An interactive, three-dimensional visual structural systems biology reconstruction of a whole *E. coli* cell

Yuxuan Xie
Rohan Grover
Winnie Shi

Abstract

No complete, three-dimensional visual reconstructions of a cell the size of prokaryotes exist on a scalable pipeline to represent proteomic, metabolomic, and genomic data in an easily accessible manner. With the availability of experimental and computational information on the structures, stoichiometry, and localization of cellular components of Escherichia coli along with the development of visualization tools such as cellPACK and chromosome simulation models, we have the "ingredients" we need to put together a model for the well-studied model.

This project integrates the necessary data mentioned above into an interactive 3D, structural biology model of an E. coli cell accessible through its website. The cell model is a still model containing structural representations of the proteins, protein complexes, cell membrane, and chromosome. The pipeline for building the model begins with gathering information from databases such as EcoCyc for the proteins or tomograms for the membranes, then computations were made to solve for pieces of information that were not already available such as structures for many of the protein complexes using homology models, and ending with the integration of all of these components using cellPACK and constraints. The DNA chromosome representation for the model was shared with us by Hacker, Li, and Elcock (2017). After backend and frontend development, the model is displayed on a website where users are allowed to click and zoom in the model to learn about and immerse themselves within the insides of an E. coli cell.

This model serves as an educational tool to a wide range of users from middle school students to researchers to help further the understanding of the E. coli cell through a spatial, visual method. Clearly, there is still a lot to learn about the components and dynamics of the E.coli cell that could be included into this model; there are also other organisms that we could create this type of representation for. Hopefully, this model can be used as a stepping stone for organizing the information available on E.coli as well as a stepping stone for developing models for other organisms at the prokaryotic scale and even the eukaryotic scale.

Executive Summary

- 3D visualizations of whole cells are valuable research tools that contribute to the understanding of structural and systems biology. These reconstructions help us:
    - identify information available and lacking, driving new experiments
    - organize and compile structural data and systems simulation data visually
    - predict complex multi-network phenotypes
- There is no 3D visual reconstruction at the scale of a whole single prokaryotic cell
- There is also a lack of organized compilation of information about *Escherichia coli*, a relatively large prokaryote, with a number of well-researched proteins, protein complexes, and metabolic networks
- This project aims to provide:
    - structural detail and protein-protein interaction involved in complex formation
    - pipeline that allows for the integration of "omics" data (genomic, proteomic, metabolomics, etc.) into such reconstruction
    - pipeline for visual reconstruction of E. coli and its expansion to other organisms
    - structural, large-scale, 3D visualization and database containing the requisite proteomic, genomic, and metabolic data
    - clean, easy-to-use website for end users to interact with the model and access *E. coli* datasets
- This project acts as a proof of concept to building even larger whole cell models for eukaryotic cells and as a stepping stone for more complex whole cell visualizations that could involve dynamic simulations in the future

I. Introduction

Cellular and molecular biologists have historically expressed great interest in visualizing the dynamic and crowded environment of the cell. Early examples of this curiosity are seen in David Goodsell's detailed paintings. His works are a beautiful attempt to recreate the cell's complex structural system (Goodsell, 2016). With the aid of computer graphics and computational software, a three-dimensional visualization of the cell has inched closer to reality (Carrera & Covert, 2015). Such a visualization could eventually enable scientists to examine the macroscopic effects of cellular perturbations, accelerate advances in biotechnology, and identify ongoing gaps in knowledge for a given organism (Purcell, Jain, Karr, Covert, & Lu, 2013).

While the application of systems biology generally limits itself to the simulation of individual cellular networks, a systems biology approach to analyze a complete biological system has been developed for many years. Traditional approaches have included the reconstruction of biological networks and the mathematical modelling and simulation of systems (Bordbar, Monk, King, & Palsson, 2014), and have advanced to involve the integration of individuals' metabolomics data to create a set of personalized kinetic parameters for simulation and analysis. Recently, researchers have been working to create a systemic, quantitative visualization of a mesoscale model, which would contain three-dimensional structural information for each molecule, protein, and component of the cellular system (Im et al., 2016). In this project, we seek to build a web-accessible visual reconstruction of the *E. coli* cell, complete with metabolic networks and protein-protein interactions. Basic criteria for success include accurately displaying the environment and dynamics of the cell, as well as clearly simulating the reactions and interactions contained within (Im et al., 2016).

Unfortunately, at the moment no complete, three-dimensional visual reconstruction of a single cell exists on a scalable pipeline to represent proteomic, metabolomic, and genomic data in an easily accessible manner, preventing students, teachers, and researchers from improving education and performing simulations with such a product. Any existing visualizations of a cell are incomplete, two-dimensional, static or purely artistic in nature (Lee, Karr, & Covert, 2013). Furthermore, several methods have allowed a systemic visualization of components of the cell, but none have allowed for the accurate, large-scale visualization of a major cell for students' and researchers' uses. Consequently, it is currently difficult to query given proteins and establish their context within a whole cell due to the lack of a pipeline to integrate available information into a visual, three-dimensional model.

Development of this pipeline would be a major boon to systems biologists and molecular biologists, who would benefit from an interactive 3D model of an *E. coli* cell to be able to see and gather the structural system and protein network of the *E. coli* cell in an organized fashion. Researchers need a platform to organize and update the information they have on *E. coli*. This would also translate to discovering what information is lacking for building a more complete model. According to the Bureau of Labor Statistics (2016), there are 85,000 biologists in the

United States—while it is impossible to estimate how many could benefit from a tool like this, it is clear that making this platform accessible to the community could greatly streamline research and development for a vast number of scientists. Additionally, building a scalable platform could extend the tool's reach. Researchers need an easy and clear way to build 3D visualizations of structural systems and protein networks such as the one we propose for *E. coli*. An organized and easy-to-use pipeline for building such a model is needed for organisms as large as *E. coli* and eventually at the scale of eukaryotic cells. Broadly, researchers who study organisms other than *E. coli* such as yeast and other bacteria can use our pipeline to create a 3D visualization of the organism of their interest.

Finally, educators and students need a 3D model to introduce the cell as a system. Teachers could better explain the scenarios in which many metabolic and molecular and macromolecular reactions occur and introduce the topic of molecular biology to students more effectively. Students could gauge a deeper understanding of the cell and hopefully be inspired to do further research the science contributing this project. High school and college level teachers and students would benefit the most educationally from this model. There are 15 million high school students enrolled in the U.S. and around 20 million college students enrolled in 2017. Even though the total number of students studying a STEM-related field may be fewer, the numbers greatly reflect the impact this model can make on inspiring curiosity towards systems biology.

II. Relevant Background Information

Many obstacles lie in the way of building a whole-cell model. Macklin, Ruggero, and Covert (2014) propose seven challenges, namely "(1) experimental interrogation, (2) data curation, (3) model building and integration, (4) accelerated computation, (5) analysis and visualization, (6) model validation, and (7) collaboration and community development." Attempts to develop such a model have resulted in two schools of thought, which can be labeled the "outside in" and "inside out" visualization approaches. The former uses tomography to build a visual model of the cell and matches metabolites and macromolecules to the acquired localization data, whereas the latter assembles individual components of the cell into a complete model. Each dogma has distinct challenges associated with it, as well as a variety of existing tools and literature to address these. Universally, experimental techniques such as cryo-electron tomography and fluorescent imaging, data from proteomics and docking, and tools such as cellPACK and CHARMM-GUI have greatly pushed forward the building of a dynamic whole-cell model (Im et al., 2016).

"Outside In" Approach

Data Sources

Whole-cell imaging techniques have greatly improved, to the point where whole *E. coli* with manganese-tagged protein have been successfully imaged using single-particle x-ray diffraction (Le Gros, McDermott, & Larabell, 2005). Tomograms of cells provide a detailed example of the internal structures of their membranes and macromolecules. The most common source of tomograms is cryo-electron tomography, in which the whole cell or thin slices are vitrified

(rapidly frozen) and then imaged using transmission electron microscopy. Cryo-electron tomography allows for remarkably detailed images, with resolutions as low as 4 nm, to be taken of bacteria cells (Milne & Subramanium, 2009). The development of fluorescent labeling has made the identification, three-dimensional visualization, and localization of many particles such as *E. coli* 100S ribosomes and the endomembrane system of *G. obsuriglobus* possible (Ortiz et al., 2010; Santarella-Mellwig et al., 2013). The "outside in" approach solves an issue brought up by Betts and Russell (2007) with respect to the "inside out" method. Even with complete proteomic data, including copy numbers and cellular locations, that proteomic data used in the latter does not necessarily indicate what proportion of each protein resides in a particular location of the cell. The use of tomography is their proposed solution to this problem, as it bridges a gap between x-ray diffraction and fluorescence tagging in terms of its high resolution and accurate localization data. The images from this technique provide valuable information and confirmation regarding the stoichiometry, structure, and location of molecules and proteins within the cell—despite limitations in capturing the cell in a dynamic state. Other limitations of tomography are its inability to capture all of the proteins, electron densities, and molecules within the cell due to constraints in labeling the components and the possibility of beam damage given thinner sections to be imaged (Keskin, Tuncbag, & Gursoy, 2016; Milne & Subramanium, 2009). Despite these restrictions, tomograms provide a valuable experimental foundation and useful validation for other approaches discussed later.

Model Visualization

The tomograms are often converted to three-dimensional models through programs such as IMOD. IMOD is a program described by Kremer, Marstronarde, and Mcintosh (1996) that is able to read the image stacks of the electron microscopy, align the images, and convert the tomograms into readable data using intensity spikes. Alternate methods of image analysis include the use of techniques such as thresholding, averaging, machine learning, and statistical approaches (Le Gros, McDermott, & Larabell, 2005; Milne & Subramanium, 2009; Ortiz et al., 2010).

"Inside Out" Approach

Information Pipeline

Genome-scale models (GEMs) are used for computational studies of metabolism, protein synthesis, and transcriptional regulation based on a specific organism (Bordbar, Monk, King, & Palsson, 2014). Recently, structural information has been mapped to those models, which are now called genome-scale models with protein structures (GEM-PROs) and provide a new view of systems biology (Brunk et al., 2016). This pipeline has expanded knowledge of biochemical network properties and can help make predictions on whole-cell phenotypes. This comprehensive information can only be visualized in two dimensions, so it is necessary to create a pipeline that generates three-dimensional models using GEM-PRO output data. One way to generate these models is by gathering the data for the system and putting it into a single "recipe" file (will be discussed further in the model assembly) and using this recipe to generate a virtual model through a computational engine with built-in assembly algorithms.

Data Sources

Protein abundance—determining how much of each species is present in the cell—can be determined using existing proteomics datasets. Betts and Russell (2007) describe the fundamental process of incorporating proteomics data into a whole-cell model. This information will also be incorporated as a major part of the "recipe" to determine how many copies of a component are placed in the model. In addition, building a three-dimensional model requires structural information for every molecule or protein in the cell system. The Protein Data Bank (PDB) provides downloadable three-dimensional protein structure files in the PDB format, which are required for the assembly process.

Protein Complex Representation

In cell systems, proteins usually do not act alone: they interact with each other and form protein complexes to carry out certain functions (Spirin & Mirny, 2003). However, a large proportion of known proteins present in *E. coli* do not have structural information available in the PDB, so finding a way to represent them is necessary. The EcoCyc database described by Karp and Riley (1999) is a resource which compiles polypeptide and protein complex data for *E. coli* in a downloadable way, potentially making large datasets describing complexes more accessible.

Additionally, predictive methods can assemble protein components into a whole complex. Krissinel and Henrick (2007) developed a method called PISA (Protein Interfaces, Surfaces and Assemblies) to infer protein interactions based on data found in the PDB. Open source softwares such as the Integrative Modeling Platform (IMP) combine electron microscopy data with proteomics data and protein structural information to model the complex structure based on their scoring mechanisms (Russel et al., 2012). The downsides of these softwares are that they work best on much smaller scales than that of a whole E. coli cell and require constraints such as cross-linking data for their modeling. Moreover, they are very computationally expensive. Another method used to map complex information is the integration of protein-protein interaction (PPI) datasets (Rajagopala et al., 2014). Abundant PPI datasets exist, containing both experimental and predicted data (Szklarczyk et al., 2016). Each binary interaction of proteins can be mapped onto the metabolic network, and the protein complex that carries out a reaction can be represented by individual components interconnected by PPI without having to be assembled together. This saves huge amounts of time, since all of the requisite data is easily obtained. Most importantly, the expansion of metabolic networks to include PPI results in emergent properties (Cusick, 2005).

In order to represent complex structures as completely as possible, a workflow is devised to generate pseudo-complex structures as shown in Figure 1. This method is used to generate structure files for all the known reaction complexes for packing purposes until a better solution is determined. First, accurate stoichiometry information for each complex is curated from literature. Structure files for each individual subunit in a complex are set as the best representative structures by GEM-PRO. The residue with minimum depth is found by using Biopython and the unique id and its last atom are stored along with structures for later use (Cock et al., 2009). For complexes with more than 3 subunits, the subunit with the least

stoichiometry coefficient is set as the “anchor”. Minimum depths then need to be calculated for each non-anchor subunit. For example, for the protein complex 'ACETYL-COA-CARBOXYLMULTI-CPLX', there are three subunits, and the stoichiometry is {"2w71": 1, "1bdo_bio1": 1, "2f9y_bio1": 1}, meaning one 2W71, one 1BDO, and one 2F9Y structure can be used to represent the complex. These IDs represent the PDB ID for the subunits. In this case, since they all have stoichiometric coefficients equal to 1, the anchor is chosen at random. Two minimum depths are calculated for the anchor and one minimum depth is calculated for the remaining non-anchor subunits. With the residue ID and atoms with the minimum depths, subunits are fused together by creating stable bonds between atoms using Pymol (Yuan, Chan, & Hu, 2017). Each non-anchor subunits is fused onto the anchor subunit.

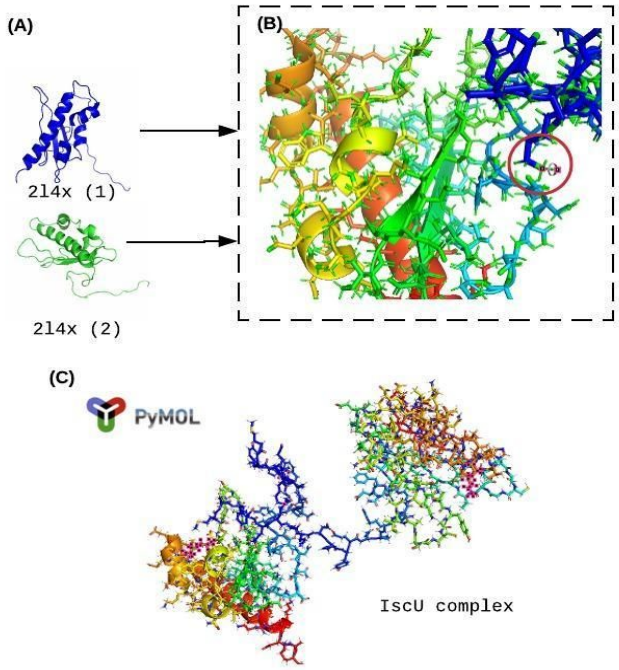


Figure 1. This describes the procedure to create dummy complexes for IscU, which contains two subunits. (A) Correct structure files for each subunit are chosen by running GEM-PRO and downloading them from the PDB. (B) For each subunit, minimum depth is detected using Biopython. (C) Using Pymol, atoms at the minimum depths are fused together to create stable bonds, thus forming a single complex structure.

Model Assembly and Visualization

Once all the data is gathered, it is used to generate the whole-cell visualization, enabling users to interact with it and query valuable information from it. A software called cellPACK has proven capable of generating large-scale biological models such as the human immunodeficiency virus (Johnson et al., 2014). Klein et al. (2017) describe cellPACK as a computationally intensive method to automatically arrange biological molecules into desired partitions, as opposed to manually placing molecules in a 3D visualization software. When

supercomputer power is available, cellPACK allows data to be organized relatively quickly into a three-dimensional mesh of a cell and displayed using the cellVIEW tool. It applies both random packing algorithms and user-defined packing methods to automatically pack the model, taking the input "recipe", a file which contains information for all the molecular components that will be present in the model. It may be possible to store structural information in the Macromolecular Transmission Format (MMTF), which loads quicker than traditional PDB files. NGL Viewer, a webGL-based application to visualize macromolecules, may also be able to visualize such a large-scale model (Rose & Hildebrand, 2015).

Cells consist of other macromolecules, metabolites, and lipids. The tool Membrane Builder from CHARMM-GUI utilizes the highly mobile membrane-mimetic (HMMM) model to simulate the lipid bilayer surrounding the membrane proteins (Im et al., 2016). C-SAC (chromosome self-avoiding chain) uses a computation strategy similar to protein structure determination to predict the folding of chromatin within the cell (Im et al., 2016). Beyond these, there are numerous molecules and intermolecular interactions between proteins, nucleic acids, ions, and metabolites yet to be explored. Due to the difference in scale between the lipid membrane and these macromolecules, as well as the extensive computations required to simulate their interactions, our proposed whole-cell model should exclude these molecules until further time, energy, and computational power have been devoted to their study.

Database Construction

A database is necessary if users wish to query information in the model for educational or research purposes. SQL (structured query language) is the traditional way to build a database. A graph database is a relatively new way to store relational data, typically as nodes and relationships (pathways). One prominent example of a graph database is neo4j, a frontend web interface that can store, query, and change data directly through a query language called Cypher (Webber, 2012).

Summary

Rather than being a setback, the lack of information available to build a complete whole-cell model illuminates the areas where further investigation is necessary to thoroughly understand the cell as a dynamic system. An interactive cell model of *E. coli*, hopefully, not only serves as a tool for exploratory purposes but can also develop into an application for predicting systemic responses of the cell. The model we propose will use an "inside out" approach paired with some techniques from the "outside in" model. Eventually, the simulation can deepen humanity's understanding of the cell and help with precise prediction of the effects of new drugs, chemicals, and other modifications (Carrera & Covert, 2015).

III. Design Goals and Constraints

Goals

A primary goal of the project was the development of a website to visualize, in three dimensions, the structures of the *E. coli* cell. Subgoals were that the website must be interactive, quick to load, and capable of allowing users to access the data used. Additionally, it was to be

both as precise and complex as possible, but did not need to be a dynamic simulation as yet. Creating a dynamic simulation would have taken too much time and computational power, and this project is currently being worked on by another team in the Systems Biology Research Group. At best, we hoped to lay the foundation for future integration of dynamic kinetic data. Another overarching goal was the ability to integrate the methods for development of this website into a pipeline for future users to replicate our work.

The reconstruction aimed to provide a three-dimensional visual reconstruction of *E. coli* by assembling atomic-level components into a cellular-scale model. This would permit researchers to look at the *E. coli* cell as a whole and summarize the known information within the proteome. This goal was of utmost importance; we assigned it a weight of 40% for the later comparison of design alternatives. Success could be measured not by the completeness of the data, but by the completeness of its annotations; that is, if certain structures were included with less certainty than others, they were to be clearly indicated as such.

A subgoal is to maximize accuracy of the representation. An accurate reconstruction depends on reliable, accurate sourcing for the stoichiometry, localization, interactions, and structures contained by proteomics datasets. The model generated was to be as accurate as possible so that people could rely on its outputs—consequently, we assigned this goal a weight of 25%. This could be measured by performing quality analysis on both experimentally and computationally determined data and matching the conditions used for data extraction. Furthermore, review of the data sources and databases used to compile information could be used to develop metrics for success.

Next, it was determined to be important to be able to convert raw data into an appropriate format for three-dimensional visualization. This was because it is important to be able to easily update the model, for example, if new data based on new experimental results is brought to light. Additionally, for other researchers to scale the platform to fit their needs, they needed to be able to insert their data into a pipeline to generate the three-dimensional model. This goal was weighted with 10% importance; its success could be measured by developing a model from new data, as well as by appending new data to the existing data and seeing if a model can successfully be built.

Finally, the speed of the model was considered essential to its success. Does the website load quickly? Will it operate on older hardware? To avoid making the tool cumbersome or unusable, these data points could be tested and the product could be optimized accordingly. We also sought to present users with a simple web interface that allows them to interact with the model and easily find the proteomics data that they need. The interface was to be intuitive and user-friendly enough that people without scientific backgrounds or technical expertise could operate it. This could be measured through user testing. Testers could have identified whether the website is aesthetically pleasing, functional, and attractive. Additionally, they could identify whether or not it is easy to tell what information is being represented. We assigned this goal a weight of 15%, as design is important, but not the investigators' priority.

Constraints

There were several constraints that could prevent completion of some of these goals. For one, it is impossible to build a perfectly accurate simulation due to the nature of the available data. Not all proteins have complete sets of experimental or computational data available; many were missing stoichiometric, interactional, or structural information. Consequently, their accuracy was compromised and the unknown variables had to be appropriately estimated or simulated. Additionally, time constraints prevented being able to generate a complete, dynamic simulation or a polished, professional website. With the rapid pace of senior design, it was difficult to form a web page with all the desired functionality and interactions. That time was better dedicated to finding and standardizing high-quality data for the simulation.

Cost was another consideration. While the software packages in use (NGL Viewer, cellPACK, ssbio, IMOD) are open source, computational costs are not insignificant. Building the whole cell model requires high computational power to pack every component into the right space, considering the massive number of components. While the Scripps Research Institute supercomputer absorbed most costs of the cellPACK algorithm in this instance, using the San Diego Supercomputer Center if necessary in the future would cost $0.025/SU for our purposes, where one SU (service unit) represents one core-hour of computing time. Luckily, this cost is also trivial for a single simulation. More importantly, the computation itself takes time; visualizing each cellPACK "recipe" takes about one to two extra weeks.

Last, the performance specifications of various devices meant that one of our goals may have worked in opposition to the rest. It was essential to balance function and functionality; in other words, if the site was feature-laden but inoperable on lower-end devices, it was not achieving its goals. We worked to make massive amounts of data quick to load and a database efficient to query for specific data points.

IV. Design Alternatives and Analysis

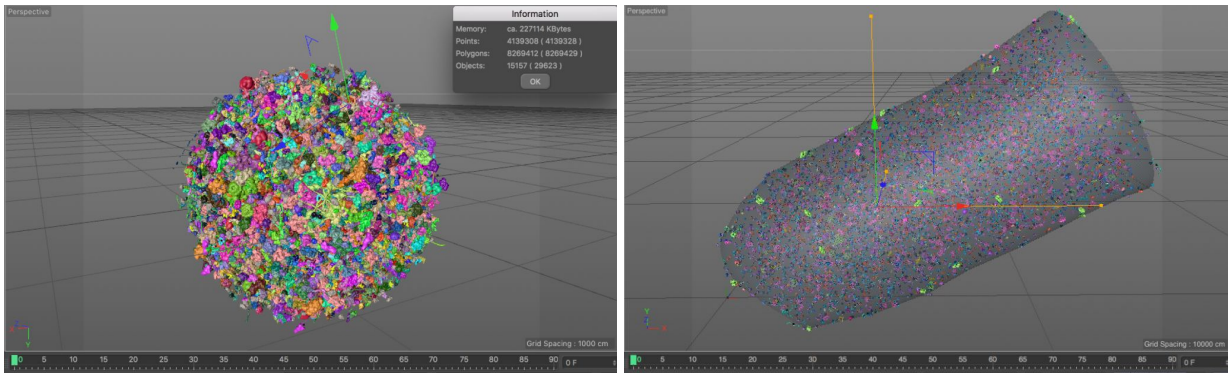
Alternative 1: cellPACK Method

This was the method investigated by the team. It took an "inside-out" approach to developing a cell-scale visual reconstruction of *E. coli*. The intent was to take existing metabolomic, proteomic, and genomic data and match it to software and tomograms to generate a three-dimensional, structurally and systematically accurate visualization of an *E. coli* cell and its components. The simulation is then transferred into a web page where it can be manipulated by end users. Numerous pitfalls exist with such an approach, but the main bottleneck was with the accuracy and quality of the data available for modeling. However, it had the advantage of easily scaling for different models and future updates to the data.

Implementation Plan

The implementation for this model began by scraping all the necessary data to start to derive an accurate model. In addition to finding protein abundances, this required awareness of individual

proteins' structures and interactions. Localization data could have been obtained for proteins by comparing their visual appearances to their appearances in tomograms, but was not due to a lack of detail in existing tromograms. Thermodynamic constraints could have been integrated into the modeling process to obtain the optimal state of the cell, as cellPACK allows a virtually limitless set of constraints. Then, the data wasa crunched by cellPACK, which outputted a three-dimensional model for each region of the cell based on the shapes and charges of molecules. The first prototype model of *E. coli*, built with the aid of our collaborators in the Scripps Research Institute molecular graphics laboratory led by Dr. Arthur Olson, is shown in Figure 2. The data from this model corresponded to individual protein shapes from the Protein Data Bank, letting each 3D shape be rendered individually as an MMTF or PLY file. These file formats are compatible with the software NGL Viewer for powerful, lightweight three-dimensional rendering of biological molecules.



(a) (b)

Figure 2. (a) A cytosol model that contains all the available structures and concentrations of cytoplasmic proteins with a sphere constraint applied. (b) The shape of the *E. coli* inner membrane, obtained from tomogram data was applied as a constraint to pack the model.

Alternative 2: Outside-In Tomography Plan

A potential alternative to the atomic model we designed could have been to take a top-down or "outside-in" approach where we could directly visualize—using 3D tomographic data—the membranes and structures present in the images. We would have to make sure that we are visualizing everything we need to see including proteins and macromolecules. With multiple tomograms, we would have multiple variations of the *E. coli* and we could potentially find an average or middle ground between the tomographic models and combine the separate images into one model with averaging or other techniques. Potential caveats would be the amount of data available and the amount of data we would be able to collect from the tomograms. Additionally, as with the proposed design alternative, it will be necessary to devise a set of standard conditions to image the *E. coli* cell at. An upside to this approach would be the accuracy relative to real conditions within the cell because the images are directly from the cell. There would be more confidence in the visualization from this model since the structures are what is imaged and not based on purely prediction data or comparison data with similar structures.

Implementation Plan

We would start out by taking tomograms of *E. coli* cells and then convert them to 3D models using special software. The software would read the image stacks, align the images, and convert the tomograms into usable data using intensity spikes. The results of such a process are displayed in Figure 3. Then, the tomogram data would be matched to protein and other macromolecule abundances (in the event of conflict, the tomogram information will override, as this alternative prefers and prioritizes tomogram data). The proteins and other structures would be rendered in three dimensions as discussed in the previous model and made available to end users online.

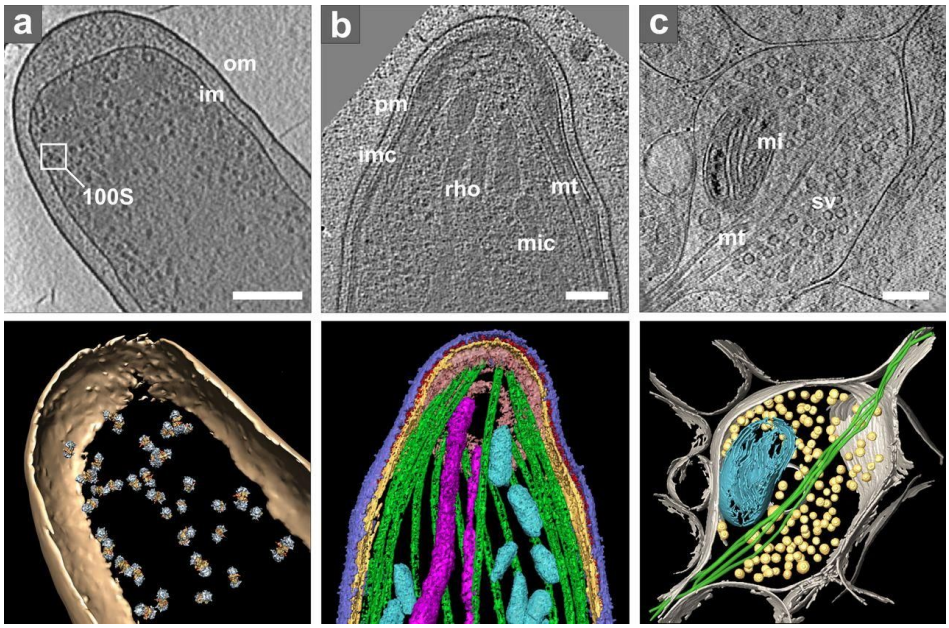


Figure 3. Tomogram slices in the top row are converted to isosurface representations shown in the bottom row to develop a model (Lučić et al., 2013).

Alternative 3: Two-Dimensional Simulation

This alternative was inspired by the online whole-cell *Mycoplasma genitalium* model depicted by Lee, Karr, and Covert (2013). That simulation uses six panels to depict the activity of a single *Mycoplasma* sample over time. This proposed alternative would use a variety of panels as well, but unlike the existing model, it would not have dynamic simulation data. However, it would have almost all of the data we seek to incorporate into our existing model, such as protein structures and networks, metabolic interactions, and genetic data. The advantages for such a model would be that it would be significantly easier to construct using existing tools, and the use of specialized software such as cellPACK and NGL Viewer would not be necessary. Its disadvantages center on the fact that localization data would be lost, meaning valuable information about intracellular interactions would not be accessible. (We see that this is a problem with the model above, where the metabolic networks depicted are not linked to any sort of locational data.)

Implementation Plan

For the most part, the steps to acquire data for visualization would be similar. In fact, tomograms would still be useful to generate a two-dimensional visualization with accurate organelle placement and general protein localization. However, because the majority of interactions would need to be generated in three dimensions, it is more important to display complexation on a graph or chart. Because these are static simulations, the data can even be displayed as a PNG, but it may be preferable to load some information through a Chart.js script. This would allow it to be annotated appropriately. Of course, ssbio and Escher will still be essential for simulating and mapping metabolic networks. Proteomic and interactomic data will be scraped from the PDB and EcoCyc, but as we will no longer have a three-dimensional depiction, PLY files and three-dimensional protein structures will not be necessary.

Alternative 4: Dynamic Simulation

This proposal is similar to both the first and second models. However, the priority would not have been on a static 3D simulation as with those, but with a dynamic rendering of the networks in the entire cell. Using known equilibrium constants for every single macromolecule, each reaction mechanism will need to be reconstructed and the subsequent reaction rate constants determined *in silico*. Once complete models have been generated, the networks will be simulated at steady state with the appropriate initial conditions. It is unclear how a live network would be rendered—this would be a major challenge with such an approach. However, should it be possible, this reconstruction would be a major coup for researchers.

Table 1. Decision Matrix weighing the advantages and shortcomings of each alternative

Alternative: Goal	**Alternative 1: cellPACK Method**	**Alternative 2: Outside-In Approach**	**Alternative 3: 2D Simulation**	**Alternative 4: Dynamic Simulation**
Goal 1: To build a 3D visual reconstruction of an *E. coli* cell (40%)	*Score: 80* *Weighted: 32* The cellPACK algorithm may run into issues with the complexities of a whole cell, but this is still one of the accepted ways to model entire cells.	*Score: 80* *Weighted: 32* The outside-in approach is the other of two accepted ways to generate a whole-cell model. Success is not guaranteed in either method.	*Score: 0* *Weighted: 0* While the simulation does qualify as a visual reconstruction, there is no chance it will ever be manipulable in 3D.	*Score: 30* *Weighted: 12* The dynamic simulation builds on the first two approaches, and adds more data making it a more complete reconstruction. However, visualizing networks over time in 3D will be so challenging that we are

				unlikely to see success.
Goal 2: A representation containing reliable and accurate data on the stoichiometry, localization, interaction, and structure of the omics data (25%)	*Score: 70* *Weighted: 17.5* Existing datasets for proteins, genes and metabolites are still incomplete, but for the most part this approach should have all the critical data conveniently compiled.	*Score: 70* *Weighted: 17.5* The end result of this will have similar information to the first approach. The main difference will be with the quality of the data in the tomogram vs. in omics sets: realistically, both will be kind of incomplete.	*Score: 65* *Weighted: 16.25* It will be difficult to render localization data in two dimensions, so this info will likely be lost. However, other information can be included that wouldn't usually render in a sole 3D model.	*Score: 50* *Weighted: 12.5* Kinetic parameters are difficult to track down and kinetomics datasets are virtually nonexistent. While *E. coli* networks have more data, it will still not be enough to generate a complete model.
Goal 3: Present data in format that is easy to update with any changes or additions needed for the model (10%)	*Score: 40* *Weighted: 4* Changes or updated information will require cellPACK to be run again with the new surfaces or stoichiometries.	*Score: 20* *Weighted: 2* Cryo-electron tomograms will be challenging to obtain, and the whole model will have to be recreated for even the most minor updates.	*Score: 70* *Weighted: 7* Each of the individual panes of such a simulation could likely be rendered separately, making updates relatively easy.	*Score: 30* *Weighted: 5* Comparable to the first model, but also having to recompile every kinetic network when new information is available.
Goal 4: Short model loading speed for the amount of data to be presented on a functional, attractive website (15%)	*Score: 80* *Weighted: 12* The data can be dropped as is into NGL Viewer or compressed into the MMTF format, making it compact and relatively quick to load.	*Score: 80* *Weighted: 12* The data would be handled the same way in this method and the previously discussed one.	*Score: 90* *Weighted: 13.5* Without surface renderings of each protein, loading the data will be a breeze. The interface will likely be intimidating for users, but also easier to operate.	*Score: 50* *Weighted: 7.5* This model has more data and requires far more server or computing power to generate and run, reducing the likelihood of success.

Weighted Totals	**65.5/100**	**63.5/100**	**36.75/100**	**37/100**

V. Design Solution

Based on the above analysis, an inside-out approach based on cellPACK was determined to be the optimal design solution to achieve our desired goal. As mentioned, we created an interactive, three-dimensional visual reconstruction of an *E. coli* cell by integrating structural data, localization data from tomograms, and macromolecular omics to create an accurate representation of the cell. Users are able to interact with the model by zooming in and out, rotating and orbiting around various axes, selecting particular structures, and exploring various layers of the model. Advanced options include fast queries for specific information (structures, subsystems, interactions and reactions) and the ability to generate users' own three-dimensional models through the pipeline.

Project Breakdown

Project 1: Create pipeline to generate "recipe"

The primary goal of this subproject was to create a pipeline that generates a recipe that the cellPACK algorithm can read. This entailed gathering *E. coli* proteomics data and mapping it to the genetic model we used. The protein complex data was updated using data from the EcoCyc and UniProt databases; for membrane proteins, the data was provided by the Orientations of Proteins in Membranes (OPM) database and the tool TMHMM, which uses hidden Markov model (a machine learning algorithm) to predict transmembrane helices of membrane proteins (Lomize, Lomize, Pogozheva, & Mosberg, 2006; Krogh, Larsson, von Heijme, & Sonnhammer, 2001). For cytoplasmic proteins and membrane proteins without detailed data, the SWISS-MODEL homology-modeling tool was found to be extraordinarily useful to predict morphologies (Biasini et al., 2014).

Figure 4 shows the two separate pipelines used to generate the cellPACK "recipe". The pipeline at the top is the structural systems biology (ssbio) pipeline for calculating and analyzing protein structural data with high quality and extensive annotations by using our open source software (Mih et al., 2018). The bottom pipeline utilizes the ME-model, or macromolecular expression model, a mathematical model developed by Palsson lab to simulate the protein complexes' formation with experimentally determined protein concentrations as constraints (O'Brien, Lerman, Chang, Hyduke, & Palsson, 2014) . These integrated high-throughput pipelines enabled us to generate data that accurately modeled *E. coli* at the genome scale and can be potentially applied to other organisms.

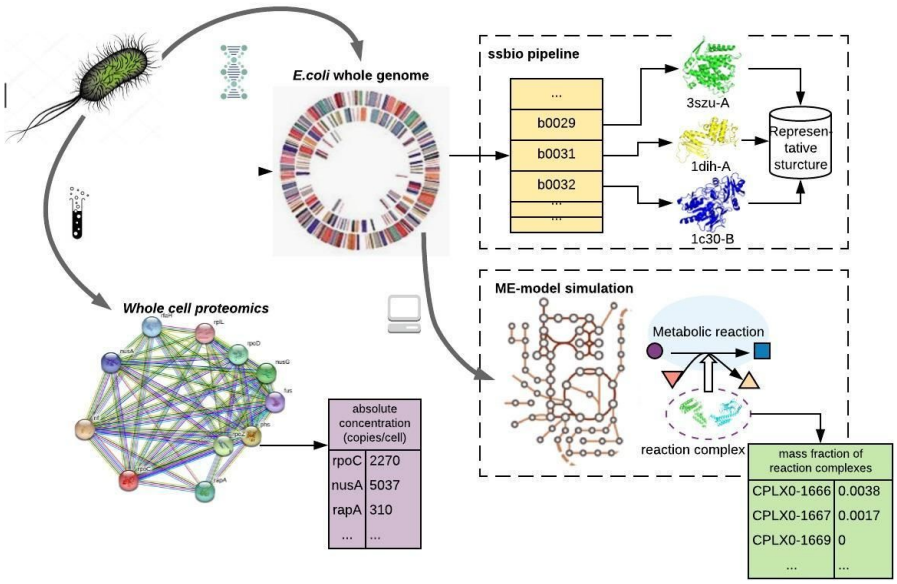


Figure 4. Pipelines to generate structural data with annotations.

Complex concentrations or copy numbers of complexes per cell were estimated using mass fractions from the ME model because there is no experimental data on complex copy numbers. The copy numbers were calculated with the assumption that there is a "limiting" complex and that the number of complexes were to be maximized. In other words, the mass fraction gave us the ratio between the complexes and we wanted to determine how to best maximize the number of complexes while maintaining the ratio. However, because some of the complexes were severely limiting such that less than 1% of proteins would be in complexes, another complex would be used as the "limit" instead. The algorithm for this calculation could be found on the project GitHub.

Project 2: Tomogram Analysis for* E. coli *Volume

Tomograms of whole *E. coli* cells must be analyzed to determine the protein localizations and membrane morphology. Each slice will be compiled to form a three-dimensional model according to existing protocols for tomographic analysis, such as those described by Delgado, Martínez, López-Iglesias, and Mercadé (2015). Due to the lack of full cell tomograms, an *E. coli* cell was modeled in the shape of a “pill” with volumetric dimensions described by Volkmer and Heinemann (2011). The tomogram model and derived model and “pill” were merged using Blender software to generate the final model as shown in Figure 5. Because the individual membrane proteins need to sit at a vertex of the mesh, the number of vertices in the mesh model was increased using MeshLab software (Cignoni et al., 2008).

Figure 5. Stills of developing membrane model with (A) the tomogram-generated model, (B) a 'pill' model, and (C) the final combined membrane model.

Project 3: cellPACK Models Displayed using NGL Viewer

The cellPACK output needed to be configured into its individual components for display using NGL Viewer. This process was automated and tested. Further, the capacity of NGL Viewer to render potentially thousands of components simultaneously and quickly on the webpage of ordinary laptops was tested. If possible, the program can be further optimized to load the model in the most resource-efficient manner. Presently, NGL Viewer is able to load large models but continues to have difficulty loading them quickly and interacting with components

Project 4: Web Interface Development

The interface for displaying the model and interactions was prototyped in Figma. The interface needs to be clean, aesthetically pleasing, and simple to understand and navigate for all levels of users. A database to link structures to their attributes and allow users to search was built and details are in the next section. The interface was designed and the final product can be seen in Figure 6.

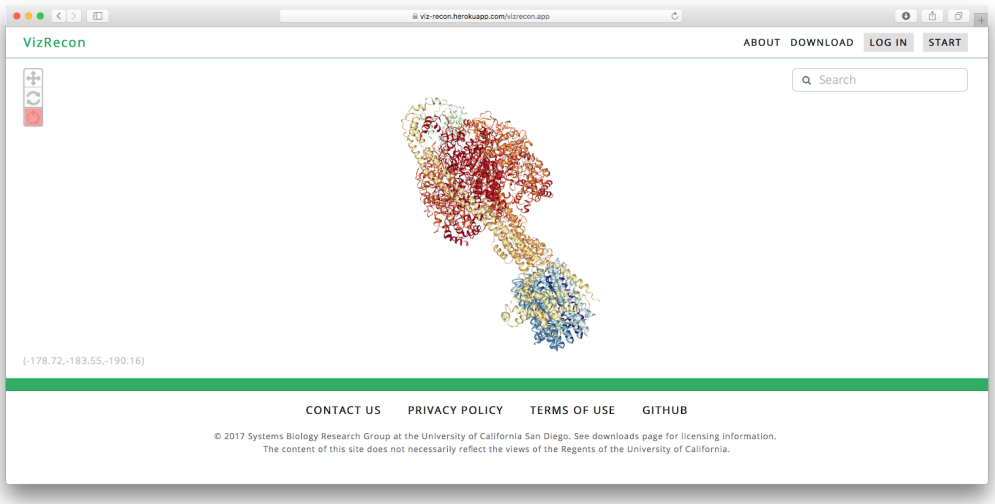


Figure 6. The web interface with search for proteins involved in the model and toolbar allowing users to zoom and interact with the protein.

Project 5: Database Construction and Application Integrations

A proposed database structure is shown in Figure 7. Necessary steps were done to construct the SQL (Postgres) database and uploaded to the online service (AWS). Pipeline to update the database has also been developed to insure the convenience of data modifications. Web applications is able to access the database and fetch information dynamically for the purpose of human-computer interaction.

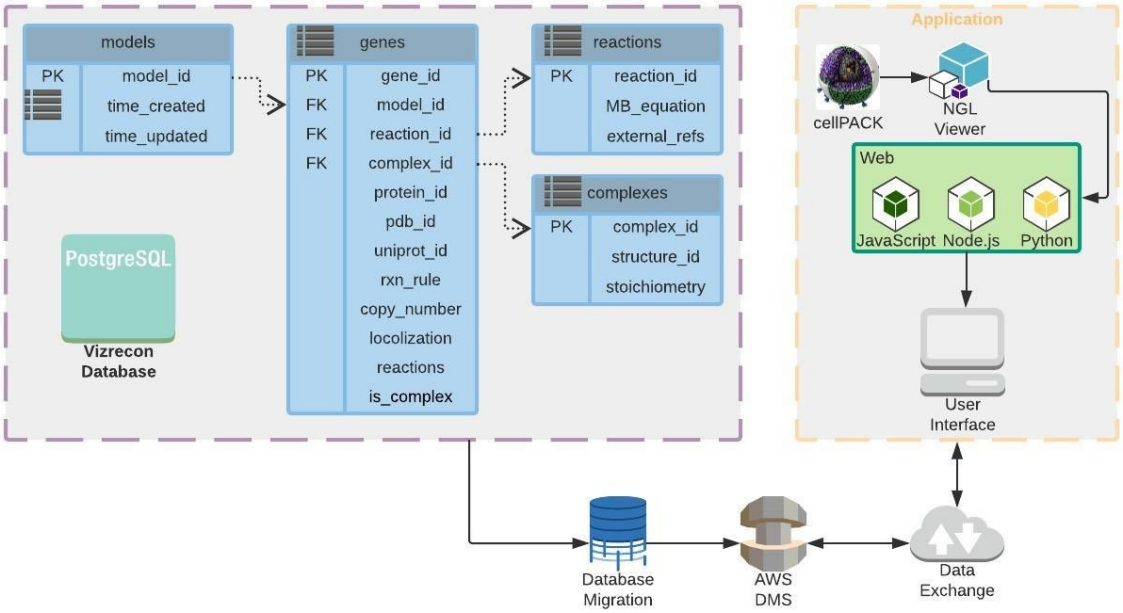


Figure 7. A schematic of the backend and frontend workflow for the displaying and accessing the model data through the website.

VI. Changes from Design Proposal

The original design proposal did not include a detailed chromosome and featured little to no protein complex representation. Hacker, Li, and Elcock shared their 500 bead per base pair chromosome model with us (2017), allowing us to incorporate an accurate, three-dimensional structural representation of the cell's genetic information. While some protein complexes could be represented using researched structures structures and homology models, unknown proteins and complexes were eventually represented with spheres of equal mass. Since complex concentrations remained unknown, ME-model complex mass fractions were used to estimate and predict the best number of complexes. This estimation was made based on the idea that there is a limiting protein and that the number of complexes are to be maximized. Other changes pertain to the model representation from the design proposal. Due to the large amount of data to be displayed on the website to be displayed, the model will first appear to the user as a series of points. The membrane and chromosome will be loaded, but the proteins and complexes will be appear as individual points rather than detailed surface structures. The user can still

access the protein structures by clicking on the points, thereby loading the protein structure. While representing the proteins as such may take away from the overall 'experience' of interacting with the cell, the sacrifice is necessary to maintain easy access to the whole cell model. This feature may be adjusted later on to enable culling for when users zoom close enough—when not as many proteins need to be loaded on the screen, higher-resolution structures can be displayed.

VII. Parts, Resources, and Costs

Below is a simple list of parts, software, and costs that were used.

Parts, Software, or Resources	Details	Costs
Individual laptops or computers	Used to develop the model and analysis	Free (already paid for)
Protein databases	Protein structures and protein complex information provided by PDB, EcoCyc, and GEM-PRO databases Unknown structures predicted using SWISS-MODEL tool Membrane structural data acquired from the OPM, TMHMM, and UniProt knowledge bases	Free
Protein interaction data	Data collected from EcoCyc and Biosystems Laboratory	Free
cellPACK	Software to pack and build the 3D models developed by researchers at Scripps	Free
IMOD	Software for tomogram analysis developed by University of Colorado in Boulder	Free
NGL Viewer	Used to display our model on the website	Free
Tomograms	Will be collected from various supplemental images from different papers and tomogram databases	Free
PostgresQL database	Used to store structural, stoichiometric, localization, PPI, and metabolic data	Free
Python	Used to automate processes and collect data	Free
Heroku	Used for website and server hosting	Free
Cinema4D (C4D), Blender, MeshLab	3D model and visualization softwares used for building membrane model and packing	Free (C4D was free with institutional licence)

(OPTIONAL) Supercomputer allocation	As of now, computations for the model were all done on personal computers or gratis at the TSRI supercomputer. The price listed is for San Diego Supercomputer Center service units. We used 0 SU.	$0.025/SU

VIII. Evaluation and Testing

The objective of this project was to develop a 3D structural visualization of an *E. coli* cell. The model was to be evaluated on its accuracy relative to the stoichiometries, structures, and localizations established in the most current experimental research and resources. Because this model was built upon resource data used to establish accuracy, evaluation of this model depended heavily on how well the predicted parts of the model adhere to the information available. However, much of the information of an *E. coli* cell is unknown; thus the evaluation of the concentration of protein complexes and the evaluation of the complex structures cannot be confirmed until further experimental results are published. Testing was done by comparing the model with the data used to build the model.

Based on this, one way to evaluate the quality of the derived model is by considering the quality of the structural data sources. For the cellPACK recipe, many of the structures used have been experimentally confirmed and stored in the Protein Data Bank; however, a number were a mix of homology models and experimental structures, and more were completely simulated. Work remains to improve the accuracy of the SWISS-MODEL pipeline (Biasini et al., 2014), so the more experimental structures that are used, the more accurate the whole cell reconstruction must be. There were 4,836,173 structures included in the final model for packing; these represented 2,159 unique proteins and protein complexes. However, 5,734 structures were collected and annotated for potential inclusion. Of these, 1829 (or 31.89%) were experimentally determined, with 3647 (63.60%) being homology models, and 85 (1.48%) being some mix of the two. For the over 4.8 million structures included, however, 3,443,327 or 70.68% were experimentally determined, meaning that the structures used in homology modeling were not necessarily significant components of the proteome. Future visualizations could color-code structures by annotation quality, making the model completeness easily visually understood.

Source	**Count**
Experimental	3443327
Homology	1425764
Mix of both	2245
Unknown	156

Table 2. The data sources for each of the structures eventually depicted in the packing algorithm. A few structures were manually curated—of these, 140 were for a single species.

The membrane model could be more accurate than the current model. The ratio of the periplasm volume to the whole cell volume for the representation was 19% whereas experimentally from literature, the value was closer to 8% (Schmidt et al., 2016). The E. coli cell model is approximately 2.5 μm by 1.2 μm wide. The general width of the periplasm around the middle of the cell (DPB) is about 30-25 nm and the widths at the ends, PPC1 and PPC2, are 260nm and 80nm, respectively. The experimental dimensions used were from Schmidt et al. (2016) and Volkmer & Heinemann (2011) for E. coli MG1655 in LB media.

The website that displays the model and allows users to access the data used to create the model will be evaluated in its speed in loading and ease of usage. Speed can be measured and tested on various computers. Currently, without the model, the final loading speed is not an issue for the web interface. Users from graduate students to teachers to the general public will be questioned to help improve the website as well as evaluate how well various users interact with website features once the model is integrated into the website.

IX. Environmental, Social, Ethical, Health & Safety Issues

Ethical Issues

The main ethical concern would be to give credit to the information and data sources used, the people who designed the software tools (NGL viewer, cellPACK, etc.), and the people who have helped us put together this project. The software products used were all open source.

Environmental, Social, Health, and Safety Issues

Due to the nature of this project, there were no health, safety, and environmental concerns in regards to chemicals, biologics, electronics, mechanics. There was the concern in regards to making sure that the data available on our website is secure for the users who are interested to download. Because the website does not require the user to submit any private or sensitive information, we should be safe in those regards.

X. Discussions and Conclusions

Our project has laid the groundwork for creating models at the scale of an *E. coli* cell. As mentioned before, there are no models with the scale or ambition of this one. The pipeline and general tools needed for generating the 'recipe' have been collected and organized such that future users and developers can utilize the information for research or for further development of other models. This model was built from the atomic level to the cellular level and the 'recipe' provides atomic scale details of the proteins to be integrated with the whole *E. coli* cell. The final model is a visual manifestation of the link between structural information of the proteome and metabolomic data. The pipeline and all of the relevant information can be found on the project website and GitHub.

The components of the model have been curated or estimated based on the best available information and most current experimental results. The reconstruction is a proteome-focused

representation and we have collected a relatively comprehensive list of proteome data. The proteome concentrations and individual protein structures, including those of protein complexes and membrane proteins, are at the highest level of detail and accuracy that could be achieved with the information provided to us at this time. As predicted in the design constraints, the development of a dynamic simulation was not feasible for this visualization considering the amount of computational power and scope of this project.

Due to the lack of full cell tomograms available for use, the membrane model had to estimated based on experimental dimensions from literature. The accuracy of membrane models could be increased with the presence of better experimental data as well as collaborations with other research groups. The chromosomal models could also improved even though the original design did not include a chromosomal model. There is a lot of research in development of a model for *E. coli* chromosome as illustrated by the nucleotide resolution model and lattice model (Hacker, Li, & Elcock, 2017; Goodsell, Autin, & Olsen, 2018). Otherwise, the model, even pending final packing, is a good foundation for future cellular systems model construction.

Another important feature of this project was making the model available on a website for users to interact with and immerse themselves in the environment of the cell. The database and database pipeline, as previously mentioned, has been made available for users. Currently, a great challenge remains in serving this huge model to the website. NGL viewer and many other viewers are built to display individual proteins. On the website, users are able to interact with specific individual proteins, but in order to display the model it was determined either the model or the means of display needed to be simplified, if not both. Currently, we are considering displaying the elements within the cell as spheres or applying culling to the model view; however, displaying the proteins as spheres does take away from getting an accurate feel of the cellular environment. Nonetheless, a website of this project was developed with Heroku and Node.js where users can easily search for the information in the database, interact with individual proteins, and learn about how to develop their own projects.

Looking at the current status of the model, a lot of improvements and future projects could be implemented, several of which are actively under consideration. The final packing of the cell and the display of the model on the website are currently the most significant obstacles for the project. Other than the final packing and NGL optimization, we have managed to achieve many of the primary objectives for this reconstruction.

XI. Recommendations and Future Directions

As mentioned in the conclusion, the most apparent remaining challenges are packing the model in a timely manner and managing to display the model on the website. Increased computational power may be needed to overcome ongoing packing challenges. However, possible solutions to the display problems would be to better optimize NGL Viewer, simplify the model, use culling when rendering the model in the viewer, or use a graphics engine as in development at Scripps.

For example, NGL Viewer could be modified to load structures sequentially or interact with all structures in a radius as opposed to requiring subpixel precision when clicking.

Improvements on the membrane model and chromosomal model would be the simplest. More constraints such as localization, thermodynamic, and kinetic may be eventually added to better represent the interactions between the proteins. To fully represent all the interactions (including but not limited to PPIs, protein and chromosomal, RNA), a lot more data and computational power would be needed but it would be an incredibly valuable addition to the understanding of the cell. This is one of the long term goals for building whole cell models. On this note, as better software for whole cell visualization is developed and as more experimental data is generated, the model can be continuously updated and "evolve".

A more immediate direction, since a pipeline has been developed, would be to develop models for other types of bacteria or even for *E. coli* under different conditions or as a different strain. Taking this a step further, we hope to integrate metabolic networks to simulate directly on the 3D visualization with the aid of molecular dynamics. This will be incredibly computationally expensive, but with the development of more powerful computing and algorithms for simulations, could come to fruition in the near future. Visualizing cellular systematic changes could be a powerful tool towards understanding the whole cell, as would developing better simplifications to the cellular system models. A visual model is a valuable tool that provides insights that a graph or a table cannot. Its potential for use in education and research is great.

XII. Appendices

A. Expanded Evaluation of Alternative Designs

Strengths of approach

A strength of this approach (in comparison to the other approaches such as the outside-in, two-dimensional, or dynamic simulation approaches mentioned above under alternative designs and in the literature review) is the availability of data and the amount of computational power needed. There is not enough data in terms of imaging to build an *E. coli* cell model from for the outside-in approach. This model also offers a much greater level of detail than the outside-in approach. Models used for dynamic simulation would require a lot of knowledge about the dynamics and quantum mechanics which would be outside the scope and time that we have for this project; it would also require a lot of computational power that we do not have. Two-dimensional modeling has its merits, but we want to take this as far as we can with the current 3D modeling and packing software technologies available to us.

This project is incredibly inexpensive and many of the tools that we need such as cellPACK, NGL viewer, and IMOD are readily available for free. Based on the project so far, all the work has been done on our personal computers, so this approach is also very feasible and likely to be completed within the time and expected skill set of the senior design project. This approach is also unique in terms of the scale. Currently, models available are of viruses and smaller systems and models of large eukaryotic cells are also in development, but there are no models of *E. coli* available like ours or models of cells at the bacteria scale. The model as mentioned in the literature review, would be a test of how feasible a model of a bacteria cell is using readily available software as well as determining what more information or softwares need to be researched to help us develop a greater understanding of the *E. coli* system. Last but not least, the model will be a great educational tool to help teachers, students, and researchers gain a greater insight on the internal crowded environment of a cell.

Weaknesses

The greatest weakness in this project is also the amount of data that we currently have on *E. coli* proteomics and having no good way of representing DNA with current modeling softwares. Out of the about 4,300 genes present in the *E. coli* cell, only about 2,000 genes have mapped abundance or stoichiometric data. Granted, some genes may not code for proteins and many proteins are produced at extremely low quantities, but there are still about half for which abundances have not been mapped. Furthermore, many structures of proteins are also not available or have not been determined and there is even less data on protein-protein interactions, especially structural data on the complexes formed. Thus, for this model, it is not that there will be a ton of empty space, but we will have to be able to systematically display the unknowns present and include homologous structures or find other means of finding and displaying that information. Furthermore, DNA structures are not modelled because there are no readily available methods or data for us to put that together and fit into the cell. Depending

on the amount of data present on the localization of the proteins, it may also be difficult for us to determine in what quantity of proteins are present on the membranes, in the periplasm, or in the cytoplasm. In other words, the abundance of the same protein in the periplasm may differ from its abundance in the cytoplasm.

In comparison to alternative methods, this model is not able to display the experimental accuracy that the outside-in experimental approaches is able to display. However, for our model, we are trying to use as much experimental data as possible. This model is also unable to display the transient interactions and time dependent properties dynamic simulations are able to provide. This project also is computationally more expensive than the outside-in and 2D approaches, but it is not so expensive that we cannot do this project entirely on our computers. There are many limitations to what the model can do because it is a static model. The model also does not have the same impact that perhaps a medical device might, but this model does have its merits in being the first of its scale and being a tool for summarizing the information we have so far and being able to be easily manipulated and updated.

B. Gantt Chart as Modified During Implementation

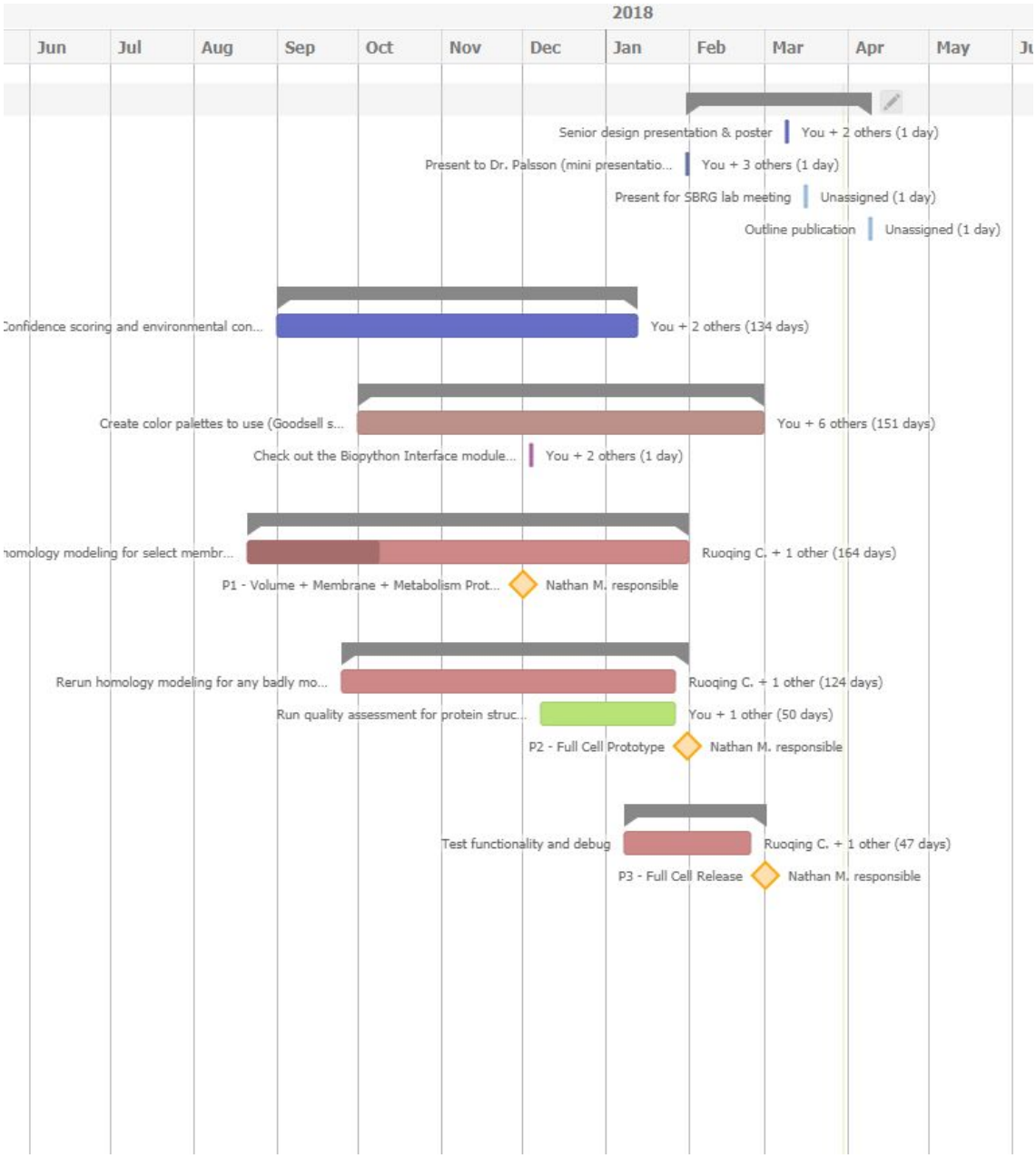


Figure B-1. A Gantt chart, depicting the timeline of most major project tasks.

The figure above is a Gantt chart depicting the schedule for the remaining tasks. Most of them deal with finding protein data and confirming the accuracy of the data retrieved. Several critical steps are dependent on each other; for example, the cytoplasm model built from the proteomics dataset must be aligned to the tomographic data compiled. Then, the membrane proteins must be aligned to the membrane surface. Finally, this data must be transferred to cellPACK and an algorithm needs to be built to transfer that to NGL Viewer. Last, a client needs to be developed

for interactive filtering and loading of files. Each of these steps is dependent on the one before it, making the first one extraordinarily critical.

Luckily, no special resources are needed except personal computers, as well as the aforementioned software. However, in the event of failure of any of these steps, we will need contingency plans. Currently, the information collected as a 'recipe' for the model has been completed. There are no major delays in moving the project along; although we are about one to two weeks behind in the website and model development as model building took about a week of two longer than expected. Features can be added to the web client as time passes, as long as the prototypes accomplish the base goals outlined previously.

C. Work Breakdown

Everyone contributed equally to the report. Certain sections or homework assignments were written by Yuxuan, Rohan, and Winnie.

D. Other Documentation

Table D-1. The components of the final cellPACK recipe.

FILES		
Component	**Filename**	**Description**
DNA	/DNA/ori_at_midcell_NM.pdb	500 basepair-per-bead resolution model...represents an orientation where oriC is at mid-cell, while the left and right chromosomal arms generally extend toward opposite cell poles
Membranes	/TOMOGRAMS/membranes7.obj	Winnie's generated membranes
Proteome - all	FINAL RECIPE - FULL (this sheet)	Marked INCLUDE for all MOLARITY>0 and any structure file available
Proteome - high quality	FINAL RECIPE - MIN (this sheet)	Marked INCLUDE for only PDBs + SWISS-MODELs + direct from OPM database
Structures	P2_structures	All PDB files for all structures in the recipe
Structures - zipped	P2_structures.zip	Archive of all PDB files for all structures in the recipe
VOLUMES + SURFACE AREAS		
Total cell volume	2.19955 um^3	
Cytoplasm	1.76970 um^3	
Periplasm	0.42985 um^3	
Outer membrane	9.00103e8 angstrom^2	
Inner membrane	7.73398e8 angstrom^2	
RECIPE - FULL		
INCLUDE	2159	Number of unique proteins (complexes + subunits) included in recipe
Copy_number2	4836173	Total number of proteins in recipe
INCLUDE: Cytosol	1674	Number of unique proteins (complexes + subunits) in cytosol
Copy_number2: Cytosol	4085588	Total number of proteins in cytosol
INCLUDE: Periplasm	144	Number of unique proteins (complexes + subunits) in periplasm
Copy_number2: Periplasm	268381	Total number of proteins in periplasm
INCLUDE: Outer_Membrane	86	Number of unique proteins (complexes + subunits) in outer membrane
Copy_number2: Outer_Membrane	387577	Total number of proteins in outer membrane
INCLUDE: Inner_Membrane	255	Number of unique proteins (complexes + subunits) in inner membrane
Copy_number2: Inner_Membrane	94627	Total number of proteins in inner membrane
RECIPE - MIN		
INCLUDE	1613	Number of unique proteins (complexes + subunits) included in recipe
Copy_number2	4156305	Total number of proteins in recipe
INCLUDE: Cytosol	1440	Number of unique proteins (complexes + subunits) in cytosol
Copy_number2: Cytosol	3815187	Total number of proteins in cytosol
INCLUDE: Periplasm	116	Number of unique proteins (complexes + subunits) in periplasm
Copy_number2: Periplasm	242573	Total number of proteins in periplasm
INCLUDE: Outer_Membrane	23	Number of unique proteins (complexes + subunits) in outer membrane
Copy_number2: Outer_Membrane	75574	Total number of proteins in outer membrane
INCLUDE: Inner_Membrane	34	Number of unique proteins (complexes + subunits) in inner membrane
Copy_number2: Inner_Membrane	22971	Total number of proteins in inner membrane

Bibliography

Acknowledgements

Dr. Nathan Mih, Edward Catoiu, Ruoqing Cheng, Liangyu Zhao, and Kritin Karkare all contributed to the development of this project. Dr. Nathan Mih, especially, served as the mentor and supervisor and as the creator of this project. Dr. Brett Barbaro guided us and helped us understand the process of building and packing the cell model. Dr. Adrian H. Elcock and William Hacker shared their chromosomal model for this project. Dr. Alexander Rose provided insights on how to use NGL viewer. Dr. Bernhard Ø. Palsson was the advisor for this project.

References

Biasini, M., Bienert, S., Waterhouse, A., Arnold, K., Studer, G., Schmidt, T., . . . Schwede, T. (2014, 04). SWISS-MODEL: Modelling protein tertiary and quaternary structure using evolutionary information. *Nucleic Acids Research, 42*(W1). doi:10.1093/nar/gku340

Bordbar, A., Monk, J. M., King, Z. A., & Palsson, B. Ø. (2014, 01). Constraint-based models predict metabolic and associated cellular functions. *Nature Reviews Genetics, 15*(2), 107-120. doi:10.1038/nrg3643

Brunk, E., Mih, N., Monk, J., Zhang, Z., O'Brien, E. J., Bliven, S. E., . . . Palsson, B. Ø. (2016, 03). Systems biology of the structural proteome. *BMC Systems Biology, 10*(1). doi:10.1186/s12918-016-0271-6

Bureau of Labor Statistics. (2016, 05). Occupational Employment Statistics. Retrieved December 14, 2017, from http://www.bls.gov/oes/

Carrera, J., & Covert, M. W. (2015, 12). Why Build Whole-Cell Models? *Trends in Cell Biology, 25*(12), 719-722. doi:10.1016/j.tcb.2015.09.004

Cignoni, P., Callieri, M., Corsini, M., Dellapiane, M., Ganovelli, F., & Ranzuglia, G. (2008). MeshLab: An Open-Source Mesh Processing Tool (V. Scarano, R. De Chiara, & U. Erra, Eds.). ERCIM News, 73, 45-46. doi:10.2312/LocalChapterEvents/ItalChap/ItalianChapConf2008/129-136

Cock, P. J., Antao, T., Chang, J. T., Chapman, B. A., Cox, C. J., Dalke, A., . . . Hoon, M. J. (2009, 03). Biopython: Freely available Python tools for computational molecular biology and bioinformatics. Bioinformatics, 25(11), 1422-1423. doi:10.1093/bioinformatics/btp163

Cusick, M. E. (2005, 10). Interactome: Gateway into systems biology. *Human Molecular Genetics, 14*(Suppl_2). doi:10.1093/hmg/ddi335

Delgado, L., Martínez, G., López-Iglesias, C., & Mercadé, E. (2015, 03). Cryo-electron tomography of plunge-frozen whole bacteria and vitreous sections to analyze the recently described bacterial cytoplasmic structure, the Stack. *Journal of Structural Biology, 189*(3), 220-229. doi:10.1016/j.jsb.2015.01.008

Goodsell, D. (2016, 02). Cellular Landscapes in Watercolor. *Journal of Biocommunication, 40*(1).

doi:10.5210/jbc.v40i1.6627

Goodsell, D. S., Autin, L., & Olson, A. J. (2018, 01). Lattice Models of Bacterial Nucleoids. The Journal of Physical Chemistry B. doi:10.1021/acs.jpcb.7b11770

Hacker, W. C., Li, S., Elcock, A. H. (2017, 06). Features of genomic organization in a nucleotide-resolution molecular model of the *Escherichia coli* chromosome. *Nucleic Acid Research, 45*(13), 7541-7554. doi: 10.1093/nar/gkx541

Im, W., Liang, J., Olson, A., Zhou, H., Vajda, S., & Vakser, I. A. (2016, 07). Challenges in structural approaches to cell modeling. *Journal of Molecular Biology, 428*(15), 2943-2964. doi:10.1016/j.jmb.2016.05.024

Johnson, G. T., Autin, L., Al-Alusi, M., Goodsell, D. S., Sanner, M. F., & Olson, A. J. (2014, 12). CellPACK: A virtual mesoscope to model and visualize structural systems biology. *Nature Methods, 12*(1), 85-91. doi:10.1038/nmeth.3204

Keskin, O., Tuncbag, N., & Gursoy, A. (2016, 04). Predicting Protein–Protein Interactions from the Molecular to the Proteome Level. *Chemical Reviews, 116*(8), 4884-4909. doi:10.1021/acs.chemrev.5b00683

Klein, T., Autin, L., Kozlikova, B., Goodsell, D. S., Olson, A., Groller, M. E., & Viola, I. (2017). Instant Construction and Visualization of Crowded Biological Environments. *IEEE Transactions on Visualization and Computer Graphics,* 1-1. doi:10.1109/tvcg.2017.2744258

Kremer, J. R., Mastronarde, D. N., & Mcintosh, J. (1996, 01). Computer Visualization of Three-Dimensional Image Data Using IMOD. *Journal of Structural Biology, 116*(1), 71-76. doi:10.1006/jsbi.1996.0013

Krogh, Anders, et al. Predicting Transmembrane Protein Topology with a Hidden Markov Model: Application to Complete Genomes. *Journal of Molecular Biology*, vol. 305, no. 3, 2001, pp. 567–580., doi:10.1006/jmbi.2000.4315.

Le Gros, M. A., Mcdermott, G., & Larabell, C. A. (2005, 10). X-ray tomography of whole cells. *Current Opinion in Structural Biology, 15*(5), 593-600. doi:10.1016/j.sbi.2005.08.008

Lee, R., Karr, J. R., & Covert, M. W. (2013). WholeCellViz: Data visualization for whole-cell models. BMC Bioinformatics, 14(1), 253. doi:10.1186/1471-2105-14-253

Lomize, M. A., Lomize, A. L., Pogozheva, I. D., & Mosberg, H. I. (2006, 01). OPM: Orientations of Proteins in Membranes database. *Bioinformatics, 22*(5), 623-625. doi:10.1093/bioinformatics/btk023

Lučić, V., Rigort, A., & Baumeister, W. (2013, 08). Cryo-electron tomography: The challenge of doing structural biology in situ. *The Journal of Cell Biology, 202*(3), 407-419. doi:10.1083/jcb.201304193

Macklin, D. N., Ruggero, N. A., & Covert, M. W. (2014, 08). The future of whole-cell modeling. *Current Opinion in Biotechnology, 28*, 111-115. doi:10.1016/j.copbio.2014.01.012

Mih, N., Brunk, E., Chen, K., Catoiu, E., Sastry, A., Kavvas, E., . . . Palsson, B. Ø. (2018, 02). ssbio:

A Python framework for structural systems biology. Bioinformatics. doi:10.1093/bioinformatics/bty077

Milne, J. L., & Subramaniam, S. (2009, 08). Cryo-electron tomography of bacteria: Progress, challenges and future prospects. *Nature Reviews Microbiology, 7*(9), 666-675. doi:10.1038/nrmicro2183

O'Brien, E. J., Lerman, J. A., Chang, R. L., Hyduke, D. R., & Palsson, B. Ø. (2014, 04). Genome-scale models of metabolism and gene expression extend and refine growth phenotype prediction. Molecular Systems Biology, 9(1), 693-693. doi:10.1038/msb.2013.52

Ortiz, J. O., Brandt, F., Matias, V. R., Sennels, L., Rappsilber, J., Scheres, S. H., . . . Baumeister, W. (2010, 08). Structure of hibernating ribosomes studied by cryoelectron tomography in vitro and in situ. *The Journal of Cell Biology, 190*(4), 613-621. doi:10.1083/jcb.201005007

Purcell, O., Jain, B., Karr, J. R., Covert, M. W., & Lu, T. K. (2013, 06). Towards a whole-cell modeling approach for synthetic biology. *Chaos: An Interdisciplinary Journal of Nonlinear Science, 23*(2), 025112. doi:10.1063/1.4811182

Rajagopala, S. V., Sikorski, P., Kumar, A., Mosca, R., Vlasblom, J., Arnold, R., . . . Uetz, P. (2014, 02). The binary protein-protein interaction landscape of Escherichia coli. *Nature Biotechnology, 32*(3), 285-290. doi:10.1038/nbt.2831

Rose, A. S., & Hildebrand, P. W. (2015, 04). NGL Viewer: A web application for molecular visualization. *Nucleic Acids Research, 43*(W1). doi:10.1093/nar/gkv402

Santarella-Mellwig, R., Pruggnaller, S., Roos, N., Mattaj, I. W., & Devos, D. P. (2013, 05). Three-Dimensional Reconstruction of Bacteria with a Complex Endomembrane System. *PLoS Biology, 11*(5). doi:10.1371/journal.pbio.1001565

Spirin, V., & Mirny, L. A. (2003, 09). Protein complexes and functional modules in molecular networks. *Proceedings of the National Academy of Sciences, 100*(21), 12123-12128. doi:10.1073/pnas.2032324100

Szklarczyk, D., Morris, J. H., Cook, H., Kuhn, M., Wyder, S., Simonovic, M., . . . Von Mering, C. (2016, 10). The STRING database in 2017: Quality-controlled protein–protein association networks, made broadly accessible. *Nucleic Acids Research, 45*(D1). doi:10.1093/nar/gkw937

Webber, J. (2012). A programmatic introduction to Neo4j. *Proceedings of the 3rd Annual Conference on Systems, Programming, and Applications: Software for Humanity - SPLASH '12*. doi:10.1145/2384716.2384777

Yuan, S., Chan, H. S., & Hu, Z. (2017, 01). Using PyMOL as a platform for computational drug design. Wiley Interdisciplinary Reviews: Computational Molecular Science, 7(2). doi:10.1002/wcms.1298

Final Project Proposal (BENG 187B)

I. Introduction

Cellular and molecular biologists have historically expressed great interest in visualizing the dynamic and crowded environment of the cell. Early examples of this curiosity are seen in David Goodsell's detailed paintings. His works are a beautiful attempt to recreate the cell's complex structural system (Goodsell, 2016). With the aid of computer graphics and computational software, a three-dimensional visualization of the cell has inched closer to reality (Carrera & Covert, 2015). Such a visualization could eventually enable scientists to examine the macroscopic effects of cellular perturbations, accelerate advances in biotechnology, and identify ongoing gaps in knowledge for a given organism (Purcell, Jain, Karr, Covert, & Lu, 2013).

While the study of systems biology generally limits itself to the simulation of individual metabolic networks, a systems biology approach to analyze a complete biological system has been developed for many years. Traditional approaches have included the reconstruction of biological networks and the mathematical modelling and simulation of systems (Bordbar, Monk, King, & Palsson, 2014), and have advanced to involve the integration of individuals' metabolomics data to create a set of personalized kinetic parameters for simulation and analysis. Recently, researchers have been working to create a systemic, quantitative visualization of a mesoscale model, which would contain three-dimensional structural information for each molecule, protein, and component of the cellular system (Im et al., 2016). In this project, we seek to build a web-accessible visual reconstruction of the *E. coli* cell, complete with metabolic networks and protein-protein interactions. Basic criteria for success include accurately displaying the environment and dynamics of the cell, as well as clearly simulating the reactions and interactions contained within (Im et al., 2016).

Unfortunately, at the moment no complete, three-dimensional visual reconstruction of a single cell exists on a scalable pipeline to represent proteomic, metabolomic, and genomic data in an easily accessible manner, preventing students, teachers, and researchers from improving education and performing simulations with such a product. Any existing visualizations of a cell are incomplete, two-dimensional, static or purely artistic in nature (Lee, Karr, & Covert, 2013). Furthermore, several methods have allowed a systemic visualization of components of the cell, but none have allowed for the accurate, large-scale visualization of a major cell for students' and researchers' uses. Consequently, it is currently difficult to query given proteins and establish their context within a whole cell due to the lack of a pipeline to integrate available information into a visual, three-dimensional model.

Development of this pipeline would be a major boon to systems biologists and molecular biologists, who would benefit from an interactive 3D model of an *E. coli* cell to be able to see and gather the structural system and protein network of the *E. coli* cell in an organized fashion. Researchers need a platform to organize and update the information they have on *E. coli*. This would also translate to discovering what information is lacking for building a more complete

model. According to the Bureau of Labor Statistics (2016), there are 85,000 biologists in the United States—while it is impossible to estimate how many could benefit from a tool like this, it is clear that making this platform accessible to the community could greatly streamline research and development for a vast number of scientists. Additionally, building a scalable platform could extend the tool's reach. Researchers need an easy and clear way to build 3D visualizations of structural systems and protein networks such as the one we propose for *E. coli*. An organized and easy-to-use pipeline for building such a model is needed for organisms as large as *E. coli* and eventually at the scale of eukaryotic cells. Broadly, researchers who study organisms other than *E. coli* such as yeast and other bacteria can use our pipeline to create a 3D visualization of the organism of their interest.

Finally, educators and students need a 3D model to introduce the cell as a system. Teachers could better explain the scenarios in which many metabolic and molecular and macromolecular reactions occur and introduce the topic of molecular biology to students more effectively. Students could gauge a deeper understanding of the cell and hopefully be inspired to do further research the science contributing this project. High school and college level teachers and students would benefit the most educationally from this model. There are 15 million high school students enrolled in the U.S. and around 20 million college students enrolled in 2017. Even though the total number of students studying a STEM related field may be fewer, the numbers greatly reflect the impact this model can make on inspiring curiosity towards systems biology.

II. Relevant Background

Many obstacles lie in the way of building a whole-cell model. Macklin, Ruggero, and Covert (2014) propose seven challenges, namely "(1) experimental interrogation, (2) data curation, (3) model building and integration, (4) accelerated computation, (5) analysis and visualization, (6) model validation, and (7) collaboration and community development." Attempts to develop such a model have resulted in two schools of thought, which can be labeled the "outside in" and "inside out" visualization approaches. The former uses tomography to build a visual model of the cell and matches metabolites and macromolecules to the acquired localization data, whereas the latter assembles individual components of the cell into a complete model. Each dogma has distinct challenges associated with it, as well as a variety of existing tools and literature to address these. Universally, experimental techniques such as cryo-electron tomography and fluorescent imaging, data from proteomics and docking, and tools such as cellPACK and CHARMM-GUI have greatly pushed forward the building of a dynamic whole-cell model (Im et al., 2016).

"Outside In" Approach

Data Sources

Whole-cell imaging techniques have greatly improved, to the point where whole *E. coli* with manganese-tagged protein have been successfully imaged using single-particle x-ray diffraction (Le Gros, McDermott, & Larabell, 2005). Tomograms of cells provide a detailed example of the internal structures of their membranes and macromolecules. The most common source of

tomograms is cryo-electron tomography, in which the whole cell or thin slices are vitrified (rapidly frozen) and then imaged using transmission electron microscopy. Cryo-electron tomography allows for remarkably detailed images, with resolutions as low as 4 nm, to be taken of bacteria cells (Milne & Subramanium, 2009). The development of fluorescent labeling has made the identification, three-dimensional visualization, and localization of many particles such as *E. coli* 100S ribosomes and the endomembrane system of *G. obsuriglobus* possible (Ortiz et al., 2010; Santarella-Mellwig et al., 2013). The "outside in" approach solves an issue brought up by Betts and Russell (2007) with respect to the "inside out" method. Even with complete proteomic data, including copy numbers and cellular locations, that proteomic data used in the latter does not necessarily indicate what proportion of each protein resides in a particular location of the cell. The use of tomography is their proposed solution to this problem, as it bridges a gap between x-ray diffraction and fluorescence tagging in terms of its high resolution and accurate localization data. The images from this technique provide valuable information and confirmation regarding the stoichiometry, structure, and location of molecules and proteins within the cell—despite limitations in capturing the cell in a dynamic state. Other limitations of tomography are its inability to capture all of the proteins, electron densities, and molecules within the cell due to constraints in labeling the components and the possibility of beam damage given thinner sections to be imaged (Keskin, Tuncbag, & Gursoy, 2016; Milne & Subramanium, 2009). Despite these restrictions, tomograms provide a valuable experimental foundation and useful validation for other approaches discussed later.

Model Visualization

The tomograms are often converted to three-dimensional models through programs such as IMOD. IMOD is a program described by Kremer, Marstronarde, and Mcintosh (1996) that is able to read the image stacks of the electron microscopy, align the images, and convert the tomograms into readable data using intensity spikes. Alternate methods of image analysis include the use of techniques such as thresholding, averaging, machine learning, and statistical approaches (Le Gros, McDermott, & Larabell, 2005; Milne & Subramanium, 2009; Ortiz et al., 2010).

"Inside Out" Approach

Information Pipeline

Genome-scale models (GEMs) are used for computational studies of metabolism, protein synthesis, and transcriptional regulation based on a specific organism (Bordbar, Monk, King, & Palsson, 2014). Recently, structural information has been mapped to those models, which are now called genome-scale models with protein structures (GEM-PROs) and provide a new view of systems biology (Brunk et al., 2016). This pipeline has expanded knowledge of biochemical network properties and can help make predictions on whole-cell phenotypes. This comprehensive information can only be visualized in two dimensions, so it is necessary to create a pipeline that generates three-dimensional models using GEM-PRO output data. One way to generate these models is by gathering the data for the system and putting it into a single "recipe" file (will be discussed further in the model assembly) and using this recipe to generate a

virtual model through a computational engine with built-in assembly algorithms.

Data Sources

Protein abundance—determining how much of each species is present in the cell—can be determined using existing proteomics datasets. Betts and Russell (2007) describe the fundamental process of incorporating proteomics data into a whole-cell model. This information will also be incorporated as a major part of the "recipe" to determine how many copies of a component are placed in the model. In addition, building a three-dimensional model requires structural information for every molecule or protein in the cell system. The Protein Data Bank (PDB) provides downloadable three-dimensional protein structure files in the PDB format, which are required for the assembly process.

Information related to protein complex poses a challenge for researchers developing a whole-cell model. The EcoCyc database described by Karp and Riley (1999) is a resource which compiles polypeptide and protein complex data for *E. coli* in a downloadable way, potentially making large datasets describing complexes more accessible.

Protein Complex Representation

In cell systems, proteins usually do not act alone: they interact with each other and form protein complexes to carry out certain functions (Spirin & Mirny, 2003). However, a large proportion of known proteins present in *E. coli* do not have structural information available in the PDB, so finding a way to represent them is necessary. The EcoCyc database described by Karp and Riley (1999) is a resource which compiles polypeptide and protein complex data for *E. coli* in a downloadable way, potentially making large datasets describing complexes more accessible.

Additionally, predictive methods can assemble protein components into a whole complex. Krissinel and Henrick (2007) developed a method called PISA (Protein Interfaces, Surfaces and Assemblies) to infer protein interactions based on data found in the PDB. Open source softwares such as the Integrative Modeling Platform (IMP) combine electron microscopy data with proteomics data and protein structural information to model the complex structure based on their scoring mechanisms (Russel et al., 2012). The downsides of these softwares are that they work best on much smaller scales than that of a whole *E. coli* cell and require constraints such as cross-linking data for their modeling. Moreover, they are very computationally expensive. Another method used to map complex information is the integration of protein-protein interaction (PPI) datasets (Rajagopala et al., 2014). Abundant PPI datasets exist, containing both experimental and predicted data (Szklarczyk et al., 2016). Each binary interaction of proteins can be mapped onto the metabolic network, and the protein complex that carries out a reaction can be represented by individual components interconnected by PPI without having to be assembled together. This saves huge amounts of time, since all of the requisite data is easily obtained. Most importantly, the expansion of metabolic networks to include PPI results in emergent properties (Cusick, 2005).

Model Assembly and Visualization

Once all the data is gathered, it will be used to generate the whole-cell visualization, enabling users to interact with it and query valuable information from it. A software called cellPACK has proven capable of generating large-scale biological models such as the human immunodeficiency virus (Johnson et al., 2014). Klein et al. (2017) describe cellPACK as a computationally intensive method to automatically arrange biological molecules into desired partitions, as opposed to manually placing molecules in a 3D visualization software. When supercomputer power is available, cellPACK allows data to be organized relatively quickly into a three-dimensional mesh of a cell and displayed using the cellVIEW tool. It applies both random packing algorithms and user-defined packing methods to automatically pack the model, taking the input "recipe", a file which contains information for all the molecular components that will be present in the model. It may be possible to store structural information in the Macromolecular Transmission Format (MMTF), which loads quicker than traditional PDB files. NGL Viewer, a webGL-based application to visualize macromolecules, may also be able to visualize such a large-scale model (Rose & Hildebrand, 2015).

Cells consist of other macromolecules, metabolites, and lipids. The tool Membrane Builder from CHARMM-GUI utilizes the highly mobile membrane-mimetic (HMMM) model to simulate the lipid bilayer surrounding the membrane proteins (Im et al., 2016). C-SAC (chromosome self-avoiding chain) uses a computation strategy similar to protein structure determination to predict the folding of chromatin within the cell (Im et al., 2016). Beyond these, there are numerous molecules and intermolecular interactions between proteins, nucleic acids, ions, and metabolites yet to be explored. Due to the difference in scale between the lipid membrane and these macromolecules, as well as the extensive computations required to simulate their interactions, our proposed whole-cell model should exclude these molecules until further time, energy, and computational power have been devoted to their study.

Database Construction

A database will be necessary if users wish to query information in the model for educational or research purposes. SQL (structured query language) is the traditional way to build a database. A graph database is a relatively new way to store relational data, typically as nodes and relationships (pathways). The most prominent example of a graph database is neo4j, a frontend web interface that can store, query, and change data directly through a query language called Cypher (Webber, 2012). It makes storing and querying relational data more efficient. If a database is ever needed for the proposed model, neo4j will be prioritized.

Onward

Rather than being a setback, the lack of information available to build a complete whole-cell model illuminates the areas where further investigation is necessary to thoroughly understand the cell as a dynamic system. An interactive cell model of *E. coli*, hopefully, will not only serve as a tool for exploratory purposes but also develop into an application for predicting systemic responses of the cell. The model we propose will use an "inside out" approach paired with some techniques from the "outside in" model. Eventually, the simulation can deepen humanity's

understanding of the cell and help with precise prediction of the effects of new drugs, chemicals, and other modifications (Carrera & Covert, 2015).

III. Design Goals and Constraints

Goals

The primary goal of the project is the development of a website to visualize, in three dimensions, the structures of the *E. coli* cell. Subgoals are that the website must be interactive, quick to load, and capable of allowing users to access the data used. Additionally, it must be both as precise and complex as possible, but does not need to be a dynamic simulation as yet. Creating a dynamic simulation will take too much time and computational power, and this project is currently being worked on by another team in the Systems Biology Research Group. At best, we can lay the foundation for future integration of dynamic kinetic data.

The reconstruction aims to provide a three-dimensional visual reconstruction of *E. coli* by assembling atomic-level components into a cellular-scale model. This will permit researchers to look at the *E. coli* cell as a whole and summarize the known information within the proteome. This goal is of utmost importance; we are assigning it a weight of 40% for the later comparison of design alternatives. Success can be measured not by the completeness of the data, but by the completeness of its annotations; that is, if certain proteins are included with less certainty than others, they should be clearly indicated as such.

A subgoal is to maximize accuracy of the representation. An accurate reconstruction depends on reliable, accurate sourcing for the stoichiometry, localization, interactions, and structures contained by proteomics datasets. The model generated must be as accurate as possible so that people can rely on its outputs—consequently, we assign this goal a weight of 25%. This can be measured by performing quality analysis on both experimentally and computationally determined data and matching the conditions used for data extraction. Furthermore, review of the data sources and databases used to compile information can be used to develop metrics for success.

Next, it is important to be able to convert raw data into an appropriate format for three-dimensional visualization. This is because it is important to be able to easily update the model, for example, if new data based on new experimental results is brought to light. Additionally, for other researchers to scale the platform to fit their needs, they need to be able to insert their data into a pipeline to generate the three-dimensional model. This goal is weighted with 10% importance; its success can be measured by developing a model from new data, as well as by appending new data to the existing data and seeing if a model can successfully be built.

Finally, the speed of the model is paramount to its success. Does the website load quickly? Will it operate on older hardware? To avoid making the tool cumbersome or unusable, these data points can be tested and the product can be optimized accordingly. We also seek to present users with a simple web interface that allows them to interact with the model and easily find the

proteomics data that they need. The interface must be intuitive and user-friendly enough that people without scientific backgrounds or technical expertise can operate it. This can be measured through user testing. Testers can identify whether the website is aesthetically pleasing, functional, and attractive. Additionally, they can identify whether or not it is easy to tell what information is being represented. We assign this goal a weight of 15%, as web design is important, but not the investigators' priority.

Constraints

There are several constraints that could prevent completion of some of these goals. For one, it is impossible to build a perfectly accurate simulation due to the nature of the available data. Not all proteins have complete sets of experimental or computational data available; many will be missing stoichiometric, interactional, or structural information. Consequently, their accuracy is compromised and the unknown variables will have to be appropriately estimated or simulated. Additionally, time constraints prevent being able to generate a complete, dynamic simulation or a polished, professional website. With only a few months remaining until the deadline, it will be difficult to form a web page with all the desired functionality and interactions. That time is better dedicated to finding and standardizing high-quality data for the simulation.

Cost is another consideration. While the software packages in use (NGL Viewer, cellPACK, ssbio, IMOD) are open source, computational costs are not insignificant. Building the whole cell model requires high computational power to pack every component into the right space, considering the massive number of components. While the Scripps Research Institute supercomputer may absorb some costs of the cellPACK algorithm, using the San Diego Supercomputer Center would cost $0.025/SU for our purposes, where one SU (service unit) represents one core-hour of computing time. Luckily, this cost is trivial for a single simulation. Furthermore, the computation itself takes time; visualizing each cellPACK "recipe" takes about one to two extra weeks.

Last, the performance specifications of various devices mean that one of our goals may work in opposition to the rest. It will be essential to balance function and functionality; in other words, if the site is feature-laden but inoperable on lower-end devices, it is not achieving its goals. We will be working to make massive amounts of data quick to load and a database efficient to query for specific data points.

IV. Design Alternatives and Analysis

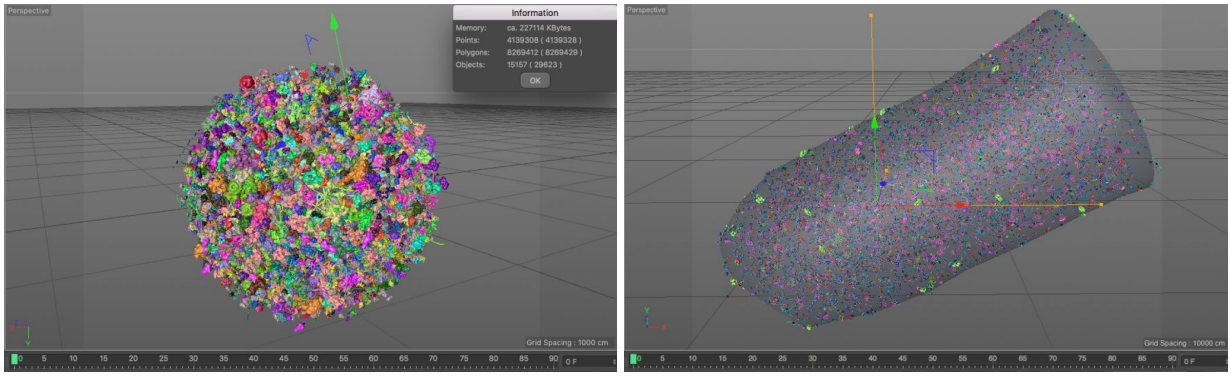
Alternative 1: cellPACK Method

This is the method currently being investigated by the team. It takes an "inside-out" approach to developing a cell-scale visual reconstruction of *E. coli*. The intent is to take existing metabolomic, proteomic, and genomic data and match it to software and tomograms to generate a three-dimensional, structurally and systematically accurate visualization of an *E. coli* cell and its components. The simulation would then be transferred into a web page where it can be manipulated by end users. Numerous pitfalls exist with such an approach, but the main

bottleneck is with the accuracy and quality of the data available for modeling. However, it has the advantage of easily scaling for different models and future updates to the data.

Implementation Plan

The implementation for this model begins by scraping all the necessary data to start to derive an accurate model. In addition to finding protein abundances, this requires awareness of individual proteins' structures and interactions. Localization data can be obtained for proteins by comparing their visual appearances to their appearances in tomograms. Thermodynamic constraints can also be integrated into the modeling process to obtain the optimal state of the cell. Then, the data would be crunched by cellPACK, which outputs a three-dimensional model for each region of the cell based on the shapes and charges of molecules. The first prototype model of *E. coli*, built with the aid of our collaborators in the Scripps Research Institute molecular graphics laboratory led by Dr. Arthur Olson, is shown in Figure 2. The data from this model could correspond to individual protein shapes from the Protein Data Bank, letting each 3D shape be rendered individually as an MMTF or PLY file. These file formats are compatible with the software NGL Viewer for powerful, lightweight three-dimensional rendering of biological molecules.



(a) (b)

Figure 2. (a) A cytosol model that contains all the available structures and concentrations of cytoplasmic proteins with a sphere constraint applied. (b) The shape of the *E. coli* inner membrane, obtained from tomogram datam was applied as a constraint to pack the model.

Alternative 2: Outside-In Tomography Plan

A potential alternative to the atomic model we are designing could be to take a top-down or "outside-in" approach where we can directly visualize—using 3D tomographic data—the membranes and such present in the images. We would have to make sure that we are visualizing everything we need to see including proteins and macromolecules. With multiple tomograms, we would have multiple variations of the *E. coli* and we could potentially find an average or middle ground between the tomographic models and combine the separate images into one model with averaging or other techniques. Potential caveats would be the amount of data available and the amount of data we would be able to collect from the tomograms. Additionally, as with the proposed design alternative, it will be necessary to devise a set of standard

conditions to image the *E. coli* cell at. An upside to this approach would be the accuracy relative to real conditions within the cell because the images are directly from the cell. There would be more confidence in the visualization from this model since the structures are what is imaged and not based on purely prediction data or comparison data with similar structures.

Implementation Plan

We would start out by taking tomograms of *E. coli* cells and then convert them to 3D models using special software. The software would read the image stacks, align the images, and convert the tomograms into usable data using intensity spikes. The results of such a process are displayed in Figure 3. Then, the tomogram data would be matched to protein and other macromolecule abundances (in the event of conflict, the tomogram information will override, as this alternative prefers and prioritizes tomogram data). The proteins and other structures would be rendered in three dimensions as discussed in the previous model and made available to end users online.

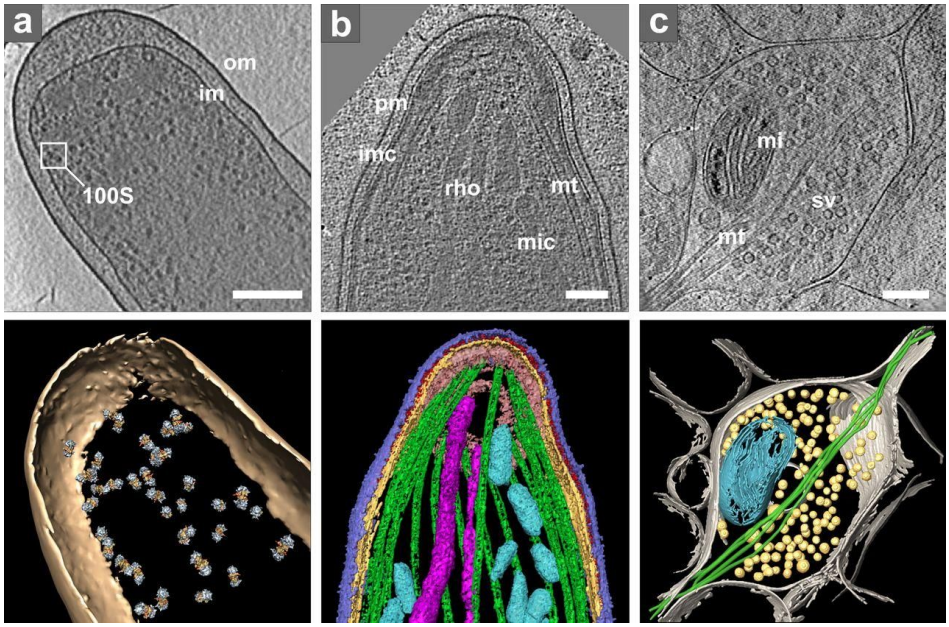


Figure 3. Tomogram slices in the top row are converted to isosurface representations shown in the bottom row to develop a model (Lučić et al., 2013).

Alternative 3: Two-Dimensional Simulation

This alternative is inspired by the online whole-cell *Mycoplasma genitalium* model depicted by Lee, Karr, and Covert (2013). That simulation uses six panels to depict the activity of a single *Mycoplasma* sample over time, as shown in Figure 1. My proposed alternative would use a variety of panels as well, but unlike the existing model, it would not have dynamic simulation data. However, it would have almost all of the data we seek to incorporate into our existing model, such as protein structures and networks, metabolic interactions, and genetic data. The

advantages for such a model would be that it would be significantly easier to construct using existing tools, and the use of specialized software such as cellPACK and NGL Viewer would not be necessary. Its disadvantages center on the fact that localization data would be lost, meaning valuable information about intracellular interactions would not be accessible. (We see that this is a problem with the model above, where the metabolic networks depicted are not linked to any sort of locational data.)

Implementation Plan

For the most part, the steps to acquire data for visualization would be similar. In fact, tomograms would still be useful to generate a two-dimensional visualization with accurate organelle placement and general protein localization. However, because the majority of interactions would need to be generated in three dimensions, it is more important to display complexation on a graph or chart. Because these are static simulations, the data can even be displayed as a PNG, but it may be preferable to load some information through a Chart.js script. This would allow it to be annotated appropriately. Of course, ssbio and Escher will still be essential for simulating and mapping metabolic networks. Proteomic and interactomic data will be scraped from the PDB and EcoCyc, but as we will no longer have a three-dimensional depiction, PLY files and three-dimensional protein structures will not be necessary.

Alternative 4: Dynamic Simulation

This proposal is similar to both the first and second models. However, the priority is not on a static 3D simulation as with those, but with a dynamic rendering of the networks in the entire cell. Using known equilibrium constants for every single macromolecule, each reaction mechanism will need to be reconstructed and the subsequent reaction rate constants determined *in silico*. Once complete models have been generated, the networks will be simulated at steady state with the appropriate initial conditions. It is unclear how a live network would be rendered—this would be a major challenge with such an approach. However, should it be possible, this reconstruction would be a major coup for researchers.

Decision Matrix

Alternative: Goal	**Alternative 1: cellPACK Method**	**Alternative 2: Outside-In Approach**	**Alternative 3: 2D Simulation**	**Alternative 4: Dynamic Simulation**
Goal 1: To build a 3D visual reconstruction of an *E. coli* cell (40%)	*Score: 80* *Weighted: 32* The cellPACK algorithm may run into issues with the complexities of a whole cell, but this is still one of	*Score: 80* *Weighted: 32* The outside-in approach is the other of two accepted ways to generate a whole-cell model. Success is	*Score: 0* *Weighted: 0* While the simulation does qualify as a visual reconstruction, there is no chance it will ever be	*Score: 30* *Weighted: 12* The dynamic simulation builds on the first two approaches, and adds more data making it a more complete

	the accepted ways to model entire cells.	not guaranteed in either method.	manipulable in 3D.	reconstruction. However, visualizing networks over time in 3D will be so challenging that we are unlikely to see success.
Goal 2: A representation containing reliable and accurate data on the stoichiometry, localization, interaction, and structure of the omics data (25%)	*Score: 70* *Weighted: 17.5* Existing datasets for proteins, genes and metabolites are still incomplete, but for the most part this approach should have all the critical data conveniently compiled.	*Score: 70* *Weighted: 17.5* The end result of this will have similar information to the first approach. The main difference will be with the quality of the data in the tomogram vs. in omics sets: realistically, both will be kind of incomplete.	*Score: 65* *Weighted: 16.25* It will be difficult to render localization data in two dimensions, so this info will likely be lost. However, other information can be included that wouldn't usually render in a sole 3D model.	*Score: 50* *Weighted: 12.5* Kinetic parameters are difficult to track down and kinetomics datasets are virtually nonexistent. While *E. coli* networks have more data, it will still not be enough to generate a complete model.
Goal 3: Present data in format that is easy to update with any changes or additions needed for the model (10%)	*Score: 40* *Weighted: 4* Changes or updated information will require cellPACK to be run again with the new surfaces or stoichiometries.	*Score: 20* *Weighted: 2* Cryo-electronic tomograms will be challenging to obtain, and the whole model will have to be recreated for even the most minor updates.	*Score: 70* *Weighted: 7* Each of the individual panes of such a simulation could likely be rendered separately, making updates relatively easy.	*Score: 30* *Weighted: 5* Comparable to the first model, but also having to recompile every kinetic network when new information is available.
Goal 4: Short model loading speed for the amount of data to be presented on a functional,	*Score: 80* *Weighted: 12* The data can be dropped as is into NGL Viewer or compressed into the MMTF	*Score: 80* *Weighted: 12* The data would be handled the same way in this method and the previously	*Score: 90* *Weighted: 13.5* Without surface renderings of each protein, loading the data will be a breeze.	*Score: 50* *Weighted: 7.5* This model has more data and requires far more server or computing

attractive website (15%)	format, making it compact and relatively quick to load.	discussed one.	The interface will likely be intimidating for users, but also easier to operate.	power to generate and run, reducing the likelihood of success.
Weighted Totals	**65.5/100**	**63.5/100**	**36.75/100**	**37/100**

V. Design Solution

Based on the above analysis, an inside-out approach based on cellPACK will continue to be the optimal design solution to achieve our desired goal. As mentioned, we will create an interactive, three-dimensional visual reconstruction of an *E. coli* cell by integrating structural data, localization data from tomograms, and macromolecular omics to create an accurate representation of the cell. Users will be able to interact with the model by zooming in and out, rotating and orbiting around various axes, selecting particular structures, and exploring various layers of the model. Advanced options will include fast queries for specific information (structures, subsystems, interactions and reactions) and eventually the ability to generate users' own three-dimensional models through the pipeline.

The manpower available extends beyond the senior design team (Grover, Shi, and Xie) and our advisor (Nathan Mih); bioinformatics students Kritin Karkare, Ruoqing Cheng, and Liyangyu Zhao as well as graduate students Eddie Catoiu and Ify Aniefuna will be contributing to different parts of the project. The overall project can be split into the following components:

Project Breakdown

Project 1: Create pipeline to generate "recipe"

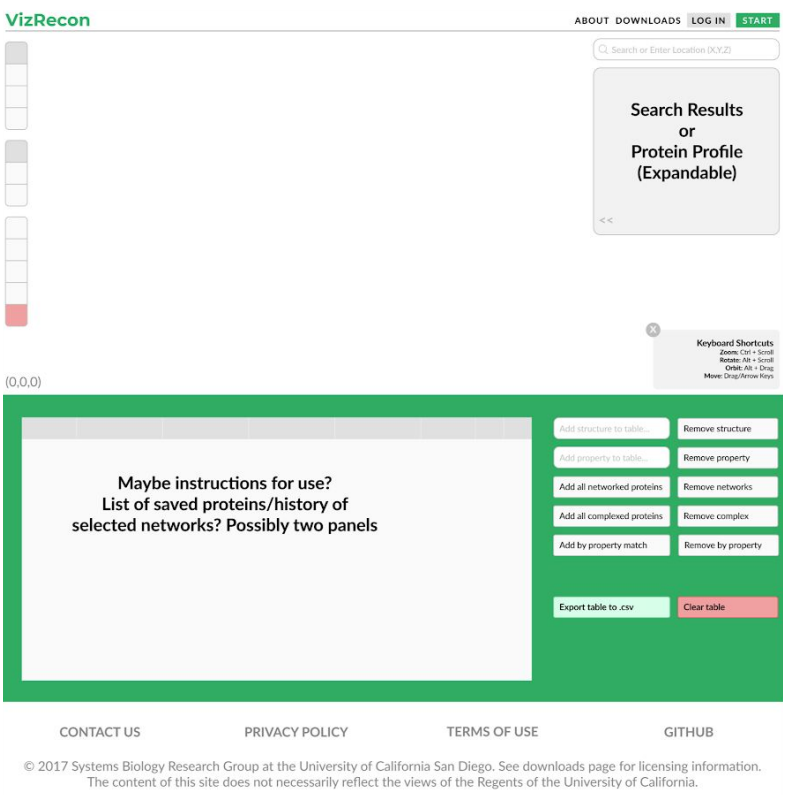
The primary goal of this subproject is to create a pipeline that generates a recipe that the cellPACK algorithm can read. This entails gathering *E. coli* proteomics data and mapping it to the genetic model we use. The protein complex data can be updated using data from the EcoCyc and UniProt databases; for membrane proteins, the data provided by the Orientations of Proteins in Membranes (OPM) database and the tool TMHMM, which uses hidden Markov model (a machine learning algorithm) to predict transmembrane helices of membrane proteins (Lomize, Lomize, Pogozheva, & Mosberg, 2006; Krogh, Larsson, von Heijme, & Sonnhammer, 2001). For cytoplasmic proteins and membrane proteins without detailed data, we will need to determine a method to represent unavailable structures. The SWISS-MODEL homology-modeling tool may be useful for this purpose (Biasini et al., 2014). The recipe compilation is being led by Xie, with Karkare focusing on orienting membrane proteins; Grover is handling complex formation data.

Project 2: Tomogram Analysis for* E. coli *Volume and Localization

Tomograms of whole *E. coli* cells must be analyzed to determine the protein localizations and membrane morphology. Each slice will be compiled to form a three-dimensional model according to existing protocols for tomographic analysis, such as those described by Delgado, Martínez, López-Iglesias, and Mercadé (2015). Shi is handling most of this analysis on her personal computer.

Project 3: cellPACK Models Displayed using NGL Viewer

The cellPACK output will need to be configured into its individual components for display using NGL Viewer. This process will need to be automated and tested. Further, the capacity of NGL Viewer to render potentially thousands of components quickly on ordinary laptops will need to be tested. If possible, the program can be optimized to load the model in the most resource-efficient manner. This subproject is being worked on by Cheng and Zhao.



Project 4: Web Interface Development

The interface for displaying the model and interactions is currently being prototyped in Figma by Grover, as shown in Figure 4. The interface needs to be clean, aesthetically pleasing, and simple to understand and navigate for all levels of users. Presently, Aniefuna has the most experience, but this task is being split between all members. A database to link structures to their attributes and allow users to search it will be likely needed. Once an interface has been designed, it will be a team effort to develop into a finalized product.

Figure 4. A draft web interface, with panels for protein analysis and three-dimensional visualization.

Project 5: Database Construction

A graph database will provide intuitive representation of gene interaction and metabolic networks to explore the effects of gene-knockout cascade. Xie has built a graph database using neo4j that takes information from GEMs and integrates available PPI data and displays the networks. The graph database (neo4j) proved to be relatively easy to learn. Traditional database

(a relational database and SQL) will take much longer time to learn and to build, but will also be more versatile and more conventionally used. Xie has been primarily working on this portion, but other members of the project will be involved in putting the database together.

VI. Parts, Resources, and Costs

Below is a simple list of parts, software, and costs that will be needed for them.

Parts, Software, or Resources	Details	Costs
Individual laptops or computers	Used to develop the model and analysis	Free (already paid for)
Protein databases	Protein structures and protein complex information will be provided by PDB, EcoCyc, and GEM-PRO databases Unknown structures could be potentially predicted using SWISS-MODEL tool	Free
Protein interaction data	Data will be collected from EcoCyc and Biosystems Laboratory	Free
cellPACK	Software to pack and build the 3D models developed by researchers at Scripps	Free
IMOD	Software for tomogram analysis developed by University of Colorado in Boulder	Free
NGL Viewer	Used to display our model on the website	Free
Tomograms	Will be collected from various supplemental images from different papers and tomogram databases	Free
SQL database	Used to build database to store structural, stoichiometric, localization, PPI, and metabolic data	Free
Python	Used to automate processes and collect data	Free
(OPTIONAL) Supercomputer Time	Depending on how much computation will be needed to put together the model, we might need some time with the supercomputer at SDSC	$0.025/SU

VII. Planning/Scheduling

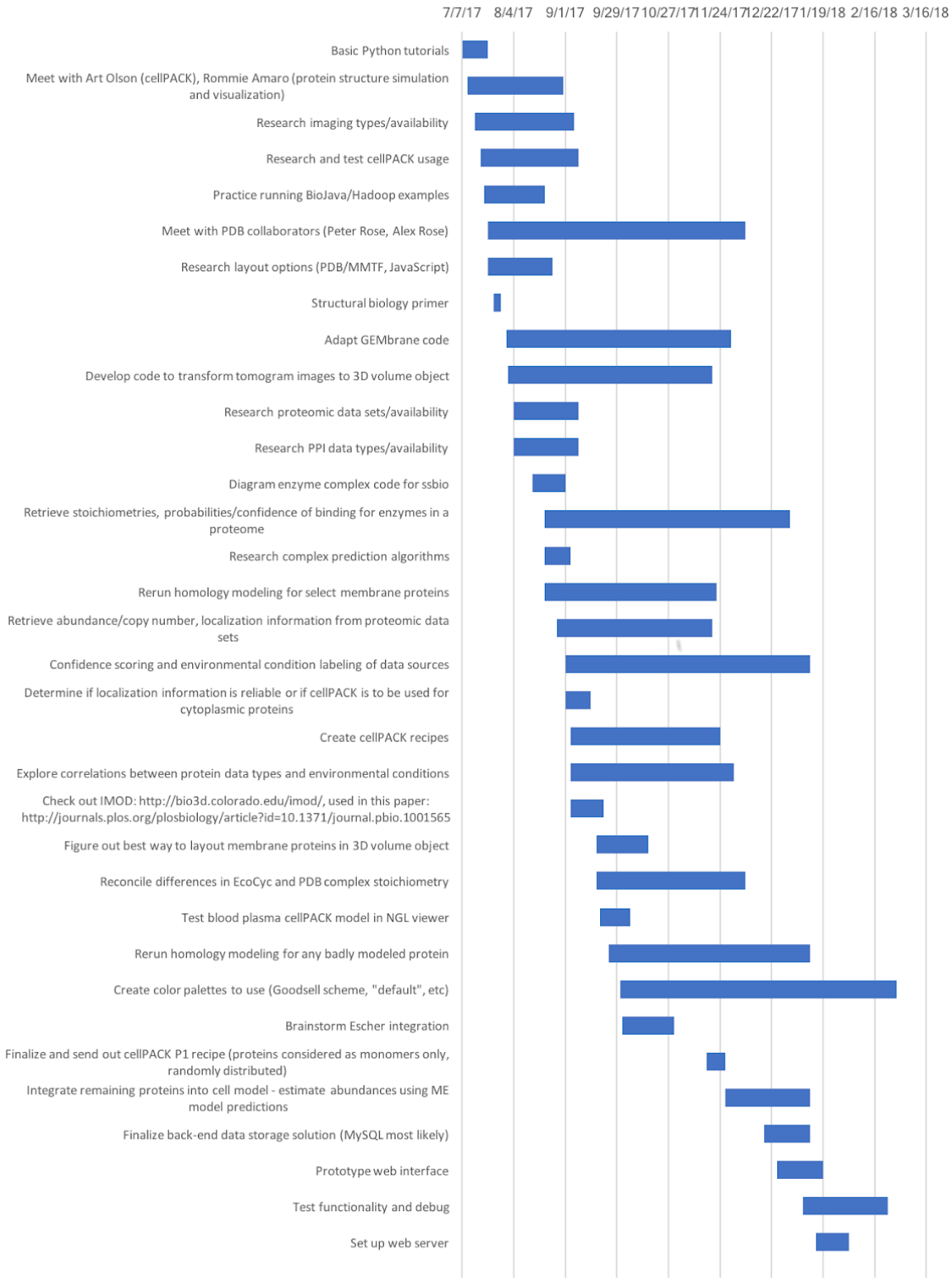


Figure 5. A Gantt chart, depicting the timeline of most major project tasks.

The figure above is a Gantt chart depicting the schedule for the remaining tasks. Most of them deal with finding protein data and confirming the accuracy of the data retrieved. Several critical steps are dependent on each other; for example, the cytoplasm model built from the proteomics

dataset must be aligned to the tomographic data compiled. Then, the membrane proteins must be aligned to the membrane surface. Finally, this data must be transferred to cellPACK and an algorithm needs to be built to transfer that to NGL Viewer. Last, a client needs to be developed for interactive filtering and loading of files. Each of these steps is dependent on the one before it, making the first one extraordinarily critical.

Luckily, no special resources are needed except high computational power and personal computers, as well as the aforementioned software. However, in the event of failure of any of these steps, we will need contingency plans. If the protein data does not necessarily fit within the tomographic constraints, we can exclude certain genes or discard tomographic data in favor of accepted dimensions. Membrane proteins can be inexactly aligned if there is difficulty obtaining normal vectors for their surface orientations or if their orientations don't match the tomogram surface. Additionally, features can be added to the web client as time passes, as long as the prototypes accomplish the base goals outlined previously.

VIII. Preliminary Assessment

Strengths of approach

A strength of this approach (in comparison to the other approaches such as the outside-in, two-dimensional, or dynamic simulation approaches mentioned above under alternative designs and in the literature review) is the availability of data and the amount of computational power needed. There is not enough data in terms of imaging to build an *E. coli* cell model from for the outside-in approach. This model also offers a much greater level of detail than the outside-in approach. Models used for dynamic simulation would require a lot of knowledge about the dynamics and quantum mechanics which would be outside the scope and time that we have for this project; it would also require a lot of computational power that we do not have. Two-dimensional modeling has its merits, but we want to take this as far as we can with the current 3D modeling and packing software technologies available to us.

This project is incredibly inexpensive and many of the tools that we need such as cellPACK, NGL viewer, and IMOD are readily available for free. Based on the project so far, all the work has been done on our personal computers, so this approach is also very feasible and likely to be completed within the time and expected skill set of the senior design project. This approach is also unique in terms of the scale. Currently, models available are of viruses and smaller systems and models of large eukaryotic cells are also in development, but there are no models of *E. coli* available like ours or models of cells at the bacteria scale. The model as mentioned in the literature review, would be a test of how feasible a model of a bacteria cell is using readily available software as well as determining what more information or softwares need to be researched to help us develop a greater understanding of the *E. coli* system. Last but not least, the model will be a great educational tool to help teachers, students, and researchers gain a greater insight on the internal crowded environment of a cell.

Weaknesses

The greatest weakness in this project is also the amount of data that we currently have on *E. coli* proteomics and having no good way of representing DNA with current modeling softwares. Out

of the about 4,300 genes present in the *E. coli* cell, only about 2,000 genes have mapped abundance or stoichiometric data. Granted, some genes may not code for proteins and many proteins are produced at extremely low quantities, but there are still about half for which abundances have not been mapped. Furthermore, many structures of proteins are also not available or have not been determined and there is even less data on protein-protein interactions, especially structural data on the complexes formed. Thus, for this model, it is not that there will be a ton of empty space, but we will have to be able to systematically display the unknowns present and include homologous structures or find other means of finding and displaying that information. Furthermore, DNA structures are not modelled because there are no readily available methods or data for us to put that together and fit into the cell. Depending on the amount of data present on the localization of the proteins, it may also be difficult for us to determine in what quantity of proteins are present on the membranes, in the periplasm, or in the cytoplasm. In other words, the abundance of the same protein in the periplasm may differ from its abundance in the cytoplasm.

In comparison to alternative methods, this model is not able to display the experimental accuracy that the outside-in experimental approaches is able to display. However, for our model, we are trying to use as much experimental data as possible. This model is also unable to display the transient interactions and time dependent properties dynamic simulations are able to provide. This project also is computationally more expensive than the outside-in and 2D approaches, but it is not so expensive that we cannot do this project entirely on our computers. There are many limitations to what the model can do because it is a static model. The model also does not have the same impact that perhaps a medical device might, but this model does have its merits in being the first of its scale and being a tool for summarizing the information we have so far and being able to be easily manipulated and updated.