Research Strategy for 2022 proposal

Instructions
https://grants.nih.gov/grants/how-to-apply-application-guide/forms-f/general/g.400-phs-398-research-plan-form.htm#3

Information from our 2017 R01 and instructions for renewal

Here is our current R01 information from NIH Reporter

https://reporter.nih.gov/search/5S6ahRaJaUGn45aiEjZhjg/project-details/9973164

It was submitted to Funding Opportunity Announcement (FOA) PA-16-160. That FOA has been reissued as current RFA PA-20-185.

https://grants.nih.gov/grants/guide/pa-files/PA-20-185.html

PA-20-185 is a general investigator initiated project R01, with no specific topic. Requests for more than $500,000 direct costs per year require special approval. However per NIGMS policy, funding for renewal grants will not exceed 3% of the funding of the final year of the current grant. For ChimeraX this means $400,000 + 3% = $412,000 in direct costs.

Review criteria include Significance, Investigators, Innovation, Approach, Environment. Also there is an overall score: "Reviewers will provide an overall impact score to reflect their assessment of the likelihood for the project to exert a sustained, powerful influence on the research field(s) involved... An application does not need to be strong in all categories to be judged likely to have major scientific impact. For example, a project that by its nature is not innovative may be essential to advance a field." "For Renewals, the committee will consider the progress made in the last funding period."

Instructions

Instructions for R01 research plan are PHS 398 given here

https://grants.nih.gov/grants/how-to-apply-application-guide/forms-f/general/g.400-phs-398-research-plan-form.htm#Intro

Research Strategy

There is a 12 page limit for the research strategy section and it includes subsections Significance, Innovation, Approach. Also renewals have a Progress Report section.

Publications List

Renewals have a publication list (not part of research strategy) that lists all publications "resulting from the project since it was last reviewed competitively". Not clear if we list all publications citing ChimeraX or just our publications.

Letters of Support

My reading of the instructions is that only partners directly participating in the work can give letters of support. "Letters of Support serve to describe terms of a collaboration or consultation and also are not de facto letters of reference from persons not actively participating in the project. Applications with letters containing such excess information may be withdrawn from the review process."

Resource Sharing Plan

This may or may not apply to us. It sounds like it is not required for us since we request less than $500K direct costs.

The PA-18-484 instructions say: "Individuals are required to comply with the instructions for the Resource Sharing Plans as provided in the SF424 (R&R) Application Guide."

The SF424 instructions say "Investigators seeking $500,000 or more in direct costs (exclusive of consortium F&A) in any budget period are expected to include a brief 1-paragraph description of how final research data will be shared,"

SPECIFIC AIMS

This project proposes continued development of advanced interactive visualization and analysis software for experimental data, including and emphasizing data from state-of-the-art (<3Å resolution) cryo-electron microscopy (cryo-EM), with the ultimate goal of understanding how cells and their molecular machinery function. The interactive visualization and analysis of structures of molecules, molecular assemblies, and protein sequence-structure relationships are critical for addressing important and highly relevant biomedical problems such as identifying the molecular basis of disease, identifying targets for drug development, designing drugs, and engineering proteins with new functions. The enormous growth in recent years in both the size and complexity of biological data sets, and especially structural data at various scales in length and time has created new and significant challenges for biomedical researchers. New and innovative software tools are vital to the successful outcomes of the myriad NIH-funded experimental research projects. There is no indication that the growth in data size and complexity will abate, and thus approaches to integrating, interpreting, and otherwise making the best use of the data require continuing, focused attention. We propose to address this challenge via the following specific aims:

Aim 1: Develop interactive visualization and analysis tools for atomic structures and associated data, including higher-order assemblies, conformational ensembles, density maps, and sequences.

Understanding the biological roles of molecular structures requires integrating a variety of associated data and analyses, with interactive visualization to link alternative views and diverse types of information with important structural elements. We will extend ChimeraX with a rich set of new features, for example: increased use of interactive 2D plots; a streamlined process for preparing structures for docking or further calculations by adding missing atoms and assigning charges; and new sequence/structure tools, including generating sequence alignments from structure superpositions. To facilitate making compelling movies for publication and teaching, we will implement a timeline animation tool. Further, we will enhance the capabilities and performance of key tools, as well as adapt to the needs of the broader structural biology community as they arise.

Aim 2: Develop interactive visualization and analysis tools for electron microscopy of molecular assemblies and cells. Profound advances in electron microscopy are extending it to atomic (1-2Å) resolution and to scales as large as the *Drosophila* brain, posing many new challenges in turning image data into biological insights. We will develop robust and widely useful analysis capabilities and make the latest innovative algorithms being developed in labs around the world accessible to the broad research community. We will provide interfaces to new machine-learning methods that predict structures, capture heterogeneous states in single-particle cryo-EM of molecular assemblies, and segment tomography of cells helping unveil mechanisms of biological systems. These interactive tools will leverage new advances in visualization such as uniform manifold approximation and projection (UMAP) for seeing patterns in higher-dimensional data, from discerning flexible modes of molecular machines to characterizing cells from observed quantitative properties. Virtual-reality visualization and lighting of tomograms will help discern new structures. Significant effort will also go into making proven tools for 3D microscopy and models accessible to the widest audience possible.

Aim 3: Provide a diverse software foundation enabling labs around the world to develop and distribute new molecular and cellular structure analysis software.

Obstacles to new structure analysis methods getting beyond journal publication and into wide use are severe. New methods require a foundation of underlying software to read complex data, visualize it, provide intuitive user interfaces, and make it easily installed by researchers on the major computer operating systems. This aim will allow other labs to easily extend ChimeraX. Steps to achieve this include design and documentation of stable public software interfaces; creating programming tutorials on all aspects of reading data, computing, and adding user interfaces such as mouse modes, toolbars, panels and commands; and utilizing online software repositories for distribution and for developers to manage new releases and collect usage statistics, bug reports, and user feedback. Simplifying the development of production-quality implementations of new analysis methods will multiply the number of new ChimeraX tools available to researchers and advance the pace of scientific discovery.

RESEARCH STRATEGY

A. SIGNIFICANCE

ChimeraX interactive visualization software and its predecessor Chimera are in use by many thousands of labs to analyze diverse experimental data and atomic models, and to present results using images and animations in journal publications. Each month, they are cited in ~350 papers (see Figure 1 for ChimeraX citations), with 500 users registering voluntarily. The exceptional impact of this software is a product of several characteristics: 1) the ability to easily combine diverse data and analysis methods, enabling a synthesis that is essential for advances in biological understanding; 2) making the latest technologies such as machine learning, virtual reality, and GPU-accelerated molecular simulation accessible to members of the broader structural biology community who are not specialists in these techniques; 3) unparalleled documentation, tutorials, and same-day support that allow researchers to overcome unique analysis problems endemic to cutting-edge research; and 4) collaborative and community development that allows labs with deep expertise to distribute their latest analysis methods as extensions to the software. The developments we propose will advance each of these strengths of the ChimeraX software to enable the next breakthroughs in understanding the molecular basis of life.

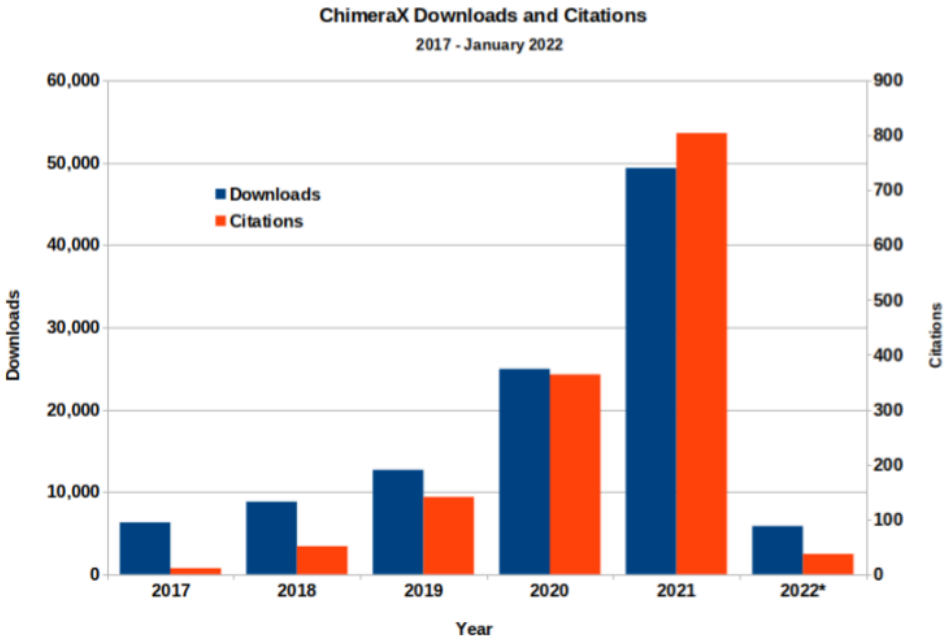


Figure 1. ChimeraX downloads and citations from 2017 through January, 2022.

Determining the structures and functions of molecular assemblies involves many types of experimental data and computational analyses, such as X-ray diffraction intensities, electron microscopy maps and segmentations, nuclear magnetic resonance distance restraints, chemical crosslinks from mass spectroscopy, sequences to assess conservation and evolutionary relationships, molecular simulations to understand dynamics, and docking to predict small-molecule binding, among many others. Integrative analysis can decipher complex assemblies such as the nuclear pore[1] (Figure 2). The ability to visualize multiple data types together is a key strength of ChimeraX, for example, to show how sequence mutations map onto a 3D atomic model and alter drug binding, or change the hydrogen bonds observed in simulations. That ChimeraX reads 79 file formats and writes 29 formats gives an indication of the diversity of data it handles. Many developments we propose will enhance integrative analysis, for example,

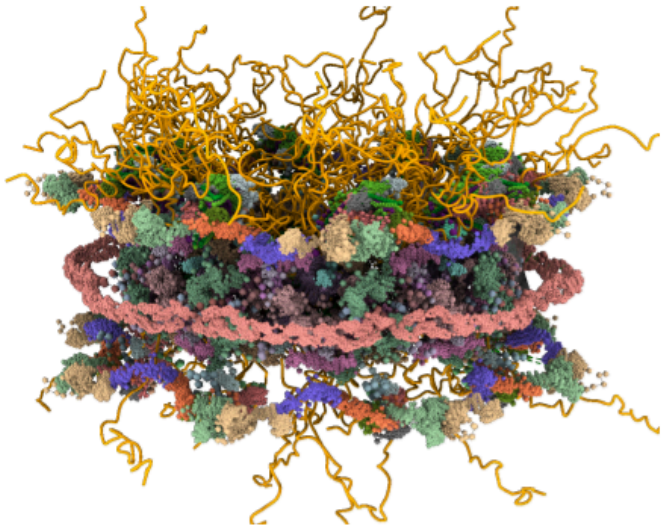


Figure 2. An integrative model of the nuclear pore complex of Saccharomyces cerevisiae solved by combining cryo-electron tomography, X-ray atomic models, comparative models, chemical cross-linking with mass spectroscopy, small angle x-ray scattering, and 3D fluorescence microscopy.

new capabilities aiding docking, simulations, sequence analysis, segmentations, crosslink and glycan visualization.

A central goal of this proposal is to make the most promising new computer technologies accessible to biology researchers. Recent advances in graphics processor units (GPUs) are leading to groundbreaking advances in all areas of science. Current affordable graphics hardware has up to 10,000 parallel processors capable of general-purpose computations, as well as dedicated neural network computational units, and offers new hardware accelerated ray-tracing for rendering. Using this commodity hardware, we can interactively simulate mutations in molecular systems to explore molecular function, untangle dynamic states of molecular machines seen in electron microscopy, and use virtual-reality headsets for immersive stereoscopic visualization of electron tomography to clearly observe cellular mechanisms. Our proposal addresses many of the technical hurdles that prevent biology researchers from using these nascent technologies. We provide more detail in the following Innovation section.

The impact of academic software is often severely limited by inadequate documentation. All ChimeraX commands, user interfaces, and methods are meticulously documented, with online tutorials created, and webinars and workshops presented for the most important analysis capabilities. This makes the tools accessible to an order of magnitude more researchers, greatly increasing their impact. All of the developments in this proposal will meet this exceptionally high documentation standard. Making researchers aware of new capabilities is also essential for maximizing impact, and we use Twitter (3700 followers) and YouTube "how-to" videos (500 subscribers) to maximize the reach of our software. About 1000 messages per year are posted to the ChimeraX mailing list, half from the development team, with same-day response time. Our feature request and bug tracking built into ChimeraX also generate about 1000 tickets per year, most often with same-day response and bug fixes within a few days. Rapid support underpins the very high utilization and impact of ChimeraX.

Our proposed developments center on creating user interfaces for new cutting-edge capabilities leveraging libraries and algorithms developed by others wherever possible. These interoperable tools allow combining many data sources in a way not possible with special-purpose analysis web services or standalone computational packages. The needs for diverse analysis methods far outstrip what the core ChimeraX development team can provide. Thus a central aim is to enable other labs to develop and distribute plugins through the ChimeraX Toolshed repository, which allows the discovery and single-click installation of new tools. To foster plugin contributions utilizing the latest technologies, we will develop standard ChimeraX interfaces to support accessing web services, support widely used machine-learning toolkits such as PyTorch and TensorFlow within ChimeraX, and offer new task-management capabilities to allow background computations with progress reporting and intermediate results for complex algorithms. ChimeraX support for plugins has been a primary goal inspired by the 50 plugins of our legacy Chimera software. With the many new elements of the ChimeraX plugin distribution system that reduce the burden on contributors and maximize the credit they receive, we anticipate 50 new plugins over the funded period, and many additional contributions over the lifetime of the software.

In summary, our vision is to enable integrative visualization and analysis across the gamut of data used to understand biological systems at the molecular and cellular level, to enable use of the latest advances in computer technology, to provide exceptional levels of documentation and support that maximize use, and to foster community contributions of new analysis methods. Successful implementation of this vision, which we have pursued through four generations of software (Midas, MidasPlus, Chimera, ChimeraX), has had a significant impact on tens of thousands of research projects. The pace of scientific discovery, the complex mix of experimental data, and advances in technology have never been greater. The proposed new capabilities of the ChimeraX integrative analysis platform will play a major role in promoting scientific discovery in the coming decade.

B. INNOVATION

This project will make several promising technologies accessible to biology researchers, including machine learning methods to analyze electron microscopy, dimensionality reduction to analyze complex patterns in molecular simulations and to extract conformational states from heterogeneous samples in cryo-EM, virtual

reality to discern new structures in electron tomography of cells, and hardware-accelerated ray-tracing to enhance contrast in electron tomography. Applications of these new methods face significant technological hurdles. Our goal is to simplify use of these next-generation techniques to make them accessible to researchers without advanced expertise.

Across many scientific disciplines, machine learning has made stunning advances in recognizing patterns and applying knowledge from large pre-existing datasets to interpret new data. Machine learning methods outperformed more traditional techniques for predicting protein structures from sequence in the latest Critical Assessment of Structure Prediction competition, CASP14[2]. Over 2600 researchers have used new ChimeraX AlphaFold prediction capabilities, and Aim 2 will allow predicting much larger assemblies and analyzing these imperfect models in combination with cryoEM data. Machine learning has been especially effective in analysis of image data, and in Aim 2 we will visualize results of algorithms such as cryoDRGN[3] and 3DFlex[4], which reconstruct multiple conformational states from 2D electron micrographs. In Aim 3, we provide the foundation for others to use neural networks in ChimeraX plugins, for example, to segment electron microscopy (see letter from Jing He). We will streamline use of hardware acceleration available in commodity graphics processors that can speed up neural network calculations 100-fold.

New algorithms for discerning patterns in high-dimensional data are being applied in biology, for instance, to categorize all cell types in the human body[5]. Dimensionality reduction techniques such as Uniform Manifold Approximation and Projection[6] (UMAP) can reduce many parameters, such as a list of the hydrogen bonds in a binding site during a molecular simulation or descriptors characterizing the conformational state of an assembly (Figure 5), to two dimensions for visualizing clusters or continuous transitions. In Aim 1, we will use UMAP and similar approaches to generate 2D plots that are interactive, where clicking the plot highlights the corresponding 3D structural pattern.

The crowding of molecular machinery in cells makes recognizing substructures in electron tomography extremely challenging. The ChimeraX virtual-reality capabilities we developed primarily for visualizing ligand binding[7,8] have been applied to electron tomography for discerning aggregate structures in Huntington disease, new virus spike morphologies, and other previously unrecognized structural features (see letter from Wah Chiu). In Aim 2, we will add virtual reality (VR) capabilities tailored to tomography such as the ability to paint structures, as well as optimization for use on very large data sets. We also will apply lighting methods developed by the Allen Institute of Cell Science (see letter from Graham Johnson) and explore the use of hardware-accelerated ray-tracing to improve visibility of cellular substructures.

In addition to the many new technologies we develop, the community plugin development enabled by Aim 3 will bring many innovative methods developed by other labs into wider use. Also beyond the innovations we foresee in this proposal, we regularly explore innovative applications of new visualization hardware to problems in biology. Recent examples include use of depth-sensing cameras to produce augmented-reality videos[9–11] explaining new science discoveries, and use of auto-stereo displays[12,13] that require no glasses for stereoscopic depth perception, combined with hand tracking[14] for manipulating molecular simulations.

C. APPROACH

Progress Report (7/1/2018 - 2/28/2022): ChimeraX has progressed from early development to become a widely used visualization package offering many new advances in analysis methods, achieved with the previous 2018-2022 R01 award. We have significantly expanded its capabilities with novel features and dramatically improved performance, as well as implemented numerous high-value tools, facilitating discoveries by thousands of researchers.

The early development of interactive ambient-occlusion lighting and curved-tube helices spurred excitement about using ChimeraX for figures and movies; subsequently, users are discovering its ease of use and the cutting-edge functionality such as AlphaFold structure prediction, resulting in a remarkable increase in ChimeraX downloads and citations over the current funding period (see Significance section). During this period, we made eight beta releases, leading up to ChimeraX version 1.0 in June 2020, a milestone sufficient to replace Chimera for many uses and to surpass it for many others. Version 1.1 was released later in 2020, 1.2 was released in May 2021, and 1.3 was released in December 2021. We published three papers about ChimeraX, two on the overall application[15,16] and another describing its use in virtual reality (VR)[7].

The original aims are similar to the current ones, covering:

Aim 1. Interactive visualization and analysis of atomic and cryo-electron microscopy data sets. Major additions in this area include map masking, interactive "volume eraser" editing, and watershed segmentation; morphing atoms and maps to highlight conformational differences; measuring map statistics, isosurface area and enclosed volume; detection of atomic contacts and clashes; molecular surface coloring by Coulombic electrostatic potential (ESP) or molecular lipophilic potential (MLP) calculated on the fly; coloring by B-factor and other properties; radial coloring; fetching biological assemblies, building unit cells, and checking crystal contacts; protein structure prediction (detailed below for aim 2); selection inspector to view and easily modify properties (bfactor, psi angle, color, *etc.*); and graphical interfaces for fitting into maps, zoning to nearby atoms, hiding "dust" (small surface bits), finding H-bonds, viewing ligand-receptor docking results, and "matchmaker" superposition based on sequence alignment. Stylized nucleotide representations (ladders, lollipops, *etc.*) and several display-style presets have been added. Color key and 2D arrow capabilities have also been added during this time, as well as support for numerous file formats of density maps, microscopy data, segmentations, atomic structures, trajectories, and sequence alignments.

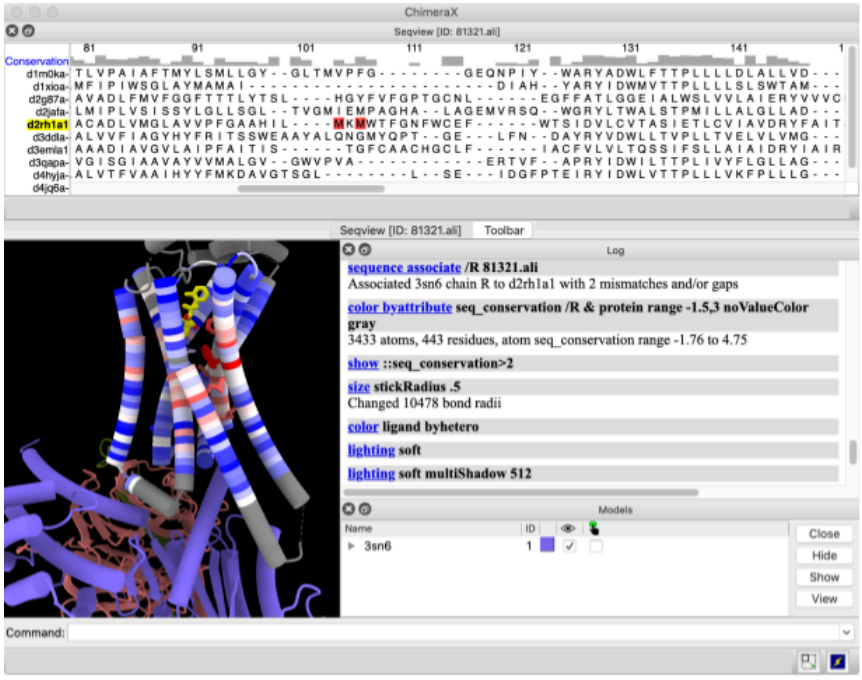


Figure 3. β2-adrenergic receptor signaling complex, as shown using the online tutorial on coloring by sequence conservation.

Simply listing features does not convey their ease of use. Many operations can be performed with a single click of an icon in the ChimeraX toolbar.

The sequence viewer now allows calculating per-column RMSDs among associated structures, as well as several metrics of sequence conservation (as provided by AL2CO[17]. These are shown in histograms above the sequences and automatically assigned as attributes of the associated residues for coloring (as in Figure 3). Sequences can easily be fetched from UniProt along with their associated annotations, such as sites of post-translational modification or disease-associated mutations. The sequence annotations are shown as colored boxes that can be clicked to select the corresponding parts of any associated structures. Multiple sequence alignments can be created or realigned using Clustal Omega[18] and MUSCLE[19] web services.

There have also been significant developments for ChimeraX with virtual reality headsets, including new modes for hand-controller manipulation and button assignments, and coordination of multi-person virtual meetings. VR has enabled insights into complex structures and interactions that would not otherwise be possible (see letters from Wah Chiu and Adam Frost), and has driven enthusiastic ChimeraX use by a dedicated set of researchers and educators.

All new ChimeraX features have been documented in the User Guide. We have also created tutorials and videos to show how they are used. Tutorials may use the click-to-execute mechanism of ChimeraX's built-in browser, making them easy to follow (for example, the tutorial used to make Figure 3). The ChimeraX home page includes quick links to the User Guide, tutorials, and videos.

Aim 2. Atomic-resolution modeling from cryo-electron microscopy. Far-reaching advances in the methods for determination of atomic structures have been introduced in ChimeraX, such as AlphaFold structure prediction for creating starting models and ISOLDE for refinement. The ISOLDE modeling suite[20] allows interactive, dynamic refinement of atomic structures modeled into cryo-EM density maps. This cutting-edge

tool is a significant driver of ChimeraX use and has been cited in several recent papers in prominent journals (in at least 16 publications in *Science*, *Nature*, and *Cell* during 2021; examples include [21–27]).

The groundbreaking machine-learning protein structure prediction method AlphaFold 2.0[28] was released in July of 2021, including code and a database of 365,000 predicted structures[29] covering the proteomes of 21 organisms. Within a month we added ChimeraX capabilities to search and fetch from the new database and predict new structures. The ChimeraX AlphaFold database search uses BLAST to provide an essential sequence search capability not available (as of January 2022) at the EBI database web site (see letter from Gerard Kleywegt). The ChimeraX prediction capability runs AlphaFold on Google Colab servers and has been used 12,000 times by 2600 users in the five months since release, and now uses the November 2021 AlphaFold 2.1 version which allows prediction of protein complexes[30].

An interface to Modeller[31,32] (running as a web service hosted by our group) has been added for filling in missing segments and for performing comparative modeling, with the ability to model both homo- and heteromultimeric protein complexes. Many other ChimeraX features for building and modifying structures have been added during the current funding period, from *de novo* peptide and nucleic acid generation to modifying atoms one by one, hydrogen addition, interactive bond rotation, analysis/replacement of amino acid sidechain rotamers, virtual mutation, and mouse modes to tug and jiggle atoms with brief OpenMM molecular mechanics/dynamics[33].

Aim 3. Facilitating community development of new methods building upon the ChimeraX platform. The ChimeraX Toolshed web repository of plugins has been implemented over this time period. A ChimeraX plugin is defined as a tool bundle that may include graphical interfaces, commands, input/output file types, specialized mouse modes, and other functions. Developers can upload their bundles to the site, and a tutorial for how to create a bundle has been added to the programmer documentation. Bundles are initially vetted by the ChimeraX team for usability, reliability and security. Many important ChimeraX APIs have been documented, we work closely with bundle developers throughout the development process, and we are highly responsive to questions via our virtual helpdesk.

The Toolshed automatically handles dependencies and platform-specific versions, and it manages plugin installation and updates from within ChimeraX. Choosing Tools... More Tools from the ChimeraX menu opens the Toolshed in the ChimeraX browser (Figure 4); navigating to a plugin of interest and simply clicking a link performs the installation. The user is later notified when updates become available for their installed plugins.

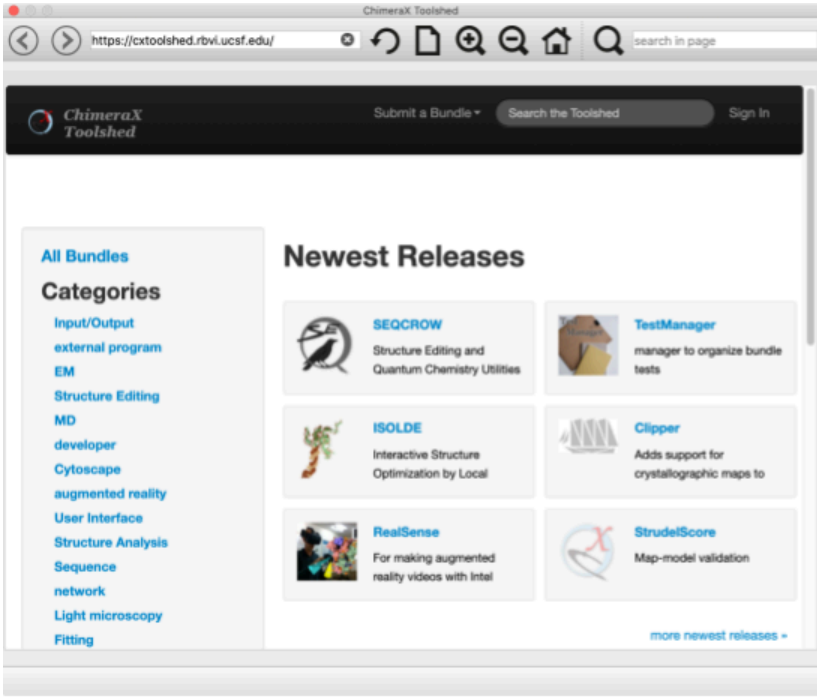


Figure 4. ChimeraX Toolshed website in the ChimeraX browser.

As of January 2022, the Toolshed includes nine bundles from six groups other than our own, with authors ranging from students to senior researchers in the U.S. and abroad. Seven bundles have been downloaded over a hundred times, five of them over a thousand (led by ISOLDE with 23,000). Additional ChimeraX plugins are under development. We are encouraged by this progress and will continue our strong support for community development and sharing of ChimeraX plugins.

Research Plan:

To address the scientific challenges and achieve the vision outlined above, we have devised a set of specific aims to focus our development efforts and to achieve maximum impact within the extensive scientific community that depends on our software for successful advancement of their research programs.

Aim 1: Develop interactive visualization and analysis tools for atomic structures and associated data, including higher-order assemblies, conformational ensembles, density maps, and sequences.

The structural biology systems being analyzed today are not only much larger than they were in previous years, but they are also "wider" in that there are more types of data, such as sequences, structures, experimental maps, and computational results, that must be brought to bear simultaneously for the most insightful outcome. The proposed developments are crucial for structure validation by repositories (RCSB PDB and integrative hybrid models, see letters from Burley and Sali); effective use of these structures in further modeling by researchers with varying levels of computational expertise; and compelling communication of findings through animations, facilitated by a graphical timeline tool (see letter from Cheng). Together, these enable research on subjects ranging from inhibitor design to the mechanisms of broadly neutralizing antibodies to the workings of the nuclear pore complex, ribosomes, and other molecular machines.

Aim 1A: Analysis capabilities

Prepare structures for docking/modeling. Frequently, atomic structures cannot be used directly as input to molecular mechanics or docking calculations because of missing side chains or hydrogens, lack of parameters, or other issues. Having to address a series of issues with multiple different programs is a significant barrier, especially for those with less experience or reduced access to software. We will implement a streamlined tool for preparing structures by: deleting unwanted solvent/ions; eliminating alternate atomic locations; changing certain modified residues to their cognate standard types (*e.g.* selenomethionine to methionine); filling out incomplete side chains using a rotamer library (see letter from Adam Frost); adding hydrogens; and assigning partial charges. Charges for non-standard residues will be computed using AmberTools[34].

Sequence-structure analysis. The power of ChimeraX sequence tools is largely in how they are bidirectionally integrated with structures, such that selections in the 2D view select the corresponding parts of a structure in 3D and vice versa, helix/strand assignments can guide sequence alignment, and properties derived from sequence (such as conservation or UniProt annotations) seamlessly map onto associated structures and vice versa (per-residue RMSDs among a set of superimposed structures can be shown as a histogram above the sequences). These further integrate with other ChimeraX applications, for example to easily evaluate whether the protein-protein interfaces of subunits fitted into a map are well conserved or subject to post-translational modification. We propose to add capabilities for:

- Adding sequences to alignments from structures, files, text, or UniProt
- Other alignment editing: sequence renaming, deletion, and reordering
    - Allowing sequence edits to make corresponding changes in associated structures
- Generating a multiple sequence alignment from a structure superposition

Also, we will improve the responsiveness of ChimeraX's sequence viewer when handling alignments of thousands of sequences. This will involve rearchitecting the display to reduce memory use and streamline rendering. As a useful alternative to showing large alignments directly, we will implement an information-dense summary view based on ProfileGrids[35].

Dimensionality reduction. It can be difficult to detect overall patterns within large sets of complex data, such as whether dataset members naturally form clusters, and how individual members or clusters relate to the others. We will incorporate dimensionality-reduction methods such as UMAP (Uniform Manifold Approximation and Projection for Dimension Reduction)[6,36] into ChimeraX so that tools can present such complex data in interactive 2D plots, where selecting a point on the plot would show the corresponding 3D structural information (modeled assembly, mutation positions, density map, *etc.* depending on the use case). Figure 5 shows an example with molecular conformations deduced from cryoEM data using machine learning. We will also add interactive 2D plots in more prosaic contexts, such as plotting measurables of a trajectory against time, or showing inter-residue contact maps.

Aim 1B: Dissemination of results

Animations. Expert users can create beautiful and complicated multistep movies with ChimeraX, as evident from recently published examples[37,38]. However, this generally requires writing a long ChimeraX command script, with several cycles of trial and error. We will simplify the process significantly by adding two major features. One is the ability to create "scenes" in which the depictions and positions of all models are

remembered, and that can easily be restored. Apart from animation, this is useful for quickly switching between different views of a system, such as for different panels of a figure. The second feature is a “timeline” tool into which scenes can be dropped and transitions smoothly interpolated (see letter from Adam Frost). The timeline will also allow inserting trajectories from molecular dynamics (MD) or morphing.

New glycan depictions. Analogous to current options for simplified nucleotide displays such as ladder rungs or “lollipops” and coloring by base type, we will implement abstracted depictions of glycan residues in accord with the community-curated standard 2D Symbol Nomenclature for for Glycans (SNFG)[39,40] but in 3D, along the lines of 3D-SNFG[41]. This will aid in rapid recognition of protein glycosylation sites and types, including complex branched-chain structures (see letter from Stephen Burley).

Improved interoperability with wwPDB repositories. With ISOLDE increasingly becoming a preferred tool for model building, it would be very useful if mmCIF files written by ChimeraX could be used directly for wwPDB deposition (see letter from Yifan Cheng). To facilitate this, we will improve the completeness of the data that ChimeraX places in the file, *e.g.* precise helix and sheet types (alpha vs. 3-10 helix, for example). We will implement a “metadata editor” for adding information such as experimental conditions and biopolymer sequence, as well as a tool to help determine the PDB 3-letter code for a ligand. Conversely, users may want to evaluate the quality of deposited structures. We will provide a tool to fetch and display the information from a wwPDB Validation Report for a structure (see letter from Stephen Burley) or Phenix validation results (see letter from Paul Adams), directly annotated on the structure itself as appropriate. Experimental crosslinks are frequently used for integrative modeling, the focus of the new PDB-Dev archive[42], and we will add support to visualize the challenging cases where crosslink end points are indeterminate because the molecular assembly has many equivalent subunits (see letters from David Agard, Stephen Burley, and Andrej Sali).

Aim 2: Develop interactive visualization and analysis tools for electron microscopy of molecular assemblies and cells.

Advances in electron microscopy methods now routinely produce atomic-resolution maps, detailed views of cellular organization, and even large-scale views of tissues. These are spurring developments in visualization and modeling, some of which we describe here, focusing on machine learning, model refinement, and virtual reality. Collaborators’ research on heat shock proteins, protein degradation, and membrane remodeling (see letters from Dave Agard, Yifan Cheng, and Adam Frost) and many others drive these new developments. In addition to these highly innovative developments, ChimeraX offers a wide range of high-performance, easy-to-use, proven visualization capabilities that we optimize to handle the ever larger experimental data sets. Our tools span very diverse microscopy imaging techniques: cryo-EM single-particle reconstruction, which produces atomic-resolution models; electron tomography, probing cell organization at lower resolutions; and blockface scanning electron microscopy (*e.g.* focused ion beam or diamond knife methods) and 3D light microscopy, covering cellular and tissue-level organization. The range of analysis needs is vast, and our goal in ChimeraX is to provide a foundation where many of the innovations can be made as plugins provided by community developers, as detailed in Aim 3 of this proposal.

Aim 2A: Developments for electron microscopy of molecular assemblies

Machine learning structure predictions. The release in 2021 of machine-learning methods such as AlphaFold[28] and RoseTTAFold[43] are groundbreaking advances in protein structure determination, for the first time offering accuracy comparable to experimental methods {PMID: 34599769}. The AlphaFold prediction service we added to ChimeraX has been used 12,000 times by 2600 researchers in a period of months and has revealed pressing needs which we plan to address. Our developments will enable predicting large complexes and using them with cryoEM data to model and validate structures of biological systems.

The first obstacle is that few research labs have the needed large-memory machine-learning GPUs (e.g. Nvidia A100 with 80 Gbytes of memory). ChimeraX currently runs AlphaFold on Google Colab virtual servers offering GPUs with only 16 Gbytes that limit structure size to about 1000 residues, inadequate for modeling most complexes. Our tests show the latest Nvidia A40 GPUs with 48 GBytes can handle up to 4000 residues, allowing most complexes to be predicted. We plan to have ChimeraX run predictions via a web service utilizing these modern GPUs on the UCSF Wynton compute cluster. Jobs will be queued and prioritized to support the current ChimeraX AlphaFold usage of 100 predictions per day.

Large complexes have multiple protein-protein interfaces, and interface predictions are often wrong (1/3 incorrect for AlphaFold[30]. A remarkable capability of AlphaFold is to reliably identify wrong interfaces by estimating distance errors between every pair of residues. Displaying a heatmap of residue-to-residue distance errors clearly reveals incorrect interfaces. We will display such heatmaps, highlight likely incorrect interfaces on 3D structures, and allow fitting the individual correctly predicted domains into cryoEM maps.

In the period of this proposal, we expect further machine-learning advances to predict nucleic acid structures and ligand binding modes. Our developments will make these revolutionary new methods of structure prediction accessible to research labs that lack the hardware and expertise to set up and run the calculations on their own.

Visualization of molecular conformational heterogeneity using machine learning. Several new cryo-EM software algorithms try to tease out multiple conformations of flexible molecular assemblies using machine learning methods (*e.g.* cryoDRGN[3], 3DFlex[4], 3D Variability Analysis[44], ManifoldEM[45]). New visualization capabilities are needed to understand the results of these methods. For example, the cryoDRGN method trains a neural network to produce a 3D map parameterized by a multi-dimensional space of parameters representing conformations and flexibility. Each of the thousands to millions of particle images seen in 2D micrographs is classified in this typically 10-dimensional space. For visualization, this 10-dimensional point cloud is projected onto a 2D scatter plot where clusters represent conformations (Figure 5). The parameters associated with a point in the plot can be fed into the neural net to produce the associated density map. We will incorporate a powerful new point projection method called Uniform Manifold Approximation and Projection[6,36] (UMAP), which we also plan to use for other analysis tools as described in Aim 1, and we will add the ability to visualize the observed conformations by computing maps for chosen points. The neural network training is typically done on a large computing cluster and will *not* be part of ChimeraX.

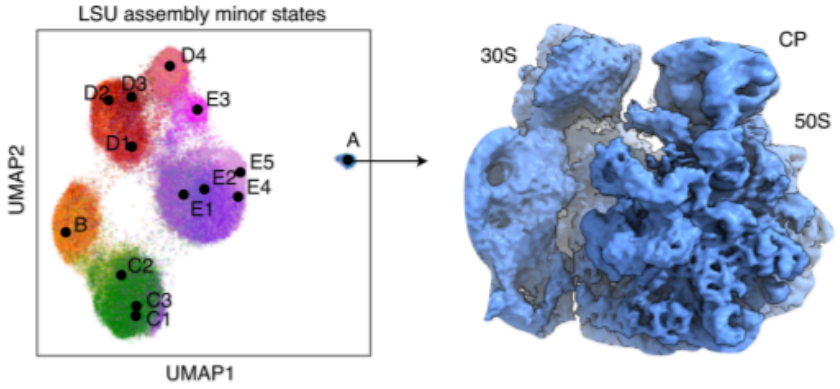


Figure 5: Visualization of ribosome conformations plotted with UMAP where each of the thousands of colored dots represents one ribosome particle seen in a 2D micrograph. At right is the 3D map computed from the cryoDRGN neural net associated with parameter values at "A".

Interactive molecular dynamics for fitting atomic models in maps. It is challenging to build correct atomic models at cryo-EM resolutions in the 2-4 Å range. The use of interactive molecular dynamics to assist in matching the density map has proved highly effective in the ISOLDE plugin to ChimeraX, and at secondary-structure resolutions (5-8 Å), in the TEMPy fitting plugin (see letters from Tristan Croll and Maya Topf). These plugins use the OpenMM library for molecular simulations on the graphics processor. We will provide these quick, localized calculations in core ChimeraX using OpenMM[33], relieving ISOLDE, TEMPy, and other plugins of the need to separately implement complex simulation code. Robust parametrization of novel ligands and modified residues is challenging, and will be accomplished with support from the OpenMM team (see letter from John Chodera).

Faster map contouring and masking. Higher resolutions have resulted in larger map sizes, with up to a billion grid points (e.g. 1024^3). We developed in ChimeraX one of the fastest contour surface calculation codes available, and plan to make it several times faster using parallel computations. Interactive model building often uses a "spotlight" mode where only the map region of current focus is shown; we will also optimize the speed of map updates as different regions are shown.

Additional building and validation tools. Modeling and validation capabilities where ChimeraX utilizes Phenix are developed with separate funding. We collaborate with Paul Adams (see letter of support) with funding from an awarded R24 grant to interface computational methods of Phenix[46,47], the most widely used modeling software, to enable fragment extension, ligand placement, and validation within ChimeraX.

Aim 2B: Developments for electron tomography of cells

Improved volumetric visualization using lighting. Ambient-occlusion lighting (where shadows are cast from all directions) greatly enhances the 3D perception of molecular models and cryo-EM map surfaces and has become widespread in publications following our interactive implementation in ChimeraX. Cellular microscopy is often limited by lack of contrast in highly crowded molecular environments, and contrast-enhancing rendering is a critical need. We will extend our ambient-occlusion lighting to transparent volumetric renderings, exploring both hardware-accelerated raycasting approaches enabled by the latest graphics cards and a method available on all graphics hardware that uses stacks of transparent planes. This is in collaboration with Allen Institute of Cell Science, which has pioneered these methods[48] (see Figure 6 and letter from Graham Johnson). Extending ambient-occlusion lighting to transparent renderings will benefit 3D imaging by both electron and light microscopy.

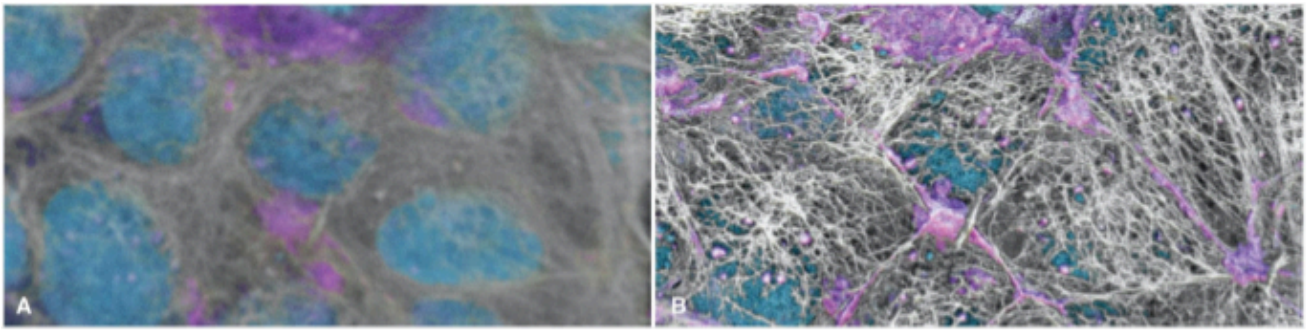


Figure 6: Comparison of unlit volumetric rendering without lighting (A) and with lighting (B) using the AGAVE photorealistic rendering method developed by the Allen Institute of Cell Science.

Visualizing, quantifying, and creating segmentations. Segmentations that identify large molecular assemblies, filaments, organelles, and diverse components within cells are a primary means of understanding cellular organization. Despite the high value of segmentations, standard file formats are still under development. ChimeraX will support the new segmentation formats being developed by the EM DataBank (see letter from Gerard Kleywegt). Quantification of segmented objects and their spatial relationships is a key need. For example, counting thousands of proteasome degradation complexes in proximity to amyloid fibrils[49] gives insight into neurodegenerative diseases. We will provide new ChimeraX capabilities to compute statistics for spatial relationships of segmented objects. These are especially needed with new machine-learning segmentation approaches to identify molecular assemblies in cells. Early steps in analyzing tomography involve visualization and often hand annotation to mark interesting regions. We will provide interactive 3D painting of structures of interest using the mouse or virtual reality.

Analysis tools for subtomogram averaging. With recent advances, the densities of multiple copies of molecular assemblies observed in tomograms can be averaged to achieve subnanometer resolutions. Such resolutions enable unique structural studies such as directly observing accessory proteins bound to microtubules (see letter from Dave Agard). Focused refinement, which requires creating a region mask, is used to optimize the resolution of specific regions in these assemblies; the separately refined map regions are then combined for visualization. We will provide tools to create masks and to combine focused refinement maps. Methods of identifying secondary structures such as alpha helices using machine learning[50] and flexibly fitting atomic models at the secondary-structure level are areas of active ChimeraX plugin development that aid subtomogram analysis (see letters from Jing He and Maya Topf).

Virtual reality visualization of tomograms. Stereoscopic visualization of tomograms using virtual reality (VR) headsets with ChimeraX has enabled remarkable discoveries such as leaf-like aggregates in neurons in Huntington's disease (see letter from Wah Chiu). The ability to naturally change viewpoints using head movements and manipulation with hand controllers shows great promise and has been under development in ChimeraX for six years and in active use for three, ranging in applications from drug-binding studies to cellular tomography. We will optimize VR visualization for the large sizes of tomogram data, allowing simple resolution control, cropping to subregions, high-performance noise reduction, and interactive annotation of structures of interest. These developments will allow researchers who are not experts in virtual reality to analyze their 3D

microscopy images, reaching many labs beyond the current early-adopters at the world's leading microscopy centers.

Aim 3: Provide a diverse software foundation enabling labs around the world to develop and distribute new molecular and cellular structure analysis software.

This aim has two major objectives: 1) to increase community development of ChimeraX plugins, and 2) to ensure ChimeraX supports cutting-edge structural biology research decades beyond the initial work of the core UCSF development team. Currently, there are about ten outside developers contributing plugins, enabling new analyses that substantially increase the impact of ChimeraX. Examples include atomic-model refinement and validation[20,51,52], quantum chemistry utilities[53], EM fitting and segmentation, and integrative modeling[54] (see letters from Tristan Croll, Gerard Kleywegt, Steven Wheeler, Maya Topf, Jing He, and Andrej Sali) enabling biological advances such as predicting drug heterocycle stacking interactions (Steven Wheeler) and modeling insulin secretion by pancreatic β-cells (Andrej Sali). We propose to simplify development of plugins and support developers outside UCSF, with the goal of building up to 100 or more outside contributors. We promote the FAIR principles[55] by making plugins Findable, Accessible, Interoperable, and Reusable through our Toolshed repository, which indexes the available plugins and allows single-click installation from within ChimeraX.

The predecessor to ChimeraX, our Chimera software, has been in productive use for 20 years, and we expect it to be fully superseded by ChimeraX within the next few years. We propose several key developments to allow ChimeraX to be used productively and continuously augmented with state-of-the-art capabilities for a 15-year lifespan. This requires empowering community developers to lead all aspects of the project by the time the core UCSF development team completes the base functionality. ChimeraX code is publicly available on GitHub, and our proposal will augment our UCSF distributions with distributions through the standard cloud-based Python package managers PyPi and Anaconda, and migrate from UCSF-centralized builds to standard cloud-based build environments such as Travis that community developers can use.

In the following we describe some of the principal developments we will undertake to maximize the impact of community developers of ChimeraX tools and assure that fruitful new developments continue beyond the funding requested in this proposal.

Aim 3A: New capabilities for plugin development

ChimeraX has facilities for reading, rendering, building, modifying, and exploring molecular data. This foundation enables the rapid development of new and innovative analysis methods as ChimeraX plugins by the research community. To further facilitate community development, we propose the following:

A standard platform for machine-learning algorithms. Machine-learning methods that use neural networks are beginning to have a dramatic impact on structural biology, for example, producing by far the best protein structure predictions from sequence in CASP14[2], and demonstrating remarkable ability to extract conformational dynamics in electron microscopy data (*e.g.* cryoDRGN[3], 3DFlex[4]). We intend to make ChimeraX a standard platform for distributing new machine-learning methods by including the standard machine-learning frameworks such as PyTorch and TensorFlow. We will provide simple programming examples so that labs developing these algorithms can produce plugins that combine neural networks with the rich visualization and user interfaces offered by ChimeraX.

Simple programming to access web services. ChimeraX uses web services[56] to perform calculations or access data on a remote server or the cloud instead of on the user's computer. Examples include performing a BLAST sequence search for structures in the EBI AlphaFold database, or predicting a protein structure from sequence using a machine learning method. Computing on the cloud allows accessing large databases, running compute-intensive jobs, and updating algorithms without the user installing any new software. We will develop simple programming interfaces to allow ChimeraX plugins to utilize web services based on the REST communication standard. This will allow submitting a job, monitoring progress, and receiving results for subsequent visualization.

Providing complex calculations with good user experience. The analysis done in ChimeraX is highly interactive, with most operations taking seconds or fractions of a second (*e.g.* computing molecular surfaces, hydrogen-bonding, or hydrophobicity, or denoising a cryo-EM map). More sophisticated calculations, web

services, and especially new algorithms in community-developed plugins that have not yet been optimized for speed can take minutes or hours to run. To avoid freezing ChimeraX, we will provide programming interfaces to manage long-running tasks and allow using ChimeraX interactively while calculations proceed. Task progress will be reported and results shown when they are available, and jobs can be stopped, for example to change the input or to reduce the input size if the estimated time to complete is prohibitive. Task-management user interfaces and parallel computing approaches will be built into ChimeraX, simplifying work for plugin developers.

Programming examples. ChimeraX community developers are primarily graduate students and postdocs rather than professional programmers. A plugin may include several common types of capabilities: reading a new file format, adding a new typed command, adding a user-interface panel and menu entry, adding a mouse mode, and adding a "preset" menu entry that sets display styles and colors. For each of these common tasks, we will provide simple programming examples that can be used as templates by community developers. ChimeraX has hundreds of functions allowing control of visualization, calculations, data reading and saving, with detailed documentation. We will develop simple programming examples that use common combinations of these extensive capabilities (https://rbvi.github.io/chimerax-recipes/), and will continue to add examples as needed. The goal is that community developers will always have example Python code to start from that allows them to make a working plugin as quickly as possible.

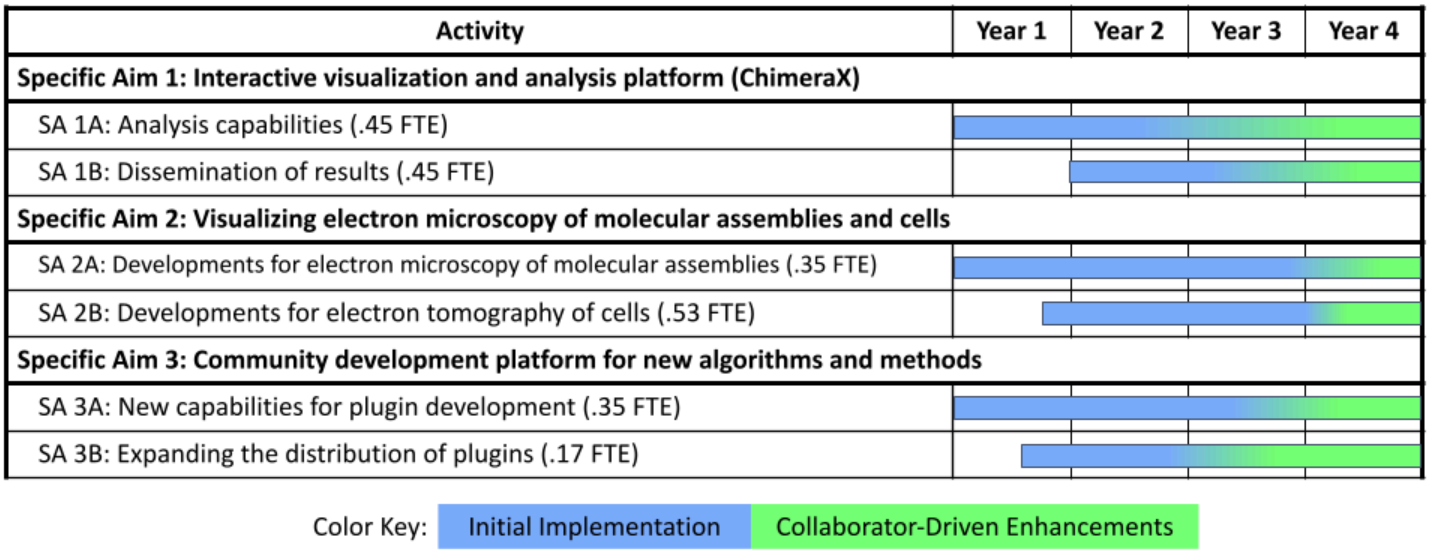
Aim 3B: Expanding the distribution of plugins

Utilizing the PyPi and Anaconda Python package managers. ChimeraX is distributed from our UCSF-hosted web site and includes the Toolshed repository of curated community-developed plugins. We will augment these approaches by also supporting distributions using the standard cloud-based Python package repositories PyPi and Anaconda. ChimeraX uses more than 50 third-party packages from these repositories, and plugins often utilize additional packages. Also providing ChimeraX through the same mechanism simplifies use by community developers, provides greater flexibility for reuse by not requiring our desktop application distribution, and provides a distribution system that is not tied to UCSF computer resources and hence can serve beyond the funded development of the UCSF team. This is crucial for maintaining a vibrant ChimeraX development environment for the expected decades-long lifetime of the software.

Cloud-based building of ChimeraX and plugins. ChimeraX is built for Windows, Mac and several Linux distributions on UCSF computer systems, requiring multiple computers and a large set of ancillary software (C++ compilers, “makefile” scripts, documentation generation tools, and 50+ third-party packages and libraries) as described in our programming documentation. It is automatically built and tested every night, both to ensure that recent code additions or changes have not broken something and to make the latest enhancements available to others via these “daily build” versions. To simplify builds, we propose migrating to standard cloud-based “continuous integration” services such as Jenkins, Travis CI, or GitLab. This ensures that community developers can more easily build ChimeraX as all system requirements are specified, facilitates testing changes, and allows plugins that use compiled C++ to be built for all operating systems without requiring the biology lab that develops the plugin to have their own suite of development machines.

Project Timeline: Most of the Specific Aims can be split roughly into two phases, Initial Implementation and Collaborator-Driven Enhancements (see table below). Implementation provides the reference functionality that may be used as a foundation for other aims. Enhancements include adding functionality requested by users, fixing bugs, and changes necessitated by new versions of platform operating systems.

Potential Problems, Alternative Strategies, Benchmarks for Success: Potential problems revolve around having adequate programmer resources for rapid development, incorporating the best available technologies, and maintaining the software beyond the period of funded development. We aim to convert as many as possible of our 38,000 registered (and many more unregistered) Chimera users to ChimeraX within the next year or two. Long-term users of complex software are notoriously reluctant to upgrade, but strong adoption so far indicates that the numerous more powerful capabilities of ChimeraX compared to Chimera are compelling and in view of the fact that we are doing *no* active development on Chimera. The ChimeraX user interface and commands preserve the best features of Chimera, making the transition easier. We will also implement an export-to-ChimeraX feature in Chimera, so that work done in Chimera will not be lost due to the transition. Our



extensive presentations of tutorials at workshops and development of online tutorial materials, now entirely focused on ChimeraX, will aid in recruiting existing Chimera users as well as graduate students new to molecular visualization. While most academic software projects are the work of a single graduate student, our software is developed by a team with deep experience in the problem domain, and we believe the scope of the work is appropriate given the human resources available. ChimeraX is based on the most widely used and stable components: the Python 3 programming language, the Qt window toolkit, and the latest GPU programming APIs of the OpenGL graphics library. We believe that these components will sustain a 15-year software lifetime (Chimera has been in use for 20 years). Community effort will be important for maintenance of the software beyond the funded development period, and a primary goal of the ChimeraX design (all of Aim 3) is to engage significant numbers of community developers.

Alternative strategies to deliver multiscale visualization to researchers include using web-based and cloud-based tools. Web browser interfaces usually offer narrow focused capabilities. In contrast ChimeraX excels in combined analysis of structure, sequence, annotations, and experimental maps, using diverse methods, often with dozens of datasets open simultaneously. Integrative analysis leads to insights that are not possible looking at one type of data at a time as commonly done with web tools. Web browsers limit graphics performance, memory use and access to local files, making them poorly suited to the integrative analysis. ChimeraX incorporates web and cloud services as described in the proposal to provide a synthesis of many data types and analysis methods.

We will use unambiguous quantitative measures of success for ChimeraX (see Significance section), which was cited 800 times in 2021 and has more than 1400 citations to date. Voluntary ChimeraX user registrations are rapidly accelerating now at 280 per month with 4200 to date, the program was downloaded over 50,000 times in 2021, and 12 plugins to ChimeraX have been written by outside developers. These numbers reflect lower bounds, as many research articles do not cite the visualization software used, and many users do not register. Our registration system asks a user to register only after 15 distinct days of use, and there are no software limitations for failure to register. This gives a conservative measure of serious users. We expect numbers of citations and registrations per month for ChimeraX to surpass Chimera levels (~300 citations and 300 registrations per month) within the next few years, when a majority of users have migrated to the new software. The stable programming APIs, Toolshed plugin store, and active support provided with ChimeraX are expected to enable hundreds of contributed tools from numerous research labs over the projected 15-year lifespan of the software, with an estimated 50 contributed tools during the proposed 4-year funding period.

References

1. Kim, S. J. *et al.* Integrative structure and functional anatomy of a nuclear pore complex. *Nature* **555**, 475–482 (2018).
2. Service, R. F. 'The game has changed.' AI triumphs at protein folding. *Science* **370**, 1144–1145 (2020).
3. Zhong, E. D., Bepler, T., Berger, B. & Davis, J. H. CryoDRGN: reconstruction of heterogeneous cryo-EM structures using neural networks. *Nat. Methods* **18**, 176–185 (2021).
4. Punjani, A. & Fleet, D. J. 3D Flexible Refinement: Structure and Motion of Flexible Proteins from Cryo-EM. *bioRxiv* (2021).
5. HuBMAP Consortium. The human body at cellular resolution: the NIH Human Biomolecular Atlas Program. *Nature* **574**, 187–192 (2019).
6. McInnes, L., Healy, J., Saul, N. & Großberger, L. UMAP: Uniform Manifold Approximation and Projection. *J. Open Source Softw.* **3**, 861 (2018).
7. Goddard, T. D. *et al.* Molecular Visualization on the Holodeck. *J. Mol. Biol.* **430**, 3982–3996 (2018).
8. ucsfpharmacy. UCSF ChimeraX pushes drug discovery into virtual reality. *YouTube* https://youtu.be/S4lDzUEUFL0 (2019).
9. Goddard, T. Mixed Reality Video Recording in ChimeraX. *ChimeraX* https://www.rbvi.ucsf.edu/chimerax/data/mixed-reality-nov2019/mrhowto.html (2019).
10. Goddard, T. Opioid Molecules Mixed Reality. *YouTube* https://www.youtube.com/watch?v=FCotNi6213w (2019).
11. Goddard, T. COVID-19. *YouTube* https://www.youtube.com/watch?v=dKNbRRRFhqY&t=92s (2020).
12. LookingGlassFactory Home Page. *LookingGlassFactory* https://lookingglassfactory.com/.
13. Goddard, T. Using a LookingGlass Display with ChimeraX. *ChimeraX* https://www.rbvi.ucsf.edu/chimerax/data/lookingglass-july2020/ (2020).
14. Goddard, T. Hand Tracking with Leap Motion in ChimeraX. *ChimeraX* https://www.rbvi.ucsf.edu/chimerax/data/leap-july2020/ (2020).
15. Goddard, T. D. *et al.* UCSF ChimeraX: Meeting modern challenges in visualization and analysis. *Protein*

Sci. (2017) doi:10.1002/pro.3235 [doi].

16. Pettersen, E. F. *et al.* UCSF ChimeraX: Structure visualization for researchers, educators, and developers. *Protein Sci.* **30**, 70–82 (2020).
17. Pei, J. & Grishin, N. V. AL2CO: calculation of positional conservation in a protein sequence alignment. *Bioinformatics* **17**, 700–712 (2001).
18. Sievers, F. *et al.* Fast, scalable generation of high-quality protein multiple sequence alignments using Clustal Omega. *Mol. Syst. Biol.* **7**, 539 (2011).
19. Edgar, R. C. MUSCLE: multiple sequence alignment with high accuracy and high throughput. *Nucleic Acids Res.* **32**, 1792–1797 (2004).
20. Croll, T. I. ISOLDE: a physically realistic environment for model building into low-resolution electron-density maps. *Acta crystallographica. Section D, Structural biology* **74**, 519–530 (2018).
21. Bunduc, C. M. *et al.* Structure and dynamics of a mycobacterial type VII secretion system. *Nature* **593**, 445–448 (2021).
22. Xu, P. *et al.* Structural insights into the lipid and ligand regulation of serotonin receptors. *Nature* **592**, 469–473 (2021).
23. Josephs, T. M. *et al.* Structure and dynamics of the CGRP receptor in apo and peptide-bound forms. *Science* **372**, (2021).
24. McCallum, M. *et al.* N-terminal domain antigenic mapping reveals a site of vulnerability for SARS-CoV-2. *Cell* **184**, 2332–2347.e16 (2021).
25. Thomson, E. C. *et al.* Circulating SARS-CoV-2 spike N439K variants maintain fitness while evading antibody-mediated immunity. *Cell* **184**, 1171–1187.e20 (2021).
26. Voss, W. N. *et al.* Prevalent, protective, and convergent IgG recognition of SARS-CoV-2 non-RBD spike epitopes. *Science* **372**, 1108–1112 (2021).
27. Williams, W. B. *et al.* Fab-dimerized glycan-reactive antibodies are a structural category of natural antibodies. *Cell* **184**, 2955–2972.e25 (2021).
28. Jumper, J. *et al.* Highly accurate protein structure prediction with AlphaFold. *Nature* **596**, 583–589 (2021).
29. Varadi, M. *et al.* AlphaFold Protein Structure Database: massively expanding the structural coverage of

protein-sequence space with high-accuracy models. *Nucleic Acids Res.* **50**, D439–D444 (2022).

30. Evans, R. *et al.* Protein complex prediction with AlphaFold-Multimer. *bioRxiv* 2021.10.04.463034 (2021) doi:10.1101/2021.10.04.463034.
31. Sali, A. & Blundell, T. L. Comparative protein modelling by satisfaction of spatial restraints. *J. Mol. Biol.* **234**, 779–815 (1993).
32. Webb, B. & Sali, A. Comparative Protein Structure Modeling Using MODELLER. *Curr. Protoc. Bioinformatics* **47**, 5.6.1–32 (2014).
33. Eastman, P. *et al.* OpenMM 7: Rapid development of high performance algorithms for molecular dynamics. *PLoS Comput. Biol.* **13**, e1005659 (2017).
34. D.A. Case, H.M. Aktulga, K. Belfon, I.Y. Ben-Shalom, S.R. Brozell, D.S. Cerutti, T.E. Cheatham, III, V.W.D. Cruzeiro, T.A. Darden, R.E. Duke, G. Giambasu, M.K. Gilson, H. Gohlke, A.W. Goetz, R. Harris, S. Izadi, S.A. Izmailov, C. Jin, K. Kasavajhala, M.C. Kaymak, E. King, A. Kovalenko, T. Kurtzman, T.S. Lee, S. LeGrand, P. Li, C. Lin, J. Liu, T. Luchko, R. Luo, M. Machado, V. Man, M. Manathunga, K.M. Merz, Y. Miao, O. Mikhailovskii, G. Monard, H. Nguyen, K.A. O'Hearn, A. Onufriev, F. Pan, S. Pantano, R. Qi, A. Rahnamoun, D.R. Roe, A. Roitberg, C. Sagui, S. Schott-Verdugo, J. Shen, C.L. Simmerling, N.R. Skrynnikov, J. Smith, J. Swails, R.C. Walker, J. Wang, H. Wei, R.M. Wolf, X. Wu, Y. Xue, D.M. York, S. Zhao, and P.A. Kollman. AmberTools21. *The Amber Project* https://ambermd.org/CiteAmber.php (2021).
35. Roca, A. I. ProfileGrids: a sequence alignment visualization paradigm that avoids the limitations of Sequence Logos. *BMC Proc.* **8**, S6–S6. eCollection 2014 (2014).
36. McInnes, L., Healy, J. & Melville, J. UMAP: Uniform Manifold Approximation and Projection for Dimension Reduction. *arXiv [stat.ML]* (2018).
37. Miletic, S. *et al.* Substrate-engaged type III secretion system structures reveal gating mechanism for unfolded protein translocation. *Nat. Commun.* **12**, 1546 (2021).
38. Abdella, R. *et al.* Structure of the human Mediator-bound transcription preinitiation complex. *Science* **372**, 52–56 (2021).
39. Varki, A. *et al.* Symbol Nomenclature for Graphical Representations of Glycans. *Glycobiology* **25**, 1323–1324 (2015).

40. Neelamegham, S. *et al.* Updates to the Symbol Nomenclature for Glycans guidelines. *Glycobiology* **29**, 620–624 (2019).
41. Thieker, D. F., Hadden, J. A., Schulten, K. & Woods, R. J. 3D implementation of the symbol nomenclature for graphical representation of glycans. *Glycobiology* **26**, 786–787 (2016).
42. Vallat, B., Webb, B., Westbrook, J., Sali, A. & Berman, H. M. Archiving and disseminating integrative structure models. *J. Biomol. NMR* **73**, 385–398 (2019).
43. Baek, M. *et al.* Accurate prediction of protein structures and interactions using a three-track neural network. *Science* **373**, 871–876 (2021).
44. Punjani, A. & Fleet, D. J. 3D variability analysis: Resolving continuous flexibility and discrete heterogeneity from single particle cryo-EM. *J. Struct. Biol.* **213**, 107702 (2021).
45. Maji, S. *et al.* Propagation of conformational coordinates across angular space in mapping the continuum of states from cryo-EM data by manifold embedding. *J. Chem. Inf. Model.* **60**, 2484–2491 (2020).
46. Adams, P. D. *et al.* PHENIX: a comprehensive Python-based system for macromolecular structure solution. *Acta Crystallogr. D Biol. Crystallogr.* **66**, 213–221 (2010).
47. Liebschner, D. *et al.* Macromolecular structure determination using X-rays, neutrons and electrons: recent developments in Phenix. *Acta Crystallogr D Struct Biol* **75**, 861–877 (2019).
48. Kroes, T., Post, F. H. & Botha, C. P. Exposure render: an interactive photo-realistic volume rendering framework. *PLoS One* **7**, e38586 (2012).
49. Guo, Q. *et al.* In Situ Structure of Neuronal C9orf72 Poly-GA Aggregates Reveals Proteasome Recruitment. *Cell* **172**, 696–705.e12 (2018).
50. Mu, Y., Sazzed, S., Alshammari, M., Sun, J. & He, J. A tool for segmentation of secondary structures in 3D cryo-EM density map components using deep convolutional neural networks. *Front. Bioinform.* **1**, (2021).
51. McNicholas, S. *et al.* Automating tasks in protein structure determination with the clipper python module. *Protein Sci.* **27**, 207–216 (2018).
52. Croll, T. I. & Read, R. J. Adaptive Cartesian and torsional restraints for interactive model rebuilding. *Acta Crystallogr D Struct Biol* **77**, 438–446 (2021).
53. Schaefer, A. J., Ingman, V. M. & Wheeler, S. E. SEQCROW: A ChimeraX bundle to facilitate quantum

chemical applications to complex molecular systems. *J. Comput. Chem.* (2021) doi:10.1002/jcc.26700.

54. Saltzberg, D. J. *et al.* Using Integrative Modeling Platform to compute, validate, and archive a model of a protein complex structure. *Protein Sci.* **30**, 250–261 (2021).
55. Wilkinson, M. D. *et al.* The FAIR Guiding Principles for scientific data management and stewardship. *Sci Data* **3**, 160018 (2016).
56. ChimeraX Team. Web Services Used by UCSF ChimeraX. *ChimeraX WebServices* https://www.rbvi.ucsf.edu/chimerax/docs/webservices.html (2021).