Pacific Northwest
NATIONAL LABORATORY

PNNL-36305

Enhancing Biopreparedness through a Model System to Understand the Molecular Mechanisms that Lead to Pathogenesis and Disease Transmission

NW-BRaVE

August 2024

PI: Margaret S. Cheung

U.S. DEPARTMENT OF ENERGY

Prepared for the U.S. Department of Energy
under Contract DE-AC05-76RL01830

DISCLAIMER

PACIFIC NORTHWEST NATIONAL LABORATORY
operated by
BATTELLE
for the
UNITED STATES DEPARTMENT OF ENERGY
under Contract DE-AC05-76RL01830

Printed in the United States of America

Available to DOE and DOE contractors from
the Office of Scientific and Technical Information,
P.O. Box 62, Oak Ridge, TN 37831-0062
www.osti.gov
ph: (865) 576-8401
fox: (865) 576-5728
email: reports@osti.gov

Available to the public from the National Technical Information Service
5301 Shawnee Rd., Alexandria, VA 22312
ph: (800) 553-NTIS (6847)
or (703) 605-6000
email: info@ntis.gov
Online ordering: http://www.ntis.gov

PNNL-36305

Enhancing Biopreparedness through a Model System to Understand the Molecular Mechanisms that Lead to Pathogenesis and Disease Transmission

NW-BRaVE

August 2024

PI: Margaret S. Cheung

Prepared for
the U.S. Department of Energy
under Contract DE-AC05-76RL01830

Pacific Northwest National Laboratory
Richland, Washington 99354

Acknowledgments

This work was supported by the U. S. Department of Energy, Office of Science, Office of Biological and Environmental Research, under **FWP 81832**. PNNL is a multi-program national laboratory operated by Battelle for the DOE under Contract DE-AC05-76RLO 1830

Contents

Figures

Tables

1.0 Introduction

The science of biopreparedness to counter biological threats hinges on understanding the fundamental principles and molecular mechanisms that lead to pathogenesis and disease transmission. Our ***vision*** to address this challenge is to create a powerful and user-friendly platform to elucidate the fundamental principles of how molecular interactions drive pathogen-host relationships and host shifts. We will enable groundbreaking discoveries by integrating a wide range of structural, genomics, proteomics, and other advanced omics measurements, along with evolutionary and artificial intelligence predictions. To ensure the system is applicable to real-world problems, we will develop it in the context of a tractable model system, the small, abundant, and accessible photosynthetic cyanobacteria and their constantly co-adapting viral pathogens, cyanophages. This will maintain the system's applicability to real-world problems and techniques, while focusing on elucidating system agnostic, general principles of detecting, assessing, and surveilling molecular interaction, adaptation, and coevolution that are extensible to other viral-host interactions.

Our overall ***objectives*** are to: (1) identify the molecular complexes that comprise the cyanobacteria redox macromolecular subsystem and how they dynamically change with bacteriophage infection *in situ*, using cryo-electron tomography; (2) profile regulatory changes during infection using proteomics, multiomics, and experimental validation, and integrate the data with *in situ* structures; (3) use genomics and metagenomics to determine environmental and population factors across time scales that impact the interactions between marine cyanobacteria and their cyanophage parasites, predicting the evolutionary origins of *in situ* structural and functional interactions, convergence and coevolution; and (4) develop a data integration and transformation platform that facilitates the integration of *in situ*, proteomic, and evolutionary measurements of molecular interactions to surveil diverse hosts and parasites in various environmental contexts. These objectives address **Focus Area 2** of the BRaVE proposal call: Reveal Molecular Interactions Across Biological Scales for Design of Targeted Interventions.

Our powerful, user-friendly platform will enhance connections between the structure, molecular phenotype and evolutionary genomics fields; all key to biopreparedness, they are often siloed and require integration (Figure 1). We will build a navigation tool that intakes globally distributed experimental data for integrated analysis and predictive modeling. We will develop, implement, and test a platform to assess host-pathogen molecular interactions, adaptation to hosts and host shifts, and coevolution between hosts and pathogens. This will revolutionize abilities to study host-pathogen interaction, encourage diverse community contributions, and reveal fundamental insights into how proteins adapt to new contexts. This will be critical for designing early interventions against future threats. Our surveillance training will deliver fair and equitable response to future pandemics and biothreats.

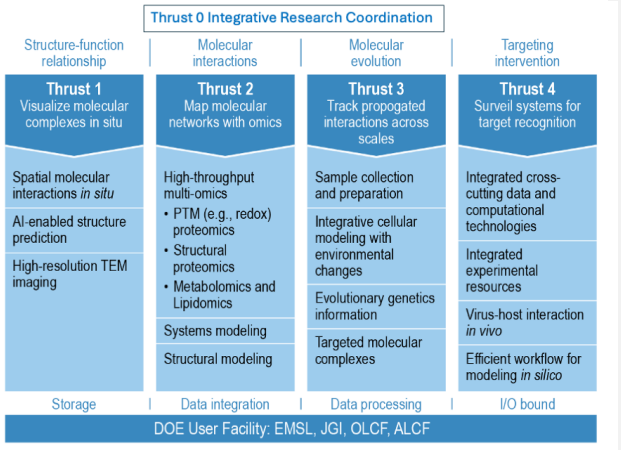

Figure 1. Structure of project thrusts.

2.0 Review of Scientific Progress toward Achieving Project Objectives

We use a thrust-based framework to build the biopreparedness molecular interaction platform. In addition to the four original thrusts, we set up **Thrust 0 (Integrative Research Coordination)** to create and coordinate a project-wide common sample culturing pipeline. A brief overview of thrust objectives and tasks are described below along with progress highlights.

2.1 Brief Review of Scientific Progress within Tasks toward the Project's Objectives/Milestones

Thrust 0 (T0) – To Coordinate, Overall Objective.

T0 enacts and harmonizes sample preparation, tracking, and provenance data management protocols across thrusts.

The main T0 tasks establish, execute, and synchronize consistent host-virus culture methods across thrusts 1-3 and works with **T4** to ensure related metadata is recorded in a consistent, reproducible, and unambiguous machine- and human-interpretable fashion. Tasks include to: 1) cultivate sufficient cells under various controlled light and infection conditions for **T2** experiments; 2) prepare infective phages and measuring associated infective titers, with sufficient quantity and proportion of intact phages for **T1** visualization; and with Salish estuary water, 3) implement growth of our main laboratory model, *Prochlorococcus marinus* strain MED4, test phage growth on MED4, and grow and isolate Salish estuarine cyanophage.

Thrust 1 (T1) – To Visualize, Overall Objective

T1 visualizes *in situ* spatial molecular interactions induced by virus infection using cryo-electron tomography (cryo-ET). **T1**'s four key tasks will unlock the intricate molecular interactions between *Prochlorococcus* and phages and are designed to advance scientific understanding and foster data transparency and accessibility, paving the way for innovative cryo-ET methods. The objective is to build a 3-D visualization and analysis platform to track morphological and compositional dynamics in the host-virus lifecycle, enabling efficient particle identification and cellular segmentation, to inform whole cell modeling and act as the roadmap for overlaying omics and other contextual data within a spatial framework.

The main tasks under **T1** develop a method to classify proteins in tomograms and probabilistically identify them by linking to genomic, proteomic, metabolomic and prediction information, to help in understanding the spatial distribution of higher-order protein assemblies and their role in host-phage interactions. Task 1.1 focuses on automated collection and reconstruction of nanoscale resolution tomograms of *MED4* with and without phage interactions. Task 1.2 focuses on mapping and classifying proteins within the reconstructed tomograms using current particle-picking and sub-volume averaging techniques. Task 1.3 focuses on developing GRIP-Tomo 2.0 to identify and map proteins more efficiently and accurately. Task 1.4 focuses on creating a publicly accessible open database of the project's MED4/phage datasets. Together, **T1** encapsulate a holistic and forward-thinking approach to advance the visualization, analysis, and understanding of complex biological interactions at the molecular level.

Thrust 2 (T2) – To Map, Overall Objective

The overall **T2** objective is to use advanced multi-omics profiling to map regulatory pathways and molecular networks to provide detailed interactome information to augment the structural base established by **T1** and experimentally complement the evolutionary genomic information from **T3**. Under **T2**, we will establish a computational pipeline to integrate multi-omics experiment data, systems modeling, and simulations with **T3's** molecular interaction and molecular evolutionary analysis, to map the molecular interactions underlying microbial phenotypes affected by viral pathogens. By incorporating the University of Illinois Urbana-Champaign's (UIUC's) supra-grained whole-cell modeling and cryo-ET data visualized from **T1**, we will expand systems modeling capabilities to spatially map interactomes with regulatory post-translational modification (PTM) specificity inside a microbe. Major molecular components and their response to controlled environmental change will be mapped by quantitative proteomics, metabolomics, lipidomics, high-throughput PTM profiling, and structural proteomics. These multi-omics profiles will be comprehensively characterized over the response of the host cell to infection and cellular oxidative stress across time and phage strains.

The main T2 tasks include to: 1) optimize multi-omics tools to ensure data generation requirements are feasible; 2) generate large-scale multi-omics, PTMs, and structural proteomics data from temporal cyano-phage infection experiments for inferring molecular interactomes and their responses to phage infections; 3) create a data-driven dynamic system model for redox regulatory pathways responsive to viral infection and replication; 4) map PTMs in the redox regulatory pathways that drive molecular evolution; 5) create a 4D whole-cell model (4DWCM) for MED4. These tasks are tightly connected to our scientific questions: How do phage infections affect host cell dynamics and function? How does redox regulation play roles in host response and phage replication? And how do interactions and regulations between phage and host cell evolve under different environmental conditions?

Thrust 3 (T3) – To Track, Overall Objective.

T3 focuses on tracking the divergence of functional interactions across scales. **T3**'s three key tasks are to sample, sequence, and analyze cyanobacteria from the coastal and estuarine environments of the Salish estuary environment along with records of environmental conditions based on ocean metadata. We are tracking genomic information, coevolution, and adaptation in wild local cyanobacteria and cyanophage by phylogenetic and molecular evolutionary analysis along with genome and metagenome datasets from JGI and NMDC. The objective is to capture the origins, conservation, and evolutionary shifts in targeted molecular interactions driven by pathogens and environments and relate them to molecular interaction visualization from **T1** and molecular network effects from **T2** of environmental conditions in lab-grown cyanobacteria.

T3's main tasks include to: 1) collect representative in situ cyanobacterial and cyanophage samples and environmental metadata for swift genomics characterization; 2) determine how cyanobacteria and bacteriophage evolution and coevolution is driven by and informs on the redox interactome, including photosynthesis; 3) integrate evolutionary information and local community with platform data annotation, biopreparedness analytics, outreach, and training.

Thrust 4 (T4) – To Surveil, Overall Objective

T4 focuses on surveilling host-pathogen interaction through multimodal data integration. **T4**'s three tasks aim to build an agnostic data integration platform for transforming data across its life cycle, from all thrusts. We are generating tools analyzing targeted molecular interaction using coarse-grained molecular simulations and AI/ML-driven molecular dynamics simulation. This platform will include the capability to interpret data from *in vivo* imaging of phage infection in lab-grown or synthesized cyanobacteria models at UIUC with structural information.

T4 main tasks include to: 1) create a platform for data integration and transformation; 2) take in molecular simulations for modeling, analysis, and interpretation; 3) provide structural insights into *in vivo* imaging analysis and synthetic biology solutions for experimental validation.

2.2 Short Vignettes of Science Highlights

Thrust 0 (T0) – To Coordinate

Over Year 1, **T0** made substantial progress working closely with thrusts 1,2, and 3 to establish reproducible and controllable host-virus experimental methodologies for performing biological experiments to address scientific questions under **T1** and **T2**. First, **T0** shipped samples of Salish estuary water from Sequim to Richland where PNNL staff established conditions suitable for the growth and maintenance of MED4 in Salish and commercial

Figure 2. Systems for production of host cells at different scales.

seawater media. **T0** deployed a cultivation system to generate enough host cells for phage reproduction (Figure 2) and for -omics experiments. Then, **T0** developed methodologies to harvest, purify, concentrate, and measure virus particle activity. Next, **T0** deployed the most probable number (MPN) counts assay to measure the infective titer (number of infective elements per mL) (Figure 3). Lastly, **T0** tested the ability of local cyanophages present in Salish water to grow on MED4, in preparation for isolation.

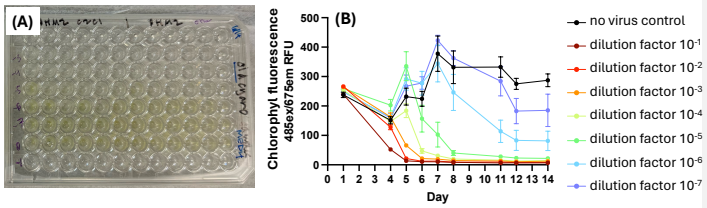

Figure 3. Most probable number (MPN) counts of virus using 96 well plate (A) and lysis curves (B).

Thrust 1 (T1) – To Visualize

During the first year of NW-BRaVE, **T1** has made *substantial progress* towards overall project goals. **T1** worked closely with **T0** to establish a reproducible and scalable cell and phage culturing pipeline that can generate all necessary samples of to visualize MED4 cells alone or with purified samples of P-SSP7 and P-HM2 phages (Figure 4). Under Task 1.1 we created and benchmarked an end-to-end automated workflow from tilt series data collection through 3D reconstruction of tomograms. We began experiments to visualize host-phage interactions (Figure 5), and identified optimal imaging parameters for these samples and validated the data analysis and output of segmented models in formats compatible with other thrusts (e.g., output matches input requirements for Lattice Microbe software from **T2**). The optimization of parameters such as tolerable electron fluence, ideal pixel size, and best reconstruction workflows and algorithms act as a foundation for planned Year 2 experiments, which will track the host-phage lifecycle with high resolution, high fidelity, and improved efficiency. Task 1.2 depends on Task 1.1 because the output tomograms are analyzed in Task 1.2 to map and

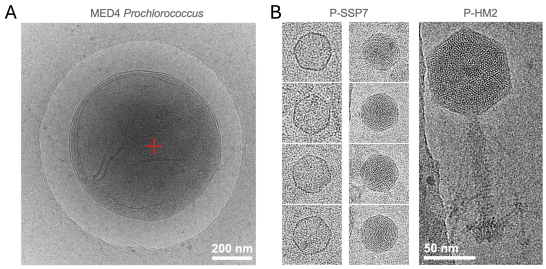

Figure 4. All samples have been successfully cultured. Example cryo-EM images of *Med4* cells (A) and the phages (B) that are the primary samples for the PNNL BRAVE project. Scalebar in (B) is the same for all phages displayed.

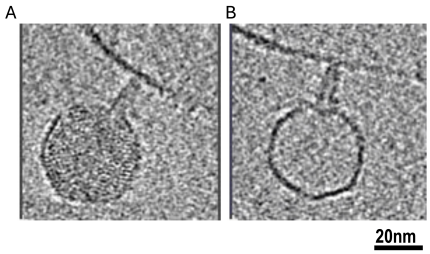

Figure 5. Cryo-electron tomography optimal imaging conditions identified. Example tomograms showing P-SSP7 infectious (DNA filled) phage (A) associated with Med4 cell membrane and a second tomogram of an empty P-SSP7 phage after injection of DNA into MED4.

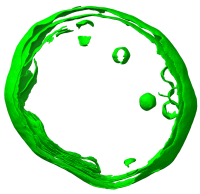

Figure 6. Optimized segmentation workflow for Med4 data identified. Example segmentation of all Med4 membranes using a combination of MemBrain-seg and ColabSeg software compatible with our automated processing pipeline. This figure was prepared with the assistance

classify proteins using current segmentation, particle-picking and sub-volume averaging techniques. We tested existing software programs for segmentation and particle picking, benchmarking them directly against MED4 tomograms generated in-house as part of Task 1.1 since no single software program provides optimal analysis for every cell type. In Year 1, we subjected the same tomogram to multiple segmentation workflows to identify the best performing software that allows reproducible and accurate segmentation of the cellular membranes while also being compatible with scripting and automation workflows. The goal was to identify at least one workflow to directly incorporate into the automated reconstruction workflows established in Task 1.1 to allow full segmentation of the membranes found in every tomogram without significant user interaction. We successfully identified that the combination of MemBrain-seg [1] and ColabSeg [2] software meets our needs and is well suited towards analyzing and detecting contrast variations in the MED4 tomograms to denote the various cellular membranes (Figure 6). We are now evaluating particle picking and subtomogram averaging software to identify the best automatable product to locate, identify and categorize numerous protein complexes with high precision. This task is pivotal to understand the intricate interactions between phage components and host cellular machinery and will be applied to the Year-2 host-phage lifecycle datasets.

While Task 1.2 leverages existing software to identify the location of proteins and complexes in tomograms, Task 1.3 advances the development of GRIP-Tomo 2.0, an enhanced version of the Graph-Identification of Proteins in Tomograms (GRIP-Tomo) software. Demonstrations of GRIP-Tomo 1.0[3] showed noise-free identification of proteins from synthetic mock tomograms. In year 1, we dramatically expanded its capability to include advanced machine learning algorithms and improved graph-based methods to identify and map proteins more efficiently and accurately from real-world experimental tomography dataset. The algorithm correctly identifies large proteins from noisy experimental single particle tomograms of mixed protein populations with high accuracy (>95%, Figure 7**, on top of the next page**) even without computationally correcting the missing wedge artifact beforehand. GRIP-Tomo 2.0 delivers faster processing times and greater scalability over GRIP-Tomo 1.0[3]; a manuscript is under development to share these results to the broader community. In Year 2 we will continue adapting and optimizing GRIP-Tomo for crowded whole cell tomograms and apply (most probable number, MPN) counts assay to measure the infective titer (number of infective elements per mL) (Figure 6). all collected MED4 datasets from Task 1.1 to accelerate protein mapping and annotation.

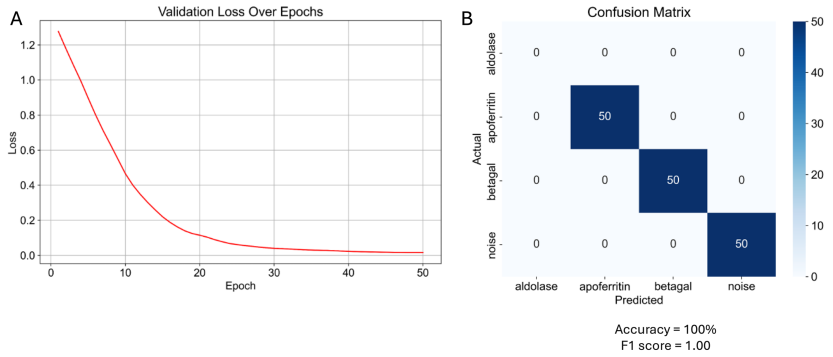

Figure 7. **Improved identification of proteins with GRIP-Tomo 2.0**. A) Validation loss of GRIP-Tomo epochs for real experimental data from a single particle tomogram containing embedded apoferritin and beta-galactosidase (betagal) protein complexes. B) Confusion matrix showing 100% accuracy in proper identification of noise, apoferritin and betagal subvolumes when trained against 4 inputs. Expanding the training with an additional 10 unique known protein structure inputs only decreases accuracy to 96%. This plot was produced by graduate student Chengxuan Li.

Thrust 2 (T2) – To Map

T2 made *substantial progress* by working closely with **T0** to assess the experimental system and optimize experimental workflows to ensure the overall workflow is ready to analyze time course samples of the infection experiments. In addition, the growth dynamics and metabolic flux profiles from modeling confirmed the feasibility of the computational approach to test the metabolic changes in the host response to viral infection in year 1. MED4's very low biomass makes it difficult to scale up. We obtained 28.7 mg and 47.8 mg of biomass (wet cell weight) respectively from100 mL and 200 mL axenic MED4 cultures under constant light. To determine required cell amounts, we subjected samples of various biomass inputs to MPlex extraction[4] for metabolomics, proteomics, and lipidomics analysis. We found that 50 mL cultures are the minimum required volume for multi-omics measurements.

We then evaluated the feasibility of an integrated multi-PTM proteomics workflow, focusing on the coverage of global protein abundance and cysteine (Cys) redox post-translational modification. Samples were collected from 50 mL axenic MED4 culture exposed to 2-hr continued light or dark conditions. The oxidation level of 2466 unique Cys sites (distributed among 965 unique proteins) were identified and quantified from in total 1273 unique proteins (Figure 8). Compared to MED4 exposed to continuous light, under dark conditions the median total oxidation stoichiometry (i.e. the ratio of the oxidized cysteines vs. total cysteines for individual proteins) only increased ~1%; however a notable number of Cys sites show significant perturbations. This implies that the overall cysteine oxidation levels might not reflect the regulations where redox modifications at specific Cys sites play important roles. In naturally

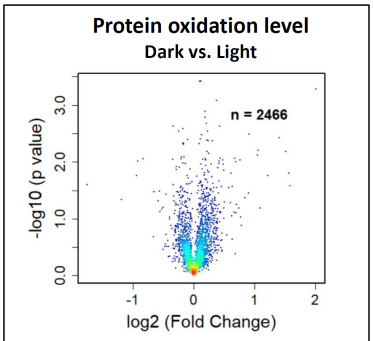

Figure 8. Pilot demonstration of multi-PTM workflow to profile protein oxidation in axenic MED4 culture exposed to 2-hr light vs. dark conditions.

occurring organisms, timing in the circadian cycle plays a large role in gene expression and regulation, and interacts with viral infection. We therefore initiated multi-omics measurements of axenic MED4 under a 12-on 12-off 24-hr circadian cycle to obtain a baseline for future viral infection experiments and to build a 4D whole-cell model of *MED4* without infection. Cultures were collected at subjective 6AM (dawn), 8AM, 10AM, 12PM, 2PM, 4PM, 6PM, 8PM, 10PM, 12AM, 2AM, 4AM. All samples are currently being analyzed.

For metabolomics assessment, we used gas chromatography-mass spectrometry (GC-MS) to identify 114 intracellular metabolites. Using liquid chromatography (LC)-MS, 112 additional metabolites were identified from the same sample-set. Coincidentally, 114 species of lipid compounds were also identified from the same sample. These results support the workflow's overall readiness for analyzing time course samples from the infection experiments.

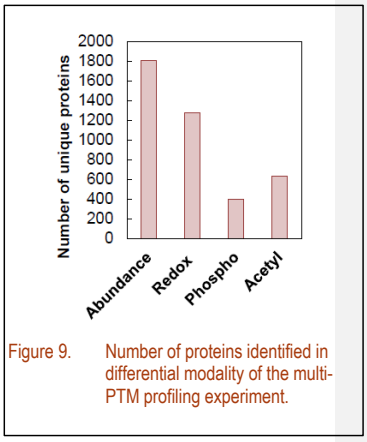

Figure 9. Number of proteins identified in differential modality of the multi-PTM profiling experiment.

We began implementing a new multi-PTM workflow to measure phosphorylation, thiol oxidation (redox), and acetylation in the MED4-phage infection system, so that we can simultaneously analyze all three PTMs along with protein abundance from the same samples, resulting in multi-modality omics data to unravel PTM-driven cellular processes. Figure 9 shows the unique proteins identified from four modalities of the multi-PTM experiment (i.e., global abundance, redox, phosphorylation, and acetylation measurements, respectively), under the two conditions. We note that both phosphorylation and acetylation events are rarely evaluated in cyanobacteria, based on published results by other labs.

We also began establishing the data transformation workflow using the published minimal cell Lattice Microbe model [5, 6]. This workflow is going to incorporate our multi-omics data into a whole-cell 4D kinetic model of MED4, with help from our UIUC collaborator. Towards such a whole-cell model of MED4, we started with a published genome-scale metabolic model (GEM) of MED4 [7] , combined with our pilot study results and our in-house annotation pipeline, to improve and gap-fill this model. We began extracting the essential metabolism for MED4 growth and transforming the genome-scale metabolic model into a kinetic model, where the kinetic rate parameters are derived from public databases, such as BRENDA [8] enzyme database and MetaCyc [9] database, and machine learning algorithms (e.g., eQuilibrator [10], DLKcat [11]).

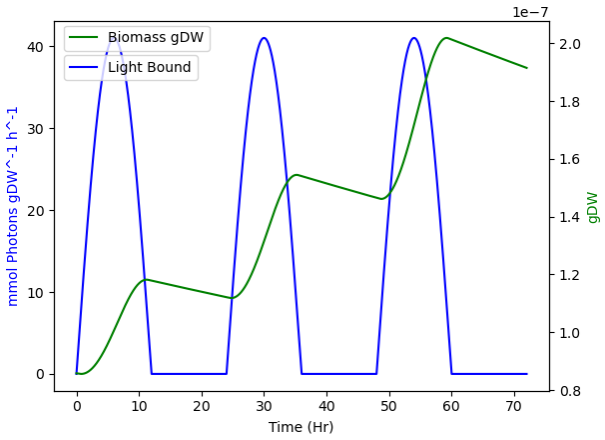

Figure 10. Growth dynamics of MED4 simulated with dynamic flux balance analysis (dFBA). Blue (green) line represents the light intensity (biomass).

Using the MED4 GEM, we successfully reproduced MED4 growth dynamics using dynamic flux balance analysis (dFBA) [12] with the COBRApy toolset [13]. Figure 10 shows growth dynamics (green) across three diel cycles, with the light-dark switch defined by a positive half sine wave function (blue line). We performed perturbation analysis to simulate the impact of auxiliary metabolic gene (AMG) [14, 15] expression on host cell metabolism under infection by two different phage strains, P-HM2 and P-SSP7 [16].

For example, Figure 11 shows the theoretical impact that expressing active CP12 [17] proteins (red solid lines), versus "Mock" without CP12 (blue dashed lines), has on growth (Biomass gDW) and metabolic fluxes of photosynthesis II reaction (PSII), ribulose-1,5-biphosphate carboxylase reaction (RuBisCO), and phosphoribulokinase reaction (PRK). The CP12 protein affects the Calvin-Benson cycle and carbohydrate metabolism by sequestering the PRK and GAPDH enzymes. However, CP12 activity is determined by the formation of disulfide bonds that are redox regulated. The photosynthesis driven redox shifting may impact how this AMG and other AMGs redirect host metabolism. Moreover, in ongoing work, we will model redox changes coupled with light-dark changes and will simulate all AMGs to understand how different phage AMGs systematically redirect metabolic fluxes during phage replication.

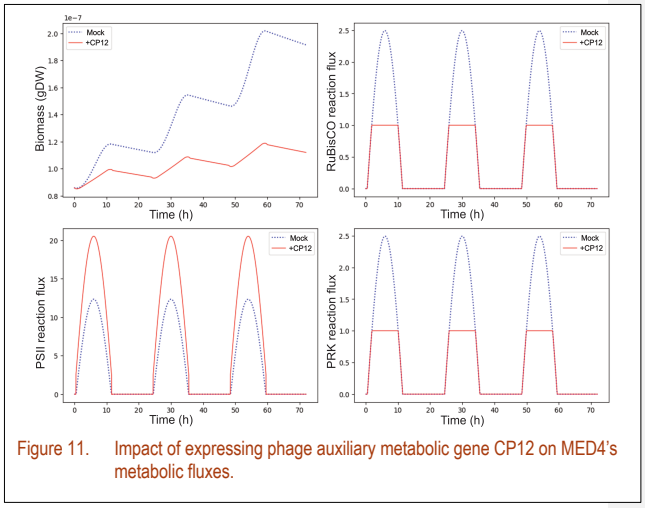

Figure 11. Impact of expressing phage auxiliary metabolic gene CP12 on MED4's metabolic fluxes.

Thrust 3 (T3) – To Track

T3 has made *substantial progress* by obtaining approvals for and implementing environmental sampling, laboratory work, and shipment of estuarine water and extracted DNA. We also developed a pipeline for sequencing estuarine DNA and performing preliminary analysis of results to inform sampling protocol development and construct relevant datasets for more detailed evolutionary analysis in year 2. In addition, we have begun analyzing existing cyanobacteria complete genomes and metagenomic data, working with CU Anschutz collaborators to automatically generate and make accessible as part of DataFed (**T4**): all cyanobacteria gene structure, domain alignments, phylogenetic trees, AlphaFold structure predictions, and for identifiably homologous domains, experimental structures and complex visualization (where available). The structure integration is being done in collaboration with **T1** and **T4**.

T3 PNNL-Sequim staff developed sampling, filtration, and DNA extraction protocols for Salish estuarine water taken from the Sequim laboratory dock, assisted by undergraduate SULI intern, Mahala Peter-Frank. As expected, low numbers of cyanobacteria were present in the Bay's surface waters during the cooler and darker fall and winter months when sampling began, but numbers dramatically improved moving into spring and summer 2024. We established a protocol to enrich picocyanobacteria in our samples, which will be especially important in future cool and dark season sampling. To sample, we use a small pump with an inlet at approximately two meters below the estuarine water surface to collect approximately 20 liters of water. Samples are filtered using a protocol designed to enrich for picocyanobacterial DNA by passing samples through a series of filters with descending pore sizes of 20, 5, and 0.2 microns. DNA is extracted from filters using a commercially available kit. Over multiple protocol development iterations, staff made dramatic improvements in both quality and quantity of DNA extracted from filters. Protocol improvements directly enabled focused high-depth sequencing. Also, sampling and environmental metadata associated with sampling were recorded in consultation with **T4**, including time of day, current/flow, tidal cycle, temperature, redox conditions.

We used a commercial product, Azenta/GeneWiz, for all DNA sequencing to date. We developed sequence analysis pipelines in tandem with iterative efforts to develop water sampling and filtering protocols. We processed raw sequence from estuarine sample reads to remove low quality reads, low quality bases at the ends of reads,

contaminating sequences such as adapters and QC spike-ins, and optical/PCR duplicate reads. A small script was written to process output to be suitable for record creation and storage on DataFed (**T4**) using an appropriate bulk upload script. The open-source software Megahit [18] was used to generate *de novo* metagenome assemblies from these reads, which were then filtered to contain only contiguous sequences of more than 10 kilobase in length (for viral contigs). Over the course of several iterations of sampling and sequencing, a target data density of 10 gigabase sequenced per sample was determined to result in greater than 50% sequence diversity and dramatically improved phage sequence recovery in metagenome assemblies. An additional pipeline was developed to provide taxonomic classification of metagenome sequence reads, using the open-source software packages Kraken2 [18] and Yacht [19], with data visualization performed using KrakenTools [20] and Krona [21]. Bracken [22] was used to provide species abundance estimates.

To perform evolutionary analysis without needing species-specific sequence alignments and the construction of phylogenetic trees (and to avoid dependency and knock-on error propagation from errors in these analyses), we are developing a one-shot evolutionary genomics analysis program called HYDROFOIL (HYper Detection ROutine FOr Identification of Lineages, riffing from Yacht). HYDROFOIL is based on the existing PClouds [23] and AnCoV [24] software, and is written in the Go programming language. It takes a Bayesian approach to first identify small clouds of evolutionarily-related kmers in sequence reads that are likely to be positively indicative of cyanobacteria or cyanophage genes or genome regions. Subsequently, the gene clouds are used to probabilistically identify the approximate taxonomic position of sequence reads, integrated within and across samples to identify probable frequency distributions of closely related organisms. Initial benchmarking suggest HYDROFOIL is orders of magnitude faster and more accurate than existing programs. Its primary motivation, however, is to produce output data for high-level probabilistic whole-genome evolutionary analysis of feature origin, evolution, and coevolution, integrated with structural and proteomic information from **T1** and **T2**, which will be a focus of year 2.

Thrust 4 (T4) – To Surveil

After the departure of Marat Valiev, co-Lead of **T4**, we recruited two new staff from PNNL to lead the task assignments. Ruonan Wu is an early career lead who coordinates with other thrusts to develop a metadata scheme. In parallel, she works with Olga Kuchar from ORNL on the data life cycle management. Amity Andersen leads the task of molecular dynamics simulations, modeling and analysis. We executed the PIER plan (please see our original proposal) and partnered with UH and TSU to recruit the students from under-represented groups to participate in the project. In addition, TSU submitted a RENEW application under the support of **T4** staff. TSU's RENEW application focuses on building future workforce of bioinformatics and genome analysis in the research area relevant to DOE BER.

To establish a computational pipeline to simulate key molecular complexes in the phage infection of MED4**, T4** focused on the molecular simulation of cyanobacterium photosystem II (PSII) assembly at the atomistic to coarse-grained levels of molecular simulation theory. It is intended to form first steps in further atomistic/coarse-grained modeling of molecular assemblies of MED4 cyanobacteria with and without post-translational modifications (PTM) from phage infection and environmental factors. In these steps, we specifically started with the atomistic modeling of the PSII subunit D1 and D2 protein structures according to the MED4 proteomics data. These structures include ligands and prosthetic groups such as metal porphyrin and metal oxide, which present a unique challenge to the atomistic modeling efforts because of the need for non-protein force-field parameters. The MED4 base chain protein D1 and D2 molecular models were initially developed with homology modeling using the cyanobacteria MIT Alphafold2[25] structure in Swiss-Model. Prosthetic groups and were added to the structure by aligning this homology model with the closely aligned mesophilic cyanobacterium, *Synechocystis* sp. PCC 6803 (PDB id: 7RCV) structures from electron microscopy. Further Alphafold2 AI, atomistic modeling, and coarse-grained modeling of the MED4 D1 and D2 PSII subunits with Alphafold2 are being planned by the University of Houston team. Once we have a good handle on the MED2 D1 and D2 PSII subunit systems, we will start looking at effect of PTMs on these subunit systems at the molecular level using the PTM-Psi package. At TSU, we have recruited five students, a postdoc, and a program coordinator to assist with student training and administrative activities. The TSU team will focus on simulating the interactions between host and viral CP12 proteins (both redox and oxidized forms) with Glyceraldehyde 3-phosphate dehydrogenase (GAPDH) and Phosphoribulokinase (PRK). The TSU team will also

conduct homology modeling upon receiving the FASTA files for all proteins and will prepare the redox and oxidized forms of CP12 using PTM-Psi[26] (upon being put through on redox and oxidized steps on PTM-Psi).

To establish infrastructure for sharing data across institutions we collaborate with nine institutions, sharing diverse data across the NW-BRaVE project. To enable our scientists to find, access, and understand the provenance of our project's data, and to enable federated scientific data management, our project is using a tool developed at ORNL called DataFed. DataFed is designed to enable scientists to collaborate across institutions on big data while providing mechanisms for enforcing standardized metadata. It is actively being hardened from a research prototype to a core operational capability as part of the data management layer of the project. During this year, development work has focused on addressing security vulnerabilities and simplifying the deployment process. ORNL provided a 1TB DataFed sandbox repository space to enable our scientists to become familiar with the DataFed tool and learn to automate and ingest their datasets. ORNL provided training for 11 researchers who were onboarded to use the tool and demonstrate the ability to store their data in a DataFed managed repository. Working with EMSL IT administrators, we established the creation of a project space on computing facility Tahoma where data can be uploaded to a DataFed repository. In preparation for facilitating interorganizational data movement and metadata management across facilities, PNNL and ORNL are working closely with IT administrators to address security concerns and understand operational issues. Additionally, at ORNL, an Integrated Research Infrastructure testbed (ACE) is being deployed that will further test our technology and do some initial complex workflows involving scientific software and federated data. Finally, the team created data flow diagrams for each thrust to capture the inter- and intra- thrust data needs, assumptions, and requirements for integration of a complex workflow.

Finally, **T4** is developing a synthetic biology solution for testing hypotheses developed in the overall project. Based on the key cyanophage genes (corresponding to phage AMG genes for CP12) identified by PNNL under **T2**, Mehta lab at UIUC made cyanobacterial DNA constructs where these genes are under a constitutive promoter in our secondary model cyanobacteria strain, *Synechococcus elongatus* PCC 7942 (Syn7942). Additionally, these DNA constructs have genomic homology regions and a selection marker. These constructs are used to make mutant Syn7942 cell lines where the cyanophage genes are constitutively expressed. We made three DNA constructs, each corresponding to a single AMG gene expression cassette. We transformed Syn7942 with these DNA constructs and have colonies under selection conditions. We isolated the genomic DNA and polymerase chain reaction (PCR) amplified the integration locus to confirm the presence of the desired genomic integration. Because Syn7942 is a polyploid organism we are currently working on generating homozygous mutant cells lines.

2.3 Future Scientific Goals, Vision, and Plans toward Meeting Project Objectives

Thrust 0 (T0) – To Coordinate

The **T0** team's future scientific goals focus on three key objectives: 1) to scale up the production of viable MED4-specific cyanophage particles that enable visualization and multi-omics studies in support of **T1** and **T2** objectives; 2) to isolate MED4-specific cyanophages from environmental samples collected by the **T3** team and investigate MED4-cyanophage co-evolution in natural settings; and 3) to initiate preliminary experiments to investigate MED4-cyanophage co-evolution under controlled environmental conditions in the laboratory. These efforts will contribute to the development of predictive understanding for the co-evolution of host-virus interactions.

Thrust 1 (T1) – To Visualize

T1 is scheduled to begin collecting time-resolved cryo-tomograms of MED4/cyanophage interactions using the established workflows (Task 1.2) and leveraging samples generated by **T0**. We will also continue to push the development of GRIP-Tomo 2.0 and begin validation experiments with tomograms containing complex distributions of proteins as part of Task 1.3. Meanwhile, Task 1.4 is largely planned and scoped for Year 3 of the project, but we have already begun due diligence work toward identifying different options for hosting the open database - including partnering with the EMSL Central Science repository as well as the EMPIAR public database. There are cases to be made for both options, and we have initiated discussions with the relevant parties to gain more information about possible implementation paths and dependencies including current required metadata standards and annotation formats working with **T4** following the Data Management Plan (from the proposal). We plan to make a fully informed

decision next year so we will be well positioned to empower the development of a robust, user-friendly online platform for hosting host-phage interaction datasets in Year 3. The creation of this open database will be a major step toward enhancing transparency and fostering a collaborative scientific environment for BRaVE related science.

Thrust 2 (T2) – To Map

Following on the successes in pilot studies and workflow optimization, our plan is to 1) generate muti-omics data on MED4's cell states across a whole cell cycle, which is equivalent to a diel cycle; 2) generate multi-omics and multi-PTM data of MED4 with phage infections, and 3) conduct multi-omics data integration and systems modeling. The data without infection will help us to establish the baseline cell states without infections and can be used as the benchmark for the subsequent multi-omics experiment with infections. Also, we will use the measured multi-omics data combined with Cryo-ET data from **T1** to build the whole-cell model of MED4 upon the genome-scale metabolic model, this whole cell model without infection will serve as the baseline model to study the host-phage interactions. In parallel, we will continue to explore the host-phage interactions at the metabolism level, mainly study the AMG's function in host growth dynamics and phage replication dynamics. We will combine our AMG studies with environmental sampling and sequencing results from **T3** to understand the evolution of phage AMGs in different environments and conditions. We will develop a coherent schema to annotate the data as federated data on DataFed under **T4.**

Thrust 3 (T3) – To Track

Following the successes in developing sample collection protocols, sequencing strategies, and pipelines for sequence analysis, we will work to identify potential hotspots/hot moments of host evolution as well as phage-host coevolution. We will measure the trajectories of cumulative genetic changes in the metagenome-assembled genomes of *Prochlorococcus* and related cyanobacteria and in the viral contigs representing the associated phages using the established evolutionary pipeline. In parallel, we will develop, apply and compare high-throughput Bayesian evolutionary analysis to allow integrated simultaneous analysis of all sequence, structure, proteomic and metabolomic data. We will integrate functional annotation and predicted functionally-important variants across cyanobacteria and cyanophage, with a special focus on large complexes (relevant to **T1**), redox proteins and auxiliary metabolic genes (AMGs) incorporated into phage genomes, informing target functions and evolutionary modifications that naturally evolved to integrate into the metabolic modeling in **T2**. This will interact synergistically with **T4** to demonstrate and create a freely available pipeline for practical evolutionary analysis of integrated structure and multi-omic data to investigate the molecular interactions and mechanisms that drive host/parasite co-evolution, and so that **T3** output at all stages can be fed to **T4** and analyzed by e.g. **T2** using molecular and metabolic modeling tools**.**

Thrust 4 (T4) – To Surveil

T4's vision is to create a complex workflow that revolves around federated biological and earth data repositories that will be embedded in the ecosystem of EMSL and JGI User Facilities, partnering university computing facilities, and ASCR facilities. Meeting **T4** deployment objectives will necessitate the following: *(1) Data:* We plan to harmonize the metadata and datatype schemas from NW-BRaVE experiment/observation or analysis and train users to update the status of the metadata on DataFed in real time. The project PI has contacted KBase experts to map out their data requirements and to connect the sequencing data type to bioimaging and structural data type useful for our project. *(2) Compute and Network:* NW-BRaVE received 1 million node hours computing allocation including NERSC, ALCF, and OLCF from an ALCC award. The complex workflow of testing, running, analyzing, visualizing, and interpreting the data based on federated data over ESnet is expected. *(3) Observation/Experiment Data Visualization and Analysis:* **T4** will reveal technical gaps in data transformation, data visualization and data analysis along the data life cycle. PI Cheung will contact ASCR experts whom she made connections at the ASCR IRI/HPDF meeting of 2024 or consult with Program Manager Resham Kulkarni about these gaps.

2.4 New Scientific Results That May Shift Current Research Focus Areas and/or Identified Project Knowledge

Based on scientific results during the first year, we project a shift in focus during years one and two towards production of knowledge and mechanisms to integrate knowledge across the scientific domains needed to understand virus and host protein interactions and coevolution. This is motivated by our increased practical understanding of the difficulty of communicating across diverse fields, and partly reflected in our incorporation of the newly created **T0** to establish, execute, and synchronize consistent host-virus culture methods across thrusts, integration of structural information into the metabolic and Lattice Microbe models, and modifications made to the focus of **T4** as we have proceeded. However, we are further motivated to focus on a tangible scientific product that will be a system of integrated workflow to make diverse knowledge easily accessible to diverse researchers. Although it is already a goal to create clear metadata schema, pointers, and versioned datasets and libraries for each set of experimental outputs to be represented in the ORNL-housed DataFed, it is our further aim to create a common underlying theoretical schema that will unify them all and serve as an organizational backbone.

This *underlying schema* will be based on mechanistic models of how data is produced, from protein expression to post-translational regulatory control affecting complex formation and location of functioning and interacting molecules withing the cells. Furthermore, because mechanisms of virus and host coevolution are at the core of NW-BRaVE research, we will be highly concerned with functional interactions that affect the *transmissibility of information* among experiments under different conditions and among divergent organisms. Such concerns are obvious when performing evolutionary analysis, but are also highly relevant in for example metabolic and structural modeling, where it is often necessary to collect experimental functional and kinetic properties, and structural complex formation, from organisms with varying levels of divergence from the target. Our expectation in preparation for future pandemics is that we will in the beginning have information from model organisms that are related to but divergent from the organism causing the pandemic, and a systematic means of transferring that information and understanding what assumptions are made during the transfer will be key. In other words, we will have model organisms to serve as knowledge bases to transfer information to the new viral system. The purpose of our model system here, cyanobacteria and cyanophages, is to hone our strategies and pipelines for making such transfers without loss of information, and minimizing language contradictions and hidden mechanistic assumptions of different subfields that may be incompatible with accurate overall understanding of the entire system.

2.5 Collaborative Research Activities with External Researchers in Pursuit of Program Objectives

- External researchers supported by the NW-BRaVE were integrated into individual thrust research activities and meetings. Progress reports and contributions from external researchers are described under each thrust in Section 2.1, 2.2, and 2.3.
- External researchers were invited to the virtual biweekly NW-BRaVE Journal Club meetings. PIs or postdocs from external institutions were invited to present their research relevant to the project objectives.
- External researchers have access to PNNL Teams, Confluence, and Gitlab. Those who are involved in the computing workflow of NW-BRaVE also have access to EMSL's computing facility Tahoma and set up Globus IDs for sharing data on DataFed.
- In March 2024, NW-BRaVE held the first hybrid Mini-Symposium at PNNL in Richland, WA. We kicked off the research activities with external researchers in 2.5 days. We mapped out the order of major experiments from each thrust and coordinated the sample preparation and production for the overall research design. Three members from the ORNL Team were on-site and dedicated a full day to interview the key members of each thrust. We surfaced the knowledge gaps in the metadata schema and data requirements necessary for capturing the provenance and the life cycle of data.

- In May 2024, PI Cheung and **T2** Early Career Lead Feng attended the in-person UIUC Quantitative Cell Biology Workshop held by Prof. Zan Luthey-Schulten to learn about the requirement of software and hardware for setting up models for Lattice Microbe and Gromacs simulations (https://github.com/Luthey-Schulten-Lab/Workshop_2024). PI Cheung gave a research presentation in the last day of the event.
- PI Cheung and Molecular Simulation Lead Andersen gave on-line presentations to UH and TSU faculty and students during recruiting and training events.

3.0 National Laboratory BRaVE Project Structure with Management and Scientific Personnel Identified

3.1 Assignments of Key Team Members to Specific Task Areas.

Please see Appendix A Table A.1 for the full list of key team members at PNNL, their roles and responsibilities.

4.0 Staffing and Budget Summary

4.1 Funding Allocation by Project Element

4.1.1 Focus on Project Deliverables / Milestones

Year-1 timelines and deliverables for each thrust, including **T0**, are in Table 1. Grey blocks denote planned activities and orange blocks denote meeting the deliverables.

Table 1: Milestones and Deliverables

Major Activities / Tasks / Milestones	23	24				25				26		
	Q4	Q1	Q2	Q3	Q4	Q1	Q2	Q3	Q4	Q1	Q2	Q3
Program Management												
All-hands kickoff meeting												
Annual meeting												
Major Technical Milestones												
Profiled host-pathogen interaction using MED4 and model virus												
Profiled host-pathogen interaction using MED4 and isolated virus from samples												
T0: Integrative Research Coordination												
Scale biomass production of MED4												
Multiplicity of Infection of model virus												
Multiplicity of Infection of isolated virus from samples												
T1: Visualize												
Document thrust data requirements												
Tomograms of MED4												
Tomograms of MED4 and model virus												
Tomograms of MED4 and isolated virus from samples												
Characterized proteins from tomograms of MED4												
Characterized proteins from tomograms of MED4 and model virus												
Characterized proteins from tomograms of MED4 and virus from isolated samples												
Annotated proteins responsive to PTM												
T2: Map												
Document thrust data requirements												
Omics of MED4												
Omics of MED4 and model virus												
Omics of MED4 and isolated virus from samples												
Lattice Microbe sample preparation												
PTM modeling with NWChem												
Systems modeling												
T3: Track												
Document thrust data requirements												
Coordinate sample collection protocols												
Environmental Sampling												
Genomic sequencing and modeling												
Send isolated samples to Richland/UIUC												
T4: Surveil												
Platform for modeling and analytics												
Platform for experimental validation												
Training and hackathon												
Data integration requirements doc.												
High-level system design of workflow												
Platform Releases												
Establish and evolve data policies												

Commented [CHM1]: I can move this to the next page, or we can keep this here and delete "on top of the next page."

Commented [mw2R1]: Yes. Thanks.

4.1.2 Present Funding

FY23 project funding was $3.150m, received September 2023. It was distributed across Overall Research Design and to each thrust at PNNL in addition to nine subcontracts (two other national labs and seven universities).

4.1.3 Document Changes in Funding Allocations to Project Elements

At PNNL: Co-PI Marat left PNNL during the Kickoff period in 2024. PI Cheung mitigated the risk by recruiting Ruonan Wu and Amity Andersen to **T4**. Ruonan Wu was promoted to Early Career Lead for **T4**. She oversees the requirement of data storage and network for data life cycle and management across thrusts. Amity Andersen is a Molecular Modeling and Simulation Task Lead. She oversees the computing requirement in a complex workflow of PTM-Psi and other software on high performance computing facilities. In addition, Dana Woodruff, key

personnel in coevolutionary sequence analysis for **T3**, retired from PNNL-Sequim in December 2023. PI Cheung mitigated the risks by recruiting Scott Edmundson from PNNL-Sequim and by creating a subcontract to support David Pollock's graduate student at the University of Colorado Anschutz.

At UIUC: Prof. Martin Gruebele retired in June 2024. As an Emeritus Professor, he is training the postdoc and students of Prof. Angad Mehta who leads the tasks of *in vivo* imaging. PI Cheung created a subcontract to support her graduate student at the University of Washington at Seattle to continue the data analysis and interpretation of bioimaging in **T4**.

4.2 Funding Allocation to External Collaborators

4.2.1 Status of External Collaborations with Universities and/or Private Sector

The project funding supports seven subcontracts at Texas Southern University (TSU) for $100K, University of Illinois, Urbana-Champaign (UIUC) for $158K, Northwest Indian College (NWIC), University of Houston (UH) for $97K, Colorado Department of Public Health and Environment (CDPHE), University of Washington (UW) for $50K and University of Colorado Denver (CUA) for $40K. Allocation of funding to NWIC and CDPHE has not yet been distributed due to contractual delays. It is estimated to have those complete by September 2024 or earlier.

4.2.2 Status of External Collaborations with Other National Laboratories

The project funding supports two subcontracts for Dr. Olga Kuchar at Oakridge National Laboratory (ORNL) and Dr. Arvind Ramanathan at Argonne National Laboratory (ANL). $400K was distributed to ORNL to provide expertise on data life-cycle management and develop data integration and transformation. $50K was distributed to ANL to provide expertise on accelerating protein identifications from tomogram using existing AI/ML software.

4.3 Personnel Actions and Procedures

PI Cheung and the remainder of the NW-BRaVE leadership followed the PIER plan (see PIER Plan from the original proposal) regarding personnel actions, hiring procedures, and procedures for encouraging young investigators. As part of career development, Early Career Leads Doo Nam Kim, Ruonan Wu, and Pavlo Bohustkyi submitted proposals to the DOE Early Career funding opportunity, while Doonam Kim and Song Feng proposed and were awarded projects with PNNL LDRD funding that were synergistic to research they are pursuing with NW-BRaVE. Each Early Career Thrust Lead also received training from PNNL as competent research mentors and proactively recruited postdocs, post masters and interns. Most impressively, we have recruited a total of seven SULI and CCI interns to the research project as new hires (Table 2). Many have converted to tech interns or post baccalaureates after the internship. In parallel, PI Cheung presented talks at UIUC, UH, and TSU for events of recruiting new students. In this process, Early Career Leads at PNNL forged close research collaboration with university partners. For example, among the **T4** cohort, Ruonan Wu and Amity Andersen participated in the RENEW application with TSU.

Table 2: Number of participants

	# Staff/ faculty	# Postdoc	# Postmaster	# Postbacc	# Graduate students	# Undergrads /SULI/CCI
PNNL	40	5	3	4		9 students with 6 SULIs and 1 CCI
ORNL	3					
ANL	2					
UIUC	3	1			2	
TSU	1	1				5
UH	1	Pending				3
UW					1	
CUA					1	
CDPHE	3					
NWIC						Pending

4.4 Capital Equipment Needs (Future)

Not Applicable.

5.0 References

1. Lamm, L., et al., *MemBrain v2: an end-to-end tool for the analysis of membranes in cryo-electron tomography.* bioRxiv, 2024: p. 2024.01.05.574336.
2. Siggel, M., et al., *ColabSeg: An interactive tool for editing, processing, and visualizing membrane segmentations from cryo-ET data.* Journal of Structural Biology, 2024. **216**(2): p. 108067.
3. George, A., et al., *Graph identification of proteins in tomograms (GRIP-Tomo).* Protein Science, 2023. **32**: p. e4538.
4. Nakayasu, E.S., et al., *MPLEx: a Robust and Universal Protocol for Single-Sample Integrative Proteomic, Metabolomic, and Lipidomic Analyses.* mSystems, 2016. **1**(3).
5. Roberts, E., J.E. Stone, and Z. Luthey-Schulten, *Lattice microbes: High-performance stochastic simulation method for the reaction-diffusion master equation.* Journal of Computational Chemistry, 2012. **34**(3): p. 245-255.
6. Thornburg, Z.R., et al., *Fundamental behaviors emerge from simulations of a living minimal cell.* Cell, 2022. **185**(2): p. 345-360.e28.
7. Ofaim, S., et al., *Dynamic Allocation of Carbon Storage and Nutrient-Dependent Exudation in a Revised Genome-Scale Model of Prochlorococcus.* Frontiers in Genetics, 2021. **12**.
8. Chang, A., et al., *BRENDA, the ELIXIR core data resource in 2021: new developments and updates.* Nucleic Acids Research, 2021. **49**(D1): p. D498-D508.
9. Caspi, R., et al., *The MetaCyc database of metabolic pathways and enzymes and the BioCyc collection of Pathway/Genome Databases.* Nucleic Acids Research, 2014. **42**(D1): p. D459-D471.
10. Flamholz, A., et al., *eQuilibrator--the biochemical thermodynamics calculator.* Nucleic Acids Research, 2011. **40**(D1): p. D770-D775.
11. Li, F., et al., *Deep learning-based kcat prediction enables improved enzyme-constrained model reconstruction.* Nature Catalysis, 2022. **5**(8): p. 662-672.
12. Henson, M.A. and T.J. Hanly, *Dynamic flux balance analysis for synthetic microbial communities.* IET Systems Biology, 2014. **8**(5): p. 214-229.
13. Ebrahim, A., et al., *COBRApy: COnstraints-Based Reconstruction and Analysis for Python.* BMC Systems Biology, 2013. **7**(1).
14. Cai, L., et al., *Biological interactions with Prochlorococcus: implications for the marine carbon cycle.* Trends in Microbiology, 2024. **32**(3): p. 280-291.
15. Crummett, L.T., et al., *The genomic content and context of auxiliary metabolic genes in marine cyanomyoviruses.* Virology, 2016. **499**: p. 219-229.
16. Liu, R., et al., *Cyanobacterial viruses exhibit diurnal rhythms during infection.* Proceedings of the National Academy of Sciences, 2019. **116**(28): p. 14077-14082.
17. Thompson, L.R., et al., *Phage auxiliary metabolic genes and the redirection of cyanobacterial host carbon metabolism.* Proceedings of the National Academy of Sciences, 2011. **108**(39).
18. Li, D., et al., *MEGAHIT: an ultra-fast single-node solution for large and complex metagenomics assembly via succinct de Bruijn graph.* Bioinformatics, 2015. **31**(10): p. 1674-6.
19. Koslicki, D., et al., *YACHT: an ANI-based statistical test to detect microbial presence/absence in a metagenomic sample.* Bioinformatics, 2024. **40**(2).
20. Lu, J., et al., *Metagenome analysis using the Kraken software suite.* Nat Protoc, 2022. **17**(12): p. 2815-2839.
21. Ondov, B.D., N.H. Bergman, and A.M. Phillippy, *Interactive metagenomic visualization in a Web browser.* BMC Bioinformatics, 2011. **12**: p. 385.
22. Lu, J., et al., *Bracken: estimating species abundance in metagenomics data.* Peerj Computer Science, 2017.
23. de Koning, A.P., et al., *Repetitive elements may comprise over two-thirds of the human genome.* PLoS Genet, 2011. **7**(12): p. e1002384.
24. Kemp, S.A., et al., *SARS-CoV-2 evolution during treatment of chronic infection.* Nature, 2021. **592**(7853): p. 277-282.

25. Jumper, J., et al., *Highly accurate protein structure prediction with AlphaFold.* Nature, 2021. **596**: p. 583-589.
26. Mejia-Rodrigue, D., et al., *PTM-Psi: A python package to facilitate the computational investigation of post-translational modification on protein structures and their impacts on dynamics and functions.* Protein Science, 2023. **32**: p. e4822.

Appendix A

Table A.1: Assignments of key team members to specific task areas from September 2023 through July 2024.

Name	Role	Total hours	FTE (over 10%)
Margaret Cheung	PI, co-Director, and Thrust lead responsible for managing all Thrust 3 subtasks and deliverables	685	.41
David Pollock	co-Director and Thrust lead responsible for managing all Thrust 3 subtasks and deliverables.	410	.24
James Evans	Thrust lead responsible for managing all Thrust 1 subtasks and deliverables.	177	.11
Weijun Qian	Thrust lead responsible for managing all Thrust 2 subtasks and deliverables.	228	.14
Doo Nam Kim	Early-career Lead in Thrust 1	547	.33
Trevor Moser	Performs cryo-electron tomography sample preparation, data collection and 3D reconstruction in Thrust 1.	192	.11
Amar Parvate	Postdoc, assists Moser in sample preparation and processes tomogram images in Thrust 1.	334	.20
Kate Baldwin	Tech Student, picks virus particles from tomograms in Thrust 1.	294	.18
Song Feng	Early-career Lead in Thrust 2 and mentor trainees.	523	.31
Xiaolu Li	Postdoc, executes the proteomic experiments in Thrust 2.	364	.23
Youngki You	Postdoc, executes the cross-linking proteomic experiments in Thrust 2.	316	.19
Owen Leiser	Early-career Lead in Thrust 3 and mentor trainees.	452	.27
Conan Johnson	Postdoc, develops evolutionary models of phage-cyanobacteria systems to assist with understanding the interactions between microbial communities and their environments in Thrust 3.	209	.13
Noelani R Boise	Contributes to sample collection from the Salish Sea at the PNNL-Sequim in Thrust 3.	403	.24
Ruonan Wu	Early Career Lead in Thrust 4 and in Integrative Research Design.	389	.23
August George	Postdoc, contributes to data transformation and visualization in Thrust 4.	612	.37

Pavlo Bohutskyi	Early-career Lead in Integrative Research Design and Coordination (Thrust 0).	556	.33
Other staff		955	.57
Admins and FCs	.	292	.17

Pacific Northwest National Laboratory

902 Battelle Boulevard
P.O. Box 999
Richland, WA 99354

1-888-375-PNNL (7665)

www.pnnl.gov