UIUC's STC QCB Retreat - January, 2026

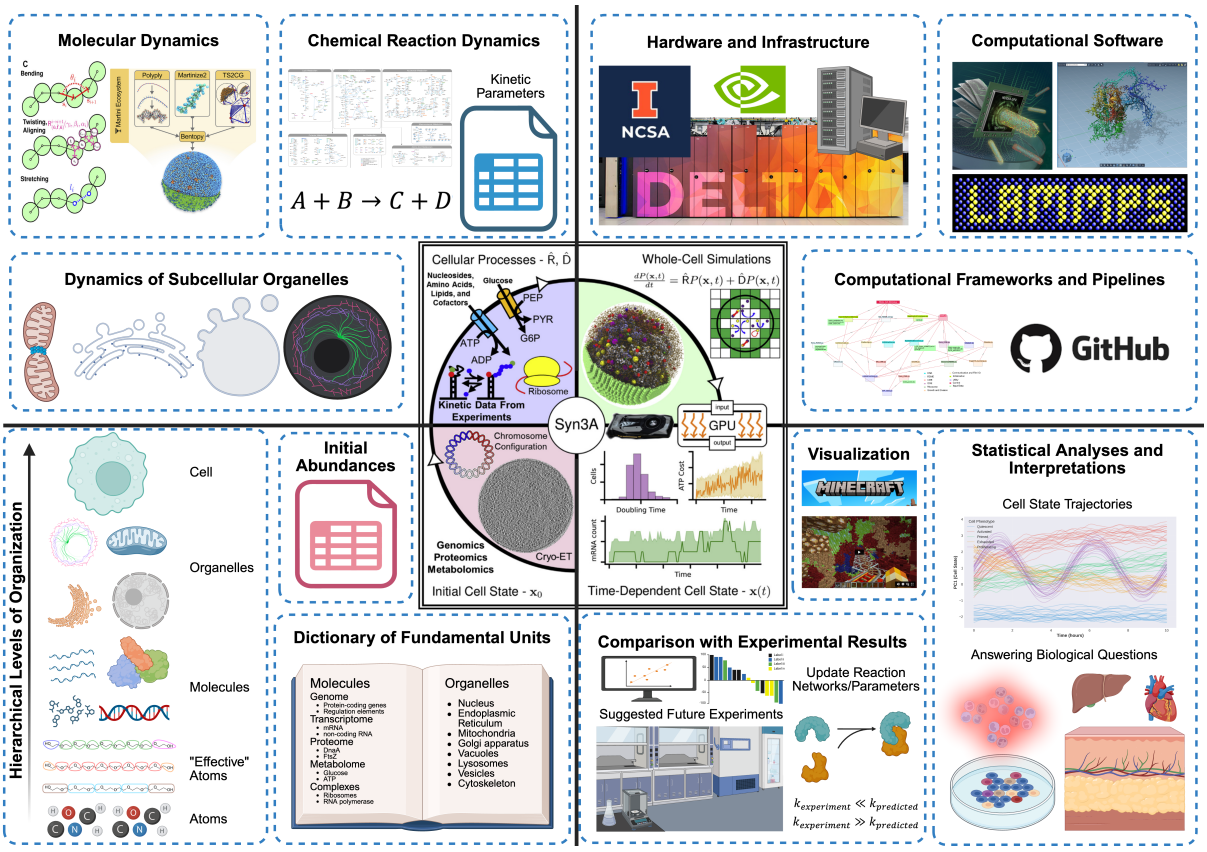

Figure 1: The Whole-Cell Model

The Whole-Cell Model: A Unifying Framework for Quantitative Cell Biology

The purpose of this GitHub repository is to help researchers affiliated with the QCB know how to make meaningful contributions toward the generation of whole-cell models. As such, the majority of this repository details the various aspects of whole-cell modeling and how experimental, computational, and theoretical labs can contribute to each of these areas. First, the overall aims of whole-cell modeling are described as well as each of the individual components of a whole-cell model. Next, each component of the whole-cell modeling process is broken up into subcomponents which can serve as areas of focus for individual labs or teams. It is assumed that readers of this repository already understand the importance of whole-cell modeling as this topic will not be explored here. Rather, this repository serves as a guide for researchers and professionals who desire to contribute to the creation of whole-cell models.

Overall Aims of the Whole-Cell Model

The overall aim of a whole-cell model is to provide a comprehensive description of the cell state across a defined period of time, often the length of the cell cycle (see Figure 1). The whole-cell model is comprised of a description of the cell state, its initial conditions, and the computational procedures that define how the cell state evolves over time. After being constructed, whole-cell models must be realized (or simulated) using a combination of hardware, software, and computational pipelines. After whole-cell models have been simulated and results have been generated, this data can be used to perform statistical analyses, answer biological questions, visualize output, and compare output with experimental results.

The Cell State and Initialization

All whole-cell models require both a definition of the cell state and initial conditions for the cell state (see Figure 2). The cell state can be thought of as a snapshot of the cell at a distinct moment in time. The cell state contains information about all aspects of the cell, and it is the cell state that evolves over time in simulations of whole-cell models. A critical part of constructing a whole-cell model is defining the cell state by determining its hierarchical levels of organization, laws of emergence that relate levels of organization of the cell state to other levels, and a dictionary of all fundamental units within each level of organization. After the cell state has been defined, it also must be initialized. Each of these requirements are explained in the following subsections.

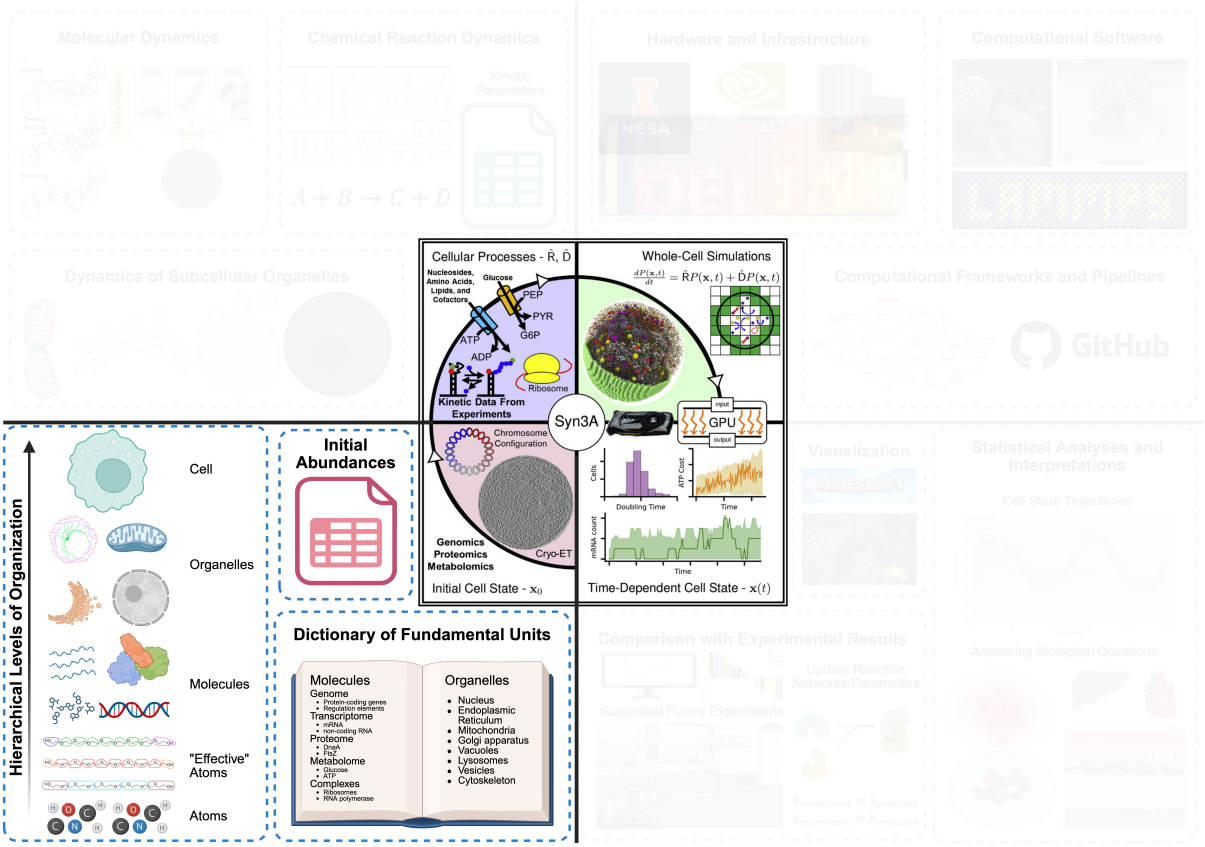

Figure 2: The Cell State and Initialization

Hierarchical Levels of Organization

Defining the cell state is accomplished first by defining hierarchical levels of organization. These levels can vary across whole-cell models, but should encompass the most fine-grained level of organization (can be atoms or "effective" atoms) through the most coarse-grained level (usually the cell itself). Not only do hierarchical levels of organization need to be defined, but so do the "laws" or computational procedures that define how fundamental units at one level of organization emerge from lower-level fundamental units. In general, these laws are well defined for lower levels of organization such as how atoms make up molecules, but they are less well defined for higher levels of organization such as how a group of individual molecules constitute a subcellular organelle.

Individual labs and teams can contribute to this aspect of whole-cell modeling by developing theoretical models of how higher levels of organization emerge from lower levels of organization. For example, what are the computational procedures that can be used to convert information about the numbers and locations of specific molecules in a cell into information about the state of the subcellular organelles? Likewise, how can

information about the organelles and molecules in a cell be used to define the cell type being studied or simulated?

Dictionary of Fundamental Units

Each hierarchical level of organization within a whole-cell model contains "fundamental units". These are the individual objects within the specified level and correspond to physical entities within the cell. For example, at the organizational level of "Molecules", fundamental units may include various metabolites, proteins, genes, RNA molecules, etc. Additionally, at the level of "Organelles", examples of fundamental units may be the plasma membrane, nucleus, mitochondria, etc. Each fundamental unit at each level of organization should have a "type" to distinguish itself from other fundamental units. Each type of fundamental unit can then be assigned an arbitrary amount of defining information such as atomic radius, charge, mass, amino acid sequence, standardized nomenclature, etc.

A critical component of defining the cell state in a whole-cell model is a dictionary of fundamental units. This dictionary contains all the information about each fundamental unit type across all hierarchical levels of organization. In essence, it is a large lookup table with which the properties of any individual fundamental unit can be accessed.

Individual labs and teams can contribute to this aspect of whole-cell modeling in a variety of ways. For example, all molecular species and molecular complexes within a whole-cell model need to be documented. If a lab is studying specific biological pathways or reactions, the lab can gather information about molecular species included in these reactions. Such information may include DNA-protein binding partners, stoichiometric ratios of protein complexes, or the identity of various RNA or protein isoforms. Similarly, if a lab is focused on the study of subcellular organelles, information about the organelle sizes, molecular compositions, or functions can be used in the dictionary of fundamental units.

Initial Conditions

The hierarchical levels of organization and dictionary of fundamental units are required to define the "cell state". The cell state is the overall description of the cell at every hierarchical level of organization, and as such, is composed of a list of all fundamental units present in the cell within each level. Each instance of a fundamental unit can also be supplied with information about its type, relationship with other fundamental units, and time-dependent properties.

The status of the cell state at the beginning of the simulation (initial conditions of the cell state) must be defined before the simulation of the whole-cell model can take place. This requires information about both the abundances and locations of each type of fundamental unit.

Individual labs and teams can contribute to this aspect of whole-cell modeling in a variety of ways. First, labs that work with high-dimensional -omics data can perform experiments to determine the abundances (absolute or relative) of a wide variety of molecules within the cell. Next, labs which perform microscopy experiments can help determine initial locations of molecules and subcellular organelles. Microscopy labs can also help determine the architecture or geometries of both organelles and of the cell as a whole. This information can be used to set the initial conditions of the cell state.

Cellular Processes

Once the cell state has been defined and initialized, computational procedures for the time evolution of the cell state are needed (see Figure 3). Here, we will refer to these procedures as "cellular processes". Essentially, cellular processes are computational algorithms that use various aspects of the most recent cell state as input and predict cell state variables at a specified time in the future. Cellular processes can be as fine- or coarse-grained as needed. Cellular processes may also receive input and produce output from multiple levels of organization of the cell state. Although not necessary, cellular processes often correspond to specific biological processes such as molecular diffusion, DNA replication, transcription, translation, signal transduction, cell proliferation, etc.

Figure 3: Cellular Processes

Molecular Dynamics

An important time-dependent property of the cell state is the location of each atom or molecule within the cell. Therefore, it is important to understand how these atomic and molecular locations change with time. A well-established variety of molecular dynamics methods can be used for such cellular processes. Not all molecules or processes within a whole-cell model require this level of granularity, but for the processes that are highly dependent on molecular locations, molecular dynamics methods offer a useful solution.

Molecular dynamics simulations have varying levels of granularity in their fundamental units. Some simulations such as "all-atom" simulations track the location of individual atoms within molecules while other simulations use coarse-grained "beads" which represent multiple atoms as their fundamental units.

Using molecular dynamics simulations in whole-cell models is an active field of research with some difficulties arising due to the discrepancy of timescales for molecular and cellular simulations. Molecular simulations require a large amount of computational power and therefore can only be performed across small timescales. However, in order to perform simulations of whole-cell models, timescales corresponding to hours are often needed.

Another benefit of molecular dynamics simulations is their ability to estimate diffusion coefficients and association/disassociation rates of molecules and molecular complexes. These simulations can be performed completely separate from a simulation of a whole-cell model, and their results can be used in creating a dictionary of the types and properties of fundamental units, as described above.

Labs and teams that have expertise in molecular dynamics can contribute to the construction of whole-cell models by developing methods for integrating molecular dynamics simulations with whole-cell model simulations. In addition, these labs and teams may contribute by determining rate constants and diffusion coefficients for specific association reactions and molecules, respectively.

Chemical Kinetics or Chemical Reaction Dynamics

Another time-dependent property of the cell state is the abundance of each molecule within the cell. The evolution of molecular abundances is governed by the kinetics of chemical reactions, where molecules of one type are converted into molecules of a different type.

Defining how molecules can be interconverted requires a variety of steps. First, researchers must construct a reaction network that defines all the possible reactions that can take place within the cell. Although much progress has been made in determining metabolic reaction networks, much work is still needed to define and refine reaction networks for the processes of DNA replication, gene regulation, transcription, translation, post-transcriptional and post-translational modifications, signal transduction, and molecular complex formation. Adding further complication to this process is the fact that many cell types have unique reaction networks, or reaction networks that are slightly different from other cell types. This is true both within and across species. Second, kinetic parameters that define the rates of these reactions must be determined. This is often done experimentally and can be both time- and labor-intensive. Other methods for determining kinetic parameters for chemical reactions include using specialized reactive force fields such as ReaxFF, quantum mechanics/molecular mechanics methods, or machine-learning methods. Third, the computational procedure for simulating chemical reactions needs to be determined. To date, there have been various methods used in whole-cell models to simulate the time evolution of chemical abundances including flux balance analysis (FBA), systems of ordinary differential equations (ODEs), the chemical master equation (CME) coupled with the Gillespie algorithm, the reaction-diffusion master equation (RDME) coupled with extensions of the Gillespie algorithm, Markov models, etc. Often, whole-cell models will use multiple methods within the same model to simulate different processes involving chemical kinetics. In these cases, a reaction network in which each reaction has a specified simulation method must be constructed.

Labs and teams can contribute to this aspect of whole-cell modeling in a variety of ways. First, labs that can delineate the sub-reactions that comprise biological processes can help generate detailed reaction networks. For example, labs that can uncover the order of reactions that lead to formation of protein-protein complexes can help construct realistic reaction networks that can be used to determine possible molecular transition pathways. Similarly, labs that specialize in studying gene regulation can construct detailed reaction networks that describe how transcription factors bind to DNA, histone modifications are performed, or DNA methylation markers are edited. Second, there exists a critical need for an enormous number of accurate kinetic parameters for a host of biological reactions. For example, of the ~16,000 biological reactions annotated for human cells in the Reactome database, only a very small minority of them have accurate kinetic parameters. Similarly, there exists a dearth of kinetic parameters for enzyme-catalyzed metabolic reactions for many cell types and organisms. Groups that have the ability to perform experiments or simulations that produce these missing kinetic parameters can work with groups that are

building whole-cell models to determine which reactions are most critical to have accurate kinetic parameters.

Dynamics of Subcellular Organelles

Whole-cell models also require information on the dynamic behaviors of subcellular organelles. Organelle geometries, locations, interactions, and molecular compositions all change as a function of time, and computational procedures that define how these processes unfold are critical for building whole-cell models. Information on organelle dynamics is important for both prokaryotic and eukaryotic cells. Although prokaryotic cells only contain a plasma membrane and cytosol, these "organelles" do change over time, and these changes are vital aspects of the cell state and cellular functioning. Furthermore, understanding organelle dynamics becomes even more important when constructing eukaryotic whole-cell models because eukaryotes contain a wide array of organelles, each with its own unique functions.

Much of the previous work on organelle dynamics has been focused on how organelles change during the process of mitosis. However, organelles also exhibit dynamic behavior within other phases of the cell cycle (i.e., G1, S, and G2). Laws governing the dynamics of organelles across the entirety of the cell cycle are needed for the generation of whole-cell models.

Included in this category is also a description of how the overall cellular architecture evolves over time. This includes both the geometry of the plasma membrane as well as the process of cell division or proliferation.

Individual labs and teams that study any aspect of organelle dynamics can contribute significantly to whole-cell modeling efforts. The molecular composition and geometries of membrane-bound organelles can be determined for various cell types by experimental efforts such as organelle isolation/fractionation and a combination of lipidomic and proteomic analyses. Microscopy methods are also invaluable in studying organelle geometries. Especially important are time-series experiments in which organelle information at multiple times across the cell cycle can be determined. Additionally, mathematical or computational models for the dynamic behavior of organelles and overall cellular architectures need to be established. Close collaboration between experimental and theoretical groups will hopefully lead to the development of dynamic models to predict organelle and overall cell dynamics.

Whole-Cell Simulations

Once the cell state is defined and initialized, and all computational procedures for the time-evolution of the cell state are in place, a whole-cell model must be realized or simulated using high-performance computing (see Figure 4). In order to perform whole-cell model simulations, hardware and infrastructure must be procured, computational software must be designed, and computational frameworks and pipelines must be established. In this area of whole-cell modeling, two main concerns are as follows: (1) reducing simulation times and costs and (2) ensuring all computational frameworks are user-friendly, reusable, and reproducible.

Figure 4: Whole-Cell Simulations

Hardware and Infrastructure

The hardware and infrastructure used for whole-cell modeling must be capable of handling vast amounts of computation and data storage. Both the information that specifies the whole-cell model and the model's time-dependent cell states must be saved. Depending on the number of cellular replicates one is simulating, the cell state storage data can accumulate quickly. Additionally, the computational processes of whole-cell models require large amounts of RAM. Groups and labs that actively research how to best utilize current hardware setups to achieve optimal performance can contribute to speeding up the process of performing whole-cell simulations.

Computational Software

While previous sections described the computational procedures that predict the time-evolution of the cell state, often these procedures are realized through the implementation of specific software. Examples of this are LAMMPS or GROMACS for molecular dynamics, Lattice Microbes for chemical kinetics, and MitoTNT for organelle dynamics. However, presently, there are many cellular processes in whole-cell modeling that do not have dedicated software. Whole-cell modeling requires the development and maintenance of both existing and new software to efficiently perform the computations required for cellular processes.

Labs and teams that have expertise in software development and maintenance can contribute to whole-cell modeling efforts. By working closely with groups involved in defining cellular processes, labs with software expertise can develop specialized software that is both useful and efficient. Additionally, the collaboration of labs and groups that work with either hardware or software can further optimize the software being constructed for specific hardware that will be used in whole-cell simulations.

Computational Frameworks and Pipelines

After computational hardware and software have been procured and developed, whole-cell models require the creation of computational frameworks or pipelines that are both efficient and user-friendly. Ideally, the field of whole-cell modeling will eventually have standard computational pipelines that all researchers can use to ensure reproducibility. However, at present, individual labs are usually individually tasked with creating and maintaining the necessary computational frameworks for whole-cell simulations. As not all labs or teams are experts in the fields of computer or data science, not all whole-cell models that have been created are user-friendly or follow standard coding guidelines.

Currently, whole-cell modeling pipelines should be both modular and easily alterable. Modular pipelines are useful because they allow individuals or research groups to work on different aspects of the code separately without needing to understand the pipeline in its entirety. Additionally, as new biology is discovered, a modular and alterable framework allows the pipeline to have additional cellular processes added with ease.

Labs and teams with experience in building and maintaining computational frameworks can contribute by offering their expertise in creating whole-cell simulation pipelines. In the future, it would be ideal to have a single pipeline that can accommodate all types of whole-cell models and simulations, regardless of cell type or organism. Increasing the efficiency and usability of whole-cell modeling pipelines will also make these invaluable tools available to the broader research community.

Model Predictions and Output

After whole-cell simulations are complete, model output should be converted into understandable and actionable information (see Figure 5). For this to take place, model output must be statistically analyzed, visualized, and compared with experimental results.

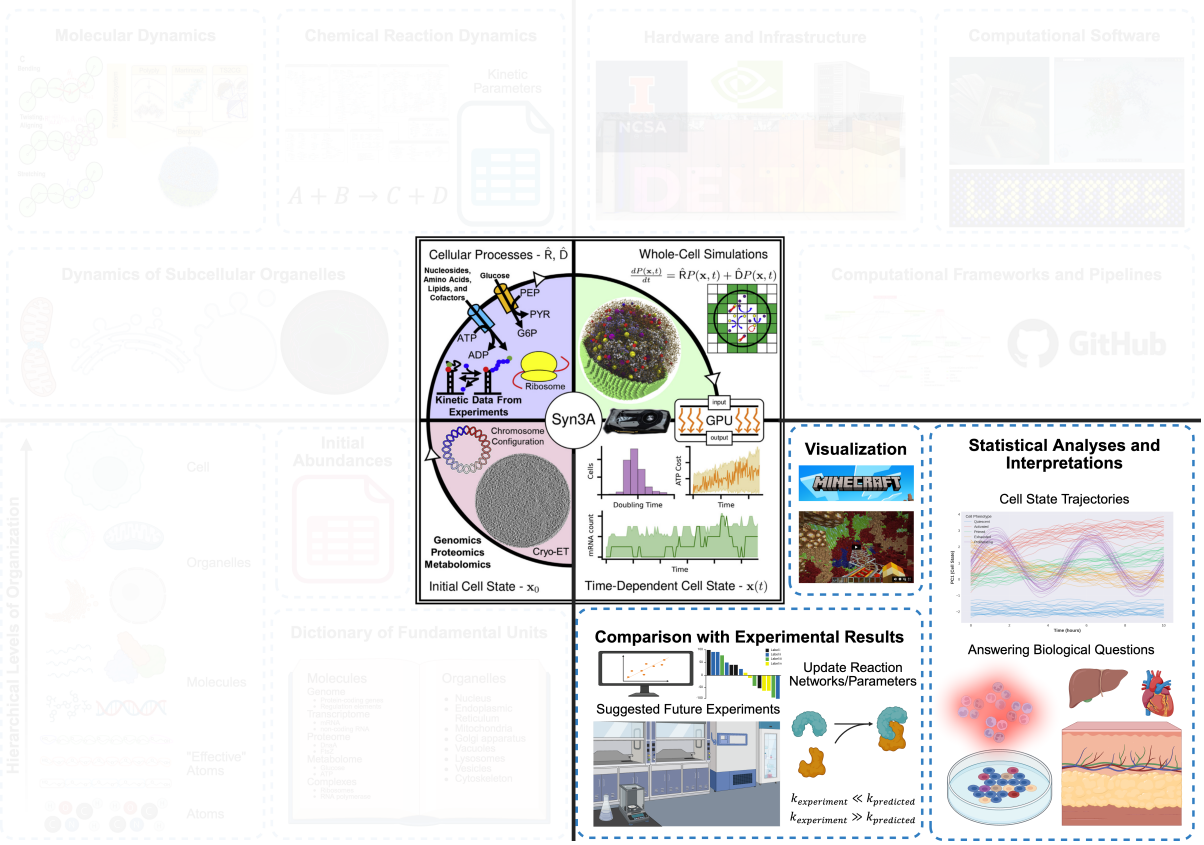

Figure 5: Model Predictions and Output

Statistical Analysis and Interpretations

The output from a whole-cell model is only useful if researchers can use this data to answer biological questions. Furthermore, appropriate statistical methods must be used in analyzing this data so that researchers can have confidence in model predictions. Output from whole-cell simulations usually takes the form of large N-dimensional matrices with information about all variables related to the cell state across the timeframe of the cell simulation. These data are usually generated in replicate (which is made possible due to the stochastic nature of certain cellular processes), meaning that the number of output data grows as sample sizes increase. Essentially, the types and number of questions that can be probed using the output from whole-cell simulations is vast. To make sense of this vast amount of information, the data must be able to be converted into a human-readable format. Additionally, both existing and novel statistical methods must be used to make conclusions about the significance of any findings.

The following is a non-comprehensive list of statistical methods that can be used to glean information from whole-cell model simulations: (1) traditional statistical methods such as linear regression, t-tests, and ANOVAs, (2) more complicated forms of regression such as multilevel-modeling, structural equation modeling, and group-based trajectory analysis, (3) Bayesian statistics, and (4) machine-learning and novel AI methods.

In addition to statistical expertise, labs with biological expertise are crucial in making use of the vast amount of data generated from whole-cell simulations. Labs studying the biology of the cell type being simulated can determine the most pressing unanswered biological questions. Working in close collaboration with the labs that generate whole-cell models, these biology-focused labs will be able to guide whole-cell modeling efforts to ensure the most useful questions are being answered. Additionally, labs and research teams that have expertise in statistical analysis can coordinate model analyses with labs that have biological expertise. The close coordination of efforts between these two types of groups will enable the field to come to important biological conclusions using rigorous quantitative methods.

Visualization

Visualizing whole-cell simulation results can be useful for both researchers and the general public. It is often said that a picture is worth a thousand words. Images and videos of whole-cell models aim to illuminate the underlying biological results produced by such models.

Clear visualizations of model out will benefit researchers and scientists. In addition to combing through vast amounts of numerical data, researchers can also intuitively understand whole-cell model predictions using accurate visualizations of model output. Additionally, researchers need to be able to communicate their models' findings to other scientists that may not be well-acquainted with the output from such models. Clear visualizations of model predictions will help scientists better understand their own results and thereby communicate them with other scientists.

Whole-cell model visualizations are also valuable because they can improve communication of biological findings with the general public. Although many individuals may be able to conjure up an image of a cell from their biology high-school class, these images are often inaccurate or not meaningful. Providing the general public with a way to visually learn about the dynamics of cell biology may help improve public scientific literacy. Additionally, visualizations of whole-cell models can be a powerful educational tool that can be used to help grade-school children better grasp difficult biological concepts.

Efforts to create accessible whole-cell model visualizations are currently in progress. A major part of these efforts is the work done to import cell architectures into the online game Minecraft. Because of this work, users are able to walk around and inspect various components of the cell by loading whole-cell models into their Minecraft games. These whole-cell modeling visualizations are made freely available to the general public. Labs and teams with experience in designing Minecraft worlds can make valuable contributions in this area of whole-cell modeling. Additionally, although Minecraft has been a focus, other visual media outlets are also available to share whole-cell modeling results. Collaboration with graphics design and visual arts departments at UIUC may also lead to an increase in artistic style of these visualizations.

Comparison with Experimental Results

Whole-cell models and simulations are only practically useful if they produce predictions that can be tested against experimental results. Experimental results are the foundation used to construct whole-cell models. Likewise, experimental results should also be the standard by which the accuracy of whole-cell models is tested. At a minimum, whole-cell models should be able to reproduce the experimental results that were used in model construction. Ideally, whole-cell models should be able to make predictions about experimental results that have not been performed yet, and as such, were not used for model construction. Models that do not capture biological results should be refined until they can make accurate predictions. Whole-cell modeling results can also provide suggestions for future experiments.

Labs and teams with expertise in performing molecular and cellular biology wet lab experiments are critical for the progress of whole-cell modeling. Additionally, labs that can use discrepancies between model output and experimental results to refine whole-cell models are necessary to help the biological community correct any inaccurate information already published.