsive environment of TMEM16 scramblase proteins. *J Comput Chem* 41, 538-551 (2020).**

4 **Cater, R. J. *et al.* Structural basis of omega-3 fatty acid transport across the blood-brain barrier. *Nature* 595, 315-319 (2021).**

5 Falzone, M. E., Malvezzi, M., Lee, B. C. & Accardi, A. Known structures and unknown mechanisms of TMEM16 scramblases and channels. *J Gen Physiol* **150**, 933-947 (2018).

6 Bevers, E. M. & Williamson, P. L. Getting to the Outer Leaflet: Physiology of Phosphatidylserine Exposure at the Plasma Membrane. *Physiol Rev* **96**, 605-645 (2016).

7 Falzone, M. *et al.* Structural basis of Ca2+-dependent activation and lipid transport by a TMEM16 scramblase. *eLIFE* **8**, pii: e43229 (2019).

8 Malvezzi, M. *et al.* Out-of-the-groove transport of lipids by TMEM16 and GPCR scramblases. *Proc Natl Acad Sci U S A* **115**, E7033-E7042 (2018).

9 Lee, B. C., Menon, A. K. & Accardi, A. The nhTMEM16 Scramblase Is Also a Nonselective Ion Channel. *Biophys J* **111**, 1919-1924 (2016).

10 Malvezzi, M. *et al.* Ca2+-dependent phospholipid scrambling by a reconstituted TMEM16 ion channel. *Nat Commun* **4**, 2367 (2013).

11 Jiang, T., Yu, K., Hartzell, H. C. & Tajkhorshid, E. Lipids and ions traverse the membrane by the same physical pathway in the nhTMEM16 scramblase. *Elife* **6** (2017).

12 Bethel, N. P. & Grabe, M. Atomistic insight into lipid translocation by a TMEM16 scramblase. *Proc Natl Acad Sci U S A* **113**, 14049-14054 (2016).

13 Stansfeld, P. J. *et al.* MemProtMD: Automated Insertion of Membrane Protein Structures into Explicit Lipid Membranes. *Structure* **23**, 1350-1361 (2015).

14 Alvadia, C. *et al.* Cryo-EM structures and functional characterization of the murine lipid scramblase TMEM16F. *Elife* **8** (2019).

15 Feng, S. *et al.* Cryo-EM Studies of TMEM16F Calcium-Activated Ion Channel Suggest Features Important for Lipid Scrambling. *Cell Rep* **28**, 567-579 e564 (2019).

16 Bushell, S. R. *et al.* The structural basis of lipid scrambling and inactivation in the endoplasmic reticulum scramblase TMEM16K. *Nat Commun* **10**, 3956 (2019).

17 Kalienkova, V. *et al.* Stepwise activation mechanism of the scramblase nhTMEM16 revealed by cryo-EM. *Elife* **8** (2019).

18 Kalienkova, V., Clerico Mosina, V. & Paulino, C. The Groovy TMEM16 Family: Molecular Mechanisms of Lipid Scrambling and Ion Conduction. *J Mol Biol*, 166941 (2021).

19 Di Zanni, E., Gradogna, A., Picco, C., Scholz-Starke, J. & Boccaccio, A. Phospholipid Scrambling Activity by TMEM16E/Ano5: Opposite Effects of Mutations Causing Bone Dysplasia and Muscular Dystrophy. *Biophys J* **116**, Supplement 1, 223A (2019).

20 Di Zanni, E., Gradogna, A., Scholz-Starke, J. & Boccaccio, A. Gain of function of TMEM16E/ANO5 scrambling activity caused by a mutation associated with gnathodiaphyseal dysplasia. *Cell Mol Life Sci* **75**, 1657-1670 (2018).

21 Marconi, C. *et al.* A novel missense mutation in ANO5/TMEM16E is causative for gnathodiaphyseal dyplasia in a large Italian pedigree. *European journal of human genetics : EJHG* **21**, 613-619 (2013).

22 Nguyen, L. N. *et al.* Mfsd2a is a transporter for the essential omega-3 fatty acid docosahexaenoic acid. *Nature* **509**, 503-506 (2014).
23 Wood, C. A. P. *et al.* Structure and mechanism of blood-brain-barrier lipid transporter MFSD2A. *Nature* **596**, 444-448 (2021).
24 Kots, E., Shore, D. M. & Weinstein, H. Adaptive Sampling using a Geometric Brownian Motion Model to Predict MD Trajectory Mobility on a Free Energy Surface. *Biophys J* 120, 78a (2021).
25 Plante, A. & Weinstein, H. Ligand-Dependent Conformational Transitions in Molecular Dynamics Trajectories of GPCRs Revealed by a New Machine Learning Rare Event Detection Protocol. *Molecules* 26 (2021).
26 Plante, A., Shore, D. M., Morra, G., Khelashvili, G. & Weinstein, H. A Machine Learning Approach for the Discovery of Ligand-Specific Functional Mechanisms of GPCRs. *Molecules* 24 (2019).
27 Carrillo-Cabada, H. *et al.* A Graphic Encoding Method for Quantitative Classification of Protein Structure and Representation of Conformational Changes. *IEEE/ACM Trans Comput Biol Bioinform* (2019).
28 Do, T. M. A. *et al.* A lightweight method for evaluating in situ workflow efficiency. *Journal of Computational Science* **48**, 101259 (2021).
29 Griffin, D. A. *et al.* Defective membrane fusion and repair in Anoctamin5-deficient muscular dystrophy. *Hum Mol Genet* **25**, 1900-1911 (2016).
30 Hicks, D. *et al.* A founder mutation in Anoctamin 5 is a major cause of limb-girdle muscular dystrophy. *Brain* **134**, 171-182 (2011).
31 Tsutsumi, S. *et al.* The novel gene encoding a putative transmembrane protein is mutated in gnathodiaphyseal dysplasia (GDD). *Am J Hum Genet* **74**, 1255-1261 (2004).
32 Jin, L. *et al.* Three novel ANO5 missense mutations in Caucasian and Chinese families and sporadic cases with gnathodiaphyseal dysplasia. *Sci Rep* **7**, 40935 (2017).
33 Lhermusier, T., Chap, H. & Payrastre, B. Platelet membrane phospholipid asymmetry: from the characterization of a scramblase activity to the identification of an essential protein mutated in Scott syndrome. . *J Thromb Haemost* **9**, 1883-1891 (2011).
34 Castoldi, E., Collins, P. W., Williamson, P. L. & Bevers, E. M. Compound heterozygosity for 2 novel TMEM16F mutations in a patient with Scott syndrome. *Blood* **117**, 4399-4400 (2011).
35 Chamova, T., Florez, L. & Guergueltcheva, V. ANO10 c.1150_1151del is a founder mutation causing autosomal recessive cerebellar ataxia in Roma/Gypsies. *J Neurol* **259**, 906-911 (2012).
36 Renaud, M. *et al.* Autosomal Recessive CerebellarAtaxia Type 3 Due to ANO10 Mutations. *JAMA Neurology* **71**, 1305-1310 (2014).
37 Vermeer, S. *et al.* Targeted next-generation sequencing of a 12.5 Mb homozygous region reveals ANO10 mutations in patients with autosomal-recessive cerebellar ataxia. *Am J Hum Genet* **87**, 813-819 (2010).
38 Wanitchakool, P. *et al.* Cellular defects by deletion of ANO10 are due to deregulated local calcium signaling. *Cellular Signalling* **30**, 41-49 (2017).
39 Balreira, A. *et al.* ANO10 mutations cause ataxia and coenzyme Q(1)(0) deficiency. *J Neurol* **261**, 2192-2198 (2014).
40 Bolduc, V. *et al.* Recessive mutations in the putative calcium-activated chloride channel Anoctamin 5 cause proximal LGMD2L and distal MMD3 muscular dystrophies. *Am J Hum Genet* **86**, 213-221 (2010).
41 Brooks, M. B. *et al.* A TMEM16F point mutation causes an absence of canine platelet TMEM16F and ineffective activation and death-induced phospholipid scrambling. *J Thromb Haemost* **13**, 2240-2252 (2015).
42 Huang, B. & Li, X. The Role of Mfsd2a in Nervous System Diseases. *Front Neurosci* **15**, 730534 (2021).

43 Wong, B. H. & Silver, D. L. Mfsd2a: A Physiologically Important Lysolipid Transporter in the Brain and Eye. *Adv Exp Med Biol* **1276**, 223-234 (2020).
44 Estrada, T. *et al.* Graphic Encoding of Macromolecules for Efficient High-Throughput Analysis. *Proc. of 9th ACM Conference on Bioinformatics, Computational Biology, and Health Informatics (ACM BCB)* (2018).
45 Heath, G. R. *et al.* Localization atomic force microscopy. *Nature* 594, 385-390 (2021).
46 LeVine, M. V. & Weinstein, H. NbIT--a new information theory-based analysis of allosteric mechanisms reveals residues that underlie function in the leucine transporter LeuT. *PLoS Comput Biol* 10, e1003603 (2014).
47 Mondal, S., Khelashvili, G., Shi, L. & Weinstein, H. The cost of living in the membrane: A case study of hydrophobic mismatch for the multi-segment protein LeuT. *Chem Phys Lipids* 169, 27-38 (2013).
48 Mondal, S., Khelashvili, G., Shan, J., Andersen, O. S. & Weinstein, H. Quantitative modeling of membrane deformations by multi-helical membrane proteins: Application to G-protein Coupled Receptors. *Biophys J* 101, 2092-2101 (2011).
49 Mondal, S., Khelashvili, G. & Weinstein, H. Not Just an Oil Slick: How the Energetics of Protein-Membrane Interactions Impacts the Function and Organization of Transmembrane Proteins. *Biophys J* 106, 2305-2316 (2014).
50 LeVine, M. V., Perez-Aguilar, J. M. & Weinstein, H. N-body Information Theory (NbIT) Analysis of Rigidi-Body Dynamics in Intracellular Loop 2 of the 5-HT2A Receptor. *Proceedings IWBBIO*, Granada 7-9 (2014).
51 Khelashvili, G. *et al.* Unusual mode of dimerization of retinitis pigmentosa-associated F220C rhodopsin. *Sci Rep* 11, 10536 (2021).
52 Morra, G. *et al.* Mechanisms of lipid scrambling by the G protein-coupled receptor opsin. *Structure* 26, 356-367 (2017).
53 Mondal, S. *et al.* Membrane driven spatial organization of GPCRs. *Sci Rep* 3, 2909 (2013).
54 LeVine, M. V. *et al.* The allosteric mechanism of substrate-specific transport in SLC6 is mediated by a volumetric sensor. *Proc Natl Acad Sci U S A* 116, 15947-15956 (2019).
55 Razavi, A. M., Khelashvili, G. & Weinstein, H. How structural elements evolving from bacterial to human SLC6 transporters enabled new functional properties. *BMC biology* 16, 31 (2018).
56 LeVine, M. V., Cuendet, M. A., Razavi, A. M., Khelashvili, G. & Weinstein, H. Thermodynamic Coupling Function Analysis of Allosteric Mechanisms in the Human Dopamine Transporter. *Biophys J* 114, 10-14 (2018).
57 Razavi, A. M., Khelashvili, G. & Weinstein, H. A Markov State-based Quantitative Kinetic Model of Sodium Release from the Dopamine Transporter. *Sci Rep* 7, 40076 (2017).
58 Brunner, J. D., Lim, N. K., Schenck, S., Duerst, A. & Dutzler, R. X-ray structure of a calcium-activated TMEM16 lipid scramblase. *Nature* **516**, 207-212 (2014).
59 Boccaccio, A., Di Zanni, E., Gradogna, A. & Scholz-Starke, J. Lifting the veils on TMEM16E function. *Channels (Austin)* **13**, 33-35 (2019).
60 Kasson, P. M. & Jha, S. Adaptive ensemble simulations of biomolecules. *Curr Opin Struct Biol* **52**, 87-94 (2018).
61 LeVine, M. V., Cuendet, M. A., Khelashvili, G. & Weinstein, H. Allosteric Mechanisms of Molecular Machines at the Membrane: Transport by Sodium-Coupled Symporters. *Chemical reviews* (2016).
62 Stolzenberg, S., Michino, M., LeVine, M. V., Weinstein, H. & Shi, L. Computational approaches to detect allosteric pathways in transmembrane molecular machines. *Biochimica et biophysica acta* (2016).

63 Stolzenberg, S. *et al.* Mechanism of the Association between Na+ Binding and Conformations at the Intracellular Gate in Neurotransmitter:Sodium Symporters. *The Journal of biological chemistry* (2015).

64 Khelashvili, G. & Weinstein, H. Functional mechanisms of neurotransmitter transporters regulated by lipid-protein interactions of their terminal loops. *Biochimica et biophysica acta* 1848, 1765 (2015).

65 Khelashvili, G. *et al.* Spontaneous Inward Opening of the Dopamine Transporter Is Triggered by PIP-Regulated Dynamics of the N-Terminus. *ACS chemical neuroscience* 6, 1825-1837 (2015).

66 Kazmier, K. *et al.* Conformational dynamics of ligand-dependent alternating access in LeuT. *Nat Struct Mol Biol* 21, 472-479 (2014).

67 Zhao, C. *et al.* Ion-controlled conformational dynamics in the outward-open transition from an occluded state of LeuT. *Biophys J* 103, 878-888 (2012).

68 Shan, J., Javitch, J. A., Shi, L. & Weinstein, H. The substrate-driven transition to an inward-facing conformation in the functional mechanism of the dopamine transporter. *PLoS One* 6, e16350 (2011).

69 Zhao, Y. *et al.* Substrate-modulated gating dynamics in a Na+-coupled neurotransmitter transporter homologue. *Nature* 474, 109-113 (2011).

70 Hamilton, P. J. *et al.* PIP2 regulates psychostimulant behaviors through its interaction with a membrane protein. *Nat Chem Biol* 10, 582-589 (2014).

71 Le, T. *et al.* An inner activation gate controls TMEM16F phospholipid scrambling. *Nat Commun* **10**, 1846 (2019).

72 Brooks, C. L., 3rd *et al.* Classical molecular dynamics. *The Journal of chemical physics* **154**, 100401 (2021).

73 Voelz, V. A., Bowman, G. R., Beauchamp, K. & Pande, V. S. Molecular simulation of ab initio protein folding for a millisecond folder NTL9(1-39). *Journal of the American Chemical Society* **132**, 1526-1528 (2010).

74 Beauchamp, K. A., McGibbon, R., Lin, Y. S. & Pande, V. S. Simple few-state models reveal hidden complexity in protein folding. *Proc Natl Acad Sci U S A* **109**, 17807-17813 (2012).

75 Kohlhoff, K. J. *et al.* Cloud-based simulations on Google Exacycle reveal ligand modulation of GPCR activation pathways. *Nature chemistry* **6**, 15-21 (2014).

76 Shukla, D., Meng, Y., Roux, B. & Pande, V. S. Activation pathway of Src kinase reveals intermediate states as targets for drug design. *Nat Commun* **5**, 3397 (2014).

77 Prinz, J. H. *et al.* Markov models of molecular kinetics: generation and validation. *The Journal of chemical physics* **134**, 174105 (2011).

78 Noe, F. & Fischer, S. Transition networks for modeling the kinetics of conformational change in macromolecules. *Curr Opin Struct Biol* **18**, 154-162 (2008).

79 Pande, V. S., Beauchamp, K. & Bowman, G. R. Everything you wanted to know about Markov State Models but were afraid to ask. *Methods* **52**, 99-105 (2010).

80 Plattner, N., Doerr, S., De Fabritiis, G. & Noe, F. Complete protein–protein association kinetics in atomic detail revealed by molecular dynamics simulations and Markov modelling. *Nature chemistry* **9**, 1005-1011 (2017).

81 Wieczorek, M. *et al.* Major Histocompatibility Complex (MHC) Class I and MHC Class II Proteins: Conformational Plasticity in Antigen Presentation. *Frontiers in immunology* **8**, 292 (2017).

82 Noe, F. & Clementi, C. Collective variables for the study of long-time kinetics from molecular trajectories: theory and methods. *Curr Opin Struct Biol* **43**, 141-147 (2017).

83 Pinamonti, G. *et al.* Predicting the Kinetics of RNA Oligonucleotides Using Markov State Models. *Journal of chemical theory and computation* **13**, 926-934 (2017).

84 Voelz, V. A., Bowman, G. R., Beauchamp, K. & Pande, V. S. Molecular Simulation of ab Initio Protein Folding for a Millisecond Folder NTL9. *Journal of the American Chemical Society* **132**, 1526–1528 (2009).

85 Beauchamp, K. A. *et al.* MSMBuilder2: Modeling Conformational Dynamics at the Picosecond to Millisecond Scale. *Journal of chemical theory and computation* **7**, 3412-3419 (2011).
86 Schwantes, C. R. & Pande, V. S. Improvements in Markov State Model Construction Reveal Many Non-Native Interactions in the Folding of NTL9. *Journal of chemical theory and computation* **9**, 2000-2009 (2013).
87 Molgedey, L. & Schuster, H. G. Separation of a mixture of independent signals using time delayed correlations. *Physical review letters* **72**, 3634-3637 (1994).
88 Naritomi, Y. & Fuchigami, S. Slow dynamics in protein fluctuations revealed by time-structure based independent component analysis: the case of domain motions. *The Journal of chemical physics* **134**, 065101 (2011).
89 Perez-Hernandez, G., Paul, F., Giorgino, T., De Fabritiis, G. & Noe, F. Identification of slow molecular order parameters for Markov model construction. *The Journal of chemical physics* **139**, 015102 (2013).
90 McGibbon, R. T. & Pande, V. S. Variational cross-validation of slow dynamical modes in molecular kinetics. *The Journal of chemical physics* **142**, 124105 (2015).
91 Dijkstra, E. W. A note on two problems in connexion with graphs. *Numerische mathematik* **1**, 269-271 (1959).
92 Deuflhard, P. & Weber, M. Robust Perron cluster analysis in conformation dynamics. *Linear algebra and its applications* **398**, 161-184 (2005).
93 Cuendet, M. A., Weinstein, H. & LeVine, M. V. The Allostery Landscape: Quantifying Thermodynamic Couplings in Biomolecular Systems. *Journal of chemical theory and computation* 12, 5758-5767 (2016).
94 Zimmerman, M. I. & Bowman, G. R. FAST Conformational Searches by Balancing Exploration/Exploitation Trade-Offs. *Journal of chemical theory and computation* **11**, 5747-5757 (2015).
95 Zimmerman, M. I. & Bowman, G. R. How to Run FAST Simulations. *Methods Enzymol* **578**, 213-225 (2016).
96 Eastman, P. *et al.* OpenMM 7: Rapid development of high performance algorithms for molecular dynamics. *PLoS Comput Biol* **13**, e1005659 (2017).
97 Balasubramanian, V. *et al.* Harnessing the power of many: Extensible toolkit for scalable ensemble applications. *IEEE International Parallel and Distributed Processing Symposium (IPDPS)*, 536-545 (2018).
98 Balasubramanian, V., Trekalis, A., Weidner, O. & S., J. Ensemble Toolkit: Scalable and Flexible Execution of Ensembles of Tasks. *In Proceedings of the 45th International Conference on Parallel Processing (ICPP), BioRxiv, arXiv:1602.00678v3* (2016).
99 Merzky, A., Santcroos, M., Turilli, M. & Jha, S. Executing Dynamic and HeterogeneousWorkloads on Super Computers. *BioRxiv , arXiv:1512.08194v4* (2016).
100 Dakka, J. *et al.* Enabling trade-offs between accuracy and computational cost: Adaptive algorithms to reduce time to clinical insight. *IEEE*, 572-577 (2018).
101 Dakka, J. *et al.* Concurrent and adaptive extreme scale binding free energy calculations. *14th IEEE International Conference on e-Science* **e-Science**, 189-200 (2018).
102 Radak, B. K. *et al.* Characterization of the three-dimensional free energy manifold for the Uracil Ribonucleoside from asynchronous replica exchange simulations. *Journal of chemical theory and computation* **11**, 377-377 (2015).
103 Shkurti, A. *et al.* Coco-md: A simple and effective method for the enhanced sampling of conformational space. *Journal of chemical theory and computation* **15**, 2587-2596 (2019).
104 Trekalis, A. *et al.* Repex: A flexible framework for scalable replica exchange molecular dynamics simulations. *IEEE*, 628-637 (2016).
105 Balasubramanian, V. *et al.* Implementing adaptive ensemble biomolecular applications at scale. *CoRR abs/1804.04736* (2018).

106 Balasubramanian, V. *et al.* Extasy: Scalable and flexible coupling of md simulations and advanced sampling techniques. *IEEE* **12th International Conference on e-Science (e-Science)**, 361-370 (2016).

107 Doktorova, M., LeVine, M. V., Khelashvili, G. & Weinstein, H. A New Computational Method for Membrane Compressibility: Bilayer Mechanical Thickness Revisited. *Biophys J* 116, 487-502 (2019).

108 Doktorova, M. & Weinstein, H. Accurate In Silico Modeling of Asymmetric Bilayers Based on Biophysical Principles. *Biophys J* 115, 1638-1643 (2018).

PERSONNEL JUSTIFICATION AND MANAGEMENT PLAN

The personnel for this project are already in place and ready to continue the work on the ongoing project as described. All are experienced with Summit and OLCC resources from work using current and past INCITE allocations. All have direct experience with all aspects of the computational project described in this application, as well as with access to, and utilization of, shared and cloud computational resources.

Harel Weinstein, D.Sc. – PI, will be responsible for designing and directing all aspects of the research in this Project, including the design of computational studies in the collaborative investigations with the experimentalist members of the NIH-supported grants and the NSF Collaborative Project for which he serves as PI. He will be coordinating the analysis and evaluation of all data and will be responsible for the reports of results and for supervising the work of his lab members.

George Khelashvili, Ph.D. – co-PI, is an Assistant Professor with an outstanding educational background and research experience in biophysics, and a strong record of publication in computational molecular biophysics. Areas of special expertise are membrane structure and dynamics, and membrane-protein interactions, and in particular systems of NSS transporters and TMEM16 scramblases. He has developed and published advanced methods for quantitative analysis of membrane protein systems, has made significant discoveries about membrane biophysics and the effects of membranes on the structure and function of membrane proteins, and has acquired ample expertise in high performance computing including on Titan, Summit (he is co-PI of the ongoing INCITE allocation).

The leadership of this INCITE proposal is composed of the PI and co-PI, and the scientific project is managed collaboratively by a team of investigators under the direction of the PI, Harel Weinstein, who is the head of the laboratory and of the academic unit. The research plan in the application was composed and written collaboratively by this team. The PI of this application and the team benefit from systematic and regular interactions with senior members of the consortia formed for the NIH-supported grants on which the PI of this application is either PI or co-PI. Consortia members include:

Alessio Accardi, Ph.D. Professor of Physiology and Biophysics in Anesthesiology, Professor of Biochemistry and Professor of Physiology and Biophysics. He will continue to collaborate directly on this project which is funded by a collaborative NIH grant (Accardi/Weinstein, MPIs). The Accardi lab is one of the leading labs in the world in structure/function studies of TMEM16 scramblases and channel proteins, and also studies other membrane protein families, such as CLC channels. The Accardi lab uses a combination of biophysical, structural (Cryo-EM), and physiological approaches to study these molecular machines and to elucidate the structural and mechanistic underpinnings of ion and lipid transport across biological cell membranes.

Filippo Mancia, Ph.D. Professor, Co-Director of Graduate Education in the Department of Physiology and Cellular Biophysics, Columbia University. He is a structural biologist with strong expertise in x-ray crystallography, single particle cryo-electron microscopy, and in production and characterization of membrane proteins for structural studies. The Mancia lab focuses on membrane protein-lipid interactions, both in terms of enzymes, which process lipid substrates, and of transporters, which mediate cellular uptake of lipidic substrates. As described in the application, he has a continuing established collaboration with the Co-PI of this application, George Khelashvili, and the approaches developed and utilized in the Mancia lab will be used in this project to address structure/function relationships in MFSD2A transporter and will serve to probe and validate the mechanistic findings from the simulations.

Shantenu Jha, PhD is the PI of the Research in Advanced Distributed Cyberinfrastructure and Applications Laboratory (RADICAL) (http://radical.rutgers.edu/) at Rutgers University. Jha is a co-PI of the NSF Molecular Sciences Software Institute (https://molssi.org/). Jha is also the Chair of the Data Driven Discovery Department at Brookhaven National Laboratory where his research interests are at the intersection of streaming data, high-performance computing, machine learning and precision medicine. Dr. Jha will continue to support this project. He will supervise his team working on the scalable adaptive ensemble-based workflows for the Project, and the continued enhancement, and testing for incorporation, of additional automated analysis tasks.

The PI of the application, Harel Weinstein, will be the contact who will provide reports, updates on the status of the work including publications, awards, and highlights of accomplishments.

Year 1		
Milestone	**Details**	**Date and Status**
1	Work on determining conditions for activation of the mammalian PLS by Ca^{2+} for experimental verification and practical use, from the mechanistic role of Ca^{2+} ions in the activation dynamics of mTMEM16F and hTMEM16E PLS. Computational probing and exploration of the hypothesis that binding of (the 3rd) distal Ca^{2+} ion allosterically triggers opening of the lipid translocation pathway. **Deliver: Residue identities in the allosteric pathway presented for collaborative experimental validation of Ca^{2+} activation of the mammalian PLS.** **Molecular constructs:** *Ca^{2+}-bound and apo wild type hTMEM16E; Apo mTMEM16F; Ca^{2+} binding site mutants in mTMEM16F and hTMEM16E: E395A, D859A (in mTMEM16F numbering).*	Begin and end in **Year 1**
2	Work on determining the structural elements responsible for the open/close regulation of the lipid pathway in mTMEM16F and hTMEM16E PLS; Based on progress in Milestone 1 and the open/close regulation, initiate the evaluation of disease mutants (DMs) in human PLS, hTMEM16E. **Deliver: (1)-The structure-based involvement of individual DMs in the minimal set of structural components required for scrambling by mammalian TMEM16; and (2) Structure-based predictions for collaborative experimental work to measure and mitigate DM-related dysfunction.** **Molecular constructs:** *Gain-of-function and inhibitory mutants in mTMEM16F: D409G, Q559K, M522P, F518A, Y563A; Disease mutants T513I, and R547Q in hTMEM16E.*	Begin in **Year 1** and Finish in **Year 2**
3	Work on determining the function of Na^+ ion in lipid transport by MFSD2A by probing first the hypotheses from structure-function experiments that (1)-binding of the Na^+ ion at E312 stabilizes an occluded state; (2) binding of Na^+ ion to D92 triggers inward opening. **Deliver: MFSD2A molecular constructs (mutants) predicted to (i)-enhance or (ii)-reduce lipid transport phenotypes for evaluation in collaborative structure-function experiments.** **Molecular constructs:** *Wild type MSFD2A and D92A, E312D, R85A mutants all in the environment of Na^+Cl^-; Wild type MFSD2A in the environment of Li^+Cl^-.*	Begin and end in **Year 1**
4	Work on determining the structural basis for the substrate specificity of MFSD2A. **Deliver: Predicted mutations that enable previously non-transported lipids to become substrates, and *vice versa* for evaluation in collaborative structure-function experiments.** **Molecular constructs:** *Wild type or mutant MFSD2A interacting with various substrate lysolipids (LPC with tail*	Begin in **Year 1** and Finish in **Year 2**

	lengths of 14 carbons or longer, LPE, and LPS) and non-transported compounds (including LPC with tail lengths between 6 and 12 carbons, LPA, and S1P).	
Year 2		
5	Complete and interpret the evaluation of disease mutants in hTMEM16E PLS. Deliver a structure-based framework for the design of modulation and repair strategies of mammalian TMEM 16 PLS. **Deliver: specific guides for initiation of biomimetic molecular system design.** **Molecular constructs:** *Disease mutants S500F, C360Y, G518E, C356G, R215G, and S555I in hTMEM16E.*	Started in **Year 1** Finish in **Year 2**
6	Complete the identification of a minimal set of structural components that are required for Na^+-dependent lipid transport in MFSD2A, and for substrate specificity. **Deliver: specific predictions for initiation of small molecule design of compounds acting as substrates, and as inhibitors, of MFSD2A transporter functions.** **Molecular constructs:** *Wild type or mutant MFSD2A interacting with various single-tailed lipids, including substrates (LPC with tail lengths of 14 carbons or longer, LPE, and LPS) and non-transported compounds (including LPC with tail lengths between 6 and 12 carbons, LPA, and S1P)*	Started in **Year 1** Finish in **Year 2**

LIST OF PUBLICATIONS RESULTING FROM PRIOR ALCC-INCITE AWARDS

Documenting the results from computational work carried out with the allocation BIP109.

RESEARCH ARTICLES (identified as RA and numbered as RA.x)

RA1. Cholesterol occupies the lipid translocation pathway to block phospholipid scrambling by a G protein-coupled receptor**.** Morra G, Razavi G, Menon AK, Khelashvili G. **Structure 2022** S0969-2126(22)00184-8, *online ahead of print.*

RA2. Phosphatidylinositol Phosphates Modulate Interactions between the StarD4 Sterol Trafficking Protein and Lipid Membranes**.** Zhang X, Xie H, Iaea D, Khelashvili G, Weinstein H, Maxfield FR. **Journal of Biological Chemistry 2022,** doi: 10.1016/j.jbc.2022.102058, o*nline ahead of print*.

RA3. Structure and mechanistic basis of MFSD2A-mediated ω-3 fatty acid transport. Cater R, Chua GL, Erramilli S, Keener J, Choy B, Tokarz P, Chin CF, Quek D, Kloss B, Pepe J, Parisi G, Wong B, Clarke O, Marty M, Kossiakoff A, Khelashvili G, Silver D, Mancia F. **Nature 2021** 595:315-319.

RA4. Localization Atomic Force Microscopy. Heath G, Kots E, Robertson J, Lansky S, Khelashvili G, Weinstein H, Scheuring S. **Nature 2021** 594:385–390.

RA5. Unusual mode of dimerization of retinitis pigmentosa-associated F220C rhodopsin. Khelashvili G, Pillai AN, Lee J, Pandey K, Payne AM, Siegel Z, Cuendet MA, Lewis TR, Arshavsky VY, Broichhagen J, Levitz J, Menon AK. **Scientific Reports 2021** 11(1):10536.

RA6. Ligand-Dependent Conformational Transitions in Molecular Dynamics Trajectories of GPCRs Revealed by a New Machine Learning Rare Event Detection Protocol. Plante A, Weinstein H. **Molecules 2021** May 20;26(10):3059.

RA7. Graphic Encoding Method for Quantitative Classification of Protein Structure and Representation of Conformational Changes. Carrillo-Cabada H, Benson J, Razavi AM, Mulligan B, Cuendet MA, Weinstein H, Taufer M, Estrada T. A **IEEE/ACM Trans Comput Biol Bioinform. 2021** 18(4):1336-1349.

RA8. Simulation of pH-Dependent Conformational Transitions in Membrane Proteins: The CLC-ec1 Cl^-/H^+ Antiporter. Kots E, Shore DM, Weinstein H. **Molecules 2021** 26(22):6956.

RA9. Ca^{2+}-dependent mechanism of membrane insertion and destabilization by the SARS-CoV-2 fusion peptide. Khelashvili G, Plante A, Doktorova M, Weinstein H. **Biophysical Journal 2021** 120:1105–19.

RA10. X-ray structure of LeuT in an Inward-Facing Occluded Conformation Reveals Mechanism of Substrate Release. Gotfryd K, Boesen T, Mortensen JS, Khelashvili G, Quick M, Terry DS, Missel JW, LeVine MV, Gourdon P, Blanchard SC, Javitch JA, Weinstein H, Loland CJ, Nissen P, Gether U. **Nature Communications**, **2020** 11(1):1005.

RA11. Membrane lipids are both the substrates and a mechanistically responsive environment of TMEM16 scramblase proteins. Khelashvili G, Cheng X, Falzone ME, Doktorova M, Accardi A, Weinstein H. **Journal of Computational Chemistry 2020** 41(6):538-551.

RA12. Exchange of water for sterol underlies sterol egress from a StARkin domain. Khelashvili G, Chauhan N, Pandey K, Eliezer D, Menon AK**. eLife 2019** 8:e53444.

RA13. The allosteric mechanism of substrate-specific transport in SLC6 is mediated by a volumetric sensor. LeVine MV, Terry DS, Khelashvili G, Siegel ZS, Quick M, Javitch JA, Blanchard SC, Weinstein H. **Proc Natl Academy Sci, US 2019**, 116(32):15947-15956.

RA14. A Machine Learning Approach for the Discovery of Ligand-Specific Functional Mechanisms of GPCRs. Plante A, Shore DM, Morra G, Khelashvili G, Weinstein H. **Molecules 2019** 24(11):2097.

RA15. How structural elements evolving from bacterial to human SLC6 transporters enabled new functional properties. Razavi AM, Khelashvili G, Weinstein H. **BMC Biology 2018** 16(1):31.

RA16. A partially-open inward-facing intermediate conformation of LeuT is associated with Na^+ release and substrate transport. Terry DS, Kolster RA, Quick M, LeVine MV, Khelashvili G, Zhou Z, Weinstein H, Javitch JA, Blanchard SC. **Nature Communications 2018** 9(1):230.

R17. Mechanisms of Lipid Scrambling by the G Protein-Coupled Receptor Opsin. Morra G, Razavi AM, Pandey K, Weinstein H, Menon AK, Khelashvili G. **Structure 2018** 26(2):356-367.e3.

R18. Thermodynamic Coupling Function Analysis of Allosteric Mechanisms in the Human Dopamine Transporter. LeVine MV, Cuendet MA, Razavi AM, Khelashvili G, Weinstein H. **Biophysical Journal 2018** 114(1):10-14.

R19. Single-molecule analysis of ligand efficacy in β_2AR-G-protein activation. Gregorio GG, Masureel M, Hilger D, Terry DS, Juette M, Zhao H, Zhou Z, Perez-Aguilar JM, Hauge M, Mathiasen S, Javitch JA, Weinstein H, Kobilka BK, Blanchard SC. **Nature 2017** 547(7661):68-73.

R20. Evolutionary Divergence of the C-terminal Domain of Complexin Accounts for Functional Disparities between Vertebrate and Invertebrate Complexins. Wragg RT, Parisotto DA, Li Z, Terakawa MS, Snead D, Basu I, Weinstein H, Eliezer D, Dittman JS. **Frontiers of Molecular Neuroscience 2017** 10:146.

R21. A Markov State-based Quantitative Kinetic Model of Sodium Release from the Dopamine Transporter. Razavi AM, Khelashvili G, Weinstein H. **Scientific Reports 2017** 7:40076.

R22. Allosteric Mechanisms of Molecular Machines at the Membrane: Transport by Sodium-Coupled Symporters. LeVine MV, Cuendet MA, Khelashvili G, Weinstein H. **Chemical Reviews 2016** 116(11):6552-87.

HAREL WEINSTEIN, D.Sc.
Weill Medical College of Cornell University, Department of Physiology and Biophysics, and Institute for Computational Biomedicine, 1300 York Ave, New York, NY 10065; (212)746-6358; haw2002@med.cornell.edu

a. Professional Preparation

Technion-Israel Institute of Technology (ITT)	Chemistry	B.S.	1962-1966
Graduate School of the Technion-IIT	Quantum Chemistry	M.Sc.	1966-1968
Graduate School of the Technion-IIT	Theoretical Physical Chemistry	D.Sc.	1968-1971

b. Appointments *(List academic & professional appts in reverse chronological order)*

2002-present	*Maxwell M. Upson Professor*, Department of Physiology and Biophysics, Weill Cornell Medical College of Cornell University, NY City (WCMC)
2002-present	Director, Institute for Computational Biomedicine, WCMC, Cornell University
2002-present	Chair, Graduate Program in Physiology, Biophysics and Systems Biology, Weill Graduate School of Medical Sciences, WCMC, Cornell University
2002-present	Tri-Institutional Professor: Rockefeller University, Sloan Kettering Institute, Cornell University
1999-2002	Director, Institute for Computational Biomedicine, Mount Sinai School of Medicine, NY City (MSSM; now Icahn Sch of Medicine at Mount Sinai)
1998-2002	*Dr. Harold and Golden Lamport Professor* of Physiology and Biophysics, MSSM
1985-2002	Professor and Chairman, Department of Physiology and Biophysics, MSSM
1979-2002	Professor of Pharmacology, MSSM
1976-1979	Associate Professor of Pharmacology, MSSM
1974-1976	Assistant Professor of Pharmacology, MSSM
1973-1974	Research Associate, Department of Chemistry, The Johns Hopkins University
2-4/74, 3-5/75	Visiting Scientist, Genetics, Stanford University School of Medicine
1971-1973	Lecturer, Chemistry, Technion, IIT
1968-1971	Senior Research Assistant, Chemistry, Technion, IIT

c. Publications Relevant to the Project (Recent, from a total of >360 publications)

1. Zhang X, Xie H, Iaea D, Khelashvili G, Weinstein H, Maxfield FR. *Phosphatidylinositol Phosphates Modulate Interactions between the StarD4 Sterol Trafficking Protein and Lipid Membranes.* **J Biol Chem. 2022** May 20: 102058. doi: 10.1016/j.jbc.2022.102058. Epub ahead of print. PMID: 35605664.

2. Plante A, Weinstein H. – *Ligand-Dependent Conformational Transitions in Molecular Dynamics Trajectories of GPCRs Revealed by a New Machine Learning Rare Event Detection Protocol***. Molecules 2021** 26(10):3059. PMCID: PMC8161244

3. Heath GR, Kots E, Robertson JL, Lansky S, Khelashvili G, Weinstein H, Scheuring S. – *Localization atomic force microscopy.* **Nature 2021** 594(7863):385-390. PMCID: PMC8697813

4. Huang Y, Wang X, Lv G, Razavi AM, Huysmans GHM, Weinstein H, et al. – *Use of paramagnetic 19F NMR to monitor domain movement in a glutamate transporter homolog.* **Nature Chem Biol 2020** 16(9):1006 -12. PMCID: PMC7442671

5. Khelashvili G, Falzone ME, Cheng X, Lee BC, Accardi A, Weinstein H. – *Dynamic modulation of the lipid translocation groove generates a conductive ion channel in Ca2+-bound nhTMEM16*. **Nature Commun. 2019** 10(1):4972. PMCID: PMC6823365

d. Research Interests and Expertise

My lab is devoted to studies in molecular and computational biophysics that address complex systems in cell and systems physiology, and to the development and application of mathematical, bioinformatics and artificial intelligence approaches to systems biology. With computational methods of biophysics we discover and quantify the structure-dynamics-function relationships of macromolecular assemblies that determine their role in cellular processes, and the mechanisms of neurotransmitters and drugs of abuse. Mechanisms emerging from the interaction of the cellular components in the membrane and the cytoplasm are followed with large-scale computational simulations and quantified with advanced methods of analysis from statistical mechanics,

information theory, and machine learning. Current themes center on molecular recognition and allosteric mechanisms of G protein-coupled receptors (5-HT and dopamine in particular), neurotransmitter transporters (including DAT, SERT, NET), multi-TM transmembrane proteins (e.g., TMEM16 scramblases), and on the biophysical properties of membranes, and protein-membrane interactions of cellular proteins involved in a variety of physiological and disease processes in drug abuse, addiction, and neurodegeneration. Since 2020 an additional area of research is the study of the molecular mechanisms of the SARS-CoV-2 to be leveraged for the design of antiviral therapy. This work includes the molecular level determination of Spike interactions with membranes through the fusion peptide region, and the mechanism of viral entry and membrane fusion. Combinations of MD simulations and AI analysis and prediction frameworks are used to design and develop specific inhibitors of the cell entry process and proliferation. This broad scholarly perspective on molecular mechanisms in cell physiology and systems biology is exceptionally well suited for mentoring graduate students and junior associates. The mechanistic focus on the etiology and mitigation of disease and biomimetic molecular engineering – enabled by novel methods of computational biology, biophysics, and machine learning – is guiding the development of productive and diverse careers in science for lab members and graduate students.

d. Synergistic Activities

I have mentored to PhD degrees a cohort of ~40 students in my lab, and have mentored and developed the scientific careers of >60 postdoctoral and junior faculty associates. As the Chair of the graduate program in Physiology, Biophysics and Systems Biology (PBSB) of the Weill Cornell Graduate School of Medical Sciences that combines the Cornell and Memorial Sloan Kettering divisions, I am responsible for the educational activities of 67 faculty members of the Program, and ~80 graduate students in this program and others. As a Tri-Institutional Professor (Weill Cornell, Rockefeller University, Sloan-Kettering Research Institute) I have founded the Tri-I Program in Computational Biology and Medicine (TPCBM) and continue to be a member of its Steering and Curriculum committees; the program includes a twice renewed T32 grant from the NIH and currently educates >45 students. Both the PBSB and TPCBM graduate programs are known for extensive outreach activities in NYC and beyond, active recruitment and inclusion of underrepresented minorities, and an outstanding graduation record. I participated in the design and scientific board leadership of the NY Structural Biology Center, a shared instrumentation facility supported by the NIH and NSF that brings together 11 public and private research institutions in the tri-State area. Other activities, past and present, devoted to the creation, integration and transfer of knowledge include: Member and Chair of review panels for **NIH, NSF** and the Canadian Research Council, The ***Wellcome*** Trust, **HHMI**, **HFS**.; **Editorial Boards and Reviewer** for: JACS, J Phys Chem, Biophys J., Nature (and Nature pub. Journals), Science, Cell, J Biol Chem, PNAS, etc.; Member (2007-2012) and Chair **(**2010-2012) **of BPNS NIH Study Section**; (2007-2010) Member, (2007-2008) **President** of The Biophysical Society; 2007-2012 Member, NIH Director's New Innovator Award Program Review Panel; **(2013-present)** Member, **NIH Director's** Review Committees for Pioneer, Transformative R01, DP2, etc.; NIH Pioneer Award (DP1) Stage 2 Panelist **(2018);** Member, **Scientific Review Committee,** Dulbecco Telethon Institute, Italy; **Boards of Directors/Scientific Advisors for**: Biophysical Society, International Society for Computational Biology, FASEB, Burke Neurological Institute; **Chair,** FASEB Data Science and Bioinformatics Subcommittee of the Science Policy Committee, etc.

e. Collaborators

In addition to colleagues in the Tri-Institutional environment, **Collaborators** include the following: **Bahar, I:** Department of Computational and Systems Biology, U Pittsburgh Med School; **Blanchard, SC**: Structural Biology, St Jude Children's Research Hospital: **Estrada, T:** Department of Computer Science, U New Mexico; **Galli, A:** Department of Surgery, U Alabama Birmingham; **Gether, U:** Department of Neuroscience, U Copenhagen, Denmark; **Javitch, JA**: Departments of Psychiatry, Columbia P&S; **Taufer, M**: Department of Electrical Engineering and Computer Science, U Tennessee.

Curriculum Vitae

GEORGE KHELASHVILI (co-PI)
Department of Physiology and Biophysics
Weill College Medical College of Cornell University (WCMC)
1300 York Avenue, Room LC501C
New York, NY, 10065, USA
Tel: 212-746-6348
Fax: 212-746-6226
Email: gek2009@med.cornell.edu

Professional Preparation

Tbilisi State University, Tbilisi, Georgia	B.Sc.	Physics	1997
Tbilisi State University, Tbilisi, Georgia	M.Sc.	Physics	1999
Illinois Institute of Technology, Chicago, IL	Ph.D	Biophysics	2005
WCMC, New York, NY	Postdoctoral Fellow	Biophysics	2005-2008

Appointments

2019-current	**Assistant Professor**, Department of Physiology and Biophysics, WCMC
2015-2019	**Assistant Professor of Research**, Department of Physiology and Biophysics, WCMC
2011-2014	**Instructor**, Department of Physiology and Biophysics, WCMC
2009-2010	**Research Associate**, Department of Physiology and Biophysics, WCMC

Five Publications Most Relevant to This Proposal

1. Khelashvili G, Menon AK – Lipid flipping mechanisms in GPCRs. **Annual Reviews in Biophysics 2021**, doi: 10.1146/annurev-biophys-090821-083030.
2. Cater R, Chua G-L, Erramilli S, Keener J, Choy B, Tokarz P, Chin C, Quek D, Kloss B, Pepe J, Parisi G, Wong B, Clarke O, Marty M, Kossiakoff A, Khelashvili G, Silver D, Mancia F – Structure and mechanistic basis of MFSD2A-mediated ω-3 fatty acid transport. **Nature 2021**, 595(7866):315-319.
3. Khelashvili G, Cheng X, Falzone M, Doktorova M, Accardi A, Weinstein H – Membrane Lipids Are Both the Substrates and a Mechanistically Responsive Environment of TMEM16 Scramblase Proteins. **Journal of Computational Chemistry 2020**, 41, 538–551
4. Khelashvili G, Falzone M, Cheng X, Lee BC, Accardi A, Weinstein H – Dynamic modulation of the lipid translocation groove generates a conductive ion channel in Ca2+-bound nhTMEM16. **Nature Communications 2019**, 10(1):4972
5. Lee BC, Khelashvili G, Falzone M, Menon AK, Weinstein H, Accardi A – Gating mechanism of the extracellular entry to the lipid pathway in a TMEM16 scramblase. **Nature Communications 2018**, 9:3251.

Research Interests and Expertise

My research aims at achieving a quantitative determination of factors and mechanisms responsible for the localization of integral signaling proteins in membrane domains and for the regulation, by membrane remodeling, of functionally relevant interactions between signaling proteins. Special attention is devoted to the mechanistic involvement in protein function of membrane components such as highly charged inositol lipids and cholesterol, as well as phosphorylation and protein scaffolding. This research is conducted in close synergy with experimental colleagues through a combination of various tools of computational biophysics, developed and utilized at the highest level of each specialty. Rigorous, quantitative computational tools I develop for the study of protein structure and function in specific cellular environment have powerful predicting capabilities and are based on multidisciplinary fields of physics and mathematics. They address a wide range of interactions at lipid interfaces. Three complementary accomplishments relevant to this project demonstrate novel insights into the roles of membranes in the function and organization of signaling proteins: 1) A theoretical framework combining continuum theory of membrane deformations around trans-membrane signaling proteins with atomistic molecular dynamics representations of protein/lipid interfacial interactions (CTMD) developed to quantify the mechanistic

involvement of membrane remodeling in the organization and function of signaling proteins; 2) A mesoscale mean-field formulation describing remodeling of lipid bilayers under the influence of peripheral membrane proteins; and 3) A thermodynamic coarse-grained model of cholesterol-enriched lipid rafts. These quantitative approaches are combined with theoretical formulations, based on Markov State Models, to evaluate kinetic properties of various physiological processes. With these integrative computational approaches, I address fundamental aspects of protein function modulated by formation and maintenance of structural membrane elements considered to play a major role in normal physiology and in a variety of diseases. The mechanistic insights I develop relate to longstanding challenges in the field of computational biophysics and have enabled major, technological advancements. My publications show that these strategies have demonstrated the power of providing systematic predictions that guide experimental biomedical research towards understanding fundamental mechanisms of cell biology and are expected to have a continuing significant impact on biomedical science.

Synergistic Activities

1. Quantitative computational tools I develop for the study of protein structure and function in specific cellular environment have powerful predicting capabilities and are based on multidisciplinary fields of physics and mathematics. They treat range of interactions at lipid interfaces and address fundamental aspects of protein function modulated by formation and maintenance of structural membrane elements considered to play a major role in normal physiology and in a variety of diseases. The mechanistic insights I develop relate to longstanding challenges in the field of computational biophysics and have enabled major, technological advancements.
2. Participated in development and curation of G-protein coupled receptor (GPCR) oligomer database (GPCR-OKB, http://filizolalab01.mssm.edu:8080/gpcr-okb/).
3. Member of the Biophysical Society (2001-present).
4. Active reviewer for scientific journals in the fields of biophysics, biochemistry and physical chemistry. These include Nature journals (Nature Communications, Nature Scientific Reports), Biophysical Journal, Journal of Chemical Physics, ACS journals (e.g. JACS, Journal of Chemical Theory and Computation, Biochemistry, Journal of Physical Chemistry), BBA, Journal of Biological Chemistry.
5. Guest Editor for Frontiers journal.

Collaborators (*past 5 years including name and current institution*)

Collaborators – Accardi A (WCMC), Andersen RC (UC), Akyuz N (WCMC), Albornoz PB (MSKCC), Altman RB (WCMC), Ammendrup-Johnsen I (U Copenhagen, UC), Andersen OS (WCMC), Arleth L (UC), Belovich AN (Vanderbilt, VAN), Bhatia V (UC), Blanchard SC (WCMC), Boudker O (WCMC), Caffrey M (Trinity College), Christensen NM (UC), Cuendet MA (WCMC), Cwiklik L (Academy of Science of Czech Rep, ASCR), Doktorova M (WCMC), Erlendsson S (UC), Erreger K (VAN), de Fabritiis G (Acellera), Falzone M (WCMC), Filizola M (Mount Sinai School of Medicine), Freed JH (WCMC), Galli A (VAN), Gether U (UC), Georgieva ER (WCMC), Gotfryd K (UC), Hamilton PJ (VAN), Harries D (Hebrew U Jerusalem), Heftberger P (U Graz), Herlo R (UC) Høiberg-Nielsen R (UC), Hof M (ASCR), Jansen AM (UC), Javitch JA (U Columbia), Johner N (U Bazel), Johnston J (Merck), Jungwirth P (Tampere UT, TUT), Jurkiewicz P (ASCR), Karlsen ML (UC), Kantola AM (U Oulu), Kollmitzer B (U Graz), Komulainen S (U Oulu), Kulig W (TUT), Lee BC (WCMC), LeVine MV (WCMC), Loland CJ (UC), Lund VK (UC), Lycas MD (UC), Madsen KL (UC), Manna M (TUT), Matthies HJ (VAN), Medina J (WCMC), Mehler EL (WCMC), Menon AK (WCMC), Mondal S (Schrodinger), Morra G (Consiglio Nazionale delle Ricerche), Noskov S (U Calgary), Olżyńska A (ASCR), Pabst G (U Graz), Perez-Aguilar JM (IBM), Plante A (WCMC), Pourmosa M (TUT), Quick M (U Columbia), Rappolt M (U Leeds), Rog T (TUT), Sahai MA (Roehampton), Saunders C (VAN), Schmidt SG (UC), Scott HL (IIT), Shan J (Schrodinger), Shi L(NIH), Shore D (WCMC), Stamou D (UC), Simonsen JB (UC), Sitte HH (Med U Viena), Stanley N (GRIB-IMIM), Stolzenberg S (Freie University), Streicher W (UC), Terry DS (WCMC), Thorsen TS (UC), Teilum K (UC), Telkki VV (U Oulu), Tian X (UC), Vazdar M (TUT), Vattulainen I (TUT), Vestergaard (UC) B, Wang H (Boyalife Wuxi), Weinstein H (WCMC), Zhao C (U Calgary), Zhou Z (WCMC), Zhao G (Shanghai Jiao Tong U).

Coeditors – Sahai MA (Roehampton), Moreira I (Coimbra); ***Graduate Advisors*** – Scott H. L. (*Illinois Institute of Technology*); ***Postdoctoral Sponsors*** – Weinstein H. (*WCMC*); ***Graduate students*** – Omar Alvarenga (*WCMC*), Margarida Rosa (*WCMC*).

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

OpenMM 7.7.0

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.

1. Marta Filizola, Departments of Pharmacological Sciences and Neuroscience, Icahn School of Medicine at Mount Sinai; e-mail: marta.filizola@mssm.edu

2. Alan Grossfield, Department of Biochemistry and Biophysics, University of Rochester Medical Center; e-mail: Alan_Grossfield@URMC.Rochester.edu

3. Davide Provasi, Department of Structural and Chemical Biology, Mount Sinai School of Medicine; e-mail:davide.provasi@gmail.commailto:davide.provasi@gmail.com

4. Lei Shi, Molecular Targets and Medications Discovery Branch, Computational Chemistry and Molecular Biophysics Unit, NIDA, Triad Technology Center; e-mail: lei.shi2@nih.gov

5. Emad Tajkhorshid, J. Woodland Hastings Endowed Chair in Biochemistry, Professor of Chemistry, Bioengineering, and Biophysics and Quantitative Biology, Director of NIH Biotechnology Center for Macromolecular Modeling and Bioinformatics, e-mail: emad@illinois.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.

additional_file.pdf
The attachment is on the following page.

N/A